

Tesis Doctoral

Mapeo por asociación en girasol: diversidad nucleotídica, desequilibrio de ligamiento e identificación de genes involucrados en la resistencia a la podredumbre húmeda del capítulo

Fusari, Corina Mariana

2010

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:

Fusari, Corina Mariana. (2010). Mapeo por asociación en girasol: diversidad nucleotídica, desequilibrio de ligamiento e identificación de genes involucrados en la resistencia a la podredumbre húmeda del capítulo. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:

Fusari, Corina Mariana. "Mapeo por asociación en girasol: diversidad nucleotídica, desequilibrio de ligamiento e identificación de genes involucrados en la resistencia a la podredumbre húmeda del capítulo". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2010.

EXACTAS UBA

Facultad de Ciencias Exactas y Naturales



UBA

Universidad de Buenos Aires



UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales

**MAPEO POR ASOCIACIÓN EN GIRASOL:
Diversidad Nucleotídica, Desequilibrio de Ligamiento e
Identificación de Genes Involucrados en la Resistencia a la
Podredumbre Húmeda del Capítulo**

Tesis presentada para optar al título de Doctor de la Universidad de
Buenos Aires en el área Ciencias Biológicas

Lic. Corina Mariana Fusari

Director de Tesis: Dra. Verónica V. Lia

Director Asistente: Dra. Norma B. Paniego

Consejero de Estudios: Dr. H. Esteban Hopp

Instituto de Biotecnología, CICVyA, INTA Castelar
Buenos Aires, 2010

MAPEO POR ASOCIACIÓN EN GIRASOL: DIVERSIDAD NUCLEOTÍDICA, DESEQUILIBRIO DE LIGAMIENTO E IDENTIFICACIÓN DE GENES INVOLUCRADOS EN LA RESISTENCIA A LA PODREDUMBRE HÚMEDA DEL CAPÍTULO

Resumen

El mapeo por asociación es una herramienta poderosa que permite la identificación de *loci* cuya contribución explica parte de la variación fenotípica observada. En general, los marcadores utilizados en estos estudios son los Polimorfismos de Nucleótido Simple (SNPs, *Single Nucleotide Polymorphisms*) y los indels (eventos de inserción, *insertion/deletion*).

Los objetivos de este trabajo comprendieron el estudio de la diversidad nucleotídica y el alcance del desequilibrio de ligamiento (DL) en girasol cultivado, y el diseño de una estrategia de mapeo por asociación para la resistencia a la Podredumbre Húmeda del Capítulo (PHC), causada por *Sclerotinia sclerotiorum*.

La frecuencia de SNPs (1/61) y la diversidad nucleotídica ($\theta=0,0067$) se determinaron sobre un panel de 31 genes y 19 líneas endocriadas de girasol. La frecuencia de SNPs en conjunto con un alcance del DL hasta 100 kb ($r^2\sim 0,1$) permitieron diseñar un estudio de asociación a través de la estrategia de genotipificación de genes candidatos. Un total de 30 genes candidatos fueron seleccionados a partir de: (1) el análisis de los perfiles transcripcionales de un genotipo de *Brassica napus* resistente a *S. sclerotiorum*, (2) una colección de transcriptos de expresión diferencial de una línea de girasol moderadamente resistente desafiada con *S. sclerotiorum* y (3) genes descritos en literatura como partícipes en los procesos de defensa frente a PHC. La elección, caracterización y genotipificación de genes candidatos implicó la generación de herramientas para la selección de los mismos y la optimización de técnicas de genotipificación de SNPs a gran escala. La detección de moléculas heterodúplex mediante: (1) corte con la enzima endonucleasa CEL1 (CEL1CH) y (2) cromatografía líquida de alto rendimiento en condiciones de desnaturalización parcial (dHPLC), demostraron ser técnicas robustas y versátiles para aumentar el número de *loci* e individuos a incluir en los estudios de mapeo por asociación. Un total de 16 genes se

genotipificaron exitosamente en la población de mapeo por asociación constituida por 134 accesos del Banco Activo de Girasol de la Estación Experimental Agropecuaria (EEA) INTA Manfredi. El análisis estadístico de asociación incluyó el uso de modelos lineales mixtos que corrigieron los problemas de estructuración y las relaciones de parentesco entre los individuos de la población de mapeo. Un alelo del gen RhoBP_B exhibió una asociación estadísticamente significativa con una menor incidencia de PHC ($p < 0,05$).

Los resultados obtenidos demuestran que es posible encontrar alelos útiles implicados en caracteres complejos a través de los estudios de mapeo por asociación en girasol.

Palabras Claves: SNPs, diversidad nucleotídica, desequilibrio de ligamiento, mapeo por asociación, girasol.

ASSOCIATION MAPPING IN SUNFLOWER: NUCLEOTIDE DIVERSITY, LINKAGE DISEQUILIBRIUM AND IDENTIFICATION OF GENES INVOLVED IN SCLEROTINIA HEAD ROT RESISTANCE

Abstract

Association mapping is a powerful tool to identify gene *loci* that may contribute to phenotypic variation. Single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) have become the markers of choice to conduct these studies.

The aims of this thesis were to study nucleotide diversity and the extent of linkage disequilibrium (LD) in cultivated sunflower to conduct association mapping for the resistance to *Sclerotinia* Head Rot disease.

SNP frequency and nucleotide diversity were assessed using a set of 31 genes and 19 sunflower inbred lines (1SNP/61 pb, $\theta=0.0067$). The relatively high frequency of SNPs found here, along with the predicted extent of LD over distances of 100 kb ($r^2\sim 0.1$) allowed the development of an association mapping strategy through candidate gene approach. A total of 30 genes were selected based on: (1) a microarray analysis from *Brassica napus* genotypes resistant to *S. sclerotiorum*, (2) differentially expressed genes from cDNA libraries from sunflower infected tissues; (3) previous literature reports. Different methods of candidate gene selection and optimization of SNP genotyping technologies for scale-up throughput were explored. CEL1 cleavage of heteroduplex (CEL1CH) and denaturing high performance liquid chromatography (dHPLC) proved to be robust and versatile techniques to increase the amount of *loci* and individuals to be included in the association panel. Sixteen genes were successfully genotyped in the association population composed by 134 inbred lines from INTA Manfredi Sunflower Germplasm Bank. Mixed linear models accounting for population structure and kinship relatedness were used for the statistical analysis of associations. One candidate gene, RhoBP_B, had significant results from association analysis, indicating that it might affect the quantitative variation in sunflower resistance to *Sclerotinia* Head Rot disease. These results demonstrate the

potential of candidate gene association mapping for complex trait dissection in sunflower.

Keywords: SNPs, nucleotide diversity, linkage disequilibrium, association mapping, sunflower.

Agradecimientos

Quisiera expresar mi agradecimiento a todas las Instituciones y personas que hicieron posible que este trabajo de tesis pudiera realizarse:

Al Instituto Nacional de Tecnología Agropecuaria (INTA) por la beca otorgada en el año 2005 para la realización de este trabajo de tesis.

Al Centro Nacional de Investigaciones Agropecuarias (CNIA), al Dr. Osvaldo Rosetti, al Ing. Agr. Raúl Ríos y a la Sra. Claudia Vallejos, por la ayuda y sugerencias recibidas en relación a la Beca otorgada por el INTA.

Al Instituto de Biotecnología y a la Dra. Elisa Carrillo por permitirme realizar este trabajo en el Instituto con libertad y confianza. A toda la gente del Instituto de Biotecnología, por su calidez y su alegría, porque hacen que valga la pena trabajar en un ambiente tan grato. Muchas gracias!

A Ruth Heinz y Esteban Hopp, por abrirme las puertas del laboratorio de Marcadores Moleculares y recibirme con tanta calidez hace ya seis años, por las enseñanzas, los viajes al IB, las charlas de trabajo y de otras cosas, y por ser guías en los primeros pasos en el laboratorio y en el Instituto.

A Norma Paniego, por pensar que tenía la capacidad de afrontar un proyecto nuevo para el grupo, por su predisposición frente a todas mis inquietudes y preocupaciones, por ser tan amable y serena, aún en mis momentos de flaqueza... ¡Muchas Gracias!

A Verónica Lia, por estar ahí siempre cuando te necesité, por compartir tu experiencia y conocimiento, y por enseñarme tantas cosas de la genética y de la vida... Por escucharme siempre y sostenerme ante las adversidades. Por creer en mi capacidad de aprendizaje... ¡y de acción! Por ser una persona en la que sé que siempre voy a poder confiar... ¡Gracias de todo corazón!

Al Banco Activo de Girasol de la EEA INTA Manfredi y al Ing. Agr. Daniel Álvarez, por facilitarme los materiales con los que se realizó este trabajo de tesis.

Al Grupo de Fitopatología de Girasol de la EEA INTA Balcarce, por la colaboración en los ensayos a campo. Especialmente quisiera agradecer a Carolina Troglia y Carla Maringolo por la gran ayuda en la evaluación fenotípica y por el apoyo en la interpretación de estos datos.

A la Unidad de Genómica del Instituto de Biotecnología (UGB) y sus integrantes: Andrea Puebla, Verónica Nishinakamasu y Pablo Vera, por la paciencia y predisposición para secuenciar y genotipificar las miles de muestras que envié a lo largo de este trabajo de tesis. Por ayudarme a resolver problemas y poner a punto nuevas técnicas de genotipificación, por todo lo que corrieron con mis últimos experimentos... Gracias!!!

A M. Valeria Moreno y Verónica Nishinakamasu por la invaluable colaboración en la genotipificación de los SSRs para estructura poblacional. Muchísimas gracias por su dedicación y buena onda!

Al Lic. MSc. Julio di Rienzo, por su asistencia en el análisis estadístico, por desasarme en la comprensión de los resultados... ¡Muchas gracias!

A Lucila Peluffo por los datos de secuencia de la colección de ADNc diferencial para la selección de los candidatos para resistencia...Gracias Lula!!!!

A Jeremías Zubrzycki por la ayuda con la construcción de los mapas de ligamiento...Gracias!!!

A Alberto Maligne por la alegría y buena voluntad en enseñarme a manejar el HPLC, por tratar de encontrar los por qué cuando las cosas no salían... por darme la confianza para poder solucionar los inconvenientes sola! ¡Gracias!

A todos los jefes del Grupo Plantas, por las charlas y consejos, porque siempre es bueno tener la visión de otros! A todos los compañeros del Grupo Plantas por estar atentos a mis avances y por amenizar los días en el laboratorio.

A toda la gente del laboratorio de Marcadores Moleculares y de todo el Grupo Plantas, por la ayuda en los experimentos, protocolos, PCRs...Muchas gracias!

A Carolina Martínez por ser el manual *in vivo* del laboratorio, por nunca decir que no cuando te pedí S.O.S, por la asistencia en la purificación de la CEL1...Gracias!

A todas las personas que me ayudaron técnica y logísticamente para que pudiera llegar a buen término con los experimentos. A los que llenaron cajas de tips a lo loco: Diego, Guille, Vero L., Jere, Seba M., gracias!!!!!!!!!!!!

A todos los que me transportaron hacia y desde el IB... No hubiera podido sin ustedes! A Fer C., Seba A., Lula, Nati, Silvio C, Oscar T., Vero L., Fabi B., Ruth, Lauka, Luigi... Gracias!!! Ari y Flor, gracias por lo incondicional que fueron durante cuatro años para llevarme y traerme del IB, para esperarme cuando me

quedaba dormida y cuando quedaba retrasada por algún experimento... Por todas las charlas matutinas que alegraban los comienzos del día... Gracias por las 8:15 que se extrañan tantoooooo, y los buques que ya no salen más!!!! Los extraño!!!

A todos los que me acompañaron estos años dentro y fuera del laboratorio, vale la pena conocer gente tan copada! a Lula, Vane, Nati, Diego, Guille, Ari... que decir... gracias por bancarme!!!

A Florencia del Viso, por permitirme formar parte del grupo Bromus y así encontrarme con una persona maravillosa, científica y humanamente, que me brindó su amistad más allá de los límites geográficos. Gracias por todas las idas y vueltas al INTA y por tantas charlas, comidas y programas que amenizaron mis soledades... Gracias por estar atenta a cada avance y aportar ideas y correcciones a pesar de los miles de km que nos separan...Se hace difícil agradecer simplemente... Muchas, muchas, muchas gracias amiga!!!!

A todos los que olvidé mencionar, pero que seguramente de una u otra manera me acompañaron en este largo proceso... ¡GRACIAS!

Dedicatorias

*A mis padres Osvaldo y Susana
y a mis hermanas Paula y Marcia,
por acompañarme y alentarme en estos años*

*A Diego,
por creer en mí*

Parte de los resultados presentados en este trabajo de tesis fueron publicados:

- **CM Fusari, VV Lia, HE Hopp, RA Heinz, NB Paniago. (2008) Identification of Single Nucleotide Polymorphisms and analysis of Linkage Disequilibrium in sunflower elite inbred lines using the candidate gene approach. *BMC Plant Biol*, 8:7.**
- **CM Fusari, VV. Lia, V Nishinakamasu, JE Zubrzycki, AF Puebla, AE Maligne, HE Hopp, RA Heinz, NB Paniago. (2010) Single nucleotide polymorphism genotyping by heteroduplex analysis in sunflower (*Helianthus annuus* L.). *Mol Breed*, DOI: 10.1007/s11032-010-9462-9.**
- **Paniago N, Heinz R, Fernández P, Lia V, Fusari C. Bioinformática aplicada a la biotecnología vegetal en: Biotecnología y Mejoramiento Vegetal II. (2010) Levitus, Echenique, Rubinstein, Hopp, Mroginski (ed). INTA. 647pp.**

“...Hay en el campo, entre todas las flores sencillas y las hierbas buenas, esa flor alegre del girasol, que es redonda y amarilla y parece que alumbra el monte...”

Anónimo



Índice

Índice	i
Índice de Figuras y Tablas	iv
Abreviaturas	vii
INTRODUCCIÓN GENERAL	1
El girasol	2
Generalidades	2
Importancia económica	3
El girasol en Argentina	4
Enfermedades que afectan el cultivo	6
Estado del conocimiento del genoma de girasol	8
Mapeo por asociación.....	9
Mapeo de QTL y mapeo por asociación	9
Desequilibrio de ligamiento (DL)	11
Posibles abordajes para el mapeo por asociación	14
Mapeo por asociación a través de la caracterización de marcadores homogéneamente distribuidos en el genoma (<i>genome-wide association, GWA</i>)	14
Mapeo por asociación a través de genes candidatos	15
Modelos estadísticos usados para el análisis de asociación en plantas	15
Antecedentes del grupo de investigación	16
Hipótesis de trabajo	18
Objetivos	18
Objetivo general	18
Objetivos específicos	18
CAPÍTULO 1	20
ESTUDIO DE LA DIVERSIDAD NUCLEOTÍDICA Y EL DESEQUILIBRIO DE LIGAMIENTO EN GIRASOL CULTIVADO	20
Introducción.....	21
Materiales y Métodos	27
Material vegetal	27
Extracción de ADN	27
Cuantificación de ADN	28
Selección de regiones genómicas para secuenciación	29
Búsqueda en colecciones de ADNc	29
Búsqueda bibliográfica	30
Búsqueda <i>in silico</i> en bases de datos de secuencias de ESTs	31
Diseño de iniciadores.....	33
Reacciones de PCR.....	41
Purificación de los productos de PCR y secuenciación automática.....	41
Identificación y análisis de SNPs	41
Estructura poblacional y distancias genéticas	42
Análisis del desequilibrio de ligamiento	44

Resultados	46
Frecuencia de SNPs y diversidad nucleotídica.....	47
Determinación de la estructura poblacional y distancias genéticas	52
Análisis del desequilibrio de ligamiento	56
Discusión	58
Frecuencia de SNPs y diversidad nucleotídica.....	58
Evaluación de la estructura poblacional y distancias genéticas	60
Determinación del desequilibrio de ligamiento	62
 CAPÍTULO 2	 65
 DESARROLLO Y VALIDACIÓN DE METODOLOGÍAS PARA LA GENOTIPIFICACIÓN DE SNPS E INDELS A GRAN ESCALA	 65
Introducción	66
Materiales y Métodos	70
Material vegetal	70
Selección de genes y diseño de iniciadores	70
Reacciones de PCR.....	72
Formación de las moléculas heterodúplex	73
Método CEL1CH	74
Purificación parcial de la enzima CEL1 endonucleasa de apio.....	74
Corte del Heterodúplex con CEL1	75
Electroforesis capilar fluorescente (ECF)	76
Método de dHPLC.....	76
Asignación de genotipos	77
Secuenciación.....	77
Construcción del mapa genético.....	78
Resultados	79
Método CEL1CH	79
Método dHPLC.....	85
Pruebas de desempeño y escalabilidad de las técnicas CEL1CH y dHPLC	90
Caracterización genotípica de accesos del banco de germoplasma mediante CEL1CH y dHPLC	92
Incorporación de marcadores funcionales al mapa de girasol cultivado.....	96
Discusión	97
Método CEL1CH	97
Método dHPLC.....	101
 CAPÍTULO 3	 106
 SELECCIÓN DE GENES CANDIDATOS PARA LA RESISTENCIA A LA PODREDUMBRE HÚMEDA DEL CAPÍTULO EN GIRASOL	 106
Introducción	107
Materiales y Métodos	111
Material vegetal	111
Selección de genes candidatos	111
(1) Selección de genes candidatos derivados de <i>Arabidopsis</i> : identificación de ortólogos mediante reconstrucción filogenética	111

(2) Selección de genes candidatos derivados de una colección de ADNc de expresión diferencial de girasol	114
(3) Selección de genes candidatos descritos en literatura con un rol empírico en los procesos de defensa frente a PHC	115
Diseño de iniciadores	117
Reacciones de PCR	118
Purificación de los productos de PCR, clonado y secuenciación automática	118
Confirmación de la identidad de los genes candidatos derivados de <i>Arabidopsis</i>	119
Identificación y análisis de SNPs y estimación del desequilibrio de ligamiento	119
Resultados	121
Selección de genes candidatos	121
Genes candidatos derivados de <i>Arabidopsis</i> : identificación y obtención de secuencias.....	121
Genes candidatos derivados de una colección de ADNc de expresión diferencial	127
Genes candidatos descritos en literatura con un rol empírico en los procesos de defensa frente a PHC	129
Caracterización de los genes candidatos en el Grupo Núcleo de líneas de girasol.....	130
Discusión	134
CAPÍTULO 4	141
MAPEO POR ASOCIACIÓN PARA LA RESISTENCIA A LA PODREDUMBRE HÚMEDA DEL CAPÍTULO EN GIRASOL	141
Introducción.....	142
Materiales y Métodos	150
Material vegetal	150
Preparación del inóculo.....	150
Evaluación fenotípica	151
Evaluación genotípica	151
Genotipificación de SSRs para análisis de la estructura poblacional	151
Genotipificación de genes candidatos para mapeo por asociación	152
Análisis estadístico.....	154
Construcción del mapa genético.....	156
Resultados	157
Evaluación fenotípica de las líneas analizadas	157
Genotipificación de SSRs para análisis de estructura poblacional.....	157
Genotipificación de genes candidatos para el análisis de asociación	160
Estructura poblacional y análisis de asociación.....	162
Caracterización del gen RhoBP_B	166
Discusión	169
DISCUSIÓN GENERAL Y CONCLUSIONES FINALES.....	176
Discusión general	177
Conclusiones finales y perspectivas	182
BIBLIOGRAFÍA.....	184
APÉNDICE	203

Índice de Figuras y Tablas

INTRODUCCIÓN GENERAL

Figura I.1. Diversidad de formas en los capítulos de la familia Asteraceae.....	3
Figura I.2. Mapas de la superficie sembrada, producción y rendimiento del cultivo de girasol en la República Argentina desde el año 1971 hasta 2005.....	5
Figura I.3. Efectos de las enfermedades causadas por hongos en girasol	6
Figura I.4. Mapeo de ligamiento y mapeo por asociación.....	10
Figura I.5. Relación entre el DL y la resolución de los estudios de asociación	12
Figura I.6. Efecto de la estructura poblacional en el DL y en las asociaciones fenotípicas.....	13
Figura I.7. Posibles abordajes del mapeo por asociación.....	14

CAPÍTULO 1: ESTUDIO DE LA DIVERSIDAD NUCLEOTÍDICA Y EL DESEQUILIBRIO DE LIGAMIENTO EN GIRASOL CULTIVADO

Figura 1.1. Identificación de SNPs <i>in silico</i> a partir de secuencias de ESTs.....	22
Figura 1.2. Escenarios extremos de Desequilibrio y Equilibrio de ligamiento.....	24
Figura 1.3. Eventos de inserción dentro de una secuencia que dan origen a distintas longitudes de indels	48
Figura 1.4. SNPs sinónimos (π_{sin}) y no sinónimos (π_{no-sin}) en las 31 regiones analizadas sobre las 19 líneas de girasol	50
Figura 1.5. Cantidad de haplotipos encontrados y diversidad haplotípica en los 31 genes de girasol analizados	51
Figura 1.6. Análisis de la estructura poblacional de los accesos de girasol caracterizados.....	53
Figura 1.7. Árbol de <i>neighbor-joining</i>	54
Figura 1.8. Análisis del desequilibrio de ligamiento (DL) en girasol.....	57
Tabla 1.1. Antecedentes de estudios de diversidad nucleotídica y alcance del DL en distintas especies vegetales	24
Tabla 1.2. Líneas endocriadas de girasol utilizadas para el desarrollo de SNPs e indels	28
Tabla 1.3. Lista de genes seleccionados para el desarrollo de SNPs e indels en girasol, estrategias de selección e iniciadores diseñados.....	34
Tabla 1.4. Descripción de los polimorfismos totales encontrados para los 31 genes secuenciados en las 19 líneas endocriadas de girasol	47
Tabla 1.5. Estimaciones de diversidad nucleotídica y polimorfismo en 31 genes de girasol	50
Tabla 1.6. Evaluación de los efectos del muestreo en las estimaciones de diversidad.....	52

CAPÍTULO 2: DESARROLLO Y VALIDACIÓN DE METODOLOGÍAS PARA LA GENOTIPIFICACIÓN DE SNPs E INDELS A GRAN ESCALA

Figura 2.1. Fundamento de la detección de SNPs a través de la formación de moléculas heterodúplex Identificación de SNPs <i>in silico</i> a partir de secuencias de ESTs.....	68
Figura 2.2. SDS-PAGE del extracto CJE (CEL1) obtenido a partir de tallos de apio	79
Figura 2.3. Optimización de la cantidad de ADN a utilizar en el ensayo de CEL1CH.....	81
Figura 2.4. Detección de SNPs en genes de girasol mediante la técnica CEL1CH	82

Figura 2.5. Efecto del caudal de la fase móvil en la corrida de dHPLC	86
Figura 2.6. Detección de SNPs en genes de girasol mediante la técnica dHPLC.....	89
Figura 2.7. Individuos heterocigotas detectados con CEL1CH y dHPLC y confirmados por secuenciación directa	94
Figura 2.8. Nuevos haplotipos detectados con dHPLC y confirmados por secuenciación directa	94
Figura 2.9. Incorporación de marcadores funcionales al mapa saturado de girasol RHA266 X PAC2	96
Tabla 2.1. Descripción de los genes evaluados para su uso en la genotipificación de SNPs mediante las técnicas de CEL1CH y dHPLC.....	71
Tabla 2.2. Programa diseñado para la formación de moléculas heterodúplex	74
Tabla 2.3. Puesta a punto del volumen de CEL1 (CJE) necesario para lograr el corte de los heterodúplex	80
Tabla 2.4. Condiciones de temperatura y gradiente de la solución B ensayadas para la detección de heterodúplex mediante dHPLC	87
Tabla 2.5. Genes seleccionados para evaluar el desempeño, la escalabilidad y el rendimiento de las técnicas CEL1CH y dHPLC.....	91
Tabla 2.6. Genotipos de las 22 líneas de girasol identificados a través de CEL1CH y dHPLC.....	93

CAPÍTULO 3: SELECCIÓN DE GENES CANDIDATOS PARA LA RESISTENCIA A LA PODREDUMBRE HÚMEDA DEL CAPÍTULO EN GIRASOL

Figura 3.1. Protocolo de identificación y selección de genes candidatos para la resistencia a la Podredumbre Húmeda del Capítulo (PHC)	120
Figura 3.2. Árboles de MP obtenidos para los genes de <i>Arabidopsis</i> seleccionados	123
Figura 3.3. Amplificación por PCR de genes candidatos para resistencia a PHC.....	129
Tabla 3.1. Genes candidatos seleccionados a partir del análisis de expresión en plantas resistentes de <i>B. napus</i> a las 24 hpi con <i>S. sclerotiorum</i>	114
Tabla 3.2. Genes candidatos seleccionados a partir de una colección de ADNc de expresión diferencial y por antecedentes experimentales de participación en los procesos de defensa frente a PHC	115
Tabla 3.3. Identificación de ortólogos mediante reconstrucción filogenética. Datos generales de las búsquedas	122
Tabla 3.4. Resultados de las búsquedas BLAST de los genes candidatos derivados de <i>Arabidopsis</i>	126
Tabla 3.5. Características de los genes candidatos para resistencia a PHC.....	132

CAPÍTULO 4: MAPEO POR ASOCIACIÓN PARA LA RESISTENCIA A LA PODREDUMBRE HÚMEDA DEL CAPÍTULO EN GIRASOL

Figura 4.1. Posibles abordajes para el mapeo por asociación.....	142
Figura 4.2. Determinación de Haplotipos a partir de sitios SNPs	147
Figura 4.3. Incidencia de PHC en las líneas de girasol que conforman la PMA.....	157
Figura 4.4. Análisis de Coordenadas Principales para las 94 líneas de girasol que conforman la PMA	163
Figura 4.5. Resultado de la inferencia de estructura poblacional utilizando el programa STRUCTURE	164

Figura 4.6. Efecto del genotipo 3:3 en la incidencia de PHC	166
Figura 4.7. Caracterización del gen RhoBP_B.....	167
Figura 4.8. Localización del gen RhoBP_B en el mapa de girasol cultivado	168
Tabla 4.1. Características de los SSRs utilizados para la evaluación de la estructura poblacional e iniciadores utilizados para su genotipificación	152
Tabla 4.2. Datos de los SSRs genotipificados en la PMA	158
Tabla 4.3. Resultado de la genotipificación de los genes candidatos en la PMA	161
Tabla 4.4. Análisis estadístico. Datos de los p-valores obtenidos para la prueba de asociación entre cada gen y la incidencia de PHC tomando en el modelo distintos valores de T	165

APÉNDICE

CAPÍTULO 1

Tabla A1.1. Funciones descritas para genes con similitud a las regiones genómicas amplificadas para la búsqueda de SNPs e indels en girasol	204
--	-----

CAPÍTULO 3

Figura A3.1. Árboles de Máxima Parsimonia (MP) obtenidos para encontrar los genes de girasol ortólogos a los genes de Arabidopsis seleccionados a partir de los datos de expresión del microarreglo realizado por Zhao y col. (2007a)	213
--	-----

Tabla A3.1. Iniciadores utilizados para la amplificación y secuenciación de los genes candidatos para resistencia a la Podredumbre Húmeda del Capítulo (PHC).....	207
--	-----

Tabla A3.2. Identificación utilizada (ID) junto con el número de acceso de GenBank para reconocer el género y la especie de la secuencia utilizada en la construcción de los árboles de Máxima Parsimonia y Máxima Verosimilitud.....	211
--	-----

CAPÍTULO 4

Figura A4.1. Análisis de los resultados del programa STRUCTURE para encontrar las K poblaciones ideales que forman la Población de Mapeo por Asociación (PMA).....	227
---	-----

Figura A4.2. Correlación entre los GL11 para distintos mapas de girasol cultivado	231
--	-----

Tabla A4.1. Material utilizado como Población de Mapeo por Asociación (PMA) para la resistencia a la Podredumbre Húmeda del Capítulo (PHC). Nomenclatura de los accesos según el Banco Activo de Girasol de la EEA INTA Manfredi (BAG-IM).....	219
---	-----

Tabla A4.2. Iniciadores utilizados para la amplificación por PCR y la genotipificación de la Población de Mapeo por Asociación (PMA).....	220
--	-----

Tabla A4.3. Matriz de datos de los ocho SSRs genotipificados para estimar la estructura poblacional y las relaciones de parentesco entre las líneas de la Población de Mapeo por Asociación (PMA)	221
--	-----

Tabla A4.4. Matriz de datos de los 16 genes candidatos genotipificados en la Población de Mapeo por Asociación (PMA).....	224
--	-----

Tabla A4.5. Autovalores del Análisis de Coordenadas Principales (ACP) determinados a partir de la distancia de Rogers (CITA), utilizando la matriz de SSRs para las 94 líneas de la Población de Mapeo por Asociación (PMA).....	227
---	-----

Tabla A4.6. Matriz de pertenencia de cada individuo con respecto a cada número de población (K) tomado como ideal.....	228
---	-----

Abreviaturas

°C	Grados Centígrados
~	Aproximadamente
AcK	Acetato de Potasio
ADN	Ácido desoxirribonucleico
ADNc	ADN copia
AE	Androesterilidad
ARN	Ácido ribonucleico
ARNm	Acido ribonucleico mensajero
BSA	Seroalbúmina bovina o albúmina sérica bovina
Cap.	Capítulo
CEL1	enzima endonucleasa 1 de apio, del inglés <i>Celery 1</i>
CEL1CH	Detección de moléculas heterodúplex con la enzima endonucleasa de apio CEL1
CJE	Extracto crudo de apio conteniendo actividad CEL1, del inglés <i>Celery Juice Extract</i>
col.	Colaboradores
com. pers.	Comunicación personal
cont.	Continuación
CTAB	Cetyl Trimethyl Ammonium Bromide, bromuro de cetiltrimetilamonio
dHPLC	Cromatografía Líquida de alto rendimiento en condiciones de desnaturalización parcial, del inglés: <i>denaturing High Performance Liquid Chromatography</i>
DL o LD	Desequilibrio de Ligamiento o <i>Linkage Disequilibrium</i>
DMSO	Dimetilsulfóxido
DNA	<i>DeoxyriboNucleic Acid</i> , ver ADN
Dr.	Doctor/a
ECF	Electroforesis Capilar Fluorescente
EDTA	Ácido etilen-diamino-tetra-acético, del inglés <i>Ethylene-diamine-tetraacetic Acid</i>
EEUU	Estados Unidos de América

EMS	Metano sulfonato de etilo, del inglés <i>ethylmethanesulfonate</i>
ESTs	Secuencias que se expresan, del inglés <i>Expressed Sequence Tags</i>
EtBr	Bromuro de Etidio
<i>g</i>	gravedad
GN	Grupo Núcleo de líneas de girasol
GWA	Estrategia de mapeo por asociación basada en la caracterización de marcadores homogéneamente distribuidos en el genoma, del inglés <i>genome-wide associations</i>
h	horas
HEPES	Ácido N-2-Hidroxietilpiperacina-N'-2'-Etanesulfónico, del inglés <i>4-(2-hydroxy-ethyl)-1-piperazine-ethane-sulfonic acid</i>
Ing. Agr.	Ingeniero Agrónomo
kb	Kilobases
KCl	Cloruro de Potasio
kDa	kiloDaltons
λ	Longitud de onda
MgCl₂	Cloruro de Magnesio
μM	Micro Molar
μg	Micro gramo
μl	Micro litro
min	Minutos
ml	Mililitro
mM	Mili Molar
MSc.	Magister Scientie
MWCO	Límite de corte de masa molecular, del inglés <i>molecular weight cut off</i>
N°	Número
nm	Nano metro
<i>pag. sig.</i>	Página siguiente
pb	Pares de bases
PCR	Reacción en cadena de la Polimerasa, del inglés <i>Polymerase Chain Reaction</i>

p. ej.	Por ejemplo
PHC	Podredumbre Húmeda del Capítulo
PMA	Población de Mapeo por Asociación
PMSF	fenil-metano-sulfonil-fluoruro, del inglés <i>phenylmethanesulfonyl fluoride</i>
QTL	<i>Loci</i> de caracteres cuantitativos, del inglés <i>Quantitative Trait loci</i>
RAPD	ADN Polimórfico Amplificado al Azar, del inglés <i>Random Amplified Polymorphic DNA</i>
RFLP	Polimorfismo en el Largo de los Fragmentos de Restricción, del inglés <i>Restriction Fragment Length Polymorphism</i>
RIL/s	Línea/s Recombinantes endocriadas, del inglés <i>recombinant inbred line/s</i>
RNA	<i>RiboNucleic Acid</i> , verARN
rpm	Revoluciones por minuto
s	Segundos
SDS- PAGE	Electroforesis en geles desnaturalizantes de poliacrilamida con SDS
SNP	Polimorfismo de Simple Nucleótido, del inglés <i>Single Nucleotide Polymorphism</i>
SSCP	Polimorfismo en la conformación de la simple hebra de ADN, del inglés <i>Single Strand Conformation Polymorphism</i>
SSRs	Microsatélites, del inglés <i>Simple Sequence Repeats</i>
UF	Unidades de Fluorescencia
USD	Dólares estadounidenses
UV	Luz ultravioleta
vs.	Versus

INTRODUCCIÓN GENERAL

El girasol

Generalidades

El girasol (*Helianthus annuus* L.) pertenece a la familia Asteraceae (=Compositae), subfamilia Asteriodae, tribu Helianthae, subtribu *Helianthinae* (Seiler y Rieseberg 1997). La familia Asteraceae es una de las más grandes y diversas dentro de las plantas con flor, conteniendo un 10% de todas las especies de angiospermas. Las compuestas se reconocen por su estructura reproductiva denominada capítulo, donde las flores se disponen en forma sésil sobre un receptáculo ensanchado (Figura I.1) (Katinas y col. 2007). Se estima que la familia Asteraceae surgió en América del Norte entre 100 y 30 millones de años atrás y sufrió una rápida y extensa diversificación (Raven y Axelrod 1974). Actualmente, más de 40 taxones económicamente importantes han sido domesticados a partir de especies silvestres pertenecientes a esta familia, incluyendo cultivos alimenticios (lechuga, achicoria, topinambur), aceiteros (girasol, cártamo), medicinales (manzanilla) y ornamentales (crisantemo, dalia, zinnia y copete).

El género *Helianthus* comprende 36 especies perennes y 12 herbáceas anuales (Schilling y Heiser 1981), entre las cuales se incluyen plantas diploides, tetraploides, y hexaploides, todas con un número cromosómico básico $x = 17$ (Heiser y Smith 1955). Su nombre deriva de las palabras griegas *Helios* que significa sol, y *Anthus* que significa flor; y le fue adjudicado por la característica que tiene la flor de mirar hacia el este hasta el período de antesis.

H. annuus esta compuesto por tres subespecies: *H. annuus* ssp. *macrocarpus* (el girasol cultivado), *H. annuus* ssp. *lenticularis* y *H. annuus* ssp. *annuus*, ambos parientes silvestres del cultivado. El girasol es una especie diploide ($2n = 2x = 34$), con un genoma haploide medio de 3500 Mpb (Baack y col. 2005) y un contenido de ADN (2C) de 7,3 pg en promedio (Jan y Seiler 2007). Su dispersión geográfica es amplia, creciendo en climas templados y subtropicales. Es una planta típicamente alógama y auto-incompatible, aunque su domesticación no sólo permitió la generación de auto-compatibilidad, sino también el aumento del tamaño, vigor y número de semillas por capítulo. Su tiempo generacional es relativamente corto, de 70 a 100 días, y el

rendimiento es de aproximadamente 1000 semillas por capítulo.



Diversas especies silvestres del género *Helianthus* han sido importantes fuentes de genes a lo largo de la domesticación y mejoramiento del girasol cultivado, contribuyendo a la identificación de genes de resistencia a enfermedades, tolerancia a sequía, esterilidad citoplasmática masculina (ECM) y otras características de interés agronómico (Jan y col. 1998). Como ejemplo de ello, los cruzamientos interespecíficos *H. annuus* x *H. petiolaris* permitieron la combinación de la ECM de *H. petiolaris* y de los genes restauradores de la fertilidad de *H. annuus*, dando lugar al desarrollo de la industria de la semilla híbrida de girasol (Fick y Zimmer 1976).

Importancia económica

El girasol es uno de los cultivos oleaginosos más importantes del mundo. La mayor proporción de la producción de girasol se encuentra dedicada a la extracción de aceite (Dorrell y Vick 1997). El aceite de girasol se encuentra en la tercera ubicación

entre los aceites comestibles vegetales más importantes a nivel mundial, con una producción cercana al 13% del total, siendo superado solamente por los aceite de palma y soja.

La semilla de girasol posee un 40% de aceite y un 25% de proteína. A partir de la semilla, se obtiene aceite comestible de alta calidad, con bajo contenido de ácidos grasos saturados (linolénico) y alto contenido de ácidos grasos mono- y di-insaturados (oleico y linoleico). El remanente, luego de la extracción de aceite, es utilizado para alimento de ganado. Otra característica de la semilla es su riqueza en fibras, con un contenido en lisina y un nivel calórico más bajo que la de soja, aunque con un contenido superior de metionina respecto a esta última (Paniego y col. 2007).

Adicionalmente, el girasol es usado en jabones y detergentes en Europa del Este, como transportador en agroquímicos, particularmente pesticidas hidrofóbicos, surfactantes, adhesivos, plásticos, suavizantes y lubricantes. Asimismo, el aceite de girasol es usado en la industria cosmética y tiene un gran potencial como insumo en la producción de biodiesel.

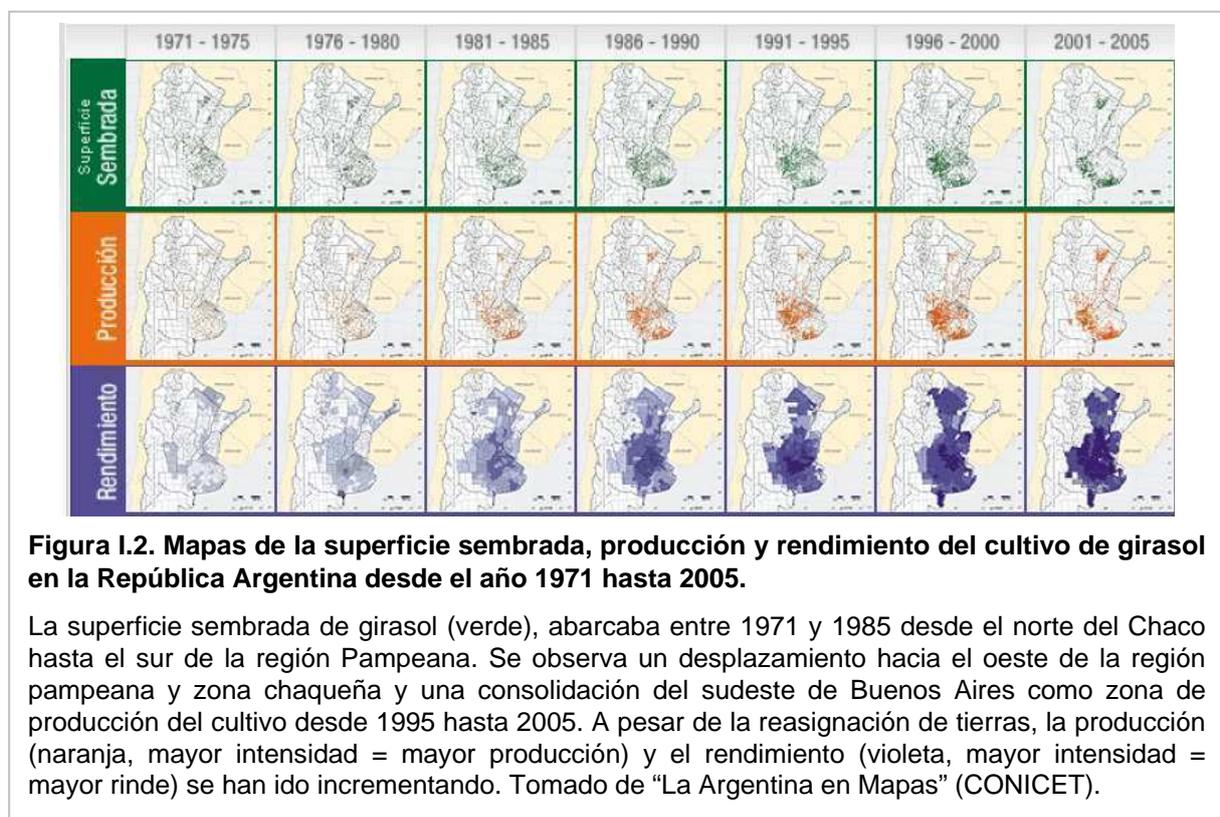
El girasol en Argentina

En la República Argentina, el cultivo se extiende entre los 24 y 38 grados de longitud sur, abarcando una amplia gama de ambientes, desde el norte en la provincia del Chaco hasta el sur de la región Pampeana. En esta zona, el girasol constituye una alternativa importante para los productores (Figura I.2) (Díaz-Zorita y col. 2003).

El cultivo de girasol es el segundo cultivo oleaginoso de importancia dentro del mercado argentino, luego de la soja. Argentina es el tercer productor mundial de girasol, luego de Rusia y Ucrania, y el primer exportador mundial de aceite de girasol, compitiendo en el segmento de aceites de mejor calidad. En la campaña 2007-08, Argentina vendió 1,45 millones de toneladas de aceite (41,4% del total de la producción), seguido por Ucrania (con el 33,4%) y Rusia (con el 12%), sumando el 87% de la oferta total.

Luego de la caída de los precios a fines de los '90, a raíz de la sobreoferta local, la situación del cultivo mejoró entre las campañas 2000-01 y 2005-06 (Figura I.2). Así, en seis años, la producción de girasol pasó de 2,93 a 3,95 millones de toneladas, lo que

implicó un 34,7% de crecimiento (es decir, un 6,2% de crecimiento anual acumulativo). El área sembrada creció un 26,6%, aunque este aumento se produjo sobre zonas de mayor riesgo agro-ecológico o con tierras de menor calidad (superficie sembrada, Figura I.2). Por ello, el rinde sólo aumentó un 6,4% (rendimiento, Figura I.2). Sin embargo, en la campaña 2007-08 se produjeron 4,51 millones de toneladas, casi un millón de toneladas más que en la anterior. En las 2,64 millones de hectáreas (ha) cosechadas (97,8% de las sembradas), el rinde promedio nacional se ubicó en 17,1 quintales por ha (Ingaramo 2005). Este aumento del rendimiento, a pesar de que las zonas de siembra no hayan sido óptimas, se debió a las mejoras tecnológicas introducidas en la producción del cultivo.



La participación de las distintas provincias en el cultivo de girasol ha ido variando desde 1986 hasta 2005 (superficie sembrada, Figura I.2). Por ejemplo, en la provincia de Buenos Aires, entre 2000 y 2005 el área aumentó un 23%. En Córdoba, en cambio, se redujo la superficie en un 68%, mientras que en La Pampa, se incrementó en un 30,5%. El Chaco y Santa Fe fueron las provincias que mayor aporte han hecho al crecimiento de la superficie de siembra. La primera ha pasado de 140.000 a

352.000 ha (152% de aumento) y la segunda, de 117.000 a 193.000 ha (un 64% más) (Figura I.2) (Díaz-Zorita y col. 2003).

Enfermedades que afectan el cultivo

La mayoría de los agentes causantes de enfermedades en girasol son hongos, y en muy pocos casos, bacterias o virus. Actualmente, las principales enfermedades del cultivo en la Argentina son: la Podredumbre Húmeda del Capítulo (PHC) causada por *Sclerotinia sclerotiorum* (Lib.) de Bary, el marchitamiento y secado anticipado, ocasionados por *Verticillium dahliae* Kleb., y el enanismo o *mildew* del girasol causado por *Plasmopara halstedii* (Farl.) Berl. & De Toni (Figura I.3).



La PHC, causada por las esporas de *S. sclerotiorum*, es la segunda enfermedad más importante del girasol en la Argentina, siendo endémica en el sudeste de la provincia de Buenos Aires, principal área de producción de girasol (Pereyra y Bazzalo 1988, Pereyra y Escande 1994). Los síntomas se manifiestan durante el llenado del grano, como lesiones en el receptáculo de consistencia blanda y color té con leche. A medida que transcurre el tiempo, puede afectar a todo el capítulo, provocando la caída total o parcial de éste, y dejando en la parte superior del tallo sólo fibras aisladas en forma de escoba (Figura I.3).

El manejo de PHC resulta difícil debido a varios factores: (1) se presentan con

frecuencia condiciones ambientales muy favorables para el desarrollo de la enfermedad, (2) existe alta persistencia del inóculo en el suelo y (3) *S. sclerotiorum* afecta otros cultivos, como soja y colza, mostrando amplitud en el rango de hospedantes (Trogia 2003). Los daños económicos más importantes por PHC son la pérdida de producción por pudrición de capítulos, la disminución de la calidad por la presencia de cuerpos extraños en los lotes de semilla producida y la acidificación del aceite que se extrae de la semilla de capítulos enfermos (Agüero y col. 2001).

En cuanto al mecanismo de acción de *S. sclerotiorum*, el ácido oxálico y numerosas enzimas poligalacturonasas se consideran los principales factores de patogenicidad del hongo. Se ha propuesto que el ácido oxálico llevaría a un debilitamiento de la planta debido a la disminución del pH del ambiente, facilitando así la invasión del hongo. Adicionalmente, se cree que estaría involucrado en el secuestro de Ca^{2+} de la pared celular, comprometiendo los mecanismos de defensa mediados por este catión y llevando a una necrosis extendida con pérdida de la integridad de la pared celular (Bateman y Beer 1965, Noyes y Hancock 1981). Asimismo, el ácido oxálico también estaría involucrado en la supresión de la maquinaria oxidativa vegetal y en la activación de la muerte celular programada (Hegedus y Rimmer 2005).

Para hacer frente a este patógeno, las respuestas de defensa de las plantas involucran diversos mecanismos, incluyendo defensas anatómicas, componentes antifúngicos preformados y compuestos post-infección (Rodríguez y col. 2004). En genotipos resistentes de varias especies afectadas por *S. sclerotiorum* se ha detectado actividad oxalato oxidasa y/o superóxido dismutasa de proteínas germinas o tipos germinas (GLPs) (p. ej. (Carter y Thornburg 2000, Dumas y col. 1993, Hu y col. 2003)). Adicionalmente, la producción de fitoalexinas y compuestos fenólicos en los capítulos de girasol pareciera correlacionar con la resistencia a la infección de *S. sclerotiorum* (Prats y col. 2007).

Dada la relevancia del cultivo de girasol, y lo devastadoras que pueden resultar las enfermedades para éste, el desarrollo e identificación de cultivares resistentes constituye una estrategia promisorio para evitar la pérdida de producción debido a las enfermedades fúngicas. En el ámbito público-gubernamental, el Grupo de

Fitopatología de Girasol de la Estación Experimental Agropecuaria INTA Balcarce (GFPG, EEA INTA Balcarce) lleva adelante el diagnóstico y monitoreo de enfermedades, estudios epidemiológicos y de patogenicidad, identificación y caracterización de resistencia genética, evaluación de fungicidas y desarrollo de modelos predictivos. En particular, el GFPG realiza la evaluación del comportamiento sanitario de cultivares ante PHC en un trabajo coordinado con la Asociación Argentina de Girasol (ASAGIR) y con el Banco Activo de Girasol de la Estación Experimental Agropecuaria INTA Manfredi (BAG-IM) (Dr. Alberto Escande, com. pers.). El BAG-IM, que funciona desde hace más de 30 años en la EEA INTA Manfredi, alberga cerca de 1200 accesos que provienen de diversos orígenes geográficos y que han sido generados a partir de diferentes metodologías de cruzamiento. Uno de los objetivos principales del BAG-IM es, justamente, contribuir al sostenimiento e incremento de la producción nacional de girasol, disminuyendo la vulnerabilidad genética a las distintas enfermedades, a través de programas de mejoramiento. En el caso de PHC, estos programas centran su objetivo en la identificación de nuevas fuentes de resistencia para incrementar la variabilidad genética del germoplasma en cultivo (Ing. Agr. Daniel Álvarez, com. pers.).

Estado del conocimiento del genoma de girasol

El grado de avance de los conocimientos públicos sobre el genoma de girasol es limitado cuando se lo compara con cultivos como el arroz, el maíz, la soja, el tomate, el trigo, la papa y la cebada, entre otros. Sin embargo, a pesar de que su genoma no está completamente secuenciado, en los últimos 10 años se ha producido un promisorio avance en los estudios genómicos y post-genómicos aplicados al girasol, lo cual ha incrementado significativamente la disponibilidad de herramientas moleculares, tales como microsatélites o SSRs (*Simple Sequence Repeats*) (Paniego y col. 2002, Tang y col. 2002, Tang y col. 2003), ESTs (*Expressed Sequence Tags*) (Fernandez y col. 2003, Gentzbittel y col. 1999), mapas genéticos de distinto nivel de saturación (Al-Chaarani y col. 2004, Lai y col. 2005, Poormohammad Kiani y col. 2007a, Talia 2008, Tang y col. 2002, Tang y col. 2003, Yu y col. 2003), mapas físicos (Talia y col. 2010), colecciones genómicas de clones de BACs (*Bacterial Artificial Chromosome*) (Bouzidi y col. 2006, Feng y col. 2006, Gentzbittel y col. 2002, Özdemir y

col. 2004) y microarreglos de ADN (Alignan y col. 2006, Fernandez y col. 2008, Hewezi y col. 2006, Lai y col. 2006).

Adicionalmente, se han localizado QTL (*Quantitative Trait Loci*) en distintos grupos de ligamiento para numerosos caracteres, entre los cuales podemos mencionar desarrollo, biomasa, ramificación, floración, tamaño de la semilla, composición y concentración de aceite en semilla, tolerancia al déficit hídrico, senescencia, y resistencia a enfermedades, entre otros (Aardse y col. 2006).

Con respecto a PHC, la mayoría de los estudios conducidos en girasol están basados en la determinación de QTL que, en general, explican un bajo porcentaje de la resistencia individualmente. La resistencia a PHC constituye un carácter de base genética compleja y altamente dependiente de las condiciones ambientales. A través de mapeo tradicional en poblaciones biparentales, se han detectado varios *loci* de carácter cuantitativo (QTL) en distintos genotipos, órganos vegetales y condiciones ambientales de crecimiento (Bert y col. 2002, Gentzbittel y col. 1998, Micic y col. 2004, Micic y col. 2005a, Micic y col. 2005b, Yue y col. 2008). En Argentina, se han localizado QTL para resistencia a PHC fundamentalmente en el grupo de ligamiento 10 del mapa genético de girasol, en líneas de buen comportamiento local (Maringolo 2007). Recientemente, se han llevado a cabo estudios básicos de transcriptómica y metabolómica orientados a la caracterización funcional de las respuestas de defensa de la planta ante el ataque de *S. sclerotiorum* (Peluffo y col. 2010a, Peluffo y col. 2010b).

Mapeo por asociación

Mapeo de QTL y mapeo por asociación

La variación fenotípica de muchos de los caracteres de importancia agronómica o evolutiva está determinada por múltiples QTL, la interacción entre éstos, el ambiente y las interacciones QTL-ambiente.

Los análisis de ligamiento en familias o poblaciones biparentales y, más recientemente, los análisis de asociación son las dos estrategias más utilizadas a la hora de disecar caracteres de base genética compleja. Ambos análisis se basan en la

coheredabilidad de los polimorfismos funcionales o variantes alélicas vecinas. La diferencia es que en los análisis de ligamiento tradicionales, son pocos los eventos de recombinación que llegan a producirse, por ejemplo, en los descendientes de un cruzamiento biparental, resultando en un mapeo de baja resolución (Figura I.4a).

En el mapeo por asociación, en cambio, la recombinación y la diversidad genética natural se analizan para lograr mapeo de alta resolución (Figura I.4b) (Zhu y col. 2008). Es decir, esta estrategia tiene el potencial de alcanzar resolución a nivel de genes individuales, dado que hace uso de los numerosos eventos de recombinación ocurridos en el seno de las poblaciones a lo largo de su historia, permitiendo así detectar aquellas formas alélicas involucradas en la variación de un determinado carácter fenotípico (Nordborg y Tavaré 2002). En el caso de las plantas, las poblaciones analizadas pueden ser verdaderas poblaciones naturales, colecciones de germoplasma, líneas o variedades elite de un programa de mejoramiento (Bressegheo y Sorrells 2006a).

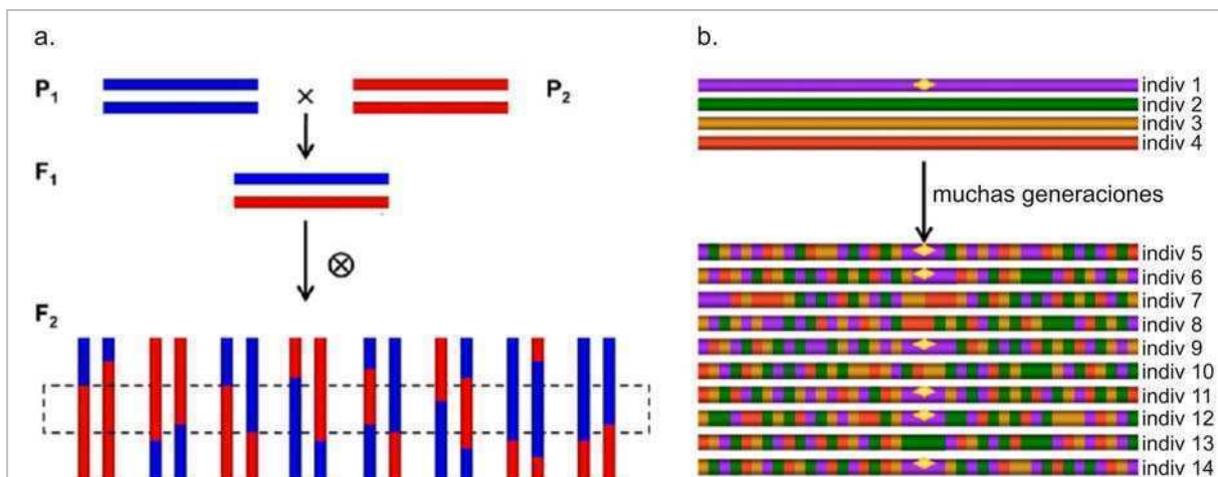


Figura I.4. Mapeo de ligamiento y mapeo por asociación.

Ambos análisis están diseñados para detectar asociaciones no aleatorias entre fenotipos y genotipos, basándose en la coheredabilidad de los polimorfismos funcionales o variantes alélicas vecinas. Sin embargo estos análisis presentan diferencias.

a. El **análisis de ligamiento** en poblaciones biparentales se realiza sobre una población (p. ej. F₂) proveniente del cruzamiento de dos parentales (P₁ y P₂) con fenotipos contrastantes para el carácter en estudio. En este caso, sólo existen pocas oportunidades de recombinación dentro de la familia (segmentos azul y rojo en los individuos F₂), resultando en un mapeo de baja resolución.

b. En el **mapeo por asociación** se intenta detectar la variante alélica responsable de un fenotipo particular (diamante amarillo) sobre una población natural (individuos 1 a 4), algunos portadores del alelo de interés. Gracias a los múltiples eventos de recombinación producidos a lo largo de muchas generaciones (individuos 5 a 14), se logra un mapeo de alta resolución. Tomado de Zhu y col. (2008).

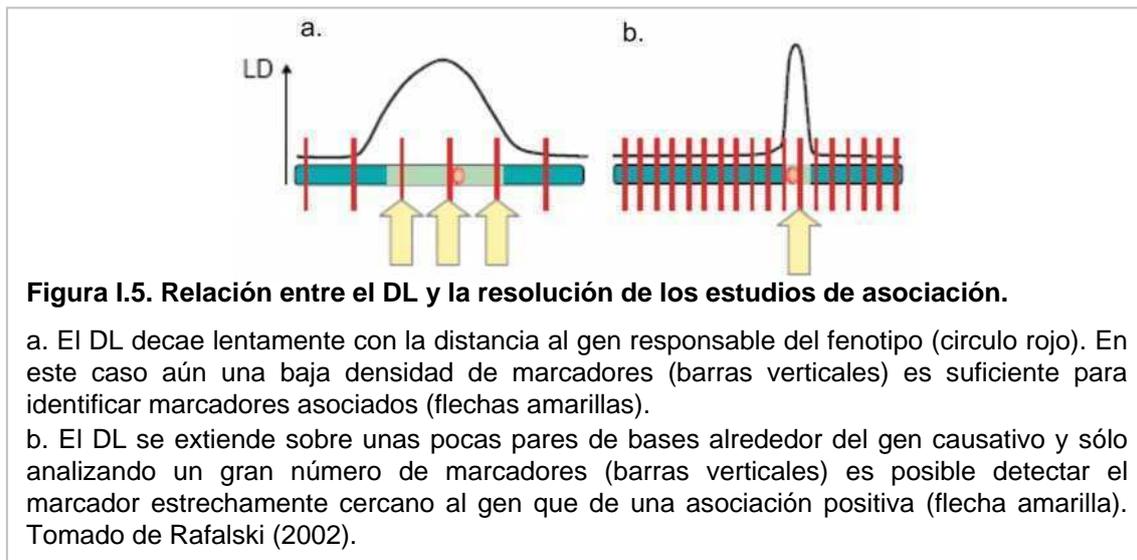
Si bien existen cientos de análisis de QTL desarrollados en numerosas especies vegetales en las dos décadas pasadas (Holland 2007, Kearsley y Farquhar 1998), sólo para un limitado número de casos ha sido posible la identificación de los genes involucrados en el QTL detectado (Price 2006). Los estudios de mapeo por asociación se presentan, entonces, como una herramienta novedosa de gran potencial para el mejoramiento agronómico, cuya aplicación involucra la interacción de disciplinas tales como genómica, genética estadística, genética de poblaciones, biología molecular y bioinformática.

En humanos, esta estrategia ha resultado muy exitosa, llevando a la detección de mutaciones puntuales vinculadas con la predisposición a enfermedades como el Alzheimer y la fibrosis quística (Corder y col. 1994, Kerem y col. 1989). A pesar de que su aplicación en cultivos y otras especies vegetales es mucho más reciente, los resultados encontrados resultan promisorios. Así, el primer caso de mapeo por asociación en plantas fue reportado por Thornsberry y col. (2001), quienes secuenciaron el gen *dwarf8* en 92 líneas elite de maíz, encontrando asociaciones con el tiempo de floración en esta especie. A partir de este trabajo surgieron otros que identificaron genes relacionados con el tiempo de floración en *Arabidopsis thaliana* (*cry*, (Olsen y col. 2004); *fri* (Hagenblad y col. 2004)) y en *Brassica napus* (*col1* (Osterberg y col. 2002)). Asimismo, la demanda de conocimiento sobre los factores que afectan los distintos caracteres de interés agronómico han posibilitado estudios de genética de asociación en numerosas especies de importancia económica como maíz, trigo, sorgo, papa, cebada y raigrás perenne (Agrama y col. 2007, Breseghello y Sorrells 2006b, Casa y col. 2008, Kraakman y col. 2006, Malosetti y col. 2007, Palaisa y col. 2004, Skot y col. 2007, Thornsberry y col. 2001, Wei y col. 2006).

Desequilibrio de ligamiento (DL)

El DL se refiere a asociaciones no aleatorias entre los alelos de distintos *loci* y constituye la base conceptual de cualquier estudio de mapeo por asociación. La distancia sobre la cual el DL persiste va a determinar el número y la densidad de marcadores, así como el diseño experimental que se ajuste mejor para el mapeo por asociación (Figura I.5). Debido a ello, es importante conocer en primera instancia el

DL y determinar el alcance del mismo en la especie de interés.



Las dos medidas comúnmente usadas para cuantificar DL son el coeficiente de disequilibrio estandarizado (D') (Lewontin 1964) y la correlación de frecuencias alélicas al cuadrado (r^2) (Hill y Weir 1994), siendo esta última la más empleada por presentar menor varianza en muestras pequeñas.

El análisis del DL ha sido materia de estudio de numerosos grupos desde hace una década, demostrando que la variación que existe entre las distintas especies, es enorme. Una amplia variedad de mecanismos, que pueden operar simultáneamente, afectan el DL, siendo la mutación, la recombinación, el sistema de apareamiento, la estructura críptica de la población y la distribución de frecuencias alélicas, los factores más importantes (Flint-García y col. 2003).

La **mutación** y la **recombinación** son los fenómenos que tienen mayor impacto. La mutación genera el material de partida para la creación sitios polimórficos que estarán en DL. En cambio, la recombinación es el principal factor que provoca una disminución del DL intra-cromosómico, y la segregación independiente disminuye el DL inter-cromosómico (Flint-García y col. 2003). El **sistema de apareamiento** también determina los niveles de DL, por ejemplo, en especies de reproducción autógena o de auto-polinización el DL tiende a ser mayor que en especies de polinización cruzada (Flint-García y col. 2003).

Una de las restricciones más importantes para la aplicación del mapeo por asociación

en plantas cultivables, es la presencia de **estructuración críptica en la población** debido a factores como la domesticación y el mejoramiento (Rafalski 2010, Thornsberry y col. 2001, Wright y Gaut 2005). El DL aumenta en una población estructurada cuando las frecuencias alélicas para dos *loci* difieren entre las subpoblaciones que la integran, independientemente del ligamiento que exista entre ellos. Un caso extremo de estructuración es el de una población “mezcla” creada a partir de dos poblaciones naturales, provocando la mezcla de cromosomas de diferentes linajes y con diferentes frecuencias alélicas (Figura I.6). Frecuentemente esto provoca que se observe DL significativo aún cuando los *loci* considerados no estén físicamente ligados, pudiendo ubicarse incluso en cromosomas distintos. Sin embargo, este DL desaparece, generalmente, luego de unas pocas generaciones de apareamiento aleatorio (Pritchard y Przeworski 2001). La presencia de estructura en la población utilizada en mapeo por asociación puede llevar a asociaciones espurias, si ésta no es tenida en cuenta en el análisis (Oraguzie y col. 2007) (Figura I.6).

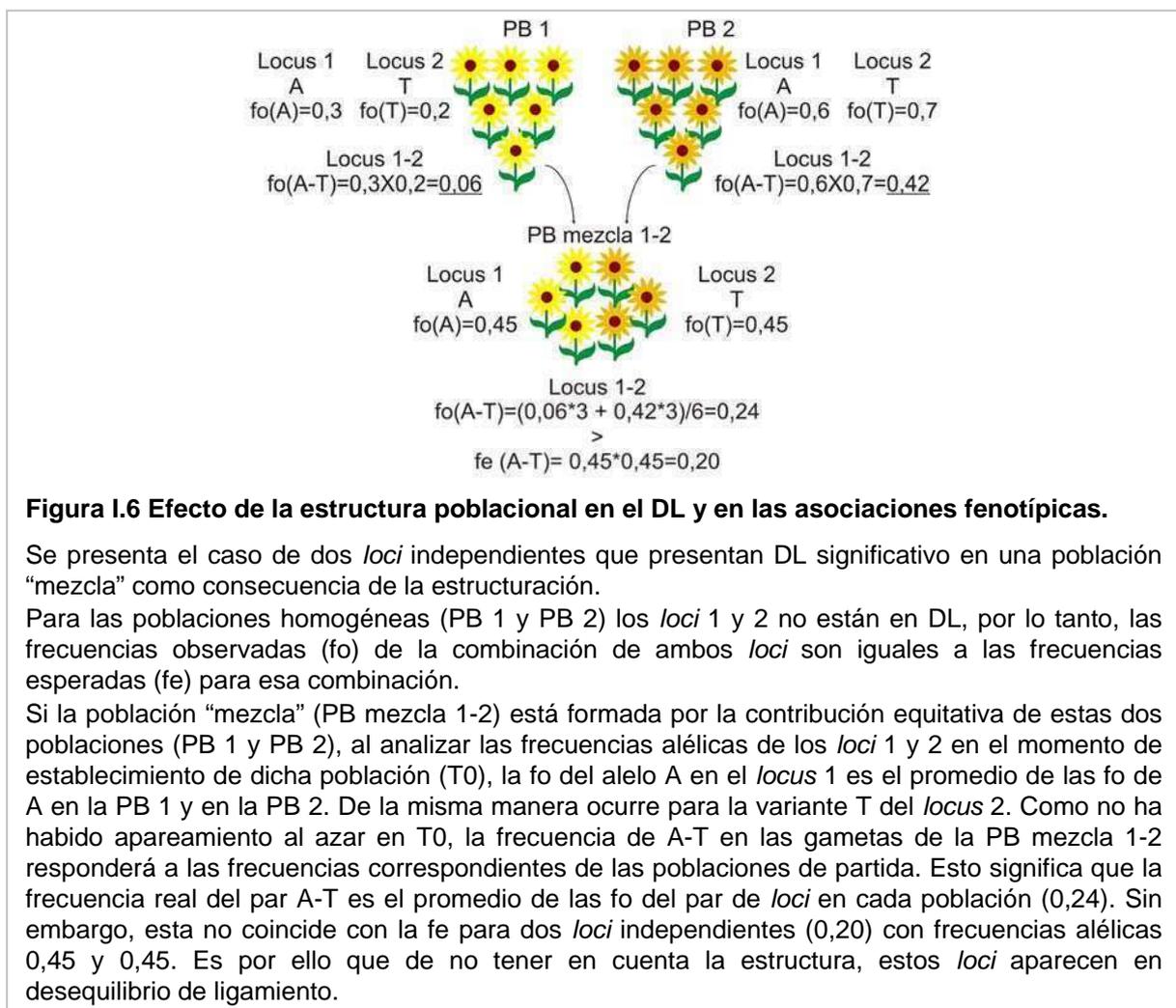


Figura I.6 Efecto de la estructura poblacional en el DL y en las asociaciones fenotípicas.

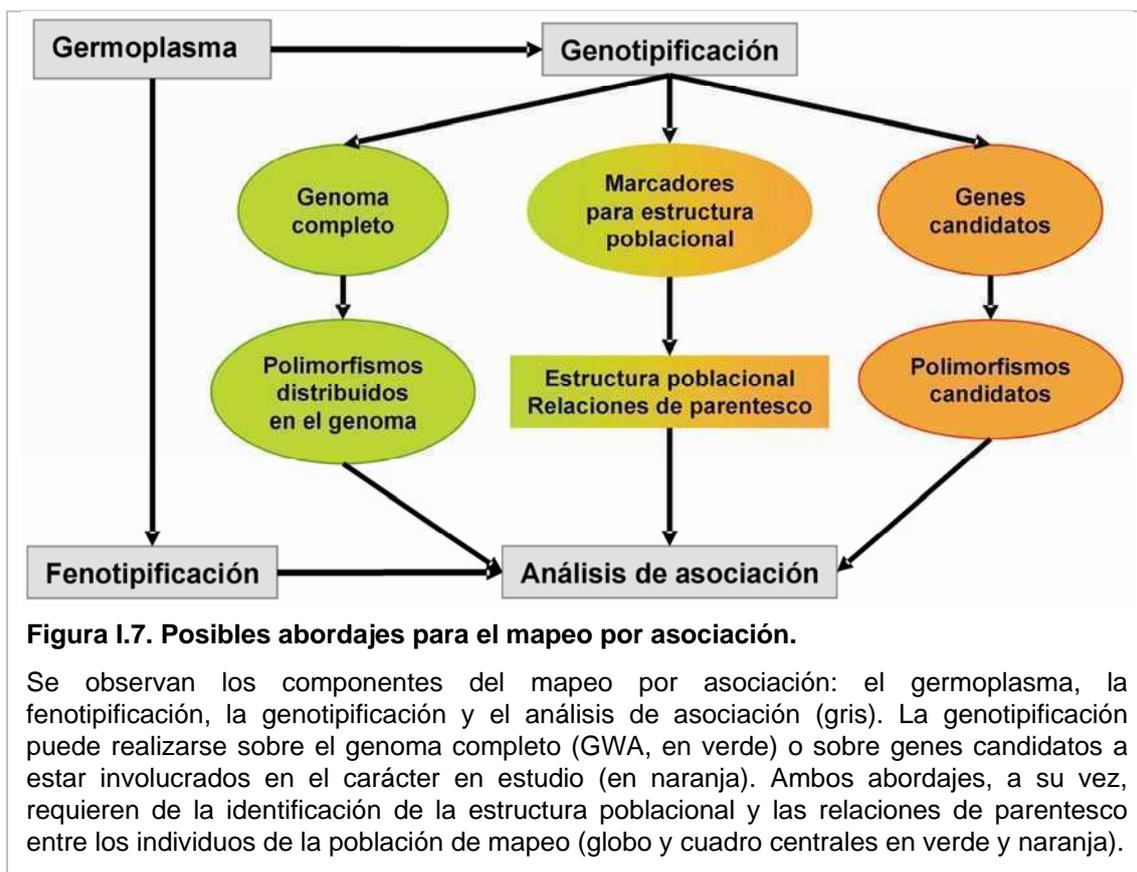
Se presenta el caso de dos *loci* independientes que presentan DL significativo en una población “mezcla” como consecuencia de la estructuración.

Para las poblaciones homogéneas (PB 1 y PB 2) los *loci* 1 y 2 no están en DL, por lo tanto, las frecuencias observadas (f_o) de la combinación de ambos *loci* son iguales a las frecuencias esperadas (f_e) para esa combinación.

Si la población “mezcla” (PB mezcla 1-2) está formada por la contribución equitativa de estas dos poblaciones (PB 1 y PB 2), al analizar las frecuencias alélicas de los *loci* 1 y 2 en el momento de establecimiento de dicha población (T_0), la f_o del alelo A en el *locus* 1 es el promedio de las f_o de A en la PB 1 y en la PB 2. De la misma manera ocurre para la variante T del *locus* 2. Como no ha habido apareamiento al azar en T_0 , la frecuencia de A-T en las gametas de la PB mezcla 1-2 responderá a las frecuencias correspondientes de las poblaciones de partida. Esto significa que la frecuencia real del par A-T es el promedio de las f_o del par de *loci* en cada población (0,24). Sin embargo, esta no coincide con la f_e para dos *loci* independientes (0,20) con frecuencias alélicas 0,45 y 0,45. Es por ello que de no tener en cuenta la estructura, estos *loci* aparecen en desequilibrio de ligamiento.

Posibles abordajes para el mapeo por asociación

El desarrollo de un estudio de mapeo por asociación en plantas involucra la evaluación fenotípica del carácter de interés (p. ej. peso del grano, contenido de aceite, días a floración) y la genotipificación de un conjunto apropiado de marcadores moleculares para cada una de las entidades a comparar (individuos de una población, líneas, variedades, etc.). De acuerdo a la escala y objetivo de cada estudio en particular, el mapeo por asociación puede ser abordado mediante dos estrategias alternativas: a través de la genotipificación del genoma completo o a través de la caracterización de genes candidatos (Figura I.7).



Mapeo por asociación a través de la caracterización de marcadores homogéneamente distribuidos en el genoma (*genome-wide association, GWA*)

Esta modalidad involucra una exploración exhaustiva del genoma para la localización de variación genética causativa. Un gran número de marcadores uniformemente distribuidos en el genoma (10.000 a 100.000) son analizados para encontrar asociaciones con varios caracteres complejos, sin requerir ningún tipo de

información previa acerca del control genético del carácter. En general, y debido a su alto costo, esta estrategia es llevada a cabo por consorcios de investigación, con experiencia complementaria y financiamiento adecuado. (p. ej.: HapMap (*Haplotype Map*) del genoma humano (Frazer y col. 2007)).

Mapeo por asociación a través de genes candidatos

La estrategia de genes candidatos se basa en la selección de genes potencialmente involucrados en el control de la variación fenotípica del carácter en estudio. Por lo tanto, en esta estrategia, el costo se reduce sustancialmente respecto de GWA. Muchos de los estudios publicados involucran la genotipificación de uno a cuatro genes (p. ej.: uno (Thornsberry y col. 2001), dos (Skot y col. 2007), tres (Brenner y col. 2010), cuatro (Haseneyer y col. 2010)), e incluso para aquellos estudios donde la genotipificación se realiza para 20 a 50 genes (p. ej.: 20 (Gonzalez-Martinez y col. 2007), 21 (Krill y col. 2010), 36 (Li y col. 2008), 51 (Ehrenreich y col. 2009)), el costo es significativamente menor que el de genotipificar 10.000 SNPs en GWA.

Dependiendo del alcance del DL, uno u otro abordaje puede ser el más adecuado. De este modo, cuando el DL decae lentamente con la distancia, la estrategia GWA es la recomendada porque la cantidad de marcadores necesarios para cubrir el genoma es menor que si el DL decae rápidamente (Figura I.5 y I.7), sin embargo, en general, este tipo de estrategia alcanza poca resolución. Por el contrario, si la disminución del DL con la distancia es relativamente rápida, la genotipificación de genes candidatos se convierte en la opción más adecuada. A pesar de ello, y debido a las limitantes en los recursos genómicos disponibles para muchos cultivos, la estrategia de genes candidatos es la elegida por la mayor parte de los grupos de investigación (Buckler y Thornsberry 2002).

Modelos estadísticos usados para el análisis de asociación en plantas

El éxito final en el mapeo por asociación dependerá en la posibilidad de separar los eventos de DL reales, de los debidos a otras causas. Es por ello que, dado que la estructura poblacional es una causa importante de detección de asociaciones espurias, la mayoría de los análisis la incorporan.

Los modelos más simples de asociación, los **modelos lineales**, asignan un factor de efecto fijo a los marcadores a evaluar y un factor de efecto aleatorio a la presencia de estructura poblacional (Brescaglio y Sorrells 2006b). Algunos estudios incorporan la información de las poblaciones ideales obtenidas mediante el método bayesiano implementado en el programa STRUCTURE (Pritchard y col. 2000), y la utilizan para asignar genotipos dentro de los análisis de asociación marcador-fenotipo (Remington y col. 2001, Simko y col. 2004b, Sköt y col. 2005). Alternativamente, la información sobre las subpoblaciones ideales obtenidas con STRUCTURE puede integrarse como un factor extra o como un conjunto de co-variables en las relaciones fenotipo-genotipo a través de un **modelo de regresión logística** (Thornsberry y col. 2001, Wilson y col. 2004). Otra forma válida del uso de STRUCTURE es a través de métodos de **análisis multivariado**. En estos casos, una matriz de distancia genética se calcula a partir de la información de marcadores moleculares y se utiliza como fuente para obtener un agrupamiento (Kraakman y col. 2006). Particularmente para colecciones de variedades o líneas elite, la información sobre las relaciones de parentesco entre las líneas suele considerarse y estimarse a partir de la información del *pedigree* o a partir de marcadores neutros y ser incorporada en el análisis como un factor de corrección de las asociaciones (Simko y col. 2004a). Finalmente, varios modelos incorporan ambos datos, tanto de estructura poblacional como las relaciones de parentesco entre individuos, en los denominados **modelos lineales mixtos** de asociación (Malosetti y col. 2007, Stich y Melchinger 2009, Yu y col. 2006).

Antecedentes del grupo de investigación

El presente trabajo de tesis se enmarca dentro de un proyecto orientado a la caracterización funcional de regiones en el genoma de girasol. En esta área, el grupo Genómica de Girasol del Instituto de Biotecnología de INTA Castelar (GG-IB INTA Castelar) ha desarrollado una colección de marcadores neutros SSRs altamente polimórficos (Paniego y col. 2002), un banco de ESTs órgano-específicos aislados de colecciones de ADN copia diferenciales (Fernandez y col. 2003) y un mapa saturado de girasol cultivado (Poormohammad Kiani y col. 2007a, Talia 2008). Los trabajos de cooperación con el GFPG-INTA Balcarce han permitido la identificación de QTL asociados a la tolerancia a PHC en poblaciones F2, obtenidas a partir del cruzamiento

de líneas elite sensibles y tolerantes (Maringolo 2007). Estudios metabólicos y transcripcionales permitieron identificar vías metabólicas y genes potencialmente involucrados con la respuesta a PHC en líneas tolerantes y sensibles (Peluffo y col. 2010b). Adicionalmente, se ha logrado establecer una plataforma bioinformática que integra herramientas dedicadas al procesamiento y análisis sistematizado de secuencias y aglutina los datos producidos por este proyecto y los datos disponibles internacionalmente para la especie.

En este trabajo de tesis se propuso profundizar la caracterización molecular de girasol a través del desarrollo de marcadores moleculares SNPs, con el objetivo de diseñar estrategias de mapeo por asociación para caracteres de interés agronómico. Este estudio involucró: la evaluación de la diversidad nucleotídica y el DL presente en líneas elite de girasol (Capítulo 1), el desarrollo de metodologías de genotipificación de SNPs e indels a gran escala (Capítulo 2) y el diseño de una estrategia de mapeo por asociación en girasol mediante genes candidatos para el carácter resistencia a PHC (Capítulo 3 y Capítulo 4).

Hipótesis de trabajo

La hipótesis que sustenta este trabajo de tesis es que a partir del conocimiento de la diversidad nucleotídica y el desequilibrio de ligamiento en regiones específicas del genoma de girasol es posible establecer asociaciones genotipo-fenotipo para caracteres complejos como la resistencia a hongos, utilizando una estrategia basada en el análisis de genes candidatos en una población no estructurada. Esta estrategia, complementaria a la identificación de QTL en poblaciones biparentales, posibilitará la identificación de variantes alélicas presentes en el germoplasma de girasol cultivado, promoviendo la exploración de la diversidad presente en los Bancos Activos y su uso en los planes mejoramiento para características relacionadas con la resistencia a hongos.

Objetivos

Objetivo general

El presente trabajo de tesis tuvo como objetivo principal el desarrollo de marcadores funcionales (SNPs e indels) y de herramientas para su eficiente genotipificación, a fin de posibilitar la implementación de un estudio de mapeo por asociación en girasol que permita encontrar variantes alélicas útiles para el mejoramiento de caracteres de importancia agronómica en el cultivo.

Objetivos específicos

- Estudiar la diversidad nucleotídica y el desequilibrio de ligamiento en girasol cultivado mediante polimorfismos SNPs e indels identificados en un conjunto de regiones genómicas y de materiales representativos de la diversidad presente en la especie cultivada en Argentina.
- Instalar y validar métodos de genotipificación de SNPs e indels a gran escala que puedan utilizarse en la caracterización eficiente de un gran número de genes y/o individuos.
- Establecer procedimientos y herramientas bioinformática para la identificación de genes candidatos a partir de distintas fuentes de información provenientes de la especie en estudio y/o de especies relacionadas que permitan la selección robusta

de genes que puedan ser analizados en el contexto del mapeo por asociación.

- Realizar un estudio de mapeo por asociación en girasol para el carácter resistencia a la Podredumbre Húmeda del Capítulo, provocada por *S. sclerotiorum*, a fin de encontrar genes o alelos útiles para el mejoramiento asistido del cultivo.

CAPÍTULO 1

ESTUDIO DE LA DIVERSIDAD NUCLEOTÍDICA Y EL DESEQUILIBRIO DE LIGAMIENTO EN GIRASOL CULTIVADO

Introducción

Los marcadores moleculares constituyen una herramienta útil para desarrollo de estudios genéticos y el mejoramiento vegetal. Los diversos sistemas de marcadores moleculares, que se han ido popularizando desde hace tres décadas, pueden ser clasificados principalmente en tres clases. La primera generación de marcadores moleculares incluye los RFLPs (*Restriction Fragment Length Polymorphisms*, polimorfismos en la longitud del fragmento digerido), los RAPDs (*Random Amplified Polymorphic DNAs*, polimorfismos en la amplificación aleatoria de ADN) y sus modificaciones. Luego aparecieron los AFLPs (*Amplified Fragment Length Polymorphisms*, polimorfismos en la longitud del fragmento amplificado) y los microsatélites o SSRs (*Simple Sequence Repeats*, secuencias repetidas simples). Finalmente, la tercera generación de marcadores moleculares, incluye a los SNPs (*Single Nucleotide Polymorphisms*, polimorfismos de nucleótido simple) y los indels (*insertions/deletions*, pequeñas inserciones) (Gupta y col. 2001).

Un SNP se define como una diferencia a nivel de base nucleotídica entre dos secuencias de ADN. En teoría, las variantes de tipo SNP pueden involucrar las cuatro posibilidades determinadas por las bases nucleotídicas (A, T, G ó C), pero en general, sólo dos se observan para un determinado *locus*. Por lo tanto, los SNPs son marcadores bialélicos, cuya información puede ser algo limitada comparada con el polialelismo de los marcadores microsatélites (SSRs). Esto, que en principio es una desventaja, se ve contrarrestado con la gran abundancia y estabilidad de los *loci* SNPs respecto de los *loci* SSRs (Oraguzie y col. 2007).

Dentro de las numerosas metodologías empleadas para encontrar variantes nucleotídicas, la búsqueda electrónica en bases de datos de colecciones de ESTs (*Expressed Sequence Tags*, secuencias que se expresan) y la secuenciación de las regiones candidatas fueron las primeras aproximaciones utilizadas en especies animales y vegetales para el desarrollo de SNPs (Cho y col. 1999, Cordeiro y col. 2006, Kota y col. 2003, Lai y col. 2005, Morales y col. 2004, Somers y col. 2003, Van y col. 2004, Van y col. 2005). Sin embargo, existen pocas herramientas disponibles para el descubrimiento de SNPs *in silico*. La mayoría puede clasificarse en alguno de los

siguientes grupos: (1) análisis manual, (2) soluciones basadas en filtros y (3) soluciones estadísticas (Figura 1.1a). Para todas ellas, como primer paso, las secuencias de ESTs que se asumen corresponden a variantes alélicas de un mismo *locus* se agrupan en *clusters* (grupos) de alta similitud. Una vez que se arman los *clusters*, las secuencias que los forman son alineadas. A partir de estos alineamientos se determina el nucleótido que tiene mayor probabilidad de ocupar cada posición y se establecen las variantes SNPs (Figura 1.1b).

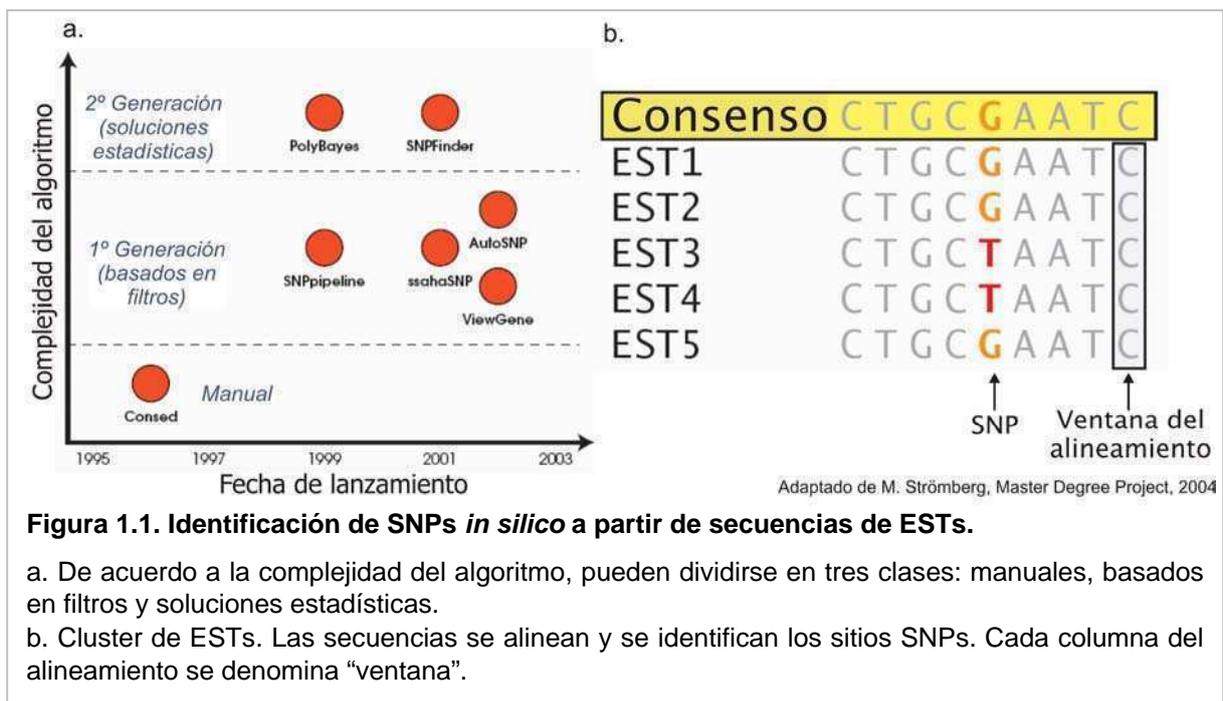


Figura 1.1. Identificación de SNPs *in silico* a partir de secuencias de ESTs.

- a. De acuerdo a la complejidad del algoritmo, pueden dividirse en tres clases: manuales, basados en filtros y soluciones estadísticas.
- b. Cluster de ESTs. Las secuencias se alinean y se identifican los sitios SNPs. Cada columna del alineamiento se denomina "ventana".

En el **análisis manual**, las secuencias de ESTs depositadas en bases de datos se agrupan y alinean, y la determinación de sitios SNPs se realiza por observación directa de los alineamientos. Las soluciones basadas en **filtros** asumen que cada sitio es polimórfico hasta que se ponga a prueba lo contrario. La mayoría de estos filtros rechazan potenciales SNPs en casos dudosos, por ejemplo, si las bases alrededor del supuesto sitio SNP tienen baja calidad o si el pico del cromatograma tiene una altura menor al 15% de la altura de los picos adyacentes. Por último, las **soluciones estadísticas** usan el criterio inverso, asumiendo que todos los sitios son conservados por defecto, hasta que se sobrepase el umbral de probabilidad de que el sitio sea variable. Dentro de las soluciones basadas en filtros (Barker y col. 2003, Kashuk y col. 2002, Ning y col. 2001), **AutoSNP** es el único programa que fundamenta la

determinación del SNP basándose en tres medidas de confianza distintas (Barker y col. 2003). Los SNPs se identifican cuando al menos dos secuencias dentro del alineamiento poseen el alelo de menor frecuencia, haciendo que sean pocos los SNPs encontrados que provengan de errores acumulados durante el proceso de secuenciación.

Los SNPs como marcadores moleculares proveen de información útil para el mapeo genético, el clonado posicional, la detección de caracteres asociados a genes a través del mapeo por asociación y la determinación de relaciones genéticas entre individuos (Syvanen 2001). En los últimos años, se han desarrollado numerosos proyectos de descubrimiento de SNPs a gran escala en especies cultivables, con el fin de realizar análisis de asociación que permitan el mejoramiento genético de los cultivos (Feltus y col. 2006, Giancola y col. 2006, Monna y col. 2006, Rostoks y col. 2005, Song y col. 2009).

El desequilibrio de ligamiento (DL) es una medida de la desviación (D) de las frecuencias alélicas observadas para los alelos en dos *loci*, respecto de las frecuencias esperadas si la combinación de alelos se diera al azar (Oraguzie y col. 2007). Por lo tanto, si D difiere significativamente de 0, se dice que existe DL. Si los *loci* 1 y 2 son ambos bialélicos, es posible la existencia de cuatro haplotipos diferentes (Figura 1.2). En presencia de DL, dos de los cuatro haplotipos estarán sobre-representados y los otro dos sub-representados. En la Figura 1.2 se muestran dos escenarios opuestos, el caso de completo **Desequilibrio de ligamiento** y el caso de **Equilibrio**. Como se mencionó previamente, las medidas más usadas para estimar DL son el coeficiente estandarizado de desequilibrio (D') (Lewontin 1964) y la correlación de la frecuencia alélica al cuadrado para pares de *loci* (r^2).

Existen diversos ejemplos de estudios sobre diversidad nucleotídica y/o DL basados en SNPs, tanto para especies vegetales modelos como para cultivos de importancia económica (Tabla 1.1). La frecuencia de los SNPs varía dentro de cada genoma. Los estudios demuestran que la distribución de los sitios polimórficos no es azarosa a lo largo del genoma o dentro de un gen, es decir, los SNPs ocurren a diferentes frecuencias en regiones codificantes y no codificantes o en distintas regiones

genómicas. Esta distribución irregular, podría deberse a diferencias en la tasa de recombinación, en la densidad génica, en los patrones de transmisión o en las fuerzas selectivas (Oraguzie y col. 2007).

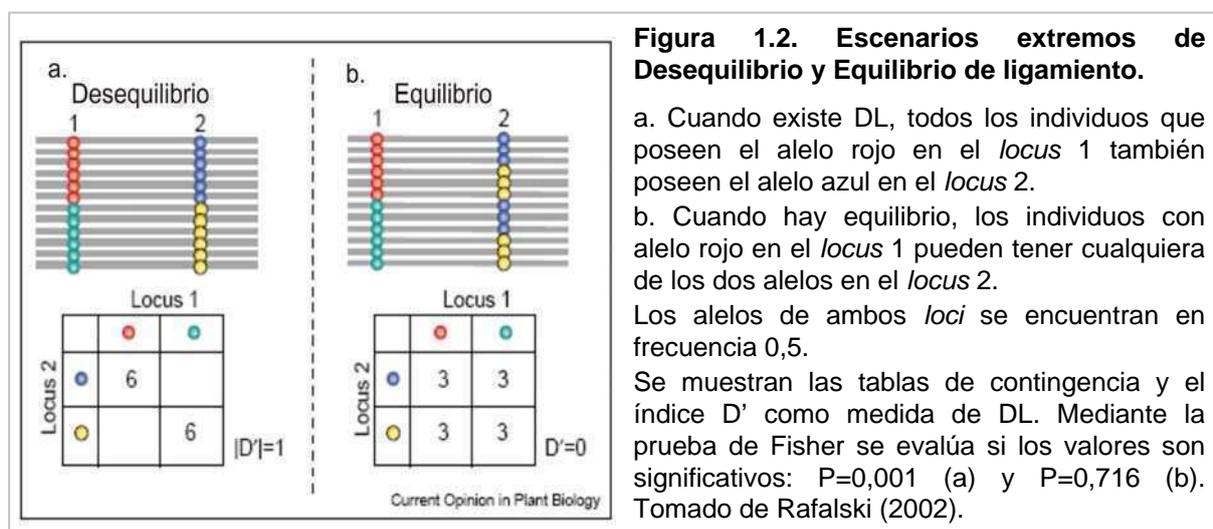


Tabla 1.1. Antecedentes de estudios de diversidad nucleotídica y alcance del DL en distintas especies vegetales.

Especie	Bases analizadas	Genes	Nº individuos	Diversidad	SNPs	Indels	Frecuencia	Alcance del DL	Cita
<i>Boechera stricta</i> y <i>B. holboellii</i>	51 kb	86	46 <i>B. stricta</i> 2 <i>B. holboellii</i>	$\theta = 0,0035$ $\pi = 0,0030$	687	-	1/74	~1 kb	(Song y col. 2009)
<i>Arabidopsis thaliana</i>	~13 kb ^a	13	20	-	83	-	-	250 kb	(Nordborg y col. 2002)
Pino	18 kb	19	32	$\theta = 0,0041$ $\pi = 0,0039$	288	30	1/63	> 2 kb	(Brown y col. 2004)
Álamo	6 kb	5	17-24	$\theta = 0,0167$ $\pi = 0,0111$	384	9	1/60	< 500 b	(Ingvarsson 2005)
Álamo negro	8 kb	9	24	$\theta = 0,0107$ $\pi = 0,0070$	-	-	1/26	300 b	(Chu y col. 2009)
Arroz	888 kb	1117	9	-	7805	-	1/113	-	(Monna y col. 2006)
	72 kb	76	21, <i>O. rufipogon</i>	$\theta = 0,0052$	1364	-	1/52	40 kb	(Mather y col. 2007)
			21, <i>O. sativa (indica)</i>	$\theta = 0,0023$	522	-	1/137	75 kb	
			18, <i>O. sativa (japonica tropical)</i>	$\theta = 0,0012$	250	-	1/288	150 kb	
		22, <i>O. sativa (japonica templada)</i>	$\theta = 0,0010$	219	-	1/328	500 kb		
Maíz	1500 kb	4/6	102/32	-	-	-	-	1500 b	(Remington y col. 2001)
	7 kb	18	32	$\pi = 0,0063$	114	55	1/61	> 600 b	(Ching y col. 2002)
	~3 kb	9	75	$\pi = 0,0078$	110	19	1/27	2 kb	(Palaisa y col. 2004)
Sorgo	29 kb	95	27	$\pi = 0,0025$	238	-	1/122	10 kb	(Hamblin y col. 2004)

Especie	Bases analizadas	Genes	Nº individuos	Diversidad	SNPs	Indels	Frecuencia	Alcance del DL	Cita
Trigo	21,5 kb	45	21	$\theta = 0,0069$	903	-	1/540	-	(Somers y col. 2003)
	21,4 kb	21	26	$\pi = 0,00055$	64	5	1/335	~2 kb	(Ravel y col. 2006)
Cebada	2,8 kb	15	24	$\pi = 0,006$	53	4	1/77	-	(Russell y col. 2004)
Raigrás perenne	37,7 kb	4	6 ^b	-	71	-	-	3 kb	(Ponting y col. 2007)
	11 kb	11	20	$\theta = 0,0306$ $\pi = 0,0314$	1095	332	1/33	500 b	(Xing y col. 2007)
Soja	76 kb	143	25	$\theta = 0,00097$	233	47	1/328	50 kb	(Zhu y col. 2003)
	33 kb	110	9	$\theta = 0,0007$	62	4	1/536	50 kb	(Van y col. 2004)
Girasol	8,2 kb	9	16 silvestres	$\theta = 0,0144$ $\pi = 0,0128$	392	31(total)	1/19	200 b	(Liu y Burke 2006)
			16 cultivados	$\theta = 0,0072$ $\pi = 0,0056$	194	-	1/39	> 700 b	
	49,4 kb	81	10 líneas 2 silvestres	$\theta = 0,0094$ $\pi = 0,0107$	1078	178	1/46	> 5,5 kb	(Kolkman y col. 2007)
Papa	25 kb	66	47	$\theta = 0,0111$ $\pi = 0,0146$	1088	57	1/23	70 kb	(Simko y col. 2006)
Uva	100 kb	230	11	$\pi = 0,0051$	1573	52	1/64	200 b	(Lijavetzky y col. 2007)

a 13 kb alrededor del locus FRI en un contexto de 250 kb

b 6 grupos de poblaciones distintas.

En girasol, hasta hace poco tiempo, los análisis de diversidad genética estaban basados en técnicas tradicionales como alozimas (Rieseberg y Seiler 1990) y SSRs (Al-Chaarani y col. 2004, Liu y col. 2003, Paniago y col. 2002). Recientemente, se publicaron dos trabajos de diversidad nucleotídica en esta especie. En 2006, Liu y col. (2006) analizaron la variación en nueve regiones genómicas sobre 32 accesos silvestres y germoplasma exótico (1 SNP/39 pb, $\theta_W=0,0072$, $\pi_T=0,0056$). Posteriormente, Kolkman y col. (2007) evaluaron la frecuencia de SNPs presentes en 81 *loci* RFLP para 10 líneas endocriadas y dos accesos silvestres (1 SNP/46 pb, $\theta_W=0,0094$, $\pi_T=0,0107$) (Tabla 1.1). Ambos trabajos revelaron valores de diversidad moderados y un alcance del DL en los accesos cultivados y líneas endocriadas similar a los valores alcanzados por especies vegetales típicamente autógamas (Tabla 1.1). Sin embargo, la necesidad de comprender el rango de la diversidad presente en girasol y el origen de la variación genética asociada a la domesticación y el mejoramiento de este cultivo, hace que sea necesario ampliar las investigaciones en esta materia. Particularmente, la elección del germoplasma a evaluar es crucial para descubrir alelos útiles, y para lograrlo, deben incluirse accesos genotípicamente diversos desde el punto de vista genético.

En resumen, el desarrollo de estudios de mapeo por asociación en girasol requiere como paso previo conocer la densidad de marcadores disponibles en el genoma del cultivo y establecer el grado de DL entre ellos. Por ello, el objetivo del presente capítulo fue estudiar la diversidad nucleotídica y el DL en girasol, caracterizando 31 regiones genómicas sobre un panel de 19 líneas endocriadas públicas que pertenecen al núcleo de mejoramiento del cultivo y representan la diversidad presente en los materiales actualmente en vigencia en nuestro país.

Materiales y Métodos

Material vegetal

Las líneas de girasol utilizadas para la caracterización de la diversidad nucleotídica y el desequilibrio de ligamiento se describen en la Tabla 1.2.

Este conjunto de 19 líneas endocriadas fue utilizado como grupo inicial para el desarrollo de SNPs e indels. Las líneas endocriadas son esencialmente homocigotas para todos los *loci*, ya que son obtenidas por cruzamientos con distintas fuentes de germoplasma y luego se produce la fijación de los alelos por autofecundaciones sucesivas. Por este motivo, se analizó solamente un individuo por línea.

Los accesos utilizados en este capítulo son públicos y fueron seleccionados porque representan la diversidad genética presente en los materiales de mejoramiento de INTA. Poseen contribuciones de germoplasma ruso, canadiense, rumano y norteamericano y de especies silvestres como *H. argophyllus* y *H. petiolaris*. Asimismo, poseen características agronómicas y morfológicas distintivas respecto a su comportamiento frente a patógenos fúngicos, estrés abiótico, número de semillas por capítulo y rendimiento de aceite. Dentro de este grupo, 15 accesos se habían utilizado previamente para el desarrollo de 550 SSRs (Paniego y col. 2002). Los cuatro accesos restantes (HA89, RHA801, RHA266 y PAC2) se incluyeron por ser genotipos de referencia y líneas parentales de poblaciones de mapeo altamente caracterizadas (Al-Chaarani y col. 2002, Al-Chaarani y col. 2004, Alfadil y col. 2007, Gentzbittel y col. 1995, Poormohammad Kiani y col. 2007a, Poormohammad Kiani y col. 2007b, Poormohammad Kiani y col. 2008).

Extracción de ADN

Se recolectaron hojas de plantas jóvenes (tres semanas de edad) en nitrógeno líquido, conservándolas a -80°C hasta su liofilización. Una vez liofilizado, se colocaron 20 mg de material vegetal en tubos de 2 ml, junto con dos bolitas de tungsteno y se lo molió con el equipo Tissue Lyser (QIAGEN, Alemania). Para la extracción de ADN se utilizó el sistema Nucleon™ Phytopure™ (GE, Healthcare Life Sciences, Argentina), de acuerdo a las especificaciones del fabricante. El resto del material molido se

conservó a -20°C para futuras extracciones. Se analizó la integridad del ADN extraído mediante electroforesis de ácidos nucleicos en geles nativos al 1% de agarosa, de acuerdo a lo descrito en Sambrook y col. (1989).

Tabla 1.2. Líneas endocriadas de girasol utilizadas para el desarrollo de SNPs e indels.

Línea	Linaje	Localización de la re-selección	Características	Número GRIN
HA52	Germoplasma Rumano ^a	Sudáfrica	Mantenedora de la androesterilidad (AE)	-
HA61	"953-88-3"/"Armavirski 3497"	EEUU	Aceitera, mantenedora AE	PI 599771
HA89	"Vniimk 8931"	EEUU	Resistente a <i>Puccinia helianthii</i>	PI 599773
HA292	"Commander"*3/"Mennonite RR" ^b	EEUU	No aceitera, mantenedora AE	PI 552937
HA303	"Voshod"	EEUU	Aceitera, mantenedora AE	PI 552941
HA369	"Teguá INTA" (Arg. 8018)	Argentina	Susceptible a <i>Plasmopara halstedii</i> (razas 1 a 5). Resistente a <i>Puccinia helianthii</i> (razas 1 a 3)	PI 534655
HA370	"RK-74-198"	Sudáfrica	Aceitera, mantenedora AE	PI 534656
HA821	"HA 300" (derivada de "Peredovik 301")	EEUU	Moderada resistencia a <i>Verticillium</i>	PI 599984
HAR2	"Impira INTA" Selección 5	Argentina	Diferencial internacional de <i>Puccinia helianthii</i>	PI 650753
HAR3	"Charata INTA" ^c selección	Argentina	Aceitera, mantenedora AE	PI 650754
HAR5	"Guayacán INTA" ^d selección	Argentina	Diferencial para <i>Plasmopara halstedii</i> y <i>Puccinia helianthii</i>	PI 650763
KLM280	"KLM" ^e selección	Argentina	Aceitera, mantenedora AE	-
PAC2	<i>H. petiolaris</i> x HA61	Francia	Resistente a <i>Sclerotinia sclerotiorum</i>	-
RHA266	<i>H. annuus</i> silvestre x Peredovik	EEUU	Restauradora resistente a <i>Plasmopara halstedii</i>	PI 599755
RHA274	("cmsPI343765"/"Ha119"/"Ha6 4-4-5")/T66006-2 ^f	EEUU	Restauradora resistente a <i>Plasmopara halstedii</i>	PI 599759
RHA293	"HA155"/"HIR34"/2/"RHA282"	EEUU	No aceitera, restauradora de la fertilidad	PI 599788
RHA374	"Arg-R43"	EEUU	Restauradora de la fertilidad	PI 560142
RHA801	Derivado de restaurador compuesto	EEUU	Restauradora de la fertilidad	PI 599768
V94 ^g	"Mp543"* h./ <i>H. argophyllus</i>	Argentina	Aceitera, mantenedora AE	-

a HA52 es un acceso originado a partir de una selección de germoplasma rumano en Potchestfrom, Transvaal, Sudáfrica.

b Tercera generación del retro-cruzamiento de "Mennonite RR" con "Commander".

c "Charata INTA" se obtuvo a partir de cruzamientos interespecíficos con germoplasma silvestre de las especies *H. annuus subsp. annuus* y *H. petiolaris*.

d "Guayacán INTA" deriva de un cruzamiento entre la variedad argentina "KLEIN" y "CM953-102", y posterior retrocruza con "KLEIN".

e "KLM" es un cruzamiento múltiple entre cultivares KLEIN X "Pergamino" (grupo de variedades locales del programa de mejoramiento de EEA INTA Pergamino que incluye "Guayacán INTA" y "Charata INTA") x "Manfredi" (grupo de variedades locales del programa de mejoramiento EEA INTA Manfredi, que incluye "Impira INTA", "Cordobés INTA" y "Manfredi INTA").

f T66006-2 proviene de Peredovik*2/953-102-1-1-41.

g "V94" es otra selección argentina de cruzamientos entre girasol cultivado ("MP543") y especies silvestres (*H. argophyllus*), "MP543" deriva de "MPRR" (mezcla precoz resistente a roya), la cual deriva de un amplio cruzamiento con especies silvestres de *Helianthus*.

Quantificación de ADN

La cuantificación de ADN se realizó por medio de espectrofotometría de

fluorescencia en placas de 96 pocillos (Fluorómetro Spectra MAX GEMINI EM, Molecular Devices, EEUU), utilizando el reactivo HOECHST 33258 (PolySciences, Cat #09460) que se intercala homogéneamente entre las hebras del ADN, de acuerdo al protocolo que se detalla a continuación. El ADN extraído (2 μ l) se mezcló con 200 μ l de la solución de trabajo TNE-Hoechst (HOECHST 100 ng/ml y TNE 1X: Tris 10 mM, NaCl 200 mM, EDTA 1 mM, pH 7,4). Para determinar las concentraciones de las muestras se construyó una curva de calibración utilizando ADN genómico de timo de cabra de concentraciones conocidas: 22,5; 45; 90; 180; 360 y 720 ng/ μ l. El espectro se ajustó a una longitud de onda de excitación (λ) de 365 nm y una longitud de onda de emisión (λ) de 468 nm. Los datos de los estándares se ajustaron a una recta y las incógnitas se determinaron por extrapolación. En algunas otras oportunidades, el ADN fue cuantificado utilizando el espectrofotómetro Nanodrop® (Thermo Scientific Nanodrop 1000).

Selección de regiones genómicas para secuenciación

Se seleccionaron 64 regiones para la identificación de SNPs, el análisis de la diversidad nucleotídica y el desequilibrio de ligamiento (DL). Las regiones fueron elegidas sobre la base de tres estrategias: (1) búsqueda en las colecciones de ADN copia (ADNc); (2) búsqueda bibliográfica; (3) búsqueda *in silico* en bases de datos de secuencias de ESTs. En la Tabla 1.3 se detallan la estrategia de selección, las regiones incluidas en el análisis, la función inferida mediante búsquedas BLASTx (Altschul y col. 1990) junto con el número de acceso del primer alineamiento, el número de acceso de la secuencia de girasol usada como molde y los oligonucleótidos iniciadores diseñados (de aquí en más iniciadores).

Búsqueda en colecciones de ADNc

El grupo Genómica de Girasol (GG-IB INTA Castelar) cuenta con colecciones de ADNc (=clonotecas) de expresión diferencial de distintos órganos y estadios florales de girasol (Fernandez y col. 2003), que fueron realizadas mediante la técnica de hibridación substractiva (Diatchenko y col. 1996). El análisis de estas clonotecas reveló la presencia de transcritos (ESTs) cuyas funciones, inferidas mediante búsquedas BLASTx, estaban relacionadas con respuestas a estrés biótico. Algunos de

estos ESTs: CP (cisteín-proteasa), LTP (*Lipid Transfer Protein*, proteína de transferencia de lípidos), PGIP1 a PGIP3 (*Poligalacturonase Inhibitor Protein*, proteína inhibidora de las poligalacturonasas) y GLP1 y GLP2 (*Germin-Like Protein*, proteína tipo germina, con probable función de oxalato oxidasa EC 1.2.3.4) fueron inicialmente caracterizados a nivel transcripcional revelando una expresión diferencial en líneas de girasol parcialmente tolerantes y susceptibles a *S. sclerotiorum* (Lic. Lucila Peluffo, com. pers.). El mensajero SFT (*sunflower* fructosiltransferasa) transcribe para la proteína del mismo nombre que intervendría en la síntesis de fructanos, polímeros de fructosa cuya función está relacionada con la protección de las plantas que los acumulan frente al frío o el congelamiento (Valluru y Van den Ende 2008). Finalmente otros mensajeros: GO (glicolato oxidasa, GO, EC 1.1.3.15), LZIP (*Leucine Zipper Protein*, proteína cierre de leucinas) y MCY (multicistatina) fueron elegidos porque la amplificación de estos productos en una prueba preliminar sobre ADN genómico de girasol evidenciaron presencia de regiones no codificantes o intrones.

Búsqueda bibliográfica

Tres factores de transcripción tipo MADS-box (MADSB-TF1, 2 y 3) fueron seleccionados a partir del trabajo de Dezar y col. (2003). El análisis de estas secuencias había revelado similitudes con los transcritos correspondientes a los factores de transcripción APETALA, AGAMOUS y PISTILLATA de *Arabidopsis thaliana*, involucrados en el desarrollo del órgano floral.

Otros cuatro genes (1-ACCO 1 a 4, oxidasas de ácidos carboxílicos) fueron seleccionados a partir de un trabajo de Liu y col. (1997). Estas enzimas están involucradas en la síntesis de etileno y poseen una localización tejido-específica y un perfil de expresión variable, determinado por estrés mecánico. De ese trabajo se seleccionaron las secuencias de girasol depositadas en la base de datos de GenBank con los números de acceso U62553, U62554, U62555 y L29405.

Otro grupo de genes fueron seleccionados a partir de los trabajos de Cellier y col. (1998, 2000), donde se describen los patrones de expresión de proteínas denominadas dehidrinas (HaDhn 1 y 2). Estos trabajos condujeron a la búsqueda de secuencias depositadas en la base de datos de GenBank utilizando como motor de búsqueda la

frase “*drought stress AND Helianthus*” (estrés hídrico y *Helianthus*). De esta búsqueda se seleccionaron las secuencias de girasol con los números de accesos BG734514, BG734516, BG734517, BG734526.

Los genes AALP (*Arabidopsis Aleurain-like protease*, proteína tipo aleuraina de *Arabidopsis*), LIM (proteína con dominio LIM, PLIM1b), ZFP (*Zinc Finger Protein*, proteína dedos de zinc), CNGC2 (*Cyclic Nucleotide Gated Ion Channel*, canal iónico abierto para nucleótidos cíclicos) y XylI (xilosa isomerasa) fueron seleccionados dentro de un conjunto de regiones con similitud de secuencia a transcritos de *Arabidopsis* que se expresan diferencialmente en la etapa de senescencia de acuerdo a lo descrito por Gepstein y col. (2003). Los iniciadores diseñados para amplificar estos transcritos en girasol fueron ensayados sobre ADN genómico, revelando un producto de PCR de mayor tamaño respecto del obtenido con ADNc como molde en la PCR. Por lo tanto, estos genes se seleccionaron no sólo por la función probable en la etapa de senescencia, sino también, por la posibilidad de encontrar SNPs dentro de intrones.

Asimismo, nueve genes analizados por Liu y Burke (2006) fueron incluidos para realizar comparaciones y evaluar discrepancias entre los patrones de polimorfismo y los niveles de diversidad nucleotídica que surgieran debido al muestreo. Estos nueve genes son: CAM (calmodulina), CHS (chalcona sintasa, EC 2.3.1.74), GAPDH (gliceraldehído 3-fosfato deshidrogenasa, EC 1.2.1.12), GIA (*Gibbelleric acid insensitive-like protein*, proteína insensible al ácido giberélico), GPX (glutación peroxidasa, EC 1.11.1.9), GST (glutación S-transferasa, EC 2.5.1.18), PGIC (*Cytosolic PhosphoGlucose Isomerase*, fosfoglucosa isomerasa citosólica, EC 5.3.1.9), SCR1 y 2 (*Scarecrow*, factores de transcripción tipo “espantapájaros” 1 y 2).

Búsqueda *in silico* en bases de datos de secuencias de ESTs

Por otro lado, se realizó un análisis *in silico* de la colección de ESTs de *H. annuus* disponible en el NCBI (hasta junio de 2006) usando una versión independiente del programa AutoSNP, diseñado para detectar polimorfismos SNPs e indels en ESTs (Barker y col. 2003). El programa usa las rutinas d2-cluster (Burke y col. 1999) y cap3 (Huang y Madan 1999) para armar grupos de secuencias de ESTs que alinean entre

sí, y utiliza la redundancia de las variantes encontradas para diferenciar SNPs reales de errores de secuenciación. Los polimorfismos candidatos son identificados cuando ocurren en un determinado número de secuencias del alineamiento. Para cada SNP candidato se calculan tres medidas de confianza: (i) la redundancia del SNP en esa posición, (ii) la co-segregación de ese candidato con otros SNPs en el alineamiento y (iii) la co-segregación pesada, que intenta ajustar los datos faltantes para que no haya un sesgo en los valores de co-segregación (Barker y col. 2003). El programa analizó 55.532 ESTs y los agrupó en 6972 grupos o *contigs*. En primer lugar se descartaron los grupos que no presentaron polimorfismos, quedando por analizar 4650 *contigs* (aproximadamente el 66%). De éstos, se seleccionaron aquellos cuyo alineamiento estaba formado por cinco o más secuencias (798 *contigs*) y con contribuciones de distintas fuentes de germoplasma (415 *contigs*). Otra restricción impuesta fue que tuvieran una función molecular conocida, asociada mediante búsquedas BLASTx, es decir, se descartaron aquellos grupos con similitud a proteínas de función “desconocidas” o “hipotética”. Esta última condición se cumplió sólo para 50 *contigs*. Finalmente, se impuso una restricción adicional: que los grupos seleccionados incluyeran secuencias de ESTs correspondientes a las colecciones de ADNc descritas por Fernández y col. (2003), reduciendo la cantidad de *contigs* a 31. Se diseñaron iniciadores sobre 28 de los 31 *contigs* para amplificar por PCR los genes correspondientes. Estos genes fueron: proteínas ribosomales (RS28, RS11, RL41 y RS16), proteínas del sistema de fotosíntesis (PS II, proteína de unión a clorofila CP26; CPS II, componente del fotosistema II de 22 kD; PSI-F, sub-unidad III del fotosistema I; LHCP, proteína captadora de luz de unión a clorofila; CPSI, sub-unidad V del fotosistema I; PSI-III-CAB, proteína de unión a clorofila tipo III del fotosistema I; CAB, proteína de unión a clorofila), distintas enzimas (PRI, inhibidor de proteasas; CPX, peroxidasa catiónica EC 1.11.1.-; CysP, proteasa de cisteínas EC 3.4.22.-; FNR, reductasa de ferredoxina-NADP EC 1.18.1.2; SAMC, S-adenosilmetionina decarboxilasa EC 4.1.1.50; NPALDP, aldolasa plastídica EC 4.1.2.-; SBP, Seudoheptulosa 1,7-bifosfatasa EC 3.1.3.37 y GCvT, proteína del sistema T de corte de glicinas EC 2.1.2.10), junto con otras proteínas de distintas funciones metabólicas (TIF, factor B04 de inicio de la traducción; ADP-RF, factor de ribosilación de ADP;

ST-LS1, proteína tejido específica inducida por luz; Ac, actina; CHP, probable chaperonina; CCLP, ciclofilina; ANT, translocador nucleotídico; NsLTP proteína de transferencia de lípidos no específica y SEM, subunidad del complejo proteasoma).

Diseño de iniciadores

Los iniciadores empleados para amplificar las regiones candidatas se describen en la Tabla 1.3. El diseño de los iniciadores se realizó con el programa Primer3 (Rozen y Skaletsky 2000), utilizando los parámetros por defecto del mismo. En algunos casos, también se utilizó el programa Vector NTI (Invitrogen, EEUU). La síntesis se realizó mediante un servicio externo (AlphaDNA, Canadá).

Mediante búsquedas BLASTx en el índice de genes de girasol (HaGI) del DFCI (Dana Farber Cancer Institute), se identificó la secuencia que mejor alineaba con cada uno de los genes candidatos seleccionados (E-valor < a 1 e-20). Las secuencias del índice de genes representan un consenso (TC, *tentative consensus*) del ensamblado de los ESTs de girasol depositados en GenBank (Benson y col. 2008). Dichos TC fueron utilizados como molde para el diseño de los iniciadores de los genes seleccionados de bibliografía o a partir de las clonotecas (Tabla 1.3). Para las secuencias pertenecientes a girasol y con un tamaño medio de 500 pb, se diseñaron iniciadores directamente sobre las mismas. Para los genes CAM, CHS, GAPDH, GPX, GST, PGIC, SCR1 y SCR2, los iniciadores se sintetizaron de acuerdo a la información del trabajo de Liu y Burke (2006) o se rediseñaron usando como molde la secuencia disponible en el Proyecto Genoma de Compuestas (http://compgenomics.ucdavis.edu/compositae_contigViewer.php). Los genes identificados a partir del análisis *in silico* fueron amplificados con iniciadores diseñados sobre el EST de mayor longitud dentro del alineamiento.

Para probar la funcionalidad, cada par de iniciadores se empleó en la amplificación por PCR utilizando como molde ADN genómico de dos de los siguiente accesos: RHA266, PAC2, HA89 o RHA801.

Tabla 1.3. Lista de genes seleccionados para el desarrollo de SNPs e indels en girasol, estrategias de selección e iniciadores diseñados.

Estrategia de selección	ID	Similitud (BLASTx)	Nº de acceso (GB)	E-valor	Iniciadores	Identificadores DFCI-TC ^a / Accesos GB ^b / Accesos CGP ^c
Búsqueda en las colecciones de ADNc del GG-IB INTA Castelar	CP	Precursor de cisteín-proteasa (<i>Solanum melongena</i>)	AAD29084	3 e-41	F: 5' GTTGGGAGTTGTTGAGTTGT 3' R: 5' GTGTTCCCAATTGTTTGATT 3'	TC9153 (TC58644)
	SFT	1,2-beta-fructan 1F-fructosiltransferasa (<i>Helianthus tuberosus</i>)	AJ009756	3 e-38	F: 5' CATGGAATGGATTGGTATTC 3' R: 5' CCATACACCCCTCTCTCC 3'	TC14607 (TC47369)
	LTP	Precursor de la proteína de transferencia de lípidos (<i>Davidia involucreta</i>)	AAL27855	2 e-66	F: 5' AGATGTAAAATGCAGAGATTCA 3' R: 5' AAGTAATGGCAGTTGTGAAGA 3'	TC10786 (TC47607)
	PGIP1	Proteína inhibidora de poligalacturonasa (<i>Vitis vinifera</i>)	AAM74142	2 e-42	IPF: 5' CCCTCTTTCTCTAACCCAACG 3' EF421R: 5' GTTCCCAAACGAGTCAGGAA 3' EF421F: 5' ACCATCAGCTGGACCAATCT 3' IPR: 5' CACCAAAGTTATATTCTTTTGTAAATG 3'	BU671906
	PGIP2	Proteína inhibidora de poligalacturonasa (<i>Actidinia deliciosa</i>)	CAA88846	2 e-66	F: 5' AATCTGAACGTAAAGGCAGT 3' R: 5' GGCAAACCTCCATAAACCT 3'	BQ971218
	PGIP3	Precursor de la proteína inhibidora de poligalacturonasas (<i>A. deliciosa</i>)	CAA88846	2 e-66	F: 5' CTTCTGCTCTCAACCAACTC 3' R: 5' GACCAATCGTCAAACATAACT 3'	TC12720 (TC41777)
	GLP1	Proteína tipo germina (<i>Chimonanthus praecox</i>)	ABV03161	8 e-75	F: 5' CAATTCAAAATGTCGATTCA 3' R: 5' AGCTTCAATCCAGAAACAAA 3'	TC13013 (TC44423)
	GLP2	Proteína tipo germina (<i>Oryza sativa</i>)	BAD05739	1 e-41	F: 5' ATCCATGGCGACACCTTTTA 3' R: 5' GGGACAACGTCTTGTGGTT 3'	BU671889

Estrategia de selección	ID	Similitud (BLASTx)	Nº de acceso (GB)	E-valor	Iniciadores	Identificadores DFCI-TC ^a / Accesos GB ^b / Accesos CGP ^c
Búsqueda en las colecciones de ADNc del GG-IB INTA (cont.)	GO	Glicolato oxidasa (<i>Zantedeschia aethiopica</i>)	AAO17067	8 e-95	F: 5' TATGCCAGAAGCTCCAAGAG 3' R: 5' CCACCATTTTTGACATTGAAGA 3'	BU671805
	LZP	Proteína cierre de leucinas (<i>Triticum aestivum</i>)	AAW66004	3 e-40	F: 5' CCAGTAACCGATCTTTTCGG 3' R: 5' AACAAACCCAGTGGTTGCAG 3'	BU671946
	MCY	Multicistatina (<i>H. annuus</i>)	BAA95416	2 e-12	F: 5' TTATGAAGCCAAGGTTTGGG 3' R: 5' CCATTACACAGACACGCGTTA 3'	BU672055
Búsqueda Bibliográfica	MADSB-TF1	Factor de transcripción MADS-box APETALA3 (<i>H. annuus</i>)	AA018231	2 e-115	F: 5' AAGCTAGAGAGAGAGAGATGG 3' R: 5' TGTACATTATGGTCAGCAC 3'	AY173070
	MADSB-TF2	Factor de transcripción MADS-box AGAMOUS (<i>H. annuus</i>)	AAO18228	4 e-127	F: 5' GAGGCTATTGAGCATGATTT 3' R: 5' AGCCATGAACTAAATCCGTA 3'	AY157724
	MADSB-TF3	Factor de transcripción MADS-box PISTILLATA (<i>H. annuus</i>)	AAN47199	7 e-91	F: 5' GGGGAGAGGAAAGATAGAGA 3' R: 5' ACCAATTTGGATACATGACC 3'	AY157725
	1ACCO 1	Oxidasa de ácido carboxílico 1-aminociclopropano (<i>H. annuus</i>)	AAB71421	9 e-94	F: 5' GTGAAAACCTGGGGATTCTTT 3' R: 5' CTCCTTTGCCTGAAACTTTA 3'	U62553
	1ACCO 2	Oxidasa de ácido carboxílico 1-aminociclopropano (ACCO2) (<i>H. annuus</i>)	AAC49824	1 e-158	F: 5' AACTGGGGATTCTTTGAGAT 3' R: 5' TCAAATCTCGGCTCCTTT 3'	U62554
	1ACCO 3	Oxidasa de ácido carboxílico 1-aminociclopropano (<i>Striga hermonthica</i>)	BAC66950	1 e-136	F: 5' ACCACTCAAACACACACACA 3' R: 5' AAAGTAGCTCCAACACAAACA 3'	TC9698 (56869)

Estrategia de selección	ID	Similitud (BLASTx)	Nº de acceso (GB)	E-valor	Iniciadores	Identificadores DFCI-TC ^a / Accesos GB ^b / Accesos CGP ^c
Búsqueda Bibliográfica (cont)	1ACCO 4	Oxidasa de ácido carboxílico 1-aminociclopropano (ACC) (<i>H. annuus</i>)	AAB71421	2 e-166	F: 5' CAAACACACACATAGCAAAAG 3' R: 5' AACTACATGATTCAAAGCTCA 3'	L29405
	LYTB	Precursor de proteína tipo LYTB (<i>Adonis palaestina</i>)	AAG21984	5 e-113	F: 5' CTTATGAAGCAAGGAAACAG 3' R: 5' GCGTCACAGATAGTGTTGAA 3'	TC11271 (TC50721)
	RL44	Proteína ribosomal L44 (<i>Gossypium hirsutum</i>)	AAB08726	3 e-58	F: 5' ATGTCTTAAACCAGCCTGT 3' R: 5' AGACTTACTGCAAGAACAAGGA 3'	TC12956 (TC51721)
	NN	Proteína hipotética (<i>Arabidopsis thaliana</i>)	CAO47819	8 e-118	F: 5' ACGTGACTATGGGATCAAAC 3' R: 5' AACAACTTAAACGGGCATA 3'	TC14260 (TC58017)
	PoIP	Probable retrotransposón (<i>Glycine max</i>)	AAO23078	4 e-44	F: 5' ATTCTCTCGGTGCTGAAAA 3' R: 5' GCATACCCACCTTAGGT 3'	TC10921 (TC49842)
	AALP	Gen asociado a senescencia, proteasa tipo aleuraina (<i>A. thaliana</i>)	AAN31820	2 e-152	F: 5' ATGAATTGAAGCATGCAGTAG 3' R: 5' ATGAGTTCTTTATCAGCCAGTAT 3'	TC9020 (TC47810)
	LIM	Proteína con dominio PLIM1b (<i>H. annuus</i>)	AAD56957	8 e-115	F: 5' TCAAGCCAGAAAGAACATTT 3' R: 5' AAGCTCTGTGGTATGCAGTT 3'	TC11684 (TC40010)
	ZFP	Proteína dedos de zinc (<i>Medicago truncatula</i>)	ABN08252	1 e-18	F: 5' AAACCTCACGCAATGAGACT 3' R: 5' CAAGCTCCCTACAAAGAAA 3'	BU025898
	CNGC2	Canal iónico de nucleótidos cíclicos (<i>G. hirsutum</i>)	AAX18166	2 e-97	F: 5' GTACTCGTTTTTCATCTTCCAAT 3' R: 5' TTGTTACATTGATGCTTTAGACA 3'	BU027548
	CAM	Calmodulina (<i>Morus nigra</i>)	ABH12274	3 e-16	F: 5' TCAGCGCAAATTACCCAAAT 3' R: 5' AGAAACTCTGGGAAGTCGATTG3'	C1037
	CHS	Chalcona sintetasa (<i>Saussurea medusa</i>)	BAA87338	3 e-169	F: 5'TTACGAAAAGTGAGCACAAGAAA 3' R: 5' AATTATGCAAACCGGAAGAA 3'	C1392 & C4492
	GAPDH	Gliceraldehído-3-fosfato deshidrogenasa (<i>Glycine max</i>)	CAA27845	1 e-86	F: 5' GGAAAGGAGACATTCCTCACTTA 3' R: 5' AAATGACTGATGACTGAAGCTCAC 3'	C1499

Estrategia de selección	ID	Similitud (BLASTx)	Nº de acceso (GB)	E-valor	Iniciadores	Identificadores DFCI-TC ^a / Accesos GB ^b / Accesos CGP ^c
Búsqueda Bibliográfica (cont.)	GIA	Proteína del tipo insensible a ácido giberélico (<i>L. sativa</i>)	ABD72473	2 e-59	F: 5' CGGTGATTCGGATACAATCT 3' R: 5' CTGAAACTCACCACCAATTCA 3'	C3138
	GPX	Glutación peroxidasa (<i>M. truncatula</i>)	AAT67997	2 e-60	F: 5' GCATGACTATGTTGGCGACT 3' R: 5' AACTGTCTGAAAGCCCTTGG 3'	C268
	GST	Glutación S-transferasa (<i>P. sativum</i>)	BAC81649	8 e-19	F: 5' TGGAATGTCTCTTTTCTTGGGA 3' R: 5' TGATGGAAGTAGAAGGCTAAGTGA 3'	C5423
	PGIC	Fosfoglucosa isomerasa citosólica (<i>Stephanomeria tenuifolia</i>)	CAI63169	2 e-17	F: 5' GGAATCCTGTCTGCTGATAAGT 3' R: 5' TTCACTTGCGAAACAAGCTC 3'	C1461
	SCR1	Factor de transcripción tipo 1 Scarecrow (<i>Castanea sativa</i>)	AAF79548	6 e-91	F: 5' TTGGAAACGGACTAAACAGTTG 3' R: 5' CGCAACCGAACAATAAACC 3'	C1850
	SCR2	Factor de transcripción tipo 2 Scarecrow (<i>O. sativa</i>)	BAF06685	3 e-49	F: 5' GGTGCCATCTCTTGAACACA 3' R: 5' ATCCAGGCTGCTGCTAATTTTGG 3'	C1874
Búsqueda <i>in silico</i> con AutoSNP	Xyl	Xilosa isomerasa (<i>Hordeum vulgare</i>)	CAA64545	6 e-38	F: 5' AAGAGACCAAATATACAGAAGCA 3' R: 5' TTTTGAATTGCTAGAAAAGAAAG 3'	CD856702
	TIF	Factor de iniciación de la traducción B04 (<i>H. annuus</i>)	AAM77753	5 e-52	F: 5' ATGGTAGGAAAAGCTTGACAACA 3' R: 5' GCATAAAACAAGAACAGAATTCAAGA 3'	BU672029
	ADP-RF	Factor de ribosilación de ADP (<i>Hyacinthus orientalis</i>)	AAT08696	5 e-34	F: 5' AAGAGAGTGAAGGCCAAGCTTAT 3' R: 5' AACATCAAGCTTCAGTGTGTGG 3'	BU671885
	PS II	Proteína de unión a clorofila CP26 en (<i>Brassica juncea</i>)	CAA65042	5 e-88	F: 5' ACAAATACGGTGCTAACTGTGGT 3' R: 5' TCTGTAGTATTCTGCACCACCAA 3'	BU671791
	CPS II	Componente cloroplastídico de 22 kDa del fotosistema II (<i>Nicotiana benthamiana</i>)	ABC59516	9 e-63	F: 5' ACAAAAAGCTGCCCTAAGAAG 3' R: 5' GTGAACCCGAATAATGGACCT 3'	BU671793

Estrategia de selección	ID	Similitud (BLASTx)	Nº de acceso (GB)	E-valor	Iniciadores	Identificadores DFCI-TC ^a / Accesos GB ^b / Accesos CGP ^c
Búsqueda <i>in silico</i> con AutoSNP (cont.)	PSI-F	Precursor cloroplastídico de la subunidad III perteneciente al fotosistema I (complejo captador de luz) (<i>Flaveria trinervia</i>)	AAA33344	3 e-42	F: 5' TGACAACATATGAAAGCAAGGAT 3' R: 5' ACCATTAACCTCCGTAGGGTTTGT 3'	BU672091
	PRI	Inhibidor de proteasas (<i>Zinnia elegans</i>)	BAC10910	3 e-13	F: 5' CGGCTACAATCTGAGAAAGAGAA 3' R: 5' CCAGCTGATACTCATACACCACA 3'	BU672016
	ST-LS1	Proteína tejido específica inducida por luz (<i>S. tuberosum</i>)	CAA28450	2 e-17	F: 5' GGTGGTGCTCTTCTGGTT 3' R: 5' GGTCGGAAGAGCCAAATGT 3'	AJ541167
	RS28	Proteína ribosomal S28 (<i>Prunus persica</i>)	CAA10104	2 e-18	F: 5' TTCGCACCATAACTTTGTGTTCT 3' R: 5' TCCTAGCTTCTATGTTGGAGTCG 3'	BU672107
	Ac	Actina 4 (<i>Torenia fournieri</i>)	BAF68888	1 e-32	F: 5' AGTAAGGAAATTACCGCGTTAGC 3' R: 5' GGTGGTCTTTAAACCATGTGAAA 3'	CD852181
	RS11	Proteína ribosomal S11 de la subunidad 40S (<i>Chenopodium rubrum</i>)	CAK22271	1 e-73	F: 5' TAAGACTCCACGAGAGGCTATTG 3' R: 5' TCAAAAACGACAAGTCAAAAACA 3'	CD853947
	NPALDP	Aldolasa plastídica (<i>N. paniculata</i>)	BAA77604	5 e-84	F: 5' GCTATGAACAAAAGCAGCAGACT 3' R: 5' TCGATACTACGTTTGAGGTTGCTA3'	CD857811
	CPX	Peroxidasa Catiónica (<i>Nelumbo nucifera</i>)	ABN46984	6 e-81	F: 5' CTCCTGGCACTAAAGGTGAGAT 3' R: 5' CATACTACTCAGACTCCATCCA 3'	BQ916123
	CysP	Cisteín proteasa (<i>Aster tripolium</i>)	BAD10859	7 e-102	F: 5' CAGAACTACATAAGGCGAGCAT 3' R: 5' GCAGGTGGTCTACAAAAGAAGA 3'	CD845834
	CHP	Chaperona hipotética (<i>A. thaliana</i>)	BAB01754	1 e-122	F: 5' CACTTAAAGCACCGGGATTTC 3' R: 5' TGTTCTGCTACCTCAACAAGATTT 3'	BQ967621
	FNR	Ferredoxina-NADP reductasa, precursor cloroplastídico (<i>Mesembryanthemum crystallinum</i>)	AAA33029	1 e-80	F: 5' CTCATTCCGCTCATTTATGT 3' R: 5' GAGATCCATCATAACCACCAAAG 3'	BU024980
	CCLP	Ciclofilina (<i>Ricinus communis</i>)	ABV26711	1 e-69	F: 5' ACAACATCCATACCTTCCACAAC 3' R: 5' CTTTACGCGGACACCAACTC 3'	CD848160

Estrategia de selección	ID	Similitud (BLASTx)	Nº de acceso (GB)	E-valor	Iniciadores	Identificadores DFCI-TC ^a / Accesos GB ^b / Accesos CGP ^c
Búsqueda <i>in silico</i> con AutoSNP (cont.)	RL41	Proteína ribosomal 60S L41 (<i>A. thaliana</i>)	-	-	F: 5' GCTGCTAGGGTTCCATAACCTAA 3' R: 5' CCTCTTCAACCTCCTCATTCTCT 3'	CD855544
	ANT	Traslocador de nucleótido adenina, precursor mitocondrial (<i>G. hirsutum</i>)	AAB72047	6 e-33	F: 5' TTGCTGGTGCTTCTTCTCTAGTT 3' R: 5' GGTACAAACCACGGTAGATGAAA 3'	BU671930
	LHCP	Precursor de proteína de unión a centros antenas clorofilas a/b (<i>Lactuca sativa</i>)	BAA03104	1 e-129	F: 5' CAAATCACACACTCAATTCAAGC 3' R: 5' CTGGAGCCCAAATCTTTAGTGAG 3'	CD849726
	RS16	Proteína ribosomal 40S S16 (<i>Euphorbia esula</i>)	AAF34799	9 e-72	F: 5' CACCAAACCTCTGTTCAATGCTTC 3' R: 5' CTCAAAACACTTCAAAACGGAAC 3'	CD854188
	NSLTP	Precursor no específico de proteína transportadora de lípidos (<i>H. annuus</i>)	CAA63340	6 e-27	F: 5' GAAAACAACCCCTGATCGTC 3' R: 5' AGGCATTACAAACCAAGAAACA 3'	BQ915289
	SEM	Probable complejo proteosoma 26 subunidad sem1-2 (<i>H. annuus</i>)	Q9XIR8	1 e-07	F: 5' TTCAAGCTCACAAGTCATAACCA 3' R: 5' TAGGTCAACAGTGGGAAGATGAT 3'	CD848681
	SAMC	S-adenosilmetionina decarboxilasa (<i>Daucus carota</i>)	AAR84406	8 e-19	F: 5' TTGAACCGGGTGAGTTTTCTATT 3' R: 5' AGGCTAAACCACCTCTTAATCCA 3'	BQ912659
	CPSI	Subunidad V del centro de reacción fotosistema I, precursor cloroplastídico (<i>Camellia sinensis</i>)	9S7N7	3 e-21	F: 5' GGGCTAAGGAGTATGTTAGCTTG 3' R: 5' ATGAAATCCAAACCACATAACCA 3'	BU672064
	PSI-III-CAB	Proteína de unión a clorofila a/b del fotosistema I tipo III (<i>A. thaliana</i>)	BAF00366	4 e-109	F: 5' AAACAGACAACCTTTGGTTTGCAT 3' R: 5' CCAACATAGCCAATCTACCATTTC 3'	BU672070

Estrategia de selección	ID	Similitud (BLASTx)	Nº de acceso (GB)	E-valor	Iniciadores	Identificadores DFCI-TC ^a / Accesos GB ^b / Accesos CGP ^c
Búsqueda <i>in silico</i> con AutoSNP (cont.)	GCvT	Proteína de clivaje de glicina sistema T, aminometiltransferasa (<i>Flaveria trinervia</i>)	P49363	4 e-54	F: 5' CAACACATAACCCCTGTGGAG 3' R: 5' TCTTCTTCAAGCACGGACTAAAC 3'	BU671872
	SBP	Sedoheptulosa-1,7-bisfosfatasa, cloroplasto (<i>A. thaliana</i>)	AAK96860	3 e-94	F: 5' ATTGTGAAAGAAAAGGGGGTTT 3' R: 5' AAACAAAGGGTTCGGGTTTTAAG 3'	BQ971804
	CAB	Proteína de unión a clorofila a/b (<i>Beta vulgaris</i>)	CAE30280	2 e-78	F: 5' AGGTTTTTACAAGCATAGGGATA 3' R: 5' ATCATGGTCTCTGGGTCACAC 3'	BQ913496

a DFCI-TC: Consenso Tentativo del índice de genes del DFCI, números de TC al momento de la búsqueda y números de TC actuales entre paréntesis (<http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=sunflower>).

b Número de acceso del GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>).

c Números de acceso de los contigs tomados del Compositae Genome Project (http://compgenomics.ucdavis.edu/compositae_contigViewer.php, Contig Viewer CGP1).

Reacciones de PCR

Para la prueba de funcionalidad de los iniciadores, la reacción de PCR se realizó en un volumen final de 12 µl con 30 ng de ADN genómico, 2 mM de MgCl₂, 0,2 mM de dNTPs, 1 U de Polimerasa Taq Platinum (Invitrogen, EEUU) y 0,25 µM de cada iniciador. Aquellos pares de iniciadores que originaron un único producto de PCR se utilizaron para amplificar las 19 líneas de girasol descritas en la Tabla 1.1, en un volumen de reacción de 25 a 50 µl, con 50 a 100 ng de ADN genómico bajo las mismas condiciones detalladas previamente. La reacción de PCR se realizó en un termociclador Mastercycler *epgradient* S (Eppendorf, Alemania) utilizando el método de disminución progresiva de la temperatura de hibridación (*touchdown*), en un rango de temperaturas desde 65°C a 58°C, con las siguientes características: 2 min a 94°C de desnaturalización inicial, 7 ciclos de (1) 40 seg a 94°C, (2) 40 seg a 65°C (-1°C/ciclo) y (3) 2 min a 72°C; seguido por 30 ciclos de (1) 40 seg a 94°C, (2) 40 seg a 58°C y (3) 2 min a 72°C; con una extensión final de 10 min a 72°C.

Purificación de los productos de PCR y secuenciación automática

Los productos de PCR de los genes se purificaron para su posterior secuenciación tratando 10 µl de la reacción de PCR con 4 µl de EXOSAP-IT (Exonuclease I & Shrimp Alkaline Phosphatase, USB, EEUU). Para los productos obtenidos en bajo rendimiento, los 50 µl de la reacción de PCR se trataron mediante QIAquick PCR Purification Kit (QIAGEN, Alemania) de acuerdo a las especificaciones del fabricante. Aquellos productos de PCR que no pudieron ser secuenciados directamente fueron clonados en el vector pGEMT-easy (Promega Biotech, EEUU) y se secuenciaron al menos dos clones con los iniciadores T7 y SP6 (cuyas secuencias están incluidas en el plásmido) para descartar errores en la amplificación. Las secuencias fueron obtenidas con el analizador automático de ADN ABI 3130xl (Applied Biosystems, EEUU). Cuando la credibilidad de las lecturas fue menor al 98%, se repitió la reacción de secuenciación.

Identificación y análisis de SNPs

Para el estudio general de diversidad nucleotídica, los archivos ABI se alinearon usando el programa ABI Prism SeqScape v2.5 (Applied Biosystems, EEUU). Este

programa provee un procedimiento para la identificación de los SNPs a partir de un organizador (*SeqScape Manager*) que permite abrir, modificar y establecer un proyecto. En primer término se creó un protocolo de análisis (*analysis protocol*) en el cual se ajustaron los parámetros: el tipo de secuenciador de los que provienen los datos (*basecaller*: KB.bpc, *dyset*: KB_3130_POP7BDTv3.mob) y el valor calidad que debe superar la base para asignar una anotación confiable ($N: > 20$), el porcentaje de aceptación en caso de dobles picos en el cromatograma ($< 20\%$), el rango de buena calidad de la secuencia (usando el valor de calidad de la secuencia: cuando cuatro de 20 bases tengan un valor de calidad < 20) y un filtro general que deben cumplir todas las secuencias (bases dobles: $< 30\%$, Ns: $< 30\%$, largo de secuencias: > 50 pb, puntaje de la muestra: > 20). Para ajustar la variabilidad entre las muestras y asegurar la calidad de los resultados, se pueden modificar los parámetros para analizar las muestras en el ajuste de análisis (*analysis settings*). Estos parámetros se usaron para todas las regiones analizadas, aunque se estableció una secuencia de referencia (*reference data group*) distinta para cada una, en donde se incorporó la información de las posiciones y las variantes SNPs a encontrar (si se poseía dicha información). Esta secuencia sirve de molde para que el programa realice los alineamientos correspondientes.

Las regiones codificantes y no codificantes de cada gen se identificaron a través de búsquedas BLASTx (Altschul y col. 1990).

Los niveles de variación genética se estimaron a través de los parámetros de polimorfismo nucleotídico (θ_w , (Watterson 1975)) y de diversidad nucleotídica (π , (Tajima 1983)). El valor θ de Watterson se basa en el número de sitios segregantes, mientras que el valor π de Tajima se basa en las diferencias entre pares de secuencias de la muestra total. Estos parámetros fueron obtenidos usando el paquete de programas DNAsp 4.10.9 (Rozas y Rozas 1999).

Estructura poblacional y distancias genéticas

El análisis de la estructura poblacional se realizó utilizando el método bayesiano implementado en el programa STRUCTURE 2.1 (Pritchard y col. 2000), sobre la base de 136 SNP provenientes de los nueve genes en común con el trabajo de Liu y Burke

(2006). A fin de contar con información adicional para la delimitación de los grupos en el conjunto de las 19 líneas secuenciadas en este capítulo, el análisis de estructura poblacional incluyó, además, los accesos utilizados en el trabajo de Liu y Burke (2006), que corresponden a 16 accesos silvestres (Arizona, wAZ; California, wCA; Colorado wCO; Iowa, wIA; Kansas wKS; México, wMEX; Missouri, wMO; Montana, wMT; Nebraska, wNE; Oklahoma, wOK; Saskatchewan-Canada, wSAS; South Dakota, wSD; Tennessee, wTN; Texas, wTX; Utah, wUT y Wyoming, wWY) y 16 accesos de girasoles primitivos y mejorados (P&M) (Arikara, cARI; Havasupai, cHaVa; Hidatsa, cHID; Hopi, Maíz de Tejas, cMDT; Maíz Negro, cMNe; Mandan, cMan; Sêneca, cSen; cmsHA89, cHA89 Júpiter, cJup; Mennonite, cMEN; Peredovik, cPERE; Sundak, cSUN; VIR 847, cVIR; VK-47, cVK47 y VNIIMK 8931, cVNI).

El método de Pritchard y col. (2000) asume que el conjunto de individuos analizados representa una mezcla proveniente de un número desconocido de poblaciones (K), y que el genoma de cada individuo puede estar conformado para la contribución de una o más de ellas. Se asume además que los *loci* son independientes y que existe panmixia dentro de cada una de las K poblaciones. El número de grupos evaluados osciló entre uno y 10. El análisis se realizó utilizando cinco réplicas por cada valor K, un largo del período de *burn-in* de 200.000 iteraciones, y un largo de la corrida de 10^5 iteraciones. No se utilizó ninguna información adicional sobre el origen de los individuos para definir los grupos. Las frecuencias alélicas se mantuvieron independientes entre los grupos, para evitar sobre-estimaciones en el número de éstos (Falush y col. 2003). La corrida que mostró la mayor probabilidad posterior de los datos se consideró para cada valor K.

Para evaluar la diversidad presente entre los accesos aquí analizados y los correspondientes al trabajo de Liu y Burke (2006), se realizó un árbol de *neighbor-joining* sobre la base de la proporción de alelos compartidos, tomando como alelos las combinaciones de SNPs observadas dentro de cada gen o haplotipos encontrados. Para ello se identificaron los distintos haplotipos de los genes CAM, CHS, GAPDH, GIA, GPX, GST, PGIC, SCR1, SCR2 para las 19 líneas elite y los 32 accesos mencionados anteriormente (números de acceso de GenBank DQ503586-DQ504161). La distancia de alelos compartidos se calculó con el programa Microsat (Minch y col.

1995-1996). El árbol se construyó mediante el algoritmo de *neighbor-joining* con el programa Phylip v3.6 (Felsenstein 1991-2004). Para la obtención de las medidas de soporte de los grupos se realizó un re-muestreo de la matriz por el método de *bootstrapping* (10.000 pseudoréplicas) implementado en el mismo programa.

Análisis del desequilibrio de ligamiento

Como se mencionó en la introducción de éste capítulo, el DL es una medida de la desviación de las frecuencias alélicas observadas para los alelos en dos *loci*, respecto de las frecuencias esperadas por azar. Esta desviación se mide como $D_{ab} = (\pi_{AB} - \pi_A\pi_B)$, donde π_{AB} es la frecuencia haplotípica y π_A , π_B , π_a y π_b son las frecuencias de los alelos A y a, B y b para los *loci* 1 y 2, respectivamente.

Una medida alternativa es D' (Lewontin 1964):

$$|D'| = \frac{(D_{ab})^2}{\min(\pi_A\pi_b, \pi_a\pi_B)} \text{ for } D_{ab} < 0;$$

$$|D'| = \frac{(D_{ab})^2}{\min(\pi_A\pi_B, \pi_a\pi_b)} \text{ for } D_{ab} > 0.$$

El coeficiente D' está basado en las frecuencias alélicas observadas de manera de tomar valores extremos de 0 y 1, aún si las frecuencias difieren entre ambos *loci* (Flint-García y col. 2003).

La otra medida es r^2 :

$$r^2 = \frac{(D_{ab})^2}{\pi_A\pi_a\pi_B\pi_b}.$$

Resulta conveniente calcular el cuadrado del coeficiente de correlación entre los dos *loci*, sin embargo, a menos que ambos *loci* se encuentren con frecuencias alélicas iguales, r^2 nunca será igual a 1.

Si bien ninguna de las dos medidas (r^2 , D') se comporta lo suficientemente bien con un tamaño de muestra pequeño y/o frecuencias alélicas bajas, r^2 presenta ventajas respecto a D' . Mientras que r^2 resume la historia de recombinación y mutación a la vez, D' mide únicamente la historia de recombinación y es afectado por muestras pequeñas, resultando en un comportamiento errático cuando se comparan *loci* con frecuencias alélicas bajas (Flint-García y col. 2003). Por lo tanto no se consideró el

valor de D' para el análisis. La medida de DL se obtuvo mediante el estadístico r^2 obtenido con DNAsp 4.10.9 (Rozas y Rozas 1999). Los niveles de significación fueron corregidos según el procedimiento de Bonferroni para comparaciones múltiples con un α global del 5%. Para graficar r^2 (a lo largo de todos los *loci*) se utilizó el programa Microcal TM Origin R v7.5 (Microcal Software, Inc., EEUU)

Resultados

Para la identificación de SNPs y el análisis de la diversidad nucleotídica se seleccionaron un total de 64 regiones genómicas (Tabla 1.3). Los iniciadores diseñados amplificaron un único producto de PCR del tamaño esperado en 40 de las 64 regiones seleccionadas (65,50%). Nueve regiones no pudieron obtenerse mediante PCR en la concentración y el volumen requeridos para la posterior reacción de secuenciación y, por lo tanto, se descartaron. Para 31 regiones (48,4%), se obtuvieron datos de secuencia de alta calidad, para las cuales pudieron estimarse la longitud de las regiones codificantes y no codificantes y los polimorfismos encontrados (Tabla 1.4). Las funciones descritas para genes con similitud a las regiones amplificadas se detallan en la Tabla A1.1 del Apéndice.

Cinco de los 10 genes seleccionados a partir de las colecciones de ADNc (Fernandez y col. 2003) pudieron ser amplificados (GO (EC 1.1.3.15), PGIP1, PGIP3 LZP y GLP). Asimismo, el factor de transcripción MADS-Box con similitud a PISTILLATA (MADSB-TF3), tres genes asociados a senescencia LIM, AALP y ZFP, y una de las enzimas 1-ACCO (1-ACCO1), que habían sido seleccionados a partir de búsqueda bibliográfica, se amplificaron exitosamente en todas las líneas de girasol descritas en la Tabla 1.1. Además, los nueve genes estudiados por Liu y Burke (Liu y Burke 2006) también pudieron amplificarse y secuenciarse sin inconvenientes (CAM, CHS, GAPDH, PGIC, GAI, GPX, GST, SCR1 y SCR2). Por último, 12 de las 31 regiones seleccionadas mediante la búsqueda *in silico* pudieron amplificarse y secuenciarse para su validación experimental. Estas secuencias incluyeron: las proteínas ribosomales L41 y S16 (RL41, RS16); enzimas como SAMC, SBP y GCvT; una subunidad del Proteasoma (SEM); proteínas de unión a clorofila a/b (CAB, PSI-III-CAB y LHCP), un precursor cloroplastídico del fotosistema I (CPSI), una proteína relacionada con patogénesis (NsLTP) y un transportador de nucleótidos (ANT). (Tabla 1.4 y Tabla A1.1).

Tabla 1.4. Descripción de los polimorfismos totales encontrados para los 31 genes secuenciados en las 19 líneas endocriadas de girasol.

Estrategia de selección	Gen ^a	SNP _T ^b	Nº Indels ^c	Longitud total (pb) ^d	Región Codificante (pb) ^d	Región no-codificante (pb) ^d
Análisis de las colecciones de ADNc del GG-IB INTA Castelar	GO	2	1 (36)	608	300	308
	PGIP1	34	5 (14)	1551	936	615
	PGIP3	3	0 (0)	676	561	115
	LZP	0	1 (8)	425	84	341
	GLP	0	3 (3)	876	648	228
Búsqueda bibliográfica	MADSB-TF3	13	11 (20)	1082	291	791
	AALP	10	2 (11)	269	189	80
	LIM	6	2 (5)	319	150	169
	1-ACCO	20	6 (20)	764	451	313
	ZFP	22	7 (51)	636	103	530
	CAM	29	6 (93)	538	117	421
	CHS	0	0 (0)	1051	978	73
	GAPDH	2	2 (3)	782	617	165
	GIA	2	1 (1)	749	504	245
	GPX	0	1 (6)	744	513	231
	GST	40	0 (0)	561	351	210
	PGIC	15	2 (4)	569	219	350
	SCR1	3	0 (0)	739	732	7
	SCR2	7	0 (0)	504	504	0
Búsqueda <i>in silico</i> con AutoSNP	RL41	3	0 (0)	100	66	34
	ANT	9	0 (0)	216	213	3
	RS16	7	0 (0)	448	405	43
	NsLTP	7	2 (13)	294	96	198
	SEM	3	0 (0)	226	87	139
	SAMC	12	1 (3)	369	189	180
	GCvT	3	0 (0)	183	180	3
	SBP	11	0 (0)	243	240	3
	LHCP	8	0 (0)	362	348	14
	CPSI	4	0 (0)	168	144	24
	PSI-III-CAB	1	1 (1)	710	387	323
	CAB	7	2 (10)	537	393	144
	Total		283	60 (302)	17.293	10.991
Promedio/locus		9,13	1,94			
Frecuencia		1/61	1/288			

a Las regiones codificantes de los genes se determinaron mediante búsquedas con BLASTx.

b SNPs totales.

c Número de indels contados como bloques de inserciones (pb de los indels).

d El largo total de las regiones codificantes y no codificantes se indican excluyendo los indels.

Frecuencia de SNPs y diversidad nucleotídica

Las secuencias obtenidas para cada uno de los 19 individuos analizados fueron depositadas en la base de datos del GenBank (<http://www.ncbi.nlm.nih.gov/genbank>) con los números de acceso EU112455 a EU113043. Se analizaron 17.293 pb de secuencia alineada por individuo. El tamaño de las regiones osciló entre 100 y 1565 pb incluyendo indels. Al observar la Tabla 1.4 se evidencia la ocurrencia de al menos un SNP en 27 de los 31 genes evaluados, con un total de 283 cambios

nucleotídicos identificados en el conjunto de genes e individuos analizados. En promedio, se detectó un SNP cada 61 pb (excluyendo los indels) y 9,13 SNPs por región. El número de sustituciones sinónimas encontrado fue de 99, casi cuatro veces más abundantes que las sustituciones no-sinónimas (28). Del total de sustituciones, un 64,66% fueron transiciones. El número de SNPs varió también entre las regiones codificantes y no codificantes, hallando 127 SNPs en 10.991 pb de zona codificante, en contraposición a los 156 SNPs encontrados en 6302 pb de secuencia intra o intergénica, de manera que la frecuencia de SNPs varió de 1 SNP/87 pb en zonas codificantes a 1/40,5 pb en zonas no codificantes.

El número de indels fluctuó entre 0 y 11, contabilizándose un total de 60 indels en el panel completo. La frecuencia de indels encontrada fue de uno cada 288 pb, con un valor promedio de 1,94 indels por región analizada. Si bien los indels por definición son pequeñas inserciones, los tamaños de estas inserciones variaron de una región a otra desde un único nucleótido hasta 52 pb en el gen CAM. En algunos casos, el número preciso de eventos de inserción que dio lugar a un bloque indel fue difícil de determinar, especialmente en aquellas regiones donde el número de pares de bases adicionadas o sustraídas para un mismo bloque era variable entre los distintos individuos, como en el caso de los genes GLP, CAM y ZFP (Figura 1.3).

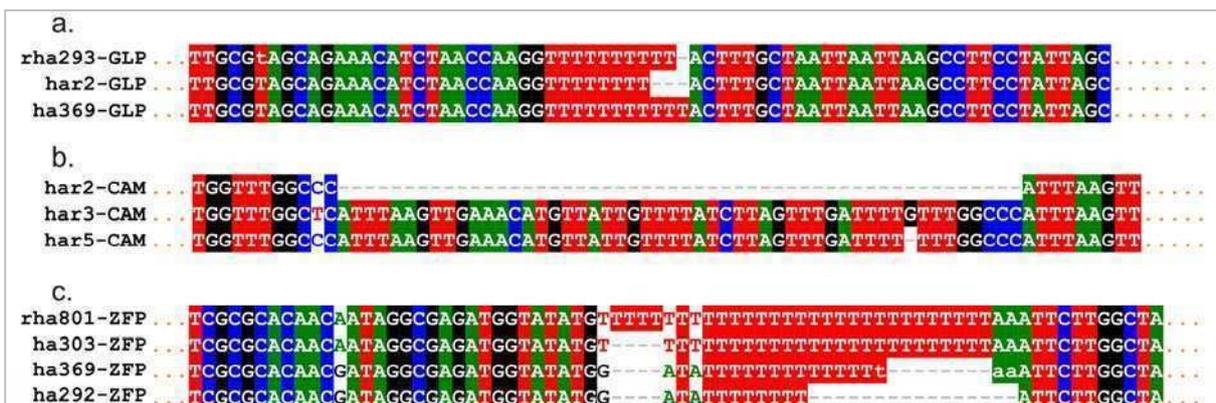


Figura 1.3. Eventos de inserción dentro de una secuencia que dan origen a distintas longitudes de indels.

Regiones de los genes GLP (a), CAM (b) y ZFP (c) con indels en intrones. Se muestran sólo las líneas representantes de los haplotipos encontrados. Se observa que el indel para las distintas líneas está formado por distinta cantidad de pares de bases adicionadas o sustraídas.

En las regiones codificantes sólo se encontraron tres indels, dos en la región MADSB-TF3 (3 pb) y uno en GADPH (1 pb). Los indels fueron excluidos de los análisis posteriores, excepto para los análisis de la cantidad de haplotipos y la diversidad haplotípica en las regiones GO, LZP, GLP y GPX (Tabla 1.5).

El análisis reveló niveles moderados de polimorfismo (Tabla 1.5). La variación genética a nivel nucleotídico fue estimada como la diversidad nucleotídica en promedio para todos los genes evaluados ($\pi_T=0,0074$), y como el número de sitios segregantes ($\theta_W= 0,0067$). Los valores de diversidad promedio de sitios silenciosos ($\pi_{sil}=0,0159$) y de sitios sinónimos ($\pi_{sin}=0,0198$) fueron mayores que los de los cambios no sinónimos ($\pi_{no-sin}=0,0014$). En 29 de las 31 regiones examinadas π_{nonsyn} fue o bien cero (0) o menor al valor de π_{syn} . Sin embargo, las regiones GO y RL41 mostraron valores de π_{no-sin} mayores a los de π_{sin} (Figura 1.4). Para GO π_{no-sin} ascendió a 0,00047, mientras que el valor de π_{sin} fue cero (0); la diferencia en estos valores se debió a una sustitución nucleotídica en la línea RHA293. En RL41 las sustituciones no sinónimas estuvieron causadas por dos sustituciones presentes únicamente en HA292 y un sitio informativo que separa las líneas HA61, HA89, HA303, KLM280, PAC2, RHA266 y RHA274 del resto de las líneas analizadas. Esta sustitución es una transversión (C/A) en la segunda posición del codón, causando el cambio de una prolina a una glutamina (es decir, el cambio de un aminoácido no polar por otro polar). No obstante, la importancia de este cambio sobre la funcionalidad de la proteína no ha sido determinada.

Si bien la frecuencia de SNPs fue mayor en las regiones no codificantes que en las codificantes, los valores promedio de polimorfismo nucleotídico y diversidad nucleotídica de las regiones no codificantes ($\theta_W=0,0057$ $\pi_T=0,0059$) fue mayor, aunque no significativamente diferente a los valores de diversidad estimados para las regiones codificantes ($\theta_W=0,0047$; $\pi_T=0,0052$).

Tabla 1.5. Estimaciones de diversidad nucleotídica y polimorfismo en 31 genes de girasol.

Gen	S _i ^a	θ _w	π _T	π _{sil}	π _{sin}	π _{no-sin}	π _{no-sin} /π _{sin}	Nº haplotipos	Diversidad Haplótica
GO	0	0,0009	0,0004	0,0003	0	0,0005	-	3	0,205
PGIP1	18	0,0067	0,0061	0,0103	0,0172	0,0012	0,0669	12	0,943
PGIP3	3	0,0013	0,0018	0,0050	0,0062	0	0	4	0,725
LZP	0	0	0	0	0	0	-	2	0,281 ^b
GLP	0	0	0	0	0	0	-	3	0,433 ^b
MADSB-TF3	5	0,0034	0,0025	0,0027	0,0159	0,0018	0,1141	9	0,801
AALP	6	0,0119	0,0117	0,0203	0,0143	0	0	4	0,661
LIM	5	0,0056	0,0076	0,0117	0,0092	0	0	4	0,579
1-ACCO	14	0,0078	0,0046	0,0069	0,0042	0,0016	0,3726	7	0,544
ZFP	22	0,0101	0,0182	0,0205	0,0266	0,0000	0,0000	2	0,515
CAM	18	0,0155	0,0137	0,0166	0,0217	0	0	6	0,801
CHS	0	0	0	0	0	0	-	1	0,000
GAPDH	1	0,0008	0,0007	0,0017	0	0	-	3	0,485
GIA	2	0,0008	0,0005	0,0011	0,0016	0	0	2	0,199
GPX	0	0	0	0	0	0	-	2	0,256 ^b
GST	31	0,0204	0,0277	0,0464	0,0636	0,0080	0,1254	9	0,772
PGIC	13	0,0081	0,0055	0,0074	0,0021	0,0012	0,5673	4	0,298
SCR1	3	0,0012	0,0018	0,0076	0,0079	0	0	3	0,649
SCR2	7	0,0040	0,0037	0,0126	0,0126	0,0009	0,0721	3	0,374
RL41	1	0,0087	0,0071	0	0	0,0145	-	3	0,556
ANT	9	0,0122	0,0225	0,0841	0,0888	0	0	2	0,526
RS16	5	0,0047	0,0066	0,0206	0,0294	0	0	3	0,573
NsLTP	5	0,0068	0,0077	0,0084	0,0380	0,0057	0,1504	3	0,433
SEM	1	0,0038	0,0018	0,0027	0	0	-	3	0,205
SAMC	7	0,0093	0,0084	0,0134	0,0357	0,0007	0,0204	5	0,684
GCvT	2	0,0047	0,0064	0,0253	0,0270	0	0	3	0,579
SBP	6	0,0142	0,0137	0,0523	0,0550	0	0	5	0,760
LHCP	8	0,0063	0,0079	0,0268	0,0313	0,0011	0,0341	3	0,602
CPSI	2	0,0068	0,0041	0,0101	0,0059	0,0010	0,1616	2	0,298
PSI-III-CAB	1	0,0004	0,0006	0,0010	0	0	-	2	0,409
CAB	7	0,0038	0,0059	0,0136	0,0203	0	0	3	0,485
Promedio	6,52	0,0067	0,0074	0,0159	0,0198	0,0014	0,0766	4	0,514

a Sitios informativos (SI) usados en las medidas de diversidad nucleotídica.

b El número de haplotipos y la diversidad haplotípica se obtuvo teniendo en cuenta los polimorfismos de tipo indel.

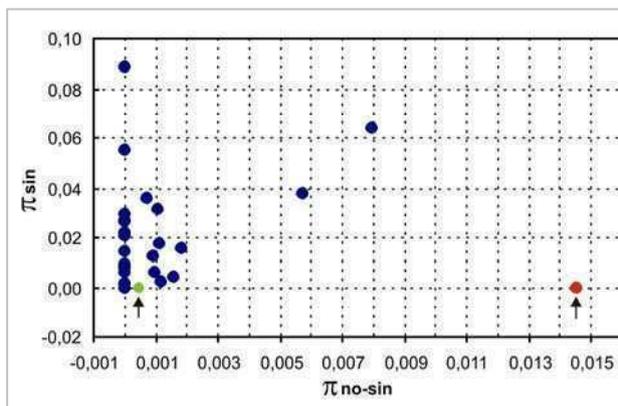
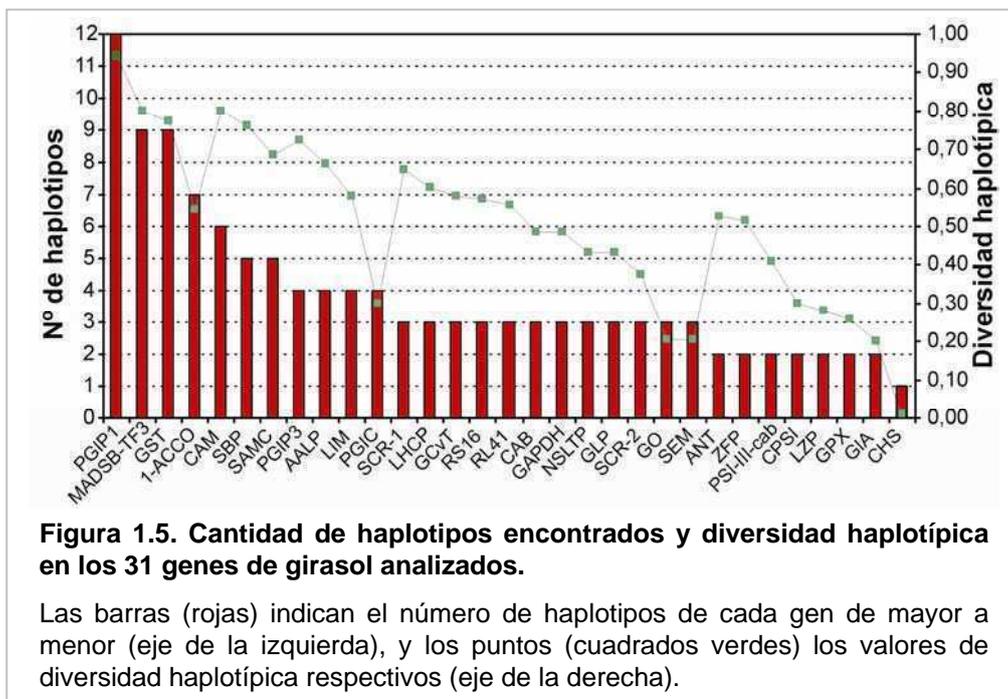


Figura 1.4. SNPs sinónimos (π_{sin}) y no-sinónimos (π_{no-sin}) en las 31 regiones analizadas sobre las 19 líneas de girasol.

Se observa la diversidad nucleotídica de SNPs sinónimos y no-sinónimos. Sólo dos regiones candidatas mostraron un valor de π_{no-sin} mayor que π_{sin}: GO (verde) y RL41 (rojo) (flechas).

La cantidad de haplotipos detectados varió de uno a 12 dependiendo de la región analizada, exhibiendo la mayoría de los genes entre dos y tres haplotipos (Figura 1.5). El promedio de diversidad haplotípica fue de 0,504; pero los valores variaron dependiendo del número de haplotipos presentes y de las frecuencias encontradas para cada uno. De esta manera, para SCR1 y SCR2 se observó la misma cantidad de haplotipos, sin embargo, la diversidad haplotípica difirió de 0,649 a 0,374 debido a las diferencias en la distribución de frecuencias de los tres haplotipos (Figura 1.5). En SCR1, tres líneas mostraron el haplotipo menos representado, mientras que los otros dos aparecieron en nueve y siete líneas, respectivamente. Por otro lado, para SCR2, los dos haplotipos minoritarios estuvieron presentes en dos líneas cada uno, mientras que el resto de las líneas exhibió el tercer haplotipo. Aunque las secuencias de los genes candidatos LZP, GLP y GPX no mostraron polimorfismos de tipo SNPs, los indels identificados permitieron determinar distintos haplotipos, obteniendo un valor de diversidad haplotípica de 0,281; 0,433 y 0,256, respectivamente. Por último, la región CHS no presentó polimorfismos de ningún tipo, por lo tanto se encontró un único haplotipo (Figura 1.5).



Para evaluar los efectos del muestreo de líneas y genes en los niveles de diversidad hallados, las estimas de diversidad obtenidas se recalcularon para las 19 líneas

secuenciadas en esta tesis en conjunto con los accesos P&M elegidos por Liu y Burke (2006), utilizando sólo el conjunto de nueve genes en común (Tabla 1.6). Los valores promedios de polimorfismo obtenidos fueron de 0,0056 para las 19 líneas, 0,0078 para el grupo de cultivares P&M y 0,0079 para el grupo total de accesos. La diversidad nucleotídica observada para este análisis fue de 0,0060, 0,0057 y 0,0069, respectivamente.

Tabla 1.6. Evaluación de los efectos del muestreo en las estimas de diversidad.

	Accesos ^a	Genes									Promedio	Dato anterior
		CAM	CHS	GAPDH	GIA	GPX	GST	PGIC	SCR1	SCR2		
θ_w	LE	0,0155	0	0,0008	0,0008	0	0,0204	0,0081	0,0012	0,0040	0,0056	0,0067^b
	P&M	0,0176	0,0005	0,0006	0,0013	0,0047	0,0190	0,0157	0,0051	0,0054	0,0078	0,0072^c
	Todos	0,0175	0,0004	0,0006	0,0015	0,0043	0,0222	0,0145	0,0046	0,0053	0,0079	-
π_T	LE	0,0137	0	0,0007	0,0005	0	0,0277	0,0055	0,0018	0,0037	0,0060	0,0074^b
	P&M	0,0138	0,0003	0,0011	0,0008	0,0021	0,0124	0,0109	0,0060	0,0042	0,0057	0,0056^c
	Todos	0,0144	0,0002	0,0010	0,0007	0,0014	0,0262	0,0090	0,0051	0,0040	0,0069	-

a Los nueve genes en común con el trabajo de Liu y Burke (2006) se re-analizaron en: (1) las líneas elite (19 alelos/19 accesos), (2) los accesos de girasol cultivado primitivos y mejorados (P&M) (32 alelos/16 accesos) y el grupo completo de accesos (51 alelos/35 accesos).

b Polimorfismo nucleotídico y diversidad nucleotídica obtenidos con el conjunto completo de 31 genes examinado en la Tabla 1.4.

c Polimorfismo nucleotídico y diversidad nucleotídica obtenidos por Liu y Burke (2006).

Determinación de la estructura poblacional y distancias genéticas

La presencia de subestructuración en la población puede llevar a asociaciones espurias y debe ser considerada en el análisis estadístico del desequilibrio de ligamiento (DL) (Flint-Garcia y col. 2003). Por lo tanto, como un paso previo al establecimiento de la extensión del DL, se analizó la estructura poblacional mediante el modelo de Pritchard y col. (2000), utilizando 136 SNPs. Esta prueba resultó útil para evaluar el alcance real del DL y para determinar si existían asociaciones espurias que pueden surgir por razones diferentes a la cercanía de los marcadores. El número de poblaciones ideales, es decir el valor de K que logró maximizar la probabilidad de los datos fue cinco (K=5).

En la Figura 1.6 se presentan los resultados obtenidos mediante el análisis con el programa STRUCTURE. Allí se observan las contribuciones de las cinco poblaciones inferidas por el modelo en el genoma de cada uno de los individuos analizados. Los

genotipos de las 19 líneas elite examinadas en este trabajo resultan, principalmente, de la contribución de dos poblaciones o acervos génicos ideales, representados en amarillo y celeste en la Figura 1.6. La pertenencia a uno u otro grupo fue mayor al 80% en la mayor parte de los individuos. Estos mismos acervos génicos constituyen la ascendencia de las entradas analizadas por Liu y Burke (2006), sólo que la composición de dichos individuos se da en la proporción inversa a la de las líneas endocriadas aquí estudiadas (Figura 1.6).

Por otra parte, la constitución de los genotipos de los individuos silvestres tiene un origen más diverso, observándose contribuciones de todas las poblaciones propuestas por el modelo (Figura 1.6). De acuerdo a este análisis se pudieron definir dos grupos para las 19 líneas estudiadas. El primer grupo (G1) compuesto por HA52, HA61, HA89, HA370, HAR3, HAR5, KLM280, PAC2, RHA266, HA274, RHA293 y RHA374 (en amarillo, Figura 1.6), y el segundo grupo (G2) incluyendo las líneas HA292, HA303, HA369, HA821, HAR2, RHA801 y V94 (en celeste, Figura 1.6).

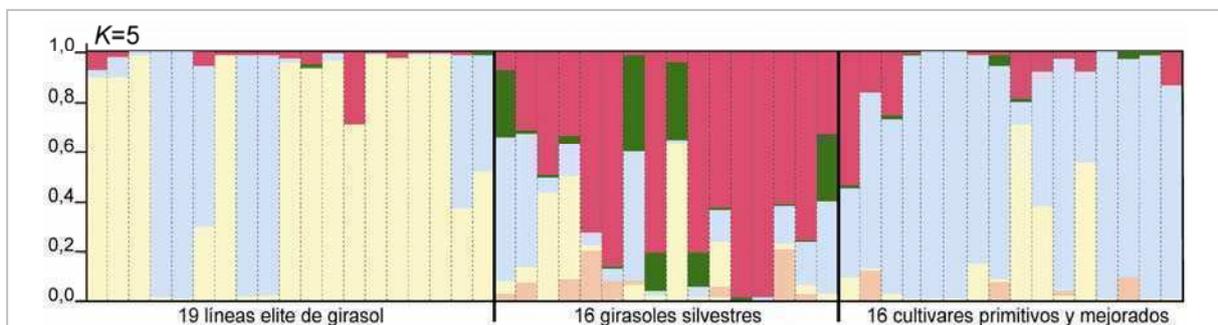


Figura 1.6. Análisis de la estructura poblacional de los accesos de girasol caracterizados.

Cada individuo (rectángulos) está coloreado en K segmentos que representan la contribución estimada de cada población. Se muestran las 19 líneas elite (de izquierda a derecha: HA52, HA61, HA89, HA292, HA303, HA369, HA370, HA821, HAR2, HAR3, HAR5, KLM280, PAC2, RHA266, RHA274, RHA293, RHA374, RHA801 y V94), los 16 accesos silvestres y los 16 accesos primitivos y mejorados estudiados por Liu y Burke (2006) separados por líneas negras.

Con el objetivo de tratar de conocer las relaciones entre las líneas elite con respecto a los demás accesos silvestres y cultivados, se realizó un árbol de *neighbor-joining* (Figura 1.7). Si bien el algoritmo produce un retículo sin raíz, se muestran representaciones enraizadas con un acceso silvestre (wWY) para facilitar la visualización de los agrupamientos obtenidos (Figura 1.7). De esta manera, puede observarse que las particiones con apoyo superior a 50% son aquellas que separan a

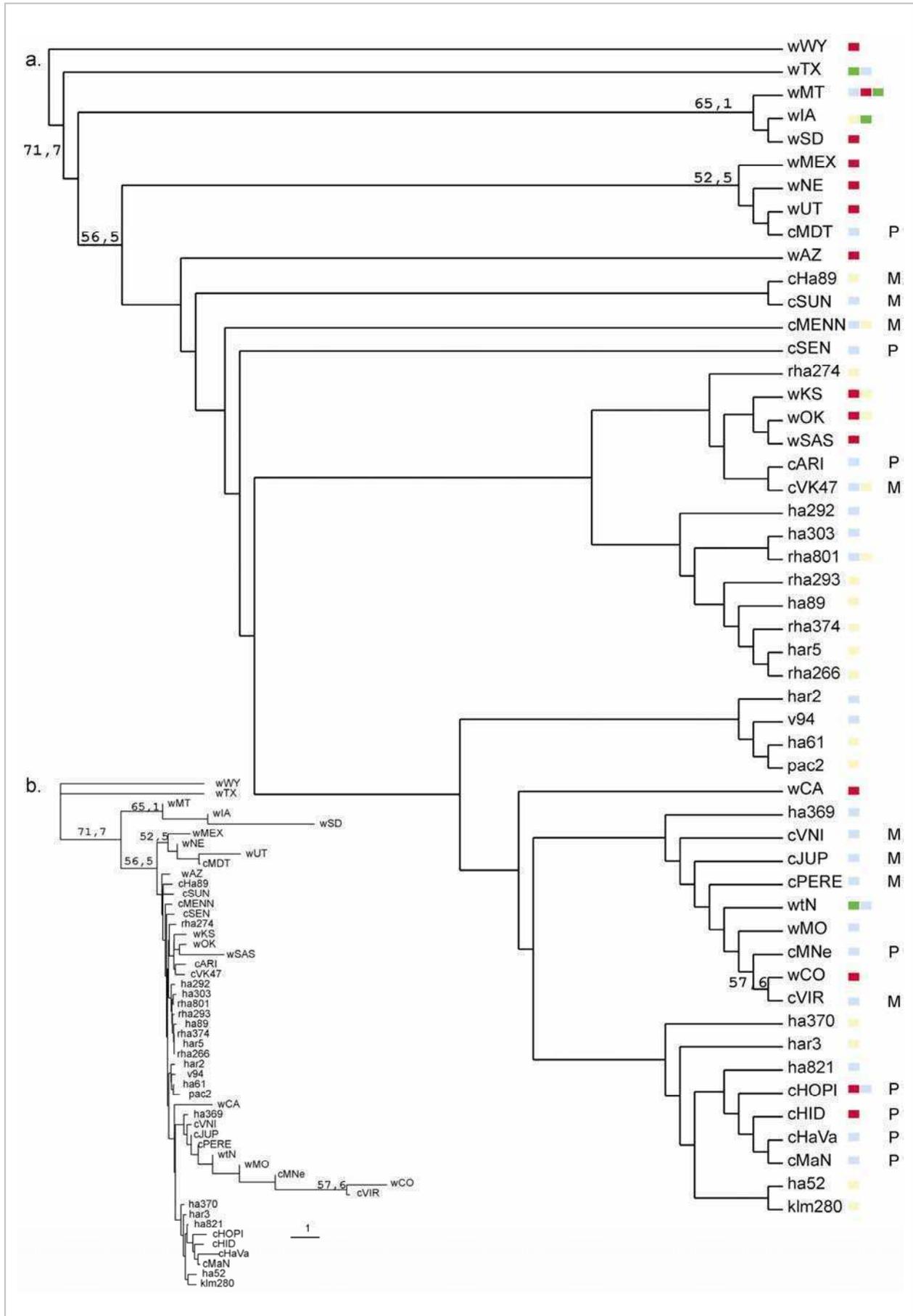
los accesos silvestres de los cultivados y de las líneas endocriadas. No obstante, los accesos silvestres wKS, wOK, wSAS, wCA, wTN, wMO quedaron distribuidos dentro del grupo formado por las líneas endocriadas y los accesos de girasol cultivado primitivo y mejorado (P&M). Los agrupamientos obtenidos no concuerdan con las categorías definidas *a priori* según el tipo de material, estando las líneas endocriadas entremezcladas con los accesos P&M.

En la Figura 1.7a se presentan, además, las pertenencias de cada individuo a las cinco poblaciones inferidas mediante el programa STRUCTURE (de la Figura 1.6). Pueden observarse individuos de la misma población agrupados en el árbol: por ejemplo, HA369, cVNI, cJUP, cPERE, wTN, wMO, cMNE, y cVIR, todos constituidos en mayor proporción por la población celeste de la Figura 1.6. Otro caso es el de las líneas HA293, HA89, RHA374, HAR5 y RHA266, todas pertenecientes a la población caracterizada como amarillo. Sin embargo, no pareciera existir una estricta correlación entre el análisis de distancias y el análisis bayesiano de estructura, probablemente debido a la menor sensibilidad del método de distancias para establecer relaciones a nivel individual (Rosenberg y col. 2001).

Figura 1.7 (pág. sig.). Árbol de *neighbor-joining*.

Se utilizó la distancia obtenida a partir de la matriz de proporción de alelos compartidos de los genes CAM, CHS, GAI, GAPDH, GPX, GST, PGIC, SCR1 y SCR2, para 16 accesos silvestres, 16 accesos cultivados primitivos (P) y mejorados (M) (Liu y Burke, 2006) y 19 líneas endocriadas.

Árbol obtenido como dendograma rectangular (a) y como filograma (b) enraizados con el taxón silvestre wWY. Se indican con colores las pertenencias obtenidas con STRUCTURE (de la Figura 1.6) y los valores de *bootstrapping* mayores a 50%.



Análisis del desequilibrio de ligamiento

De acuerdo al análisis de la estructura poblacional el G1 y G2 están caracterizados por diferentes conjuntos de frecuencias alélicas (Figura 1.6). Por esta razón, los pares de estimaciones del DL (r^2) se calcularon para: (1) el grupo completo de líneas endocriadas (Figura 8a, 8c) y (2) para conjunto de líneas pertenecientes a G1 (Figura 8b, 8d). El G2 no se incluyó en el análisis debido a que el número de líneas que involucraba era menor. La Figura 1.8 muestra los gráficos de dispersión de los valores de r^2 versus la distancia física entre todos los pares de SNPs dentro de un gen, graficados en conjunto para todas las regiones polimórficas incluidas en este trabajo. Como todas las regiones tienen longitudes menores a dos kpb, este análisis expone los patrones de DL a distancias cortas. Se observan los resultados obtenidos para el panel completo de genes polimórficos teniendo en cuenta las 19 líneas y las 12 líneas del G1 (Figura 8a y b, respectivamente). Para estos dos análisis, la línea de tendencia del DL decae muy lentamente, debido a los valores r^2 (1^{***}) de tres regiones (1-ACCO, PGIP1 y ZFP) que presentaron los pares de *loci* más distantes entre sí (673 pb, 875 pb y 615 pb entre uno y otro SNP, respectivamente). Teniendo en cuenta que este comportamiento se observó sólo para 3/31 regiones, los datos de LD se recalcularon, excluyendo estos tres genes. Para el grupo completo de genotipos, la línea de tendencia decae lentamente, alcanzando el valor de 0,64 a las 643 pb (Figura 8c). Por el contrario, cuando el gráfico del DL incluye únicamente el grupo G1, la línea de tendencia decae más rápidamente, llegando a un valor de $r^2=0,48$ para la misma distancia (Figura 8d). En concordancia con las predicciones teóricas acerca de los factores que afectan el DL, la inclusión del grupo completo de líneas produjo un sesgo del DL hacia mayores distancias. El análisis *interlocus* no mostró DL entre *loci* (datos no mostrados).

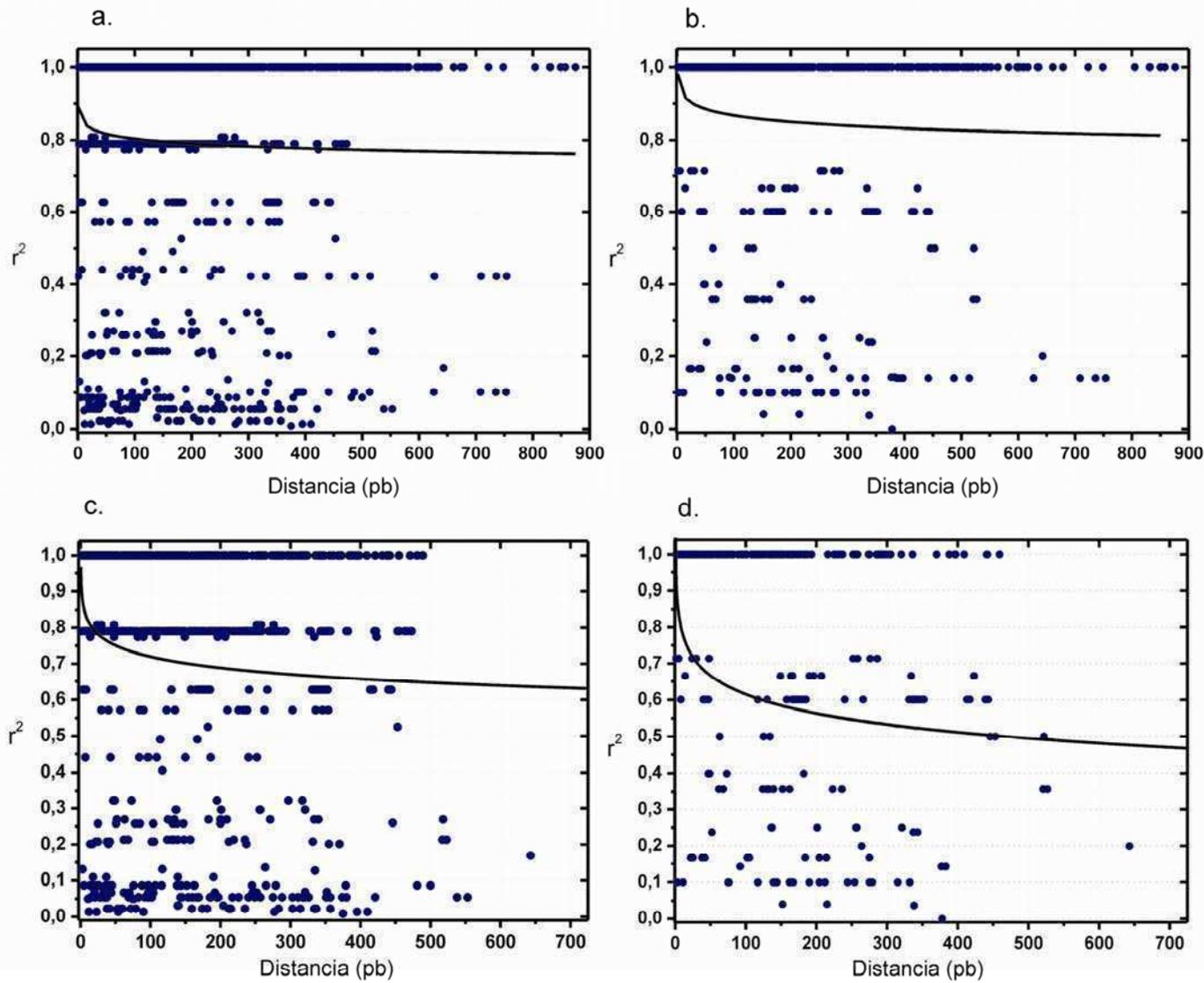


Figura 1.8. Análisis del desequilibrio de ligamiento (DL) en girasol.

Gráfico de dispersión del DL de las 19 líneas elite analizadas para el conjunto de 31 genes secuenciados (a) y para 28 genes excluyendo 1-ACCO, PGIP1 y ZFP (c).

Gráfico de dispersión del DL para el grupo homogéneo G1 identificado mediante el análisis con STRUCTURE utilizando la información de los 31 genes secuenciados (b) y sólo de los 28 genes excluyendo a 1-ACCO, PGIP1 y ZFP (d).

Discusión

Frecuencia de SNPs y diversidad nucleotídica

El estudio de la diversidad nucleotídica en girasol se realizó a través de la selección de un conjunto de genes para el desarrollo de SNPs e indels. Los genes evaluados se seleccionaron mediante tres estrategias: la búsqueda de secuencias en las colecciones de ADNc desarrolladas por el grupo GG-IB INTA Castelar, la búsqueda bibliográfica y la búsqueda de SNPs *in silico* a través del análisis de secuencias de ESTs depositadas en GenBank (Tabla 1.3). Todas las secuencias fueron elegidas por su función, ya sea por comportarse como secuencias control o regiones neutras o debido a su probable rol en mecanismos de respuesta a estrés biótico y abiótico. De un total de 64 genes, 31 pudieron ser secuenciados en las 19 líneas elite de girasol usadas en programas de mejoramiento, revelando diferentes patrones de polimorfismos. La frecuencia de SNPs encontrada en el conjunto de accesos evaluados fue de 1 SNP/61 pb. Este valor es comparable al obtenido por Ching y col. (2002) para líneas endocriadas de maíz (1 SNP/60,8 pb), pero es mayor que el reportado por Tenaillon y col. (2001) también para maíz (1 SNP/104 pb). Sin embargo, las diferencias encontradas en estos dos estudios pueden explicarse por las diferencias en los accesos muestreados. Por otro lado, la abundancia de SNPs encontrada en girasol es también comparable a la descrita para *Pinus taeda*, el cual exhibió 1 SNP/63 pb (Brown y col. 2004). Por el contrario, otros cultivos de interés agronómico como sorgo (*Sorghum bicolor*) (1 SNP cada 123 pb) (Hamblin y col. 2004), soja (*G. max*) (1 SNP/328 y 1 SNP/536 pb) (Van y col. 2005, Zhu y col. 2003) y arroz (*O. sativa*) (1 SNP/113 pb y 1 SNP/100 pb) (Garris y col. 2003, Monna y col. 2006) presentaron una frecuencia de SNPs menor que las líneas analizadas en este capítulo. La ocurrencia de SNPs en girasol fue reportada en 2006 por Liu y Burke para 16 accesos de girasol primitivos y mejorados (1 SNP/39 pb, $\theta_W=0,0072$, $\pi_T=0,0056$) y en 2007 por Kolkman y col. para 10 líneas elite (1 SNP/46 pb, $\theta_W=0,0094$, $\pi_T=0,0107$) (Kolkman y col. 2007, Liu y Burke 2006). La diferencia entre estos valores y las estimas obtenidas para las 19 líneas de girasol puede explicarse por: (1) las diferencias esperadas en la diversidad genética de los accesos analizados, es decir, accesos de girasol cultivados primitivos y mejorados versus líneas elite puras; (2) las

diferentes fuentes de variabilidad consideradas (p. ej. la definición de indels) y (3) las diferencias en la cantidad de regiones o en el criterio de selección de las mismas. Con respecto a este último punto, de los 31 genes seleccionados en el presente capítulo, 22 son regiones nuevas no caracterizadas previamente, que incluyen proteínas relacionadas con respuesta a estrés y otras proteínas seleccionadas al azar, con lo cual representan una buena colección para estimar los patrones de SNPs existentes en el genoma de girasol. Para determinar si los efectos de la varianza *interlocus* (o el muestreo de genes) pueden distorsionar los valores de diversidad nucleotídica (θ_W y π_T), se realizó un segundo análisis con los datos de secuencia provenientes de los nueve genes compartidos con el trabajo de Liu y Burke (2006) para las 19 líneas elite y para los 16 accesos de girasol cultivado primitivo y mejorado (P&M) (Tabla 1.6). Al igual que en el análisis de los 31 genes, el promedio del polimorfismo nucleotídico θ_W , presentó un valor menor ($\theta_W=0,0056$) sobre la población de líneas puras que el valor obtenido para los accesos P&M ($\theta_W=0,0078$). Del mismo modo, el valor promedio de diversidad nucleotídica para las 19 líneas también mostró para el subconjunto de nueve genes, el mismo comportamiento que para el total de 31 genes, resultando mayor para las 19 líneas ($\pi_T=0,0060$) que para los P&M ($\pi_T=0,0057$) (Tabla 1.6). Estos datos avalan los resultados obtenidos para la totalidad de los genes analizados en las 19 líneas de girasol. Además, los valores promedios de θ_W y π_T para los nueve genes, obtenidos sobre el conjunto de los 35 accesos (19 líneas más 16 P&M) fueron mayores que los valores obtenidos para cualquiera de los dos grupos por separado. En consecuencia, estas discrepancias en los valores de diversidad no son causadas por el muestreo y por lo tanto reflejan diferencias genuinas en los niveles de polimorfismo para distintos grupos de individuos. Si bien θ y π expresan esencialmente lo mismo, ambas estimas se obtienen por distintas vías, reflejando a su vez, diferentes datos. De esta manera, θ se obtiene como medida del número de sitios segregantes y su valor se ve influenciado por la presencia de alelos raros, mientras que π es una medida de las diferencias alélicas medidas de a pares de secuencias. En este contexto, las discrepancias entre los valores obtenidos para el grupo de 19 líneas y el grupo de 16 accesos P&M resulta entendible, ya que la falta de alelos raros se observa cuando existen cuellos de botella pronunciados como los que dieron origen a

las líneas puras, y asimismo, el incremento de las diferencias de a pares de secuencias puede ser consecuencia de la naturaleza divergente de los accesos seleccionados para este estudio. Los valores obtenidos en conjunto afirman este supuesto, pesando tanto la presencia de alelos raros en los accesos P&M (aumento de la estima θ_W) como la naturaleza divergente de las líneas elite (aumento de la estima π_T). Los accesos de girasol silvestre analizados en el trabajo de Liu y Burke (2006) evidenciaron una frecuencia de 1 SNPs/19 pb, y valores de diversidad y polimorfismo nucleotídico ($\theta_W=0,0144$; $\pi_T=0,0128$) mayores que los valores obtenidos para las 19 líneas puras. Estos resultados son concordantes con la historia de selección, recombinación y mejoramiento que sufrieron tanto los girasoles cultivados analizados en ese trabajo como las 19 líneas elite analizadas en este capítulo.

El análisis de los SNPs en regiones codificantes (1/87 pb) y regiones no codificantes (1/40,5 pb) evidencia la mayor conservación de secuencia de las regiones codificantes con respecto a las otras. Más aún, del total de SNPs en regiones codificantes el 78% (99 SNPs) fueron cambios sinónimos. Además, los valores promedio de la diversidad nucleotídica de sitios silenciosos ($\pi_{sil}=0,0140$) y de sitios sinónimos ($\pi_{sin}=0,0174$) fueron mayores que el promedio de la diversidad de sitios no-sinónimos ($\pi_{no-sin}=0,0013$), sin embargo, dos *loci* mostraron un π_{no-sin} mayor que el π_{sin} . Estos genes son: GO ($\pi_{no-sin} = 0,00047$; $\pi_{sin}=0$) y RL41 ($\pi_{no-sin} = 0,0145$ y $\pi_{sin} = 0$). En particular para el último gen, la única sustitución no sinónima es causada por un sitio informativo que cambia la secuencia de aminoácidos. Este tipo de cambios ocurren frecuentemente en líneas endocriadas que estuvieron sujetas a selección artificial, por ejemplo, para arroz se observaron cambios sin sentido en sitios clásicamente invariables en cultivares, probablemente a consecuencia de la selección artificial durante el proceso de domesticación (Sato y col. 2001).

Evaluación de la estructura poblacional y distancias genéticas

El modelo de Pritchard y col. (2000) basado en agrupamientos reveló que las líneas endocriadas aquí examinadas están constituidas principalmente por dos acervos génicos, conformando las sub-poblaciones G1 (amarillo) y G2 (en celeste) (Figura 1.6). Al comparar este resultado con el árbol de *neighbor-joining* (Figura 1.7), las líneas pertenecientes al G1, también formaron un grupo, aunque relacionadas con

individuos pertenecientes a otros acervos génicos (rojo y celeste, Figura 1.7). Al parecer, la resolución del árbol no alcanza a discriminar la estructuración de la muestra, que sí logra encontrar el método bayesiano. No obstante, el conocimiento de las relaciones genéticas entre estos y otros materiales de mejoramiento resulta esencial para poder establecer los cruzamientos adecuados en los procesos de selección de nuevas variedades. El análisis de la diversidad genética en las colecciones de germoplasma puede además, proporcionar nuevas herramientas para la clasificación de los accesos y la identificación de accesos núcleo con objetivos de mejoramiento de caracteres específicos. En este sentido, el árbol de *neighbor-joining* permitió evidenciar que las líneas elegidas para el desarrollo de SNPs contemplan la diversidad presente en la especie y esto muestra correlación con los datos del *pedigree* de la Tabla 1.2 (Figura 1.7).

Si bien todas las líneas elite quedaron agrupadas con gran parte de los accesos cultivados, algunas líneas parecieran compartir más alelos con los cultivados primitivos (p. ej. HA370, HAR3, HA821, HA52, KLM280), y con algunos accesos silvestres (p. ej. RHA274), que con los cultivados mejorados (Figura 1.7). Esto podría deberse a que los procesos de introgresión y cruzamiento llevados a cabo para la obtención de las líneas endocriadas, se realizaron en mayor medida con estos materiales (primitivos y silvestres) que con los cultivados mejorados. De hecho, los agrupamientos observados en el árbol con los accesos silvestres (wKS, wOK, wSAS, wCA, wTN, wMO, wCO), podrían reflejar introgresiones genuinas de estos haplotipos en las líneas endocriadas, de acuerdo a la información del *pedigree* (Tabla 1.2).

Por otro lado, tanto el acceso cultivado primitivo cMDT, como el cultivado mejorado cVIR comparten alelos con otros accesos silvestres (wUT-wMEX-wNE y wCO, respectivamente). Éstas son las únicas particiones con accesos de distinta naturaleza (silvestres y cultivados) con un soporte mayor al 50%. De este modo, probablemente estos agrupamientos sí estén reflejando fehacientemente el origen de esas introgresiones, que dieron lugar a la diversidad presente en el germoplasma cultivado de la especie (Figura 1.7).

Determinación del desequilibrio de ligamiento

El equilibrio y el desequilibrio de ligamiento son términos que la genética de poblaciones usa para describir la probabilidad de co-ocurrencia de alelos en distintos *loci* en una población determinada. Generalmente, el término ligamiento se refiere a la co-segregación de los alelos de dos *loci* diferentes, mediada por una conexión física en el cromosoma en el que se encuentran (Oraguzie y col. 2007). La subdivisión de la población y la estructuración de la población son factores que incrementan los valores de DL, sin embargo, el efecto depende del número de poblaciones, la velocidad de intercambio entre éstas y la velocidad de recombinación (Gaut y Long 2003).

Los estudios de asociación basados en el desequilibrio de ligamiento se comenzaron a implementar hace relativamente poco tiempo en plantas, encontrando una cierta resistencia debido, por un lado, a los efectos poco claros de la estructuración poblacional, y por el otro, a la falta de conocimiento general de la estructura del DL en la mayoría de las especies vegetales (Flint-Garcia y col. 2005).

El estudio del DL en girasol no escapa a estos factores que deben ser tenidos en cuenta, ya que la compleja historia de selección y obtención de las líneas endocriadas, y la consecuente estratificación del germoplasma podrían conducir a sobre-estimaciones del alcance del DL. Por consiguiente, esto extendería las correlaciones a *loci* no ligados físicamente, provocando, de esta manera, el fracaso del mapeo por asociación. La inclusión del estudio de estructura poblacional y, consecuentemente, la determinación de dos sub-poblaciones (G1 y G2), fueron fundamentales para el correcto análisis del alcance del DL en las líneas endocriadas (Figura 1.8).

Por otro lado, se evidenció un factor adicional a tener en cuenta en la evaluación del DL: la presencia de genes o regiones con comportamiento excepcional respecto del total de genes muestreados. De esta manera, los gráficos del DL se realizaron para las 19 líneas para la totalidad de los pares de SNPs posibles (Figura 8a), y para el G1 para el total de SNPs agrupados (Figura 8b), evidenciando un decaimiento mucho más tenue (al graficar la línea de tendencia) que al excluir los tres genes de comportamiento excepcional (Figura 8c y 8d). Es posible que para estos genes

(PGIP1, ZFP y 1-ACCO1) la presión de selección sea un factor determinante en la extensión del DL hacia *loci* más distantes, de modo que excluyéndolos del estudio pudo obtenerse una estimación del DL ($r^2=0,48$ a las 643 pb) en concordancia con los resultados obtenidos por otros autores para accesos cultivados (Kolkman y col. 2007, Liu y Burke 2006).

Sin duda el análisis de estructura permitió obtener un panorama más exacto del DL en girasol cultivado, ya que los resultados obtenidos para el G1 (Figura 1.8d), coinciden con los resultados obtenidos en el trabajo de Kolkman y col. (2007). Es decir, a pesar de que la longitud de los genes analizada sólo permitió obtener los patrones de desequilibrio a distancias cortas, la línea de tendencia alcanza un valor de $r^2=0,32$ a las 5500 pb, coincidiendo con el valor obtenido por estos autores para otras líneas endocriadas. La caída del DL es claramente diferente en materiales silvestres comparado con materiales cultivados. En el primer grupo, el DL es fuerte sólo a distancias menores a 200 pb (Liu y Burke 2006), mientras que en las líneas endocriadas el DL persiste al menos hasta las 700 pb (Figura 1.8). Estos mismos patrones se observaron tanto para el grupo de líneas endocriadas estudiado por Kolkman y col. (2007) como para los accesos cultivados P&M estudiados por Liu y Burke (2006).

Al comparar con otras especies, los patrones de DL varían de una a otra. Para líneas endocriadas de maíz, el DL no decae al menos en las 600 pb analizadas por los autores (Ching y col. 2002), coincidiendo con lo observado para las líneas de girasol. Sin embargo, Tenaillon y col. (2001) analizaron el DL en el cromosoma 1 sobre otro grupo de líneas endocriadas y con razas nativas de maíz, encontrando que el DL comienza a disminuir a las 200-300 pb. Por otro lado, el DL evaluado en otros *loci* y otro grupo de individuos evidenció un valor despreciable ($r^2<0,1$) a las 1500 pb (Remington y col. 2001), reflejando el hecho de que el alcance del DL en especies de polinización cruzada es menor que en especies autóгамas (Oraguzie y col. 2007). Para el caso de otra especie de fecundación cruzada como *Solanum tuberosum*, el DL muestra un comportamiento intermedio ($r^2=0,21$ a 1 kb; $r^2=0,14$ alrededor de las 70 kb), probablemente debido a que la propagación vegetativa que se realiza en este cultivo contrarresta el efecto de fecundación cruzada (Simko y col. 2006). Por el

contrario, en las especies con autofecundación, las estimas del DL se hacen despreciables a mayores distancias. Tal es el caso de soja, con una extensión del DL de más de 50 kb (Zhu y col. 2003). También se puede mencionar a *A. thaliana*, con DL evidente hasta las 150 kb (Nordborg y col. 2002) y finalmente a arroz, con DL hasta por lo menos las 100 kb (Garris y col. 2003). De manera análoga para sorgo, el DL comienza a decaer a las 10 kb (Hamblin y col. 2004). Todos estos cultivos poseen patrones de DL comparables a lo observado para el girasol cultivado.

Los resultados hasta aquí obtenidos permiten inferir que, al menos para una primera aproximación, el mapeo de asociación en girasol podría realizarse prescindiendo de la gran densidad de marcadores que se reporta como necesaria en literatura.

CAPÍTULO 2

DESARROLLO Y VALIDACIÓN DE METODOLOGÍAS PARA LA GENOTIPIFICACIÓN DE SNPS E INDELS A GRAN ESCALA

Introducción

Los polimorfismos SNPs (*Single Nucleotide Polymorphism*) e indels (*insertion/deletion*) se han convertido en los marcadores moleculares de elección desde comienzos de Siglo XXI. Su uso está cada vez más asociado a las aplicaciones que usualmente estaban circunscriptas a los microsatélites, es decir, la identificación de cultivares, la construcción de mapas genéticos y el mejoramiento asistido por marcadores moleculares (Lijavetzky y col. 2007). Una de las razones por las cuales el uso de SNPs se ha vuelto tan popular, es la posibilidad de establecer asociaciones con caracteres fenotípicos. Sin embargo, para establecer y validar asociaciones claras entre genotipos y variantes alélicas es necesaria la genotipificación de miles de SNPs sobre cientos de individuos, para lo cual se requieren metodologías que permitan el procesamiento de múltiples *loci* SNPs de manera rápida, sencilla y a costos relativamente moderados (Ganal y col. 2009).

Si bien existen numerosas alternativas metodológicas para la genotipificación masiva de SNPs (Comai y col. 2004, Lin y col. 2009, Makridakis y Reichardt 2001, Ren y col. 2004, Tsuchihashi y Dracopoli 2002, Wang y col. 2005, Xiao y Oefner 2001), la mayoría excede el presupuesto disponible en los laboratorios académicos o instituciones gubernamentales dedicadas a programas de mejoramiento de cultivos. En el contexto de los estudios de mapeo por asociación, muchos laboratorios de biotecnología vegetal se han inclinado por la caracterización de SNPs en genes candidatos en lugar de la estrategia de genoma completo (*genome-wide association*), con el objeto de controlar el número de *loci* a genotipificar y, de esta manera, los costos (Dracatos y col. 2008, Ehrenreich y col. 2009, Pajerowska-Mukhtar y col. 2009). Sin embargo, aún en proyectos de mapeo por asociación más modestos, se requiere contar con tecnologías de genotipificación precisas, simples, y que puedan aplicarse a muestras individuales o a aquellas formadas por conjuntos de individuos.

En respuesta a tales necesidades, se han desarrollado técnicas que involucran la detección de SNPs de manera eficiente y económica, a través de la formación de moléculas heterodúplex. La detección de SNPs mediante estas estrategias implica la comparación de dos o más alelos de un *locus* dado (A y B, Figura 2.1), cuyos productos de amplificación específicos se mezclan y someten a un proceso de

desnaturalización y lenta re-naturalización de las hebras de ADN. A partir de este proceso, por apareamiento aleatorio de las hebras complementarias se producen dos tipos de moléculas: (1) las moléculas homodúplex, que son idénticas a las moléculas originales (Homo, Alelos A y B, Figura 2.1), y (2) las moléculas heterodúplex, que resultan del apareamiento de una hebra proveniente del alelo A y la complementaria del alelo B, y viceversa (Hetero, Moléculas Heterodúplex, Figura 2.1). Estas últimas, poseen apareamiento imperfecto en los sitios variables (SNPs o indels). Este principio de detección de SNPs ha sido utilizado en la práctica, a través de distintas aproximaciones experimentales, incluyendo hibridación de oligonucléidos (Pease y col. 1994, Southern 1996), ensayos Taq-Man® (Higuchi y col. 1993, Livak y col. 1995), *molecular beacons* (Barreiro y col. 2009, Wang y col. 2005), ensayos electrónicos de *dot blot* (Gilles y col. 1999), cromatografía líquida de alto rendimiento en condiciones de desnaturalización parcial (dHPLC) (Oefner y Underhill 1998) y corte con enzimas endonucleasas específicas (Oleykowski y col. 1998).

En particular, el análisis de heterodúplex mediante corte específico con la enzima CEL1 (CEL1CH) se ha utilizado con éxito para identificar mutaciones puntuales derivadas del tratamiento de plantas con el mutágeno químico EMS (metanosulfonato de etilo) denominándose TILLING (*Targeted Induce Local Lesions IN Genomes*) (Cooper y col. 2008, Perry y col. 2003, Till y col. 2007, Uauy y col. 2009, Wang y col. 2008, Xin y col. 2008). Posteriormente, esta técnica de genética inversa se adaptó para la caracterización de diversidad natural, recibiendo el nombre de EcoTILLING (Comai y col. 2004, Galeano y col. 2009, Gilchrist y col. 2006, Nieto y col. 2007). La enzima CEL1 (endonucleasa 1 de apio, *Apium graveolens* L.) es miembro de la familia de las endonucleasas vegetales, las cuales reconocen apareamientos imperfectos (*mismatches*) en las dobles hebras de ADN, y cortan ambas hebras en el extremo 3' de la región distorsionada (Figura 2.1). La detección de mutaciones o SNPs con CEL1 ha sido desarrollada comercialmente por Transgenomics Inc. (EEUU) y adaptada a diferentes plataformas de identificación de los fragmentos resultantes de la digestión enzimática, incluyendo el sistema Wave®, electroforesis estándar en geles de agarosa o poliacrilamida, electroforesis capilar fluorescente (ECF) y el sistema LICOR® (Surveyor Mutation Kit, Transgenomic Inc., EEUU). No obstante, la

enzima CEL1 es de fácil obtención a partir de tallos de apio fresco, lo que la convierte en una técnica de factible implementación en los laboratorios de bajo presupuesto.

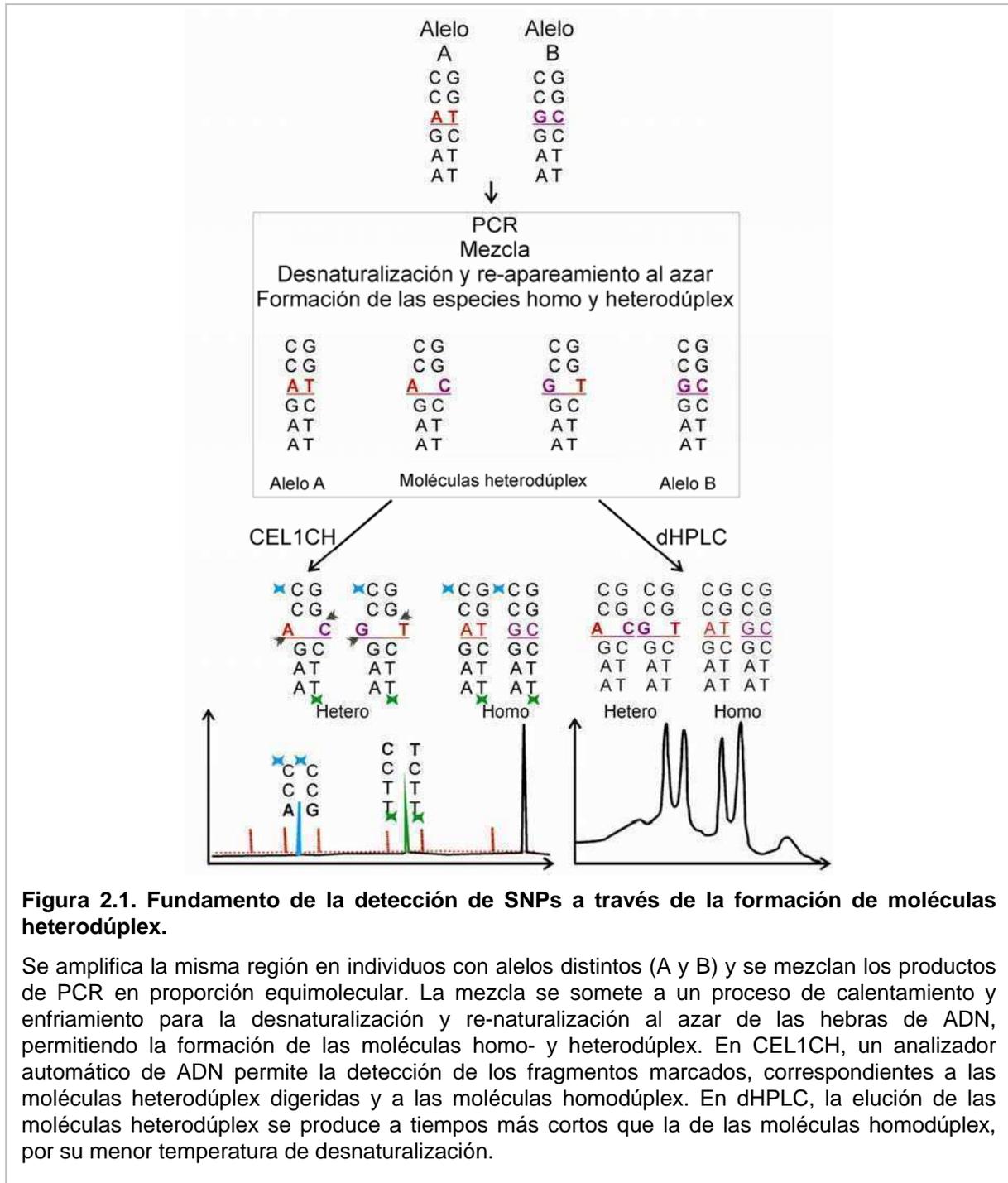


Figura 2.1. Fundamento de la detección de SNPs a través de la formación de moléculas heterodúplex.

Se amplifica la misma región en individuos con alelos distintos (A y B) y se mezclan los productos de PCR en proporción equimolecular. La mezcla se somete a un proceso de calentamiento y enfriamiento para la desnaturalización y re-naturalización al azar de las hebras de ADN, permitiendo la formación de las moléculas homo- y heterodúplex. En CEL1CH, un analizador automático de ADN permite la detección de los fragmentos marcados, correspondientes a las moléculas heterodúplex digeridas y a las moléculas homodúplex. En dHPLC, la elución de las moléculas heterodúplex se produce a tiempos más cortos que la de las moléculas homodúplex, por su menor temperatura de desnaturalización.

La técnica de dHPLC (*denaturing High Performance Liquid Chromatography*) es otra alternativa para el análisis de heterodúplex, ya que revela la presencia de la variación genética mediante la retención diferencial de las moléculas homo- y heterodúplex al ser inyectadas en una columna iónica de fase reversa en condiciones de

desnaturalización parcial (Xiao y Oefner 2001). Tradicionalmente, las moléculas heterodúplex interaccionan con la columna más débilmente y presentan una elución más rápida que las moléculas homodúplex (Figura 2.1). Este método se ha utilizado en numerosas oportunidades mostrando alta sensibilidad y eficiencia (Giordano y col. 1999, Kota y col. 2001, Kota y col. 2008, Lai y col. 2005, Randall y col. 2005, Wagner y col. 1999).

A pesar de que el estudio de la diversidad nucleotídica en girasol ha generado un amplio número de marcadores funcionales (1805 SNPs y 269 indels, Cap. 1, (Kolkman y col. 2007, Liu y Burke 2006)), no se han explorado aún alternativas de genotipificación de mediano a alto rendimiento además de la secuenciación directa de fragmentos de PCR. La utilización de SNPs e indels en la caracterización de germoplasma y en futuros estudios de genética de asociación requiere de métodos alternativos de detección para estos marcadores.

Por lo tanto, se decidió evaluar la posibilidad de uso de las técnicas CEL1CH y dHPLC en la determinación de SNPs e indels en girasol, con el objeto de generar nuevas herramientas que posibiliten la caracterización genotípica de un número mayor de individuos y la incorporación de marcadores funcionales a los mapas genéticos disponibles para girasol cultivado.

Materiales y Métodos

Material vegetal

Se seleccionaron nueve líneas endocriadas para las cuales se disponía de la secuencia correspondiente a 31 genes descritos en el Cap. 1 (HA61, HA89, HA292, HAR2, HAR5, KLM280, PAC2, RHA266 y RHA801). Éstas se utilizaron como líneas control en la evaluación de las metodologías CEL1CH y dHPLC. Un segundo grupo constituido por 22 líneas de genotipo desconocido, fue utilizado para evaluar el rendimiento y la escalabilidad de los ensayos con CEL1CH y dHPLC. Estas líneas son: MP557, HA336, F164B, PAC1, P94, CM307, L156, R432, A71, A59, R419, R417, R423, HA64, HAR4, RK416, RHA275, HA853, HA441, RHA440, RHA439 y RHA274b. Todas las líneas pertenecen al Banco Activo de Girasol-INTA Manfredi (BAG-IM, Córdoba, Argentina). Adicionalmente, se utilizó una población proveniente del cruzamiento entre las líneas RHA266 y PAC2, que consta de 90 líneas recombinantes endocriadas (RILs, *recombinant inbred lines F8*), para el mapeo de las regiones que resultaron polimórficas entre las líneas parentales (Al-Chaarani y col. 2004). Se realizó la extracción de ADN de acuerdo a lo descrito en el Cap. 1.

Selección de genes y diseño de iniciadores

De los 31 genes secuenciados sobre las 19 líneas en el Cap. 1, 24 fueron seleccionados para ser evaluados mediante las técnicas CEL1CH y/o dHPLC. Los siete genes restantes no fueron evaluados debido a la ausencia de SNPs (GLP, CHS), el reducido tamaño de la región a analizar (< 200 pb) (RL41) o la presencia de numerosos SNPs, hecho que dificulta la detección de cada uno de ellos mediante estas técnicas (SBP, SAMC, ZFP, GST). La identidad de los genes y la función probable, junto con la longitud de las regiones, los iniciadores diseñados para amplificarlas y los polimorfismos a identificar se detallan en la Tabla 2.1.

Todos los iniciadores fueron diseñados mediante el programa Primer3 (Rozen y Skaletsky 2000). Para las regiones en las que se utilizó la técnica de CEL1CH, los iniciadores se diseñaron de acuerdo a las especificaciones del sistema comercial de identificación de SNPs mediante corte de los heterodúplex con la enzima Surveyor

(Surveyor Mutation Kit™, Transgenomics Inc., EEUU), marcando los extremos 5' de los mismos con fluoróforos FAM (f) o Hex (h) (Tabla 2.1). Para los genes evaluados mediante dHPLC, se utilizaron iniciadores previamente diseñados para la secuenciación de los genes (Tabla 1.3, Cap. 1), que se ajustaran a las sugerencias de Oefner y Underhill (1998), o se diseñaron iniciadores nuevos. Para regiones de longitud superior a las 850 pb, se rediseñaron iniciadores nuevos para acotar la identificación a regiones entre 200 y 850 pb.

Tabla 2.1. Descripción de los genes evaluados para su uso en la genotipificación de SNPs mediante las técnicas de CEL1CH y dHPLC.

Gen	Longitud SNPs ^a	Indels ^b	Iniciadores diseñados ^c / Criterio de rechazo	Método	
CPSI Subunidad V de Fotosistema I, precursor	183	2	-	< 200 pb	CEL1CH
SEM Subunidad sem 1-2 del complejo proteasoma	236	3	-	SNPs cerca de los extremos de la secuencia	CEL1CH
GcVT Proteína del sistema T de corte de Glicina	241	3	-	SNPs cerca de los extremos de la secuencia	CEL1CH
AALP Proteasa tipo Aleuraína	287	7	1 (5)	SNPs cerca de los extremos de la secuencia	CEL1CH
SCR2 Factor de transcripción tipo 2 <i>Scarecrow</i>	504	7	-	SNPs cerca de los extremos de la secuencia	CEL1CH
LIM Proteína con dominio LIM	204	3	1 (4)	Ff: 5' CTCAGAGTGCAAACAGACTCTCAAG 3' Rh: 5' GGTATGCAGTCCATCCACCTTAAC 3'	CEL1CH
MADSB-TF3 Factor de transcripción MADS-box	371	1	-	Ff: 5' CGTAAGTGTGGTGCTTCTTCTCGT 3' R: 5' TTGTTATCTATGGCTCGTCTGG 3'	CEL1CH
PGIP1 Proteína inhibidora de poligalacturonasa	388	4	-	Ff: 5' TGGACCAGTAACCTCCGACTGCTGTA 3' R: 5' TTCCTGACTCGTTTGGGAAC 3'	CEL1CH
	511	2	2 (3)	F: 5' TTCCTGACTCGTTTGGGAAC 3' R: 5' CTTACTGGTGGGTCCCATGT 3'	dHPLC
	301	1	-	F: 5' TTTGCCAGAGACGTTGACAG 3' R: 5' CACCAAAGTTATATTCTTTGTAATG 3'	dHPLC
PGIP3 Proteína inhibidora de poligalacturonasa	516	2	-	Ff: 5' TCATTTGGGAAGTTCAAAGTCTCG 3' Rh: 5' GCACATACTTAGATGGAACAAAGAAGCAC3'	CEL1CH
	604	2	-	F: 5' AAACCTACCGGCTCTATTCC 3' R: 5' TTAGATGGAACAAAGAAGCACA 3'	dHPLC
CAB Proteína de unión a clorofila a/b	509	5	2 (10)	Ff: 5' GATACTGAATGTCCCAAGTGGTATGAC3' Rh: 5' CCCTGACATAGTTTGGACTATGGTGTGT3'	CEL1CH
	191	2	-	F: 5' GGATACTGAATGTCCCAAGT 3' R: 5' GACTTCATTTGGAGGCAAAC 3'	dHPLC
GIA Proteína insensible a ácido giberélico	569	2	-	Ff: 5' GTGACTACCCAAATCAAACCACACC 3' Rh: 5' CTCAGCACAGGCCATAAGAGTGTGTA 3'	CEL1CH
	365	1	-	F: 5' CAAATCAAACCACACCCTCTTCC 3' R: 5' CGTACCTACTAACGAGCCAGCGATA 3'	dHPLC
	308	1	-	F: 5' ACCAAAACCCATCCGATCTATCTTC 3' R: 5' AGCACAGGCCATAAGAGTGTGTACC 3'	dHPLC
GPX Glutación peroxidasa	584	-	1 (6)	Ff: 5' CTGTTGGGGTTATCCTGTGATGATGT 3' Rh: 5' AACTGTCTGAAAGCCCTTGGGAAAC 3'	CEL1CH
	345	-	1(6)	F: 5' GACTCATTACCATCATCTTCTCTC 3' R: 5' CTTCTGCACCAACATCACAATAGTA 3'	dHPLC

Gen	Longitud SNPs ^a Indels ^b		Iniciadores diseñados ^c / Criterio de rechazo		Método
PSI-III-CAB	612	1	-	F: 5' AACAGACAACCTTTGGTTTGCAT 3'	CEL1CH
Proteína tipo III de unión a clorofila a/b del Fotosistema I	336	1	-	Rh: 5' AAGGGTCTGCCAGTAGTCGTATGT 3' F: 5' TTTCTTACCTGGATGGAAGGT 3' R: 5' ACCTGGGAGACTGTAAACATCAAAG 3'	dHPLC
SCR1	625	2	-	Ff: 5' GCTGATAAGTTGAAAGCGTTGCAGTC 3'	CEL1CH
Factor de transcripción tipo 1 <i>Scarecrow</i>	812	3	-	Rh: 5' CATTCTCACGTTCTATCGTCGCATC 3' F: 5' GGAATCCTGTCTGCTGATAAGT 3' R: 5' TTCACCTGCGAAACAAGCTC 3'	dHPLC
ANT	216	12	2 (2)	SNPs cerca de los extremos de la secuencia	CEL1CH
Translocador de adenin-nucleótido	216	12	2 (2)	Idénticos a Tabla 1.3	dHPLC
LHCP	424	8	-	SNPs cerca de los extremos de la secuencia	CEL1CH
Precursor de la proteína de unión a clorofila a/b receptora de luz	424	8	-	Idénticos a Tabla 1.3	dHPLC
NsLTP	415	7	2 (13)	Numerosos SNPs e indels	CEL1CH
Proteína de transferencia de lípidos no-específica	415	7	2 (13)	Idénticos a Tabla 1.3	dHPLC
CAM	642	20	6 (91)	Numerosos SNPs e indels	CEL1CH
Calmodulina	634	17	1 (4)	Idénticos a Tabla 1.3	dHPLC
RS16	301	2	-	Ff: 5' AGATTCTCCGTTACAAGGCGTTTCG 3'	CEL1CH
Proteína ribosomal S16 de la subunidad 40S	301	2	-	Rh: 5' TACGACTTCTGGAACCTAGCACGAG 3' No Evaluado	dHPLC
GO	669	2	1 (36)	Ff: 5' GAAGAACTTTGAGGGCTTGGACCTT3'	CEL1CH
Glicolato oxidasa	669	2	1 (36)	Rh: 5' CCACCATCTAAGAACACGGGAACAC 3' No evaluado	dHPLC
LZP	274	-	1 (8)	Ff: 5' TTGCTAAAGGGGAAGTGGTCAAAT 3'	CEL1CH
Proteína cierre de leucinas	274	-	1 (8)	Rh: 5' TCCTAAGTCCAATGCCAAGTTGAAG 3' No evaluado	dHPLC
1-ACCO	450	5	1 (1)	Ff: 5' ACCAAGGTTAGCAACTACCCACCAT 3'	CEL1CH
Oxidasa del ácido 1 aminociclopropano-1-carboxílico	450	5	1 (1)	Rh: 5' GGGTACGATTGTCTTTCTCTGTTGG 3' Secuencia con indeterminaciones	dHPLC
PGIC	204	3	1 (1)	SNPs cerca de los extremos de la secuencia	CEL1CH
Fosfoglucosa isomerasa citosólica	204	3	1 (1)	F: 5' GTCTTATCCGTTGACATTTTTCAGG 3' R: 5' AGAGCTTGAGTGTAAGGCAGGATA 3'	dHPLC
GAPDH	309	2	1 (2)	La posición de los SNPs impide el reconocimiento de los fragmentos luego de la digestión con CEL1	CEL1CH
Glicerol 3-fosfato deshidrogenasa	309	2	1 (2)	F: 5' CTCCTTTCGTCAAAGTCATCG 3' R: 5' AGGTTGGGACAATGTTTAGTGCT 3'	dHPLC

a Número de SNPs en la región a amplificar.

b Número de indels (pb de los indels)

c Iniciadores Sentido (F) y Antisentido (R) marcados con FAM (f) o HEX (h).

Reacciones de PCR

Las reacciones de PCR se realizaron en un volumen de 50 µl como se detalló previamente. Para aquellos casos en los cuales la cantidad de producto de PCR obtenido fue escasa, se ensayó la reacción de PCR con un programa *touchdown* en un rango de 60°C a 55°C. El programa fue el siguiente: 2 min a 94°C de desnaturalización inicial, cinco ciclos de (1) 40 seg a 94°C, (2) 40 seg a 60°C (-1°C/ciclo) y (3) 2 min a 72°C; seguido por 32 ciclos de (1) 40 seg a 94°C, (2) 40 seg a 55°C y (3) 1 min a 72°C; con una extensión final de 10 min a 72°C. Para los productos de PCR obtenidos mediante iniciadores marcados con fluoróforos, los geles de agarosa fueron visualizados utilizando la técnica estándar descrita en el Cap. 1 y en

un escáner de fluorescencia (Typhoon Trio, GE Healthcare Bio-Sciences, Inglaterra), para corroborar la incorporación de la marca fluorescente en los mismos.

Formación de las moléculas heterodúplex

Para la puesta a punto e instalación de las técnicas CEL1CH y dHPLC en la genotipificación de SNPs, se trabajó con las nueve líneas de secuencia conocida. Debido a que las líneas endocriadas son esencialmente homocigotas, las moléculas heterodúplex debieron crearse artificialmente, mezclando los productos de PCR de dos líneas de haplotipo distinto para cada gen, como se muestra en la Figura 2.1. Estas líneas se usaron luego como controles en la prueba de desempeño y escalabilidad de ambas técnicas con 22 accesos de girasol de genotipo desconocido. Los productos de PCR de cada una de las 22 líneas se mezclaron equimolecularmente con el producto de PCR de cada línea control, representativa de los haplotipos hallados.

Para la genotipificación de las 90 RILs de la población de mapeo, el producto de PCR de cada RIL se mezcló con el producto de PCR de una de las líneas parentales (RHA266).

Para descartar heterocigosis, contaminación de la muestra heterodúplex y/o presencia de haplotipos distintos a los evaluados, se realizaron controles homodúplex para todas las líneas. Para ello, este control se desarrolló con un solo producto de PCR de la línea en cuestión, ya sea, de las líneas de genotipo conocido (controles), de las de genotipo desconocido, de los parentales de la población de mapeo o de las RILs (ver más detalles en la sección asignación de genotipos de este capítulo).

La cuantificación de los productos de PCR para realizar las mezclas heterodúplex se llevó a cabo mediante espectrometría de fluorescencia usando el reactivo de HOECHST de acuerdo a lo descrito en el Cap. 1. En algunos casos, se recurrió a la cuantificación de los productos de PCR por comparación de los datos de volumen e intensidad de las bandas visualizadas en el gel de agarosa, utilizando un patrón de peso molecular de concentración conocida (100 pb, New England BioLabs, EEUU), mediante el programa Image Quant (Amersham, Inglaterra).

Todas las muestras se sometieron a un proceso de desnaturalización y lenta re-

naturalización en el termociclador Mastercycler *epgradient S* (Eppendorf, Alemania), utilizando el programa descrito en la Tabla 2.2, de acuerdo a lo sugerido en la versión comercial de CEL1 (Surveyor Mutation Kit™, Transgenomic, EEUU). Posteriormente, las muestras fueron analizadas mediante HPLC o incubadas con la enzima CEL1 endonucleasa para poder detectar los polimorfismos correspondientes.

Tabla 2.2. Programa diseñado para la formación de moléculas heterodúplex.

Temperatura	Tiempo o velocidad	Ajuste de la rampa de descenso de temperatura
95°C	2 min	100% (4,5°C/seg)
95°C hasta 85°C	-2°C/seg	45%
85°C hasta 25°C	-0,1°C/seg	2%
4°C	Infinito	100%

Método CEL1CH

Purificación parcial de la enzima CEL1 endonucleasa de apio

El extracto crudo de apio conteniendo actividad CEL1 (CJE, *Celery Juice Extract*) se obtuvo a partir de dos protocolos disponibles en la literatura (Till y col. 2003, Till y col. 2004). Todos los pasos se efectuaron a una temperatura de 4°C. Se lavaron, secaron y procesaron 0,5 kg de tallos de apio fresco. Posteriormente, este extracto (360 ml en total) se ajustó a la Solución Amortiguadora A (0,1M Tris-HCl (pH=7,7), 100 µM PMSF y se centrifugó a 2600 g durante 20 min. Se retuvo el sobrenadante (400 ml) y se llevó al 25% de saturación con sulfato de amonio (NH₄)₂SO₄, agregando 144 g de sal por litro de solución muy lentamente. Se mantuvo en agitación durante 30 min. Se centrifugó la solución (440 ml) a 13.000-16.200 g por 40 min. Se retuvo nuevamente el sobrenadante y se llevó al 80% de saturación con (NH₄)₂SO₄, agregando 390 g de sal por litro de solución. Esta suspensión se mantuvo en agitación durante 30 min y luego se centrifugó a 13.000-16.200 g durante 90 min. Se descartó el sobrenadante y se conservó el pellet -80°C hasta el día siguiente. Posteriormente, se resuspendió en Solución Amortiguadora A (0,1 del volumen inicial). El concentrado se transfirió a una membrana de diálisis Spectra/Por® (límite de corte 12-14.000 MWCO), y se dializó contra la Solución Amortiguadora A (32 l), en tandas de 8 l renovados cada 1 h. El dializado se dividió en alícuotas de 500 µl, conservándolas a -20°C, denominándolo CJE (extracto de enzima CEL1, del inglés

celery juice extract). Se comprobó que la actividad de las alícuotas permanecía intacta por un periodo de al menos tres etapas de descongelado sucesivas.

Se realizó el análisis de proteínas por electroforesis en geles de poliacrilamida en condiciones desnaturalizantes (SDS-PAGE), de acuerdo a lo informado por Laemmli (1970).

Corte del Heterodúplex con CEL1

Para ajustar la cantidad de ADN necesaria para una digestión eficiente de las moléculas heterodúplex (sitios no apareados) por parte de CJE, se ensayaron dos muestras. Por un lado se utilizaron los controles FAM-G y FAM-C del sistema comercial Surveyor™ (Transgenomic, EEUU) y por el otro, se amplificó mediante PCR la región RS16 sobre dos genotipos controles de haplotipo distinto. Se ensayaron muestras de 100 a 500 ng de ADN total para los heterodúplex FAM-G/FAM-C con distintos volúmenes de CJE (0,2 µl; 0,5 µl; 1 µl; 2 µl y 5 µl). Por otro lado, se analizaron muestras de 250, 300, 400 y 500 ng de ADN total para FAM-G/FAM-C y RS16-Haplotipo1/RS16-Haplotipo2 con 0,2 µl de CJE. A cada muestra (18 µl) se le agregó, además de CJE, 2 µl de Solución Amortiguadora CJE 10X (5 ml 1M MgSO₄, 5 ml 1 M HEPES, pH 7,5, 2,5 ml 2 M KCl, 0,1 mL 10% Triton® X-100, 5 µl 20 mg/ml BSA; 37,5 ml agua de-ionizada). Las muestras se incubaron a 45°C por 15 min y la reacción enzimática se detuvo por agregado de 5 µl de Solución de Punto Final (*stop solution*, 0,15 M EDTA pH 8,0). Luego, el ADN se precipitó con 2,5 volúmenes de etanol absoluto a -20°C por 30 min, se centrifugó a 3600 g durante 45 min y se resuspendió en 5 µl de agua de-ionizada. Los controles FAM-G y FAM-C son fragmentos de PCR de 632 pb, cuya secuencia varía en la posición número 217 [C/G] desde el extremo 5'. La actividad de CJE se evidenció cuando se observaron tres picos de fluorescencia a las 217, 415 (corte desde el extremo 5' de la hebra complementaria) y 632 pb (largo de la molécula homodúplex). Para RS16 (301 pb) la actividad de CJE se confirmó cuando se observó un pico de fluorescencia a las 206 pb, correspondiente al sitio SNP entre los dos haplotipos mezclados. Paralelamente, se realizaron los controles de corte del heterodúplex con el sistema comercial (Surveyor™) de acuerdo a las especificaciones del fabricante (Surveyor™, Transgenomic, EEUU).

Electroforesis capilar fluorescente (ECF)

De manera similar a la genotipificación de microsatélites (SSRs), el producto de la digestión con CJE (5 μ l) se diluyó con 10 μ l de Hi-Di formamida (Applied Biosystems, EEUU) y 0,25 μ l de estándar de tamaño GeneScan 500 (-250) ROX (Applied Biosystems, EEUU). Las muestras se incubaron a 95°C por 5 min y posteriormente en hielo durante 5 min. Cada mezcla se inyectó en el analizador ABI 3130xl (Applied Biosystems, EEUU) y se separaron los fragmentos en capilares de 50 cm de largo conteniendo polímero POP-7 (Applied Biosystems, EEUU). Los resultados fueron analizados utilizando el programa GeneMapper v3.7 (Applied Biosystems, EEUU).

Método de dHPLC

El análisis por dHPLC se realizó con un sistema Agilent Serie 1100 (Agilent Technologies Inc., EEUU) equipado con un sistema de bio-compatibilidad, una unidad de de-gasificación de solvente, una bomba binaria, un módulo con control de temperatura para la inyección de muestra automática, un horno para la columna y un detector UV-Visible de arreglo de diodos.

La separación de las moléculas homo- y heterodúplex se realizó utilizando una columna para ADN Helix™ Varian (3X50 mm, Varian Argentina LTD, Argentina). La cromatografía se ensayó con caudales de 0,45 y 0,9 ml/min y con un gradiente binario lineal de las Soluciones Helix A y B (Varian Argentina LTD, Argentina). Las moléculas homo- y heterodúplex de ADN se detectaron a $\lambda=260$ nm. Se ensayaron las temperaturas de desnaturalización parcial de acuerdo a la predicción realizada por el programa *dHPLC Melt Program* disponible en el sitio de Internet de la Universidad de Stanford (<http://insertion.stanford.edu/melt.html>). Para detectar los SNPs en cada gen, la temperatura se optimizó empíricamente ensayando las muestras heterodúplex obtenidas al mezclar dos líneas control de haplotipo distinto. El sistema dHPLC se controló antes de cada serie de corridas inyectando el estándar pUC18 digerido con la enzima Hae III (Varian Argentina LTD, Argentina o Sigma Aldrich, EEUU) y el estándar de detección de heterodúplex DYS271 (Varian Argentina LTD, Argentina), los cuales fueron analizados a 50°C y 56°C, respectivamente, siguiendo las recomendaciones del fabricante de la columna. Se siguieron las corridas de

dHPLC y se detectaron los picos mediante el programa HPCHEM (Agilent Technologies Inc., EEUU), utilizando los parámetros de integración por defecto del programa.

Asignación de genotipos

Dado que el producto de PCR de cada línea de genotipo incógnita se mezcló con cada uno de los productos de PCR de las líneas de genotipo conocido (líneas control), las mezclas control/incógnita revelaron un perfil cromatográfico de tipo heterodúplex en todos los casos, excepto para la mezcla con el producto de PCR de la línea control que tuviera el mismo haplotipo que la línea de genotipo incógnita. De esta manera, cuando la mezcla control/incógnita mostró un perfil de tipo homodúplex, se asignó al genotipo incógnita el mismo genotipo que la línea control.

Para el análisis de las RILs, la observación de un perfil de tipo heterodúplex en las mezclas RIL/RHA266 se atribuyó a la presencia del haplotipo derivado del parental PAC2. En caso contrario, si la mezcla RIL/RHA266 mostraba un perfil homodúplex, se le asignó a la RIL el haplotipo derivado de RHA266.

Los controles homodúplex sirvieron para determinar presencia de nuevos haplotipos o individuos heterocigotas, como se detalla a continuación. La presencia de haplotipos distintos a los presentes en las líneas control se determinó cuando las mezclas control/incógnita no mostraron un patrón de homodúplex para ninguno de los haplotipos ensayados, pero los controles homodúplex mostraron el patrón esperado de tipo homodúplex. Los individuos heterocigotas se determinaron cuando tanto las mezclas heterodúplex como el control homodúplex dieron un patrón de tipo heterodúplex.

Tanto para los individuos heterocigotas, como para los que presentaron haplotipos nuevos, los productos de PCR se secuenciaron como se describe más adelante.

Secuenciación

Los productos de PCR se purificaron y secuenciaron de acuerdo a lo descrito en el Cap. 1. Los cromatogramas se analizaron con el programa BioEdit Sequence

Alignment Editor v7.0.9.0 (Hall 1999). Las posiciones heterocigotas se determinaron por análisis de los cromatogramas.

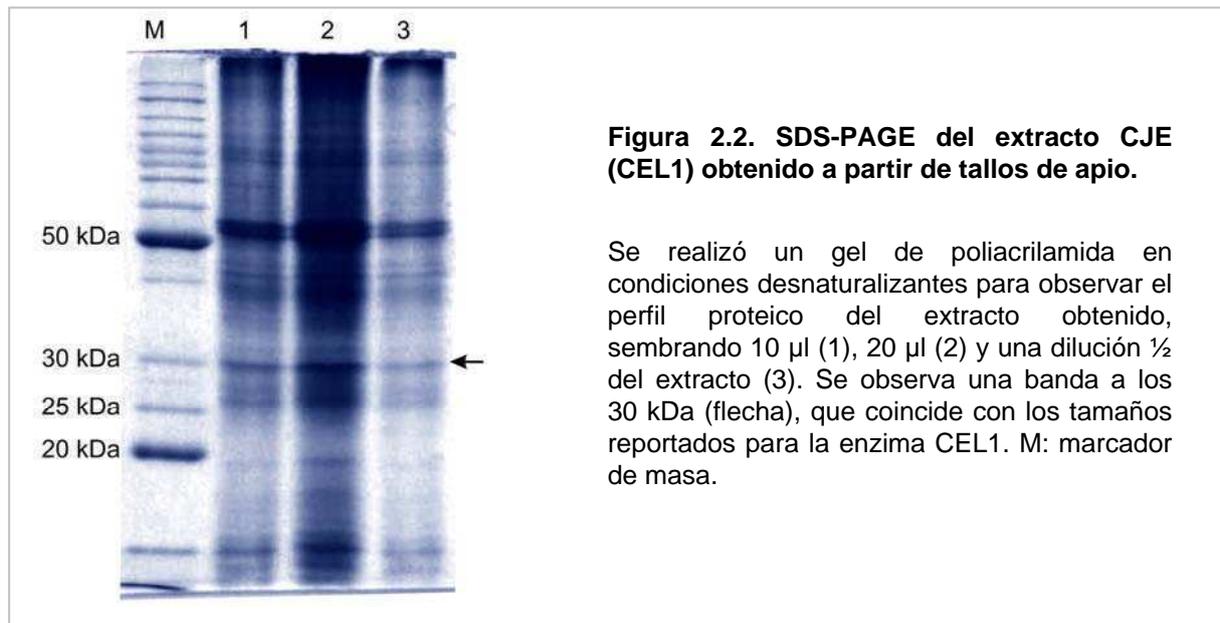
Construcción del mapa genético

Los polimorfismos genotificados en la población de 90 RILs mediante dHPLC se agregaron al mapa disponible, derivado del cruzamiento entre RHA266 X PAC2 (Poormohammad Kiani y col. 2007a), usando el programa CarthaGene 0.999 (Schiex y Gaspin 1997). Se utilizó la prueba de Chi-cuadrado para el análisis de la segregación de cada *locus*. La función de mapeo de Kosambi (1944) se utilizó para calcular las distancias del mapa (cM) a partir de las frecuencias de recombinación. Las frecuencias de recombinaciones estimadas en las poblaciones de mapeo RILs fueron corregidas por el factor que considera meiosis múltiples $r=R/(2-2R)$, donde $R=nr/(nr + nn)$ es la frecuencia de recombinación y nr y nn son el número de genotipos recombinantes y no recombinantes, respectivamente (Haldane y Waddington 1931). Los *loci* fueron agrupados dentro de cada grupo usando un valor mínimo de LOD (*likelihood odds*) score de 4,0 y una frecuencia de recombinación máxima de 0,35. Una vez definidos los grupos de ligamiento que contenían los marcadores funcionales, el orden específico de cada marcador dentro de cada grupo de ligamiento fue analizado de manera individual usando el programa Mapmaker 3.0 (Lander y col. 1987). El programa Mapchart v2.1 (Voorrips 2002) se usó para realizar la representación gráfica de los grupos de ligamiento y las ubicaciones de los marcadores funcionales.

Resultados

Método CEL1CH

La enzima endonucleasa CEL1 utilizada en este trabajo se obtuvo en el laboratorio, como extracto crudo (CJE) a partir de tallos de apio frescos. El CJE obtenido se corrió en un gel de poliacrilamida para proteínas en condiciones desnaturalizantes (SDS-PAGE) para evaluar la composición del mismo (Figura 2.2).



La optimización del método CEL1CH consistió en probar, por un lado, la cantidad mínima de ADN que permitiera observar el patrón de fragmentos con una buena señal, y por el otro, el volumen de CJE necesario para lograr un corte eficiente y específico de las moléculas heterodúplex. La detección de las moléculas homo- y heterodúplex se evaluó cualitativamente (Tabla 2.3, Figura 2.3). Se ensayó la cantidad total de ADN, incubando 100, 200 y 500 ng de la mezcla de los productos de PCR FAM-G/FAM-C con distintos volúmenes de CJE (0,2 a 5,0 µl, Tabla 2.3). Los patrones de digestión esperados se obtuvieron para las mezclas de 500 ng de ADN incubadas con 0,2; 0,5 y 1 µl de CJE. Para la mezcla de 100 ng de ADN incubada con 0,2 µl de CJE también se obtuvo el patrón de fragmentos correcto, pero la señal de los picos fue menor. Mayores volúmenes de enzima (2,0 y 5,0 µl de CJE) no digirieron correctamente la muestra (Tabla 2.3).

Tabla 2.3. Puesta a punto del volumen de CEL1 (CJE) necesario para lograr el corte de los heterodúplex.

CJE (μ l)	ADN total (ng)	Controles heterodúplex FAMC-FAMG		Detección del homodúplex
		5' 217 pb	5' 415 pb	632 pb
0,2	100	++	++	++
	200	+	-	+
	500	+++	+++	+++
0,5	100	+	-	-
	200	++	-	+
	500	+++	++	+++
1	100	-	-	-
	200	+	-	-
	500	++	+	++
2	100	-	-	-
	200	-	-	-
	500	+	-	-
5	100	-	-	-
	200	+	-	-
	500	-	-	-

El signo más (+) indica presencia es de picos. Una mayor señal se indica con cantidades mayores de signos (+). El signo menos (-) indica ausencia de picos. Se observan las condiciones que dieron el patrón de digestión esperado (en rojo).

Para examinar más exhaustivamente la cantidad de ADN a utilizar, habiendo determinado el volumen óptimo de CJE en 0,2 μ l, se incubaron mezclas de FAM-C/FAM-G y de RS16-Haplotipo 1/RS16-Haplotipo 2, conteniendo 250, 300, 400 y 500 ng de ADN (Figura 2.3). Si bien todas las cantidades probadas mostraron el patrón de digestión esperado, se eligió la cantidad de 300 ng de ADN porque fue la menor cantidad de ADN cuya digestión fue tan eficiente como la lograda usando la enzima comercial, mostrando la mejor relación señal/ruido (Figura 2.3). Por lo tanto, esta condición fue la utilizada para la posterior genotipificación con el método CEL1CH.

Un total de 13 *loci* correspondientes a los genes LIM, MADB-TF3, PGIP1, PGIP3, CAB, GIA, GPX, PSI-III-CAB, SCR1, RS16, LZP, GO y 1-ACCO cumplieron con los requerimientos necesarios para ser genotipificados mediante CEL1CH, de acuerdo al criterio establecido por el sistema comercial Surveyor Mutation Kit™ (Transgenomics Inc., EEUU). Estas consideraciones incluyen largo del fragmento, distancia entre los iniciadores y los SNPs, y distancia entre SNPs. Las regiones restantes no pudieron ser evaluadas mediante esta técnica porque no satisficieron alguno de dichos requerimientos.

La optimización de la técnica se realizó sobre el conjunto de las nueve líneas control

de genotipo conocido, amplificando las regiones mediante PCR con iniciadores marcados en el extremo 5' con los fluoróforos FAM y/o HEX. Todos los pares de iniciadores produjeron un único y abundante producto de PCR, excepto para los *loci* MADSB-TF3, PGIP1, PSI-III-CAB, y CAB. Para estos cuatro *loci* se realizó una segunda ronda de PCR utilizando sólo uno de los iniciadores con marca fluorescente.

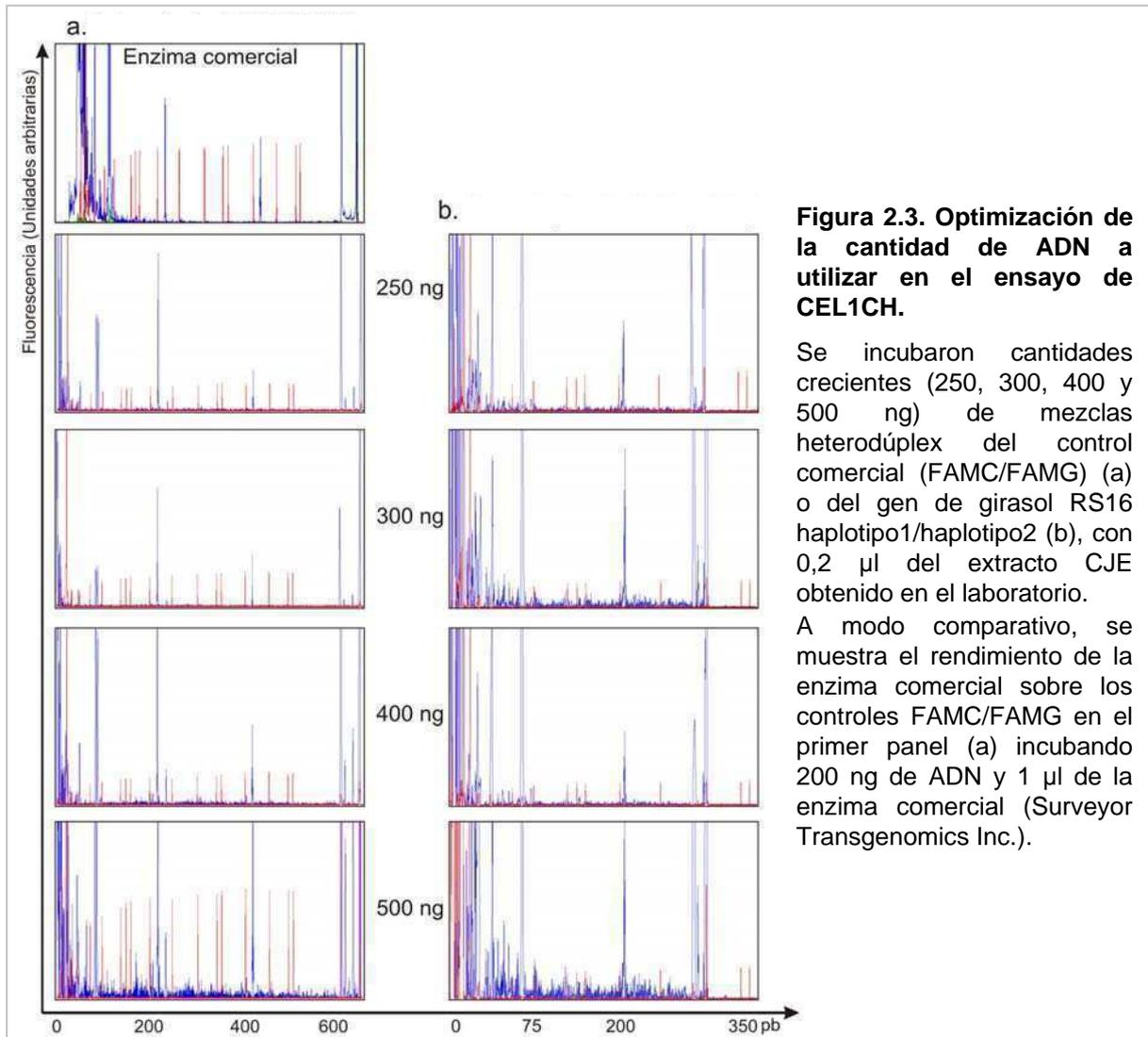


Figura 2.3. Optimización de la cantidad de ADN a utilizar en el ensayo de CEL1CH.

Se incubaron cantidades crecientes (250, 300, 400 y 500 ng) de mezclas heterodúplex del control comercial (FAMC/FAMG) (a) o del gen de girasol RS16 haplotipo1/haplotipo2 (b), con 0,2 μ l del extracto CJE obtenido en el laboratorio.

A modo comparativo, se muestra el rendimiento de la enzima comercial sobre los controles FAMC/FAMG en el primer panel (a) incubando 200 ng de ADN y 1 μ l de la enzima comercial (Surveyor Transgenomics Inc.).

Se obtuvieron producto de PCR específicos y con buen rendimiento para MADSB-TF3, PGIP1 y PSI-III-CAB; sin embargo, la amplificación de CAB fue escasa, por lo tanto se excluyó este *loci* del análisis subsiguiente.

De este modo, la técnica de CEL1CH fue capaz de detectar los polimorfismos satisfactoriamente en 10 de los 12 genes testados, pudiendo caracterizar los genotipos de todas las líneas control (Figura 2.4). Estos genes, constituyen un conjunto representativo de las diferentes situaciones que pueden encontrarse en el

genoma de girasol, esto es, regiones con un único SNP (MADSB-TF3 y PSI-III-CAB), dos SNPs (PGIP3, RS16, GIA y SCR1), un indel de ocho pb (LZP) o un patrón de polimorfismos complejo (PGIP1, 1 ACCO, LIM) (Figura 2.4).

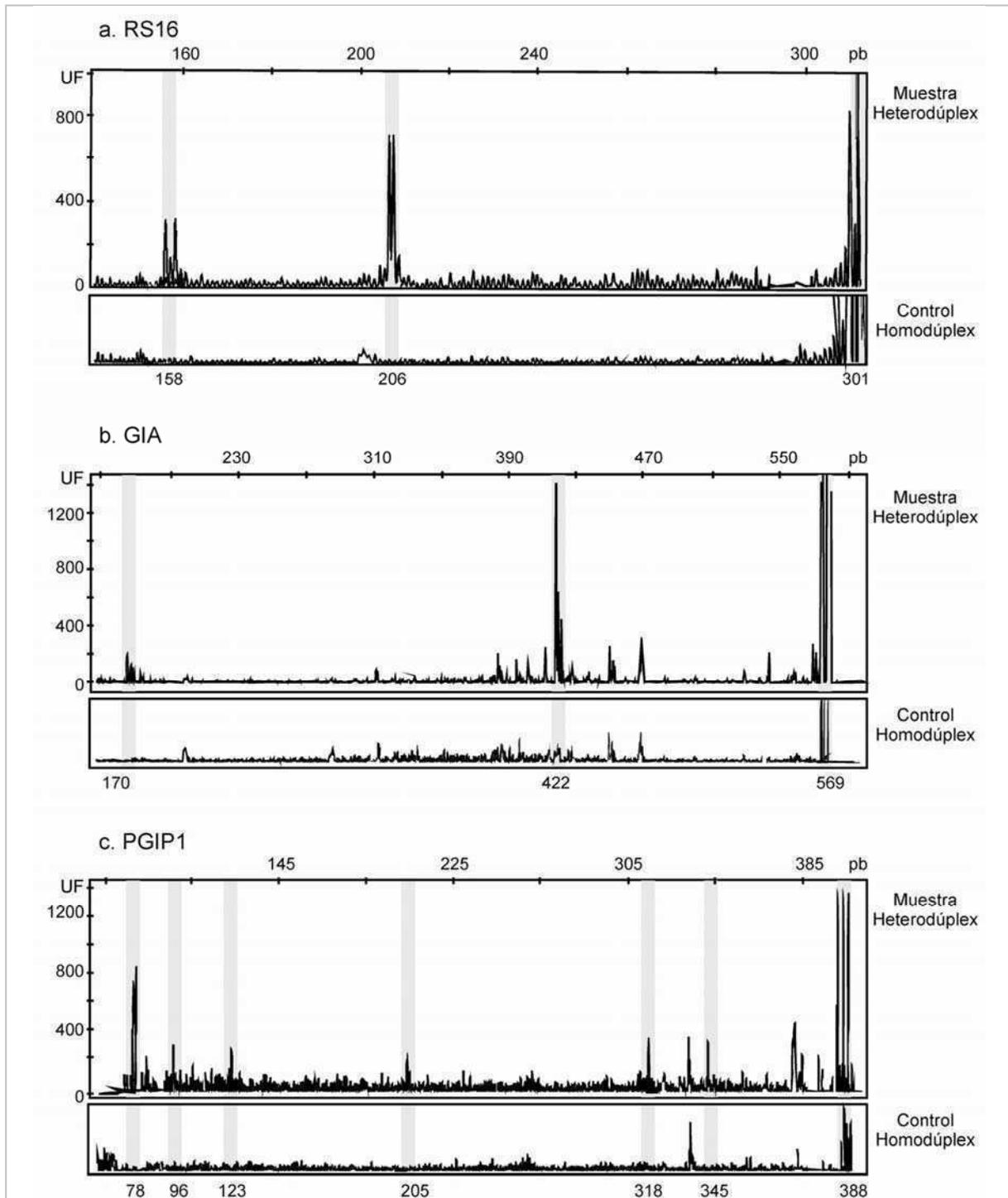


Figura 2.4. Detección de SNPs en genes de girasol mediante la técnica CEL1CH.

Perfiles electroforéticos para los genes RS16 (a), GIA (b), PGIP1 (c), PGIP3 (d), SCR1 (e), LZP (f), PSI-III-CAB (g), LIM (h), 1-ACCO (i) y MADSB-TF3 (j). Para cada región se observa el perfil de la muestra heterodúplex en el panel superior, y del control homodúplex en el panel inferior. Las barras grises indican los tamaños de los fragmentos marcados esperados por corte de los heterodúplex en los sitios variables (SNPs) y/o el tamaño de los fragmentos homodúplex. El eje Y indica unidades de fluorescencia (UF), el eje X indica distancias en pb.

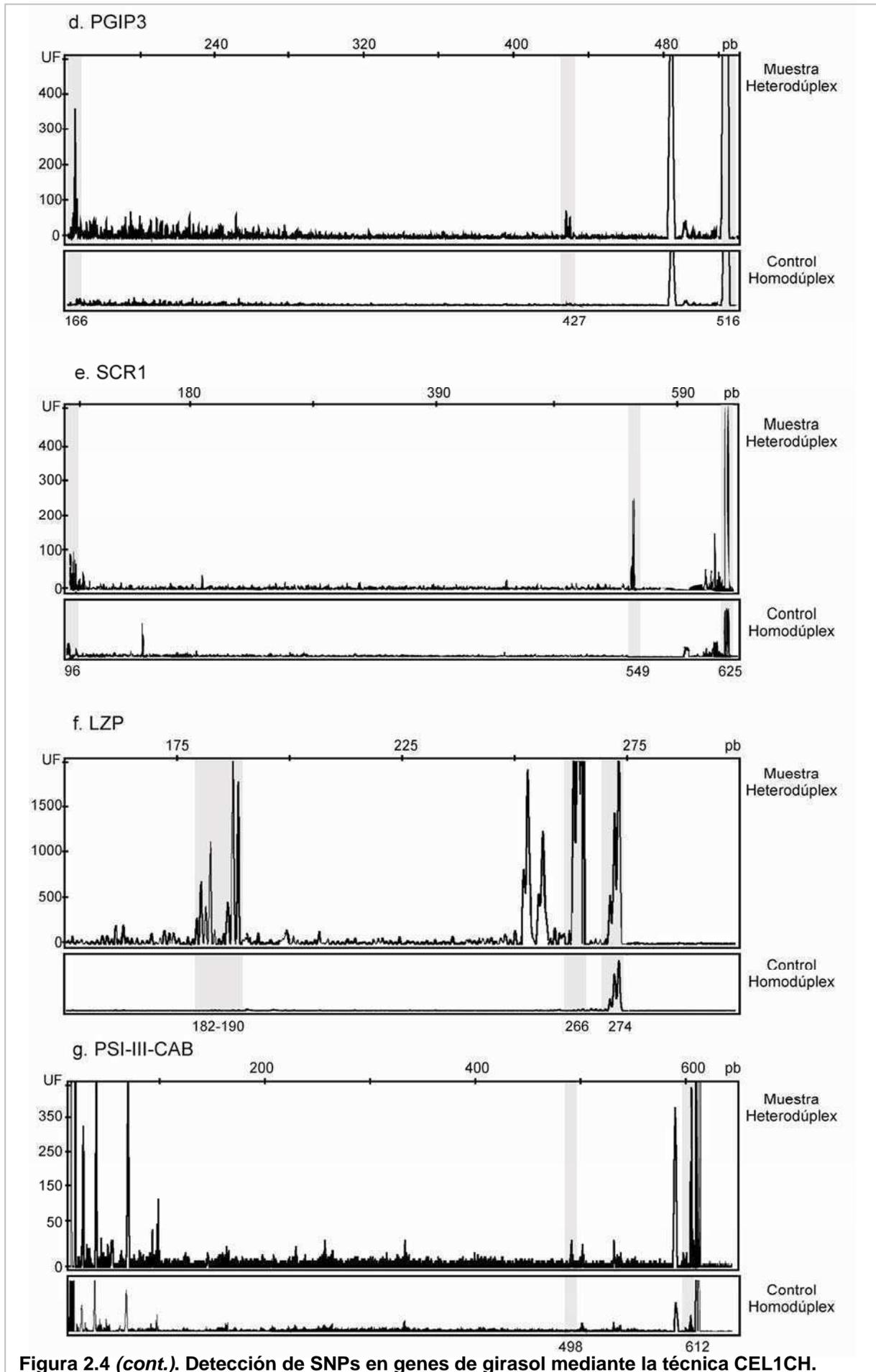
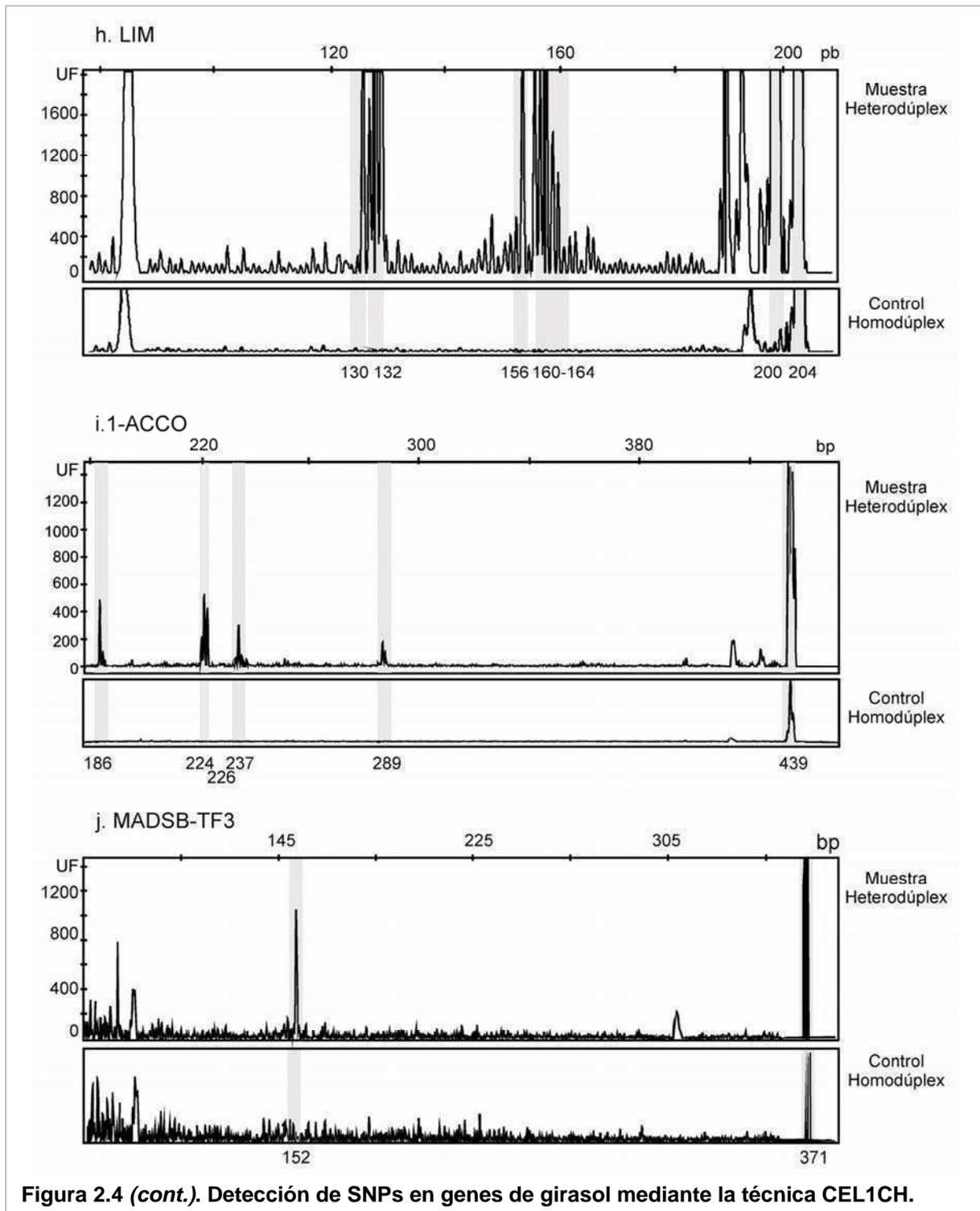


Figura 2.4 (cont.). Detección de SNPs en genes de girasol mediante la técnica CEL1CH.



Es interesante mencionar que si bien CJE digirió adecuadamente los heterodúplex para el *locus* GO, el polimorfismo detectado para este gen consiste en una inserción de 36 pb, lo que permite genotipificar este *locus* a través de la discriminación directa del tamaño de fragmentos mediante electroforesis capilar, de manera similar a la genotipificación de SSRs. Por otro lado, para la región GPX, la incubación con CJE no

mostró el patrón esperado de fragmentos, siendo el único *locus* para el cual la CJE falló en detectar una inserción de seis pb.

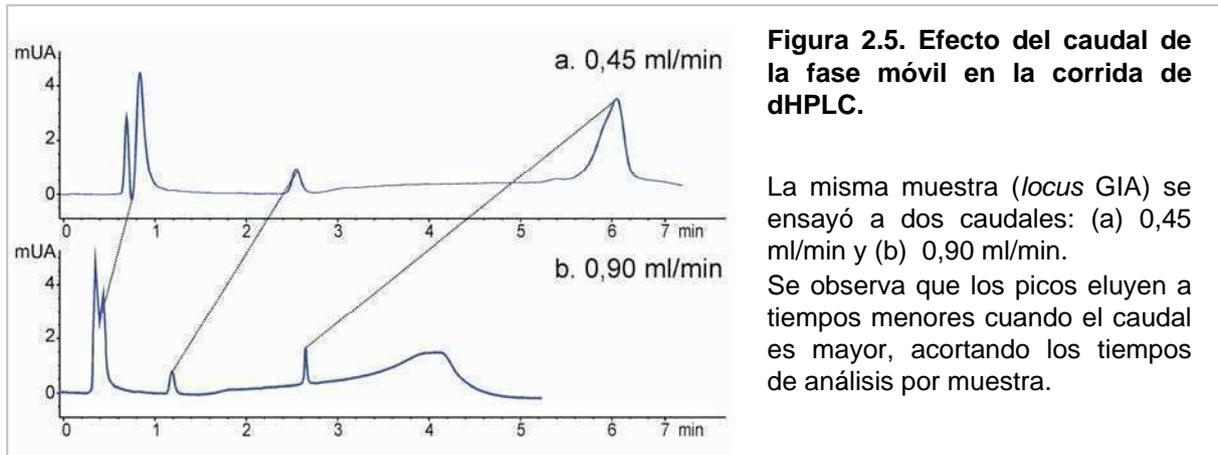
El largo de los fragmentos identificados luego de la electroforesis capilar fue el esperado de acuerdo a las posiciones de los sitios no apareados. Los controles homodúplex, por su parte, sólo mostraron un pico que se correspondió con el largo total del fragmento (Tabla 2.1, Figura 2.4). Aunque para la mayoría de los casos se utilizaron ambos iniciadores marcados con fluoróforos, algunas regiones fueron amplificadas utilizando sólo un iniciador marcado, detectando claramente los polimorfismos correspondientes.

Método dHPLC

El método de dHPLC fue evaluado sobre un total de 20 genes de los 24 seleccionados en primera instancia (Tabla 2.1). Los genes RS16, LZP no fueron incluidos en el análisis debido a que ya habían sido evaluados y genotipificados satisfactoriamente mediante CEL1CH. Para el gen 1-ACCO, las condiciones de temperatura y gradiente no pudieron predecirse por la presencia de indeterminaciones en la secuencia, mientras que, como se mencionara previamente, el gen GO pudo caracterizarse por una técnica más simple que el análisis de heterodúplex. Todos los pares de iniciadores ensayados tuvieron un buen rendimiento, excepto uno de los pares del *locus* PGIP1 (fragmento de 511 pb, Tabla 2.1). Por lo tanto, el *locus* PGIP1 sólo se evaluó mediante el fragmento de 301 pb (Tabla 2.1).

Para el análisis cromatográfico, se utilizó un equipo de HPLC de las características descritas anteriormente. Se ensayaron dos caudales de fase móvil para determinar el tiempo de corrida óptimo. La primera serie de experimentos se corrieron a un caudal de 0,45 ml/min. Estas corridas duraron 11 min y la elución de los fragmentos se evidenció a partir de los 6-7 min (Figura 2.5). Posteriormente se aumentó el caudal de la fase móvil a 0,9 ml/min, esto condujo a una elución más rápida, sin pérdida de resolución. En consecuencia, los tiempos de corrida se redujeron de 11 a 5 min, incluyendo el tiempo de equilibrio de la columna para una nueva corrida (Figura 2.5). El protocolo de optimización en dHPLC también involucró el ajuste de la concentración de ADN y el volumen de muestra a inyectar en la columna. Este

análisis permitió definir 100 ng como la cantidad mínima de ADN a inyectar, en un volumen de 5 a 10 μl (alrededor de 20 ng/ μl).



La temperatura y el gradiente óptimos para lograr una buena separación difirieron para cada *locus* ensayado de acuerdo a la composición nucleotídica de cada uno, pero todos fueron optimizados ajustando la temperatura entre los 54°C y 61°C (Tabla 2.4). La combinación de parámetros evaluados permitió la discriminación de las moléculas homo- y heterodúplex para los genes CPSI, GcVT, AALP, SCR2, LIM, MADSB-TF3, CAB, SCR1, LHCP, CAM y PGIC (Tabla 2.4). Al igual que para CEL1CH, los genes genotipificados mediante dHPLC presentaron distintas cantidades de SNPs y/o indels, es decir, un SNP (SCR2, MADSB-TF3), dos SNPs (CPSI, CAB), tres SNPs (GcVT, SCR1) y SNPs e indels (AALP, LIM, PGIC y CAM). Para los nueve genes restantes, no fue posible distinguir diferencias en los tiempos de retención de los dos tipos de moléculas con ninguna de las condiciones ensayadas (SEM, PSI-III-CAB, GIA, GPX, PGIP1, PGIP3, ANT, NsLTP and GAPDH). Por otro lado, tanto para SCR1 como para CAM, se precisaron dos combinaciones de gradiente y solución B para revelar todos los genotipos incluidos en el análisis.

Tabla 2.4. Condiciones de temperatura y gradiente de la solución B ensayadas para la detección de heterodúplex mediante dHPLC.

GEN	Temperatura de disociación de la molécula (°C)	Temperatura de dHPLC (°C)	%B en el tiempo (0,0-0,5-3,0 min) ^a	Caudal (ml/min)
CPSI	55,5	53	50,0-52,7-58,1	0,9
		54	50,0-53,7-59,1	0,9
		54	50,0-53,2-58,6	0,9
		55	50,0-51,4-56,8	0,9
SEM	57,0	58	50,0-51,8-57,2	0,9
		58	50,0-54,2-56,6	0,9
GcVT	61,3	60	50,0-52,3-57,7	0,9
		59	50,0-53,7-59,1	0,9
		58	50,0-54,9-60,3	0,9
AALP	59,5	60	50,0-52,8-58,2	0,9
SCR2	59,5	54	50,0-59,7-65,1	0,9
		56	50,0-54,5-59,9	0,9
LIM	56,0	58	50,0-51,7-57,1	0,9
		52	50,0-58,8-64,2	0,9
		57	50,0-54,6-60,0	0,9
MADSB TF3	56,0	57	50,0-53,2-58,6	0,9
		56	50,0-55,7-61,1	0,9
PGIP1 (301 pb)	57,0	56	50,0-54,3-59,7	0,9
		54	50,0-58,7-64,1	0,9
		58	50,0-53,2-58,6	0,9
		56	50,0-54,6-60,0	0,9
PGIP3	57,0	52	50,0-61,4-66,8	0,9
		54	50,0-67,8-73,2	0,9
		57	50,0-53,2-58,6	0,9
		56	50,0-65,0-70,4	0,9
CAB	59,3	59	50,0-60,9-66,3	0,9
		59	50,0-50,1-55,5	0,9
		57	50,0-52,8-58,2	0,9
GIA (365 pb)	60,5	56	50,0-54,2-59,6	0,9
		60	50,0-56,2-61,6	0,45
		58	50,0-58,6-64,0	0,45
		59	50,0-57,2-62,6	0,45
GIA (308 pb)	59,0	62	50,0-53,0-58,4	0,45
		60	50,0-51,8-57,2	0,9
		59	50,0-53,2-58,6	0,9
		58	50,0-54,6-60,0	0,9
GPX	56,0	56	50,0-57,3-62,7	0,9
		60	44,2-49,2-54,6	0,9
		58	50,0-53,3-58,7	0,9
PSI-III-CAB SCR1	51,0	56	50,0-56,0-61,4	0,9
		52	50,0-55,6-61,0	0,9
		60	50,0-67,4-72,8	0,45
LHCP	60,0	59	50,0-67,4-72,8	0,45
		59	50,0-64,2-68,6	0,9
		61	50,0-61,4-65,8	0,9
LHCP	60,0	60	50,0-56,7-62,1	0,9
		58	50,0-59,4-64,8	0,9

GEN	Temperatura de disociación de la molécula (°C)	Temperatura de dHPLC (°C)	%B en el tiempo (0,0-0,5-3,0 min) ^a	Caudal (ml/min)
ANT	59,5-59,8	60	50,0-50,7-56,1	0,9
		58	50,0-53,4-58,8	0,9
		57	50,0-54,8-60,2	0,9
		58	50,0-54,8-60,2	0,9
NsLTP	53,5	61	42,7-47,7-53,1	0,9
		58	50,0-51,1-56,5	0,9
		55	50,0-55,9-61,3	0,9
		56	50,0-54,5-59,9	0,9
		57	50,0-53,2-58,6	0,9
		54	50,0-57,3-62,7	0,9
CAM	55,5	54	50,0-65,3-70,7	0,9
		55	50,0-64,2-69,6	0,9
		56	50,0-62,6-68,0	0,9
		57	50,0-61,4-66,8	0,9
		60	50,0-58,5-63,9	0,9
PGIC	55,0	54	50,0-52,4-57,8	0,9
		55	50,0-54,9-60,3	0,9
		56	44,6-49,5-55,0	0,9
		53	50,0-52,1-57,5	0,9
GAPDH	57,5 - 58,0	54	50,0-59,4-64,8	0,9
		56	50,0-55,2-60,6	0,9
		58	50,0-53,9-59,3	0,9
		60	50,0-51,1-56,5	0,9

a Condiciones óptimas de detección diferencial de las moléculas homo- y heterodúplex (negrita y subrayado).

Con respecto a los perfiles de elución de las muestras heterodúplex, para LIM y CPSI pudieron observarse los patrones esperados de cuatro picos (dos picos heterodúplex y dos picos homodúplex, Figura 2.6a y b). Las muestras correspondientes a SCR1 y GCvT resolvieron tres picos (Figura 2.6c y d). Para el resto de las regiones se observaron sólo dos picos para las muestras heterodúplex (PGIC, AALP, CAM, CAB, PGIC, LHCP, Figura 2.6e-i). De acuerdo al fundamento de la técnica, los picos que eluyen primero corresponden a las moléculas heterodúplex, mientras que los que eluyen más tarde se corresponden con las moléculas homodúplex. De hecho, los perfiles cromatográficos de los controles homodúplex mostraron un solo pico, cuyo tiempo de retención se correspondió con el de los picos que eluyeron a tiempos mayores en las muestras heterodúplex. Particularmente para SCR2 y MADB-TF3, los perfiles de los controles homodúplex resultaron inesperados, presentando cada uno un pico adicional (Figura 2.6j y k). El tiempo de retención de este pico fue menor que el tiempo de retención de los picos heterodúplex, por lo tanto no fue interpretado como una molécula heterodúplex.

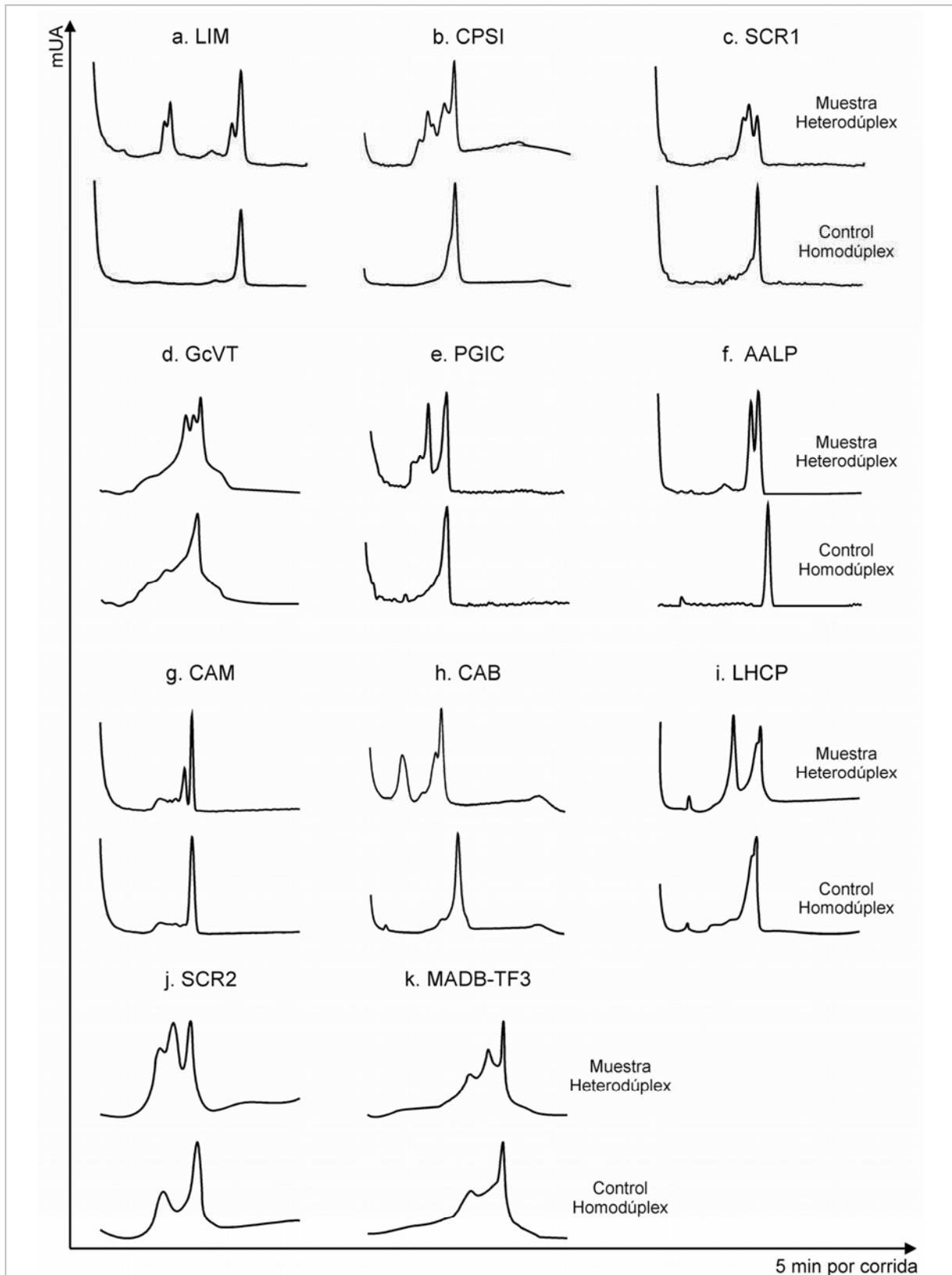


Figura 2.6. Detección de SNPs en genes de girasol mediante la técnica dHPLC.

Cromatogramas óptimos de los genes: LIM (a), CPSI (b), SCR1 (c), GcVT (d), PGIC (e), AALP (f), CAM (g), CAB (h), LHCP (i), SCR2 (j) y MADSB-TF3 (k). Muestras heterodúplex para cada gen (cromatograma superior) y controles homodúplex (cromatograma inferior). El eje Y se expresa en miliunidades de Absorbancia (mUA) y el eje X en tiempo (min).

Pruebas de desempeño y escalabilidad de las técnicas CEL1CH y dHPLC

Para realizar una prueba de desempeño, escalabilidad y rendimiento de ambas técnicas, se seleccionaron cinco *loci* para CEL1CH (RS16, 1-ACC0, PGIP1, MADSB-TF3 y GIA) y siete *loci* para dHPLC (CPSI, CAB, AALP, SCR1, LIM, PGIC y CAM). Estos 12 *loci* se utilizaron para:

- (1) caracterizar un grupo de 22 accesos de girasol cultivado (líneas endocriadas de genotipo desconocido),
- (2) genotipificar la población de RILs e incorporar marcadores funcionales al mapa saturado de girasol cultivado.

En la Tabla 2.5 se indican los genes evaluados, las longitudes de los productos de amplificación, las posiciones variables de acuerdo al fragmento de PCR amplificado y la identificación de los haplotipos.

Si bien en las secuencias nucleotídicas de las líneas control se encontraron, en algunos casos, sitios polimórficos adicionales (RS16: dos SNPs, 1-ACCO: cinco SNPs, PGIP1: cuatro SNPs, Tabla 2.1), se muestran sólo aquellos SNPs y/o indels caracterizados en los accesos de genotipo desconocido (Tabla 2.5).

Sólo la técnica de dHPLC fue seleccionada para la incorporación de marcadores funcionales al mapa de girasol cultivado (SCR1, LIM, PGIC, CAM), y los haplotipos de las líneas parentales se identifican como 1 (RHA266) y 2 (PAC2) (Tabla 2.5).

Tabla 2.5. Genes seleccionados para evaluar el desempeño, la escalabilidad y el rendimiento de las técnicas CEL1CH y dHPLC.

Método	Largo ^b	Haplotipos evaluados ^c	ID Haplotipos
CEL1CH			
RS16	301	T A 206	1 2
1ACC0	450	T..G..T..C A...G..T 224 237 226 289	1 2
PGIP1	388	G..C..G C..T..A 78 345 318	1 2
MADSB-TF3	371	C G 152	1 2
GIA	569	T..C C..G 170 422	1 2
dHPLC			
CPSI	183	T..C C..G 107 150	1 2
CAB	191	C..C T..T T..C 87 101	1 2 3
AALP	287	A..C..T..A..A..T..T..GGTTA..AATGTG..C..G..G G..C..C..G..G..T..C..GGTTA..AATGTG..C..G..G A..C..T..A..A..T..T..-----AATGTG..T..G..G G..T..T..A..A..C..T..GGTTA..-----C..A..C 47 95 137 155 193-198 201 65 129 151 161-165 200 206	1 2 3 4
SCR1	812	G..A..C A..T..C A..A..A 108 751 561	1 ^b 2 3
LIM	324	G..T..-..C..A..A..G..---- G..G..-..T..G..T..A..---- A..T..T..C..G..T..A..TAAT 33 87 241 267 76 194 243 269-272	1 2 3
PGIC	204	A..A..A..T G..C..A..- A..C..G..- 109 151 117 169	1 2 3
CAM	642	T..C..A..G..G..G..A..G..T..T..T..T..G..C..T..TTAC..C C..T..G..A..A..A..C..A..C..A..A..C..C..T..G..C..----..A 21 64 140 148 165 277 361 437 467-480 58 136 141 161 212 354 413 453 560	1 2

a Los números debajo de las bases nucleotídicas indican la posición de los sitios polimórficos en el alineamiento de todos los haplotipos evaluados.

b Para los *loci* SCR1, LIM, PGIC y CAM, el haplotipo 1 corresponde al presentado por la línea parental RHA266 y el haplotipo 2 corresponde al de la línea parental PAC2, los otros haplotipos descriptos se encontraron dentro del conjunto de las 9 líneas control secuenciadas en el Cap. 1.

Caracterización genotípica de accesos del banco de germoplasma mediante CEL1CH y dHPLC

Los resultados correspondientes a las variantes alélicas de los 22 accesos de girasol analizados se exponen en la Tabla 2.6. Pudieron determinarse un total de 110 genotipos mediante CEL1CH y 131 genotipos mediante dHPLC, habiendo quedado indeterminado el genotipo de HAR4 para el *locus* LIM.

Si se observa en detalle la Tabla 2.6, se aprecia que las 22 líneas caracterizadas para el *locus* RS16 sólo presentaron un sitio variable en la posición 206, mientras que los sitios polimórficos en las líneas control secuenciadas fueron dos (Tabla 2.1). En este caso, uno de estos SNPs resultó monomórfico para los 22 accesos genotipificados. Por otro lado, a pesar de que las regiones PGIP1 y 1-ACCO poseían cuatro y cinco sitios variables de acuerdo a los datos de secuencia de las líneas control, los 22 accesos sólo pudieron ser caracterizados para tres y cuatro variantes respectivamente, debido a una menor eficiencia en el corte de un sitio no apareado para cada uno de estos *loci*.

Tanto la técnica de CEL1CH como la técnica de dHPLC fueron capaces de detectar individuos heterocigotas (HA336, HAR4, R432 y HA853), ya que en ellos se forman moléculas heterodúplex sin necesidad de mezclar el producto de PCR con el de otro individuo (Tabla 2.6). Las moléculas correspondientes a los individuos heterocigotas fueron digeridas por la enzima CJE, mostrando un patrón similar al de las muestras heterodúplex (RS16, MADSB-TF3). En el caso de dHPLC, los individuos heterocigotas exhibieron un perfil de tipo heterodúplex sin haber sido mezclados con ningún control (SCR1 y PGIC). Todos los genotipos heterocigotas fueron confirmados por secuenciación directa y análisis de los cromatogramas, observando dobles picos en las posiciones de los SNPs respectivos (Figura 2.7).

Por otro lado, los controles homodúplex de las líneas MP557, P94, F164B y HA64 para los genes CAB, LIM y SCR1 mostraron el perfil esperado, pero sus genotipos no pudieron determinarse, ya que formaron moléculas heterodúplex con todos los haplotipos enfrentados. Estos casos también fueron analizados por secuenciación directa, encontrándose nuevos haplotipos (Figura 2.8). Estos SNPs y/o indels inéditos aportan nuevos datos sobre la variabilidad de estos genes en el germoplasma analizado.

Tabla 2.6. Genotipos de las 22 líneas de girasol identificados a través de CEL1CH y dHPLC.

Método	CEL1CH									dHPLC																																				
	Genes	RS16	1-ACCO			PGIP3	MADS B-TF3	GIA	CPSI	CAB	AALP							SCR1			LIM				PGIC																					
Posición en pb	206	224	226	237	289	78	315	345	152	170	422	107	150	87	101	47	65	95	129	137	151	155	161-165	163-198	200	201	206	108	561	751	33	76	87	194	241	243	267	269-272	109	117	151	169				
MP557	T	T	G	T	C	G	C	G	C	C	G	T	C	t	c ^a	A	C	T	A	A	T	T	D ^b	I	T	G	G	G	A	C	G	T	D	C	A	A	G	D	A	C	G	T				
HA336	<u>AT</u> ^c	T	G	T	C	C	T	A	G	T	C	T	C	C	C	A	C	T	A	A	T	T	I	I	C	G	G	A	T	C	G	G	D	T	G	T	A	D	A	A	A	T				
F164B	T	A	-	G	T	G	C	G	C	C	G	C	G	T	C	A	C	T	A	A	T	T	D	I	T	G	G	G	A	C	g	t	d	c	a	a	a	i	A	A	A	T				
PAC1	T	T	G	T	C	C	T	A	G	C	G	T	C	T	T	A	C	T	A	A	T	T	I	I	C	G	G	a	t	c	G	T	D	C	A	A	G	D	A	A	A	T				
P94	A	T	G	T	C	C	T	A	G	C	C	T	C	T	T	G	C	C	G	G	T	C	I	I	C	G	G	A	T	C	G	G	D	T	G	T	A	D	A	A	A	T				
CM307	T	A	-	G	T	G	C	G	G	C	G	T	C	T	T	A	C	T	A	A	T	T	D	I	T	G	G	G	A	C	G	T	D	C	A	A	G	D	A	C	G	T				
L156	T	T	G	T	C	G	C	G	G	C	G	T	C	C	C	A	C	T	A	A	T	T	I	I	C	G	G	G	A	C	A	T	I	C	G	T	A	I	A	A	A	T				
R432	T	T	G	T	C	G	C	G	C	T	C	T	C	T	T	A	C	T	A	A	T	T	I	I	C	G	G		<u>T/A</u>		G	G	D	T	G	T	A	D	G	C	A	D				
A71	A	T	G	T	C	C	T	A	G	C	G	T	C	T	T	A	C	T	A	A	T	T	I	I	C	G	G	G	A	C	G	G	D	T	G	T	A	D	A	A	A	T				
A59	T	T	G	T	C	G	C	G	C	C	G	C	G	T	T	G	T	T	A	A	C	T	I	D	C	A	C	A	A	A	A	T	I	C	G	T	A	I	A	A	A	T				
R419	A	T	G	T	C	C	T	A	G	C	G	T	C	T	T	A	C	T	A	A	T	T	I	I	C	G	G	G	A	C	G	G	D	T	G	T	A	D	A	A	A	T				
R417	T	T	G	T	C	G	C	G	G	C	G	T	C	T	T	A	C	T	A	A	T	T	I	I	C	G	G	G	A	C	G	T	D	C	A	A	G	D	A	A	A	T				
R423	A	T	G	T	C	C	T	A	C	C	G	T	C	T	T	A	C	T	A	A	T	T	I	I	C	G	G	G	A	C	G	T	D	C	A	A	G	D	G	C	A	D				
HA64	T	T	G	T	C	G	C	G	C	C	G	C	G	C	C	G	C	C	G	G	T	C	I	I	C	G	G	g	a	c	G	T	D	C	A	A	G	D	A	A	A	T				
HAR4	A	T	G	T	C	C	T	A	<u>C/G</u>	C	G	C	G	T	T	A	C	T	A	A	T	T	I	I	C	G	G		<u>T/A</u>													^d	A	A	A	T
RK416	A	A	-	G	T	G	C	G	C	C	G	T	C	T	T	A	C	T	A	A	T	T	D	I	T	G	G	G	A	C	A	T	I	C	G	T	A	I	A	A	A	T				
RHA275	T	T	G	T	C	C	T	A	G	C	G	T	C	C	C	A	C	T	A	A	T	T	D	I	T	G	G	G	A	C	G	T	D	C	A	A	G	D	A	A	A	T				
HA853	A	T	G	T	C	G	C	G	C	C	G	T	C	T	T	A	C	T	A	A	T	T	I	I	C	G	G	A	T	C	A	T	I	C	G	T	A	I	<u>G/A</u>	<u>C</u>	<u>A</u>	<u>D</u>				
HA441	T	T	G	T	C	C	T	A	G	C	G	T	C	T	T	A	C	T	A	A	T	T	I	I	C	G	G	G	A	C	G	T	D	C	A	A	G	D	A	A	A	T				
RHA440	A	T	G	T	C	C	T	A	G	C	G	T	C	T	T	G	C	C	G	G	T	C	I	I	C	G	G	A	A	A	G	G	D	T	G	T	A	D	A	A	A	T				
RHA439	T	T	G	T	C	C	T	A	C	T	C	T	C	T	T	A	C	T	A	A	T	T	D	I	T	G	G	A	A	A	G	G	D	T	G	T	A	D	A	A	A	T				
RHA274 ^b	T	T	G	T	C	C	T	A	G	C	G	T	C	T	T	A	C	T	A	A	T	T	D	I	T	G	G	A	A	A	G	T	D	C	A	A	G	D	A	A	A	T				

a Las letras minúsculas indican que se encontraron otros haplotipos para esas líneas en esos *loci*.

b Los indels fueron reemplazados por I (presencia de inserción) y D (ausencia de inserción), de acuerdo al haplotipo correspondiente descrito en la Tabla 2.5.

c Los genotipos subrayados identifican los *loci* heterocigotas para los accesos correspondientes.

d El *locus* LIM para la línea HAR4 fue el único haplotipo indeterminado.

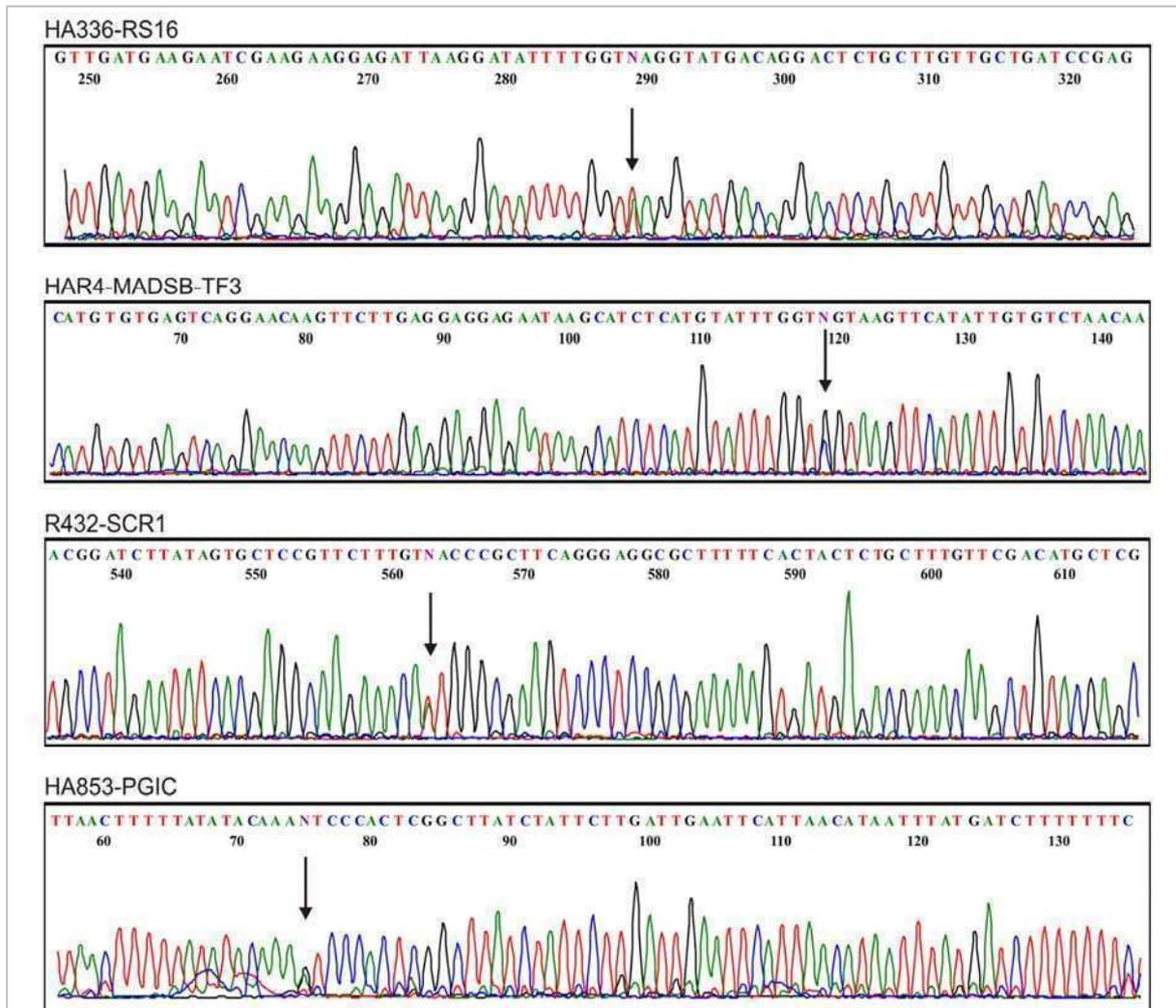
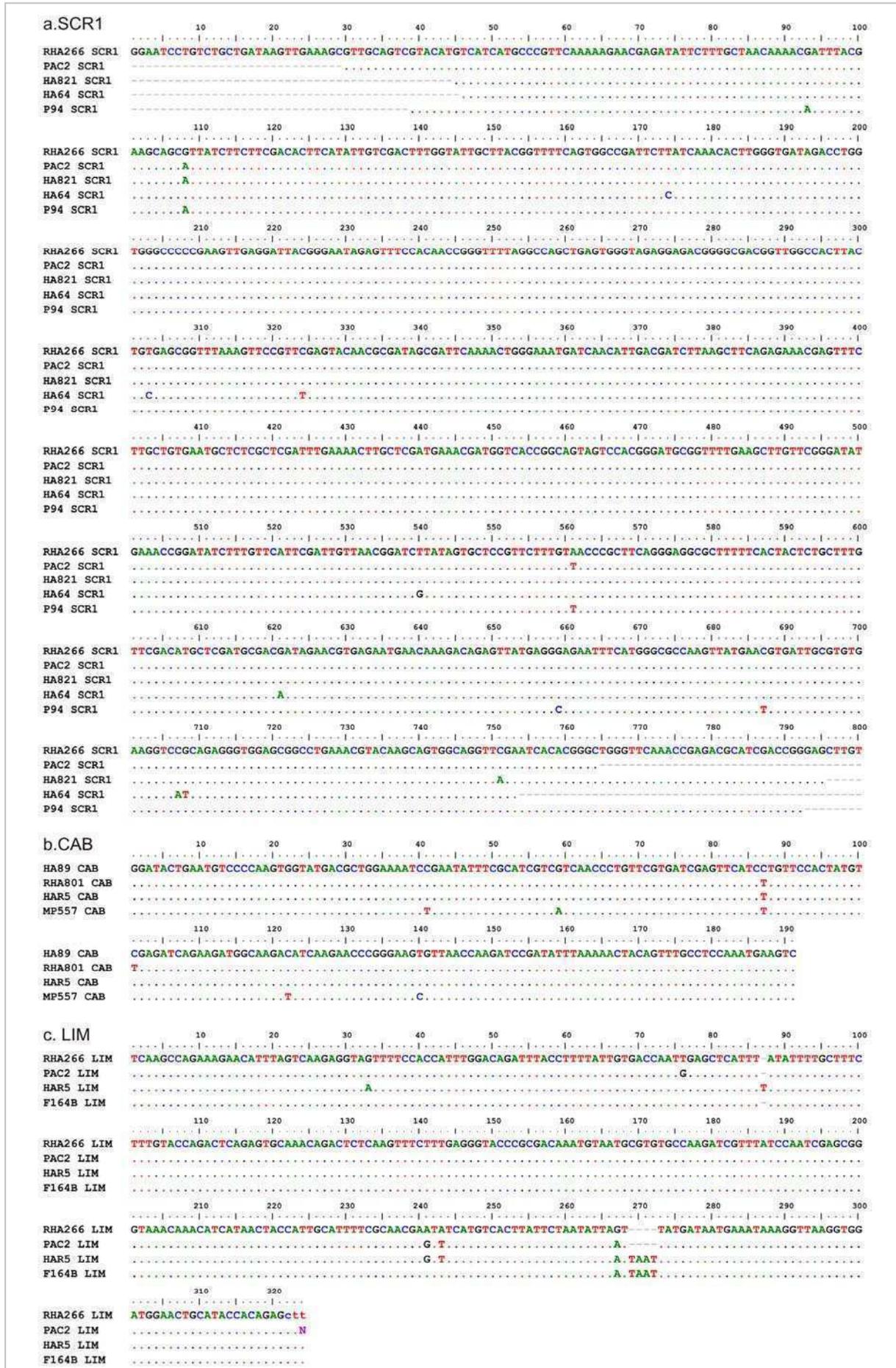


Figura 2.7. Individuos heterocigotas detectados con CEL1CH y dHPLC y confirmados por secuenciación directa.

Los productos de PCR de los individuos heterocigotas fueron secuenciados directamente. Electroferogramas obtenidos con el iniciador sentido (F) para los individuos HA336, HAR4, R432 y HA853 que se habían detectado como heterocigotas mediante CEL1CH (RS16 y MADSB-TF3) y dHPLC (SCR1 y PGIC). Las flechas indican las posiciones indeterminadas debido a la presencia de dos alelos para esos sitios.

Figura 2.8 (pág. sig.). Nuevos haplotipos detectados con dHPLC y confirmados por secuenciación directa.

Genotipos de las líneas control para los genes SCR1 (a), CAB (b) y LIM (c), junto con las líneas que evidenciaron un haplotipo distinto a los conocidos. Los puntos identifican posiciones idénticas o secuencia conservada. Se destacan las posiciones variables, aquellos SNPs no presentes entre las líneas control corresponden a los nuevos sitios encontrados.



Incorporación de marcadores funcionales al mapa de girasol cultivado

Cuatro genes candidatos analizados mediante dHPLC (SCR1, LIM, PGIC y CAM) y polimórficos para las líneas RHA266 y PAC2 se genotipificaron en 90 RILs de la población de mapeo y se incorporaron como marcadores funcionales al mapa saturado de girasol cultivado descrito previamente (Poormohammad Kiani y col. 2007a). Estos cuatro *loci* se ubicaron en distintos grupos de ligamiento (GL), como se muestra en la Figura 2.9. SCR1 pudo localizarse en el GL6, a 3,4 y 1,3 cM de los marcadores flanqueantes ORS381 y ORS1233, respectivamente. El *locus* PGIC pudo asociarse a la porción distal del GL9, a 5,8 cM del marcador ORS844. Por su parte LIM se insertó entre los marcadores HA3204 y HA2579 del GL10, a 1,2 y 2 cM, respectivamente. Finalmente, CAM se ancló en el GL13, entre dos marcadores relativamente distantes entre sí: E35M49_10 y HA4208, ubicándose a 6,2 y 7,5 cM de los mismos.

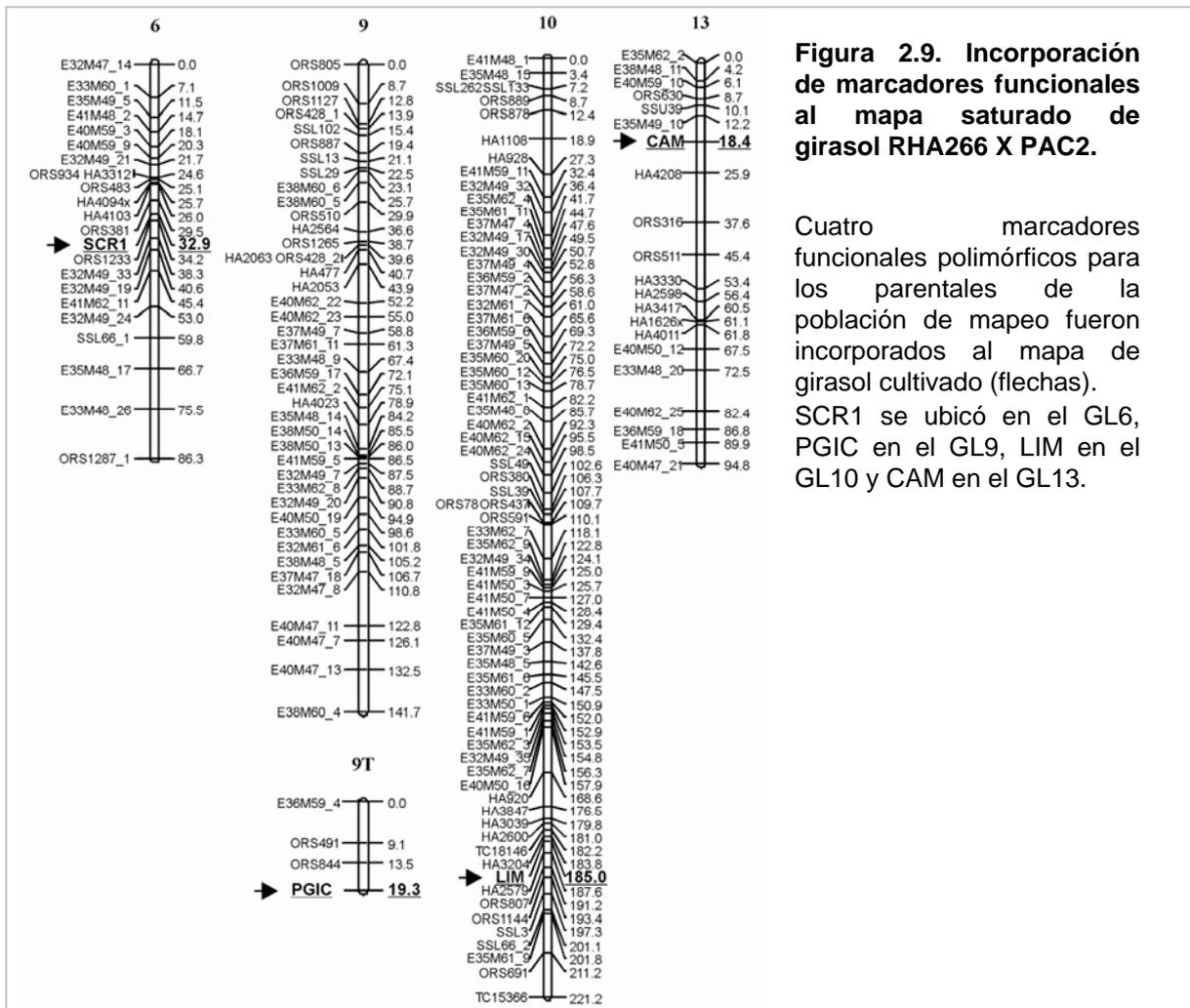


Figura 2.9. Incorporación de marcadores funcionales al mapa saturado de girasol RHA266 X PAC2.

Cuatro marcadores funcionales polimórficos para los parentales de la población de mapeo fueron incorporados al mapa de girasol cultivado (flechas). SCR1 se ubicó en el GL6, PGIC en el GL9, LIM en el GL10 y CAM en el GL13.

Discusión

Las técnicas de genotipificación CEL1CH y dHPLC presentadas en este capítulo han demostrado ser capaces de proporcionar, a los laboratorios de bajo presupuesto dedicados al mejoramiento molecular, una herramienta para incrementar la capacidad de genotipificación de SNPs en animales y plantas a un costo razonable.

Estas técnicas comenzaron a utilizarse hacia fines de los años 90, y con mayor intensidad en el último lustro, para estudios moleculares de plantas a través de la estrategia de “genética inversa”. Sin embargo, al momento de la realización de este trabajo, no existían datos sobre la aplicación de CEL1 en girasol, y solamente se encontraron dos trabajos en donde se utilizó la técnica de dHPLC para genotipificar SNPs o indels en este cultivo (Lai y col. 2005, Lexer y col. 2003). En los últimos años, el número de marcadores funcionales disponibles para la caracterización del germoplasma y para estudios de asociación ha ido creciendo ininterrumpidamente (Kolkman y col. 2007, Liu y Burke 2006).

En teoría, dado un determinado gen candidato, varias regiones pueden ser blanco para el análisis con CEL1CH o dHPLC. En la práctica, sin embargo, la elección de la técnica a utilizar estará determinada por los requerimientos, las potencialidades y las desventajas de cada una. Generalmente, las consideraciones a tener en cuenta incluyen el número de sitios variables dentro de la región, el tamaño del producto de amplificación, el conocimiento de la secuencia completa y los costos económicos de cada método.

Método CEL1CH

La posibilidad de obtener la enzima CEL1 endonucleasa en el laboratorio reduce considerablemente el costo de la técnica CEL1CH, sin pérdida de sensibilidad del método. Según Galeano y col. (2009), los tallos de apio son una muy buena fuente de la enzima porque tienen baja cantidad de pigmentos interferentes. El CJE obtenido en este trabajo se utilizó para cortar las bases no apareadas de las moléculas heterodúplex, probando una eficiencia de magnitud similar a la evidenciada por su contraparte comercial (Tabla 2.3, Figura 2.3). La efectividad de CJE coincide con lo

encontrado por Oleykowski y col. (1998), quienes demostraron que CEL1 escindía las bases no apareadas del ADN, cortando en el extremo 3' de las mismas. Si bien CEL1 también posee actividad exonucleasa 5'-3' y podría digerir el producto de PCR por completo comenzando por el extremo marcado con fluoróforos (Barkley y Wang 2008), el tiempo de incubación de 15 min aquí determinado, evitó que la actividad CEL1 exonucleasa prevaleciera por sobre la actividad endonucleasa, impidiendo la pérdida de la señal fluorescente.

Otro punto importante a tener en cuenta en CEL1CH es el costo de los aparatos utilizados para la detección de los fragmentos fluorescentes, como el analizador de ADN LICOR®, cuyo uso es uno de los más difundidos. En este caso, los experimentos se realizaron utilizando una plataforma de electroforesis capilar automática, originalmente establecida para el análisis y genotipificación de SSRs a gran escala (Unidad de Genómica, IB-INTA Castelar). A pesar de que la metodología propone amplificar las regiones con ambos iniciadores marcados con fluoróforos, fue posible detectar los SNPs e indels con sólo un iniciador marcado, reduciendo el costo de esta etapa. Si bien la electroforesis capilar tiene un costo por corrida de 2,5 dólares estadounidenses (USD), éste puede reducirse mediante la estrategia *multiplex*, combinando la amplificación de regiones marcadas con fluoróforos distintos, que a su vez, al ser digeridos por CJE den patrones de corte diferenciales para cada una. La posibilidad de combinación dependerá, en última instancia, de las características moleculares de las regiones a estudiar.

A pesar del gran impacto que ha tenido el uso de CEL1 en las estrategias de “genética inversa”, se la ha utilizado muy poco en la evaluación de la diversidad presente en poblaciones naturales debido, probablemente, a las restricciones de la técnica en cuanto al tamaño de los fragmentos a amplificar, el número de SNPs por región y a las exigencias en el diseño de los oligonucleótidos iniciadores. No obstante, ha sido empleada con éxito en el análisis de la diversidad presente en ecotipos de *Arabidopsis thaliana* para cinco genes (Comai y col. 2004), haciendo que la metodología se expanda a unas pocas especies vegetales como melón, álamo, poroto y tomate (Gilchrist y col. 2006, Nieto y col. 2007, Yang y col. 2004). Por otro lado, la efectividad en la identificación de SNPs mediante la digestión de heterodúplex para

arroz (*Oryza sativa*) fue menor que la lograda por la técnica de PCR-RF-SSCP (PCR-Restriction Fragment - Single Strand Conformation Polymorphism), si bien Shirawasa y col. (2007) utilizaron una fuente distinta a la de apio para obtener la endonucleasa.

De los 31 genes que habían sido secuenciados en las 19 líneas de girasol (Cap. 1), 13 cumplieron con los requerimientos de la técnica CEL1CH (Tabla 2.1). Luego del ensayo utilizando las líneas control, pudieron establecerse las condiciones óptimas de identificación para 10 de estas regiones (Figura 2.4). El desempeño de CEL1CH fue satisfactorio, permitiendo la identificación de genotipos que variaban desde un sitio polimórfico en el gen MADSB-TF3, hasta un patrón de polimorfismos más complejo, como en el caso de los genes 1-ACCO, PGIP1 y LIM (Figura 2.4). Aunque la utilización de CEL1 hasta el momento se encuentra generalmente circunscripta a la detección de una única mutación sobre el gen de interés (Cooper y col. 2008, Nieto y col. 2007, Uauy y col. 2009), en este trabajo pudieron establecerse las condiciones para la detección simultánea de hasta cinco SNPs en una misma región, incluyendo indels de hasta ocho pb.

Barkley y Wang (2008) sugieren que las limitaciones impuestas sobre la longitud de los productos de amplificación y la dificultad de genotipificar SNPs ubicados en los extremos de las regiones de interés, pueden superarse analizando regiones que se solapan entre sí para asegurar la detección de todos los sitios polimórficos dentro del gen blanco. Aunque todos los genes fueron examinados para establecer regiones o iniciadores adecuados para esta técnica, algunos debieron ser descartados, aún después de analizar exhaustivamente todas las posibilidades. Sin embargo, para los genes MADSB-TF3 y PGIP1, que tenían más de 1000 pb de longitud, se diseñaron iniciadores en una región interna a fin de amplificar 371 y 488 pb, respectivamente. Estas regiones pudieron genotipificarse sin problemas. Por otro lado, algunas regiones se descartaron debido a que el tamaño del fragmento original era menor o muy cercano al límite inferior sugerido para la técnica (<200 pb, CPSI, RL41, PGIC). Otros cinco genes (SEM, GcVT, AALP, SCR2, ANT y LHCP) fueron descartados porque la distancia entre los iniciadores y los SNPs adyacentes era menor a la sugerida (75 pb como mínimo). En otros casos, si bien no era una condición limitante de la técnica, se consideró la distancia entre SNPs (GAPDH) y el gran número de

SNPs en la región (NsLTP, CAM) como un impedimento en el reconocimiento de todos los fragmentos digeridos por la enzima CJE. Por ello, estas regiones también fueron desestimadas para su genotipificación mediante esta técnica.

El método CEL1CH tradicional involucra la marcación fluorescente de ambos oligonucleótidos iniciadores, ya que la visualización de los fragmentos provenientes del corte de ambas hebras sirve como control y confirmación dentro del mismo experimento. No obstante, las amplificaciones realizadas con sólo un iniciador marcado no comprometieron la detección de los fragmentos digeridos, tal como se evidencia para los genes MADSB-TF3, PGIP1 y PSI-III-CAB (Figura 2.4).

La principal ventaja del método CEL1CH es que puede detectar la posición específica del SNP sin la utilización de compuestos químicos tóxicos. CEL1CH parecer ser una técnica excepcionalmente poderosa cuando la importancia radica en localizar la posición de los sitios variables más que la identificación de los haplotipos (conjunto de SNPs), como lo reportaron Hung y col. (2008) en la detección de las mutaciones puntuales involucradas en la enfermedad hereditaria beta talasemia.

A pesar de que la cantidad de ADN necesaria para realizar el experimento es una clara desventaja respecto a otras técnicas, un trabajo reciente describe la optimización de la cantidad de ADN para TILLING y EcoTILLING en papa, con cantidades que varían entre 25 y 0,5 ng (Elias y col. 2009).

Como se mencionó anteriormente, CEL1CH fue una metodología eficiente en la detección de los genotipos para 10 *loci*, y además, fue suficientemente sensible como para detectar individuos heterocigotas para los genes RS16 y MADSB-TF3 (Tabla 2.6, Figura 2.7). Es interesante mencionar que si bien las secuencias disponibles para los genes PGIP1 y 1-ACCO incluían cuatro y cinco SNPs en las regiones a amplificar de las líneas control, en los 22 accesos de girasol de genotipo desconocido sólo fue posible detectar tres y cuatro SNPs para cada gen, respectivamente. No obstante, los resultados demuestran que CEL1CH permite genotipificar independientemente los SNPs pertenecientes a un mismo fragmento, aún cuando alguno de los fragmentos digeridos no presente buena señal fluorescente como para ser asignado a un sitio SNP inequívocamente. Esto no es posible para otros métodos de genotipificación

como dHPLC, cuyo fundamento se basa en la detección de los sitios como un conjunto, es decir como un haplotipo. En este sentido, CEL1CH fue exitoso en la caracterización genotípica de los genes PGIP1 y 1-ACCO.

Método dHPLC

La técnica de dHPLC, ha sido utilizada con éxito en la determinación de mutaciones involucradas en algunas enfermedades de origen genético (Giordano y col. 1999, Wagner y col. 1999), la determinación de resistencia a los antibióticos en bacterias y comunidades microbianas (Randall y col. 2005), la evaluación de la diversidad nucleotídica en ecotipos de *A. thaliana* (Spiegelman y col. 2000), la caracterización de genes candidatos en girasol (Lexer y col. 2003) y el mapeo de SNPs en cebada y girasol (Kota y col. 2001, Kota y col. 2008, Lai y col. 2005).

El método de dHPLC presenta ventajas tales como la independencia del uso de iniciadores fluorescentes, la adaptabilidad a sistemas semi-automatizados, la alta reproducibilidad y sensibilidad y, fundamentalmente, brinda un criterio objetivo para la clasificación de la muestra (Abbas y col. 2004).

Varios estudios demuestran la mayor efectividad y especificidad de la técnica con respecto a otras técnicas de discriminación alélica como PCR-RFLP (*PCR-Restriction Fragment Length Polymorphism*), secuenciación directa y SSCP (Abbas y col. 2004, Jones y col. 1999, Yu y col. 2005). Además, Costabile y col. (2006) revisaron distintas metodologías utilizadas para el diagnóstico de la inmunodeficiencia primaria, enfatizando la alta precisión que posee dHPLC. Como se mencionó anteriormente, el método dHPLC identifica las variantes en su conjunto, es decir los haplotipos en lugar de SNPs individuales, lo que para algunos casos como los estudios de asociación genotipo-fenotipo, puede resultar mucho más informativo que los datos de marcadores aislados (Rafalski 2002). De los 31 genes secuenciados, 20 cumplían las condiciones para ser genotipificados por dHPLC, habiendo podido optimizarse la detección de 11 genes mediante esta técnica (Figura 2.5). Más aún, siete genes fueron seleccionados para caracterizar los genotipos de 22 accesos de girasol y/o agregar marcadores funcionales al mapa saturado de girasol cultivado, como pruebas de desempeño y escalabilidad de la técnica (Tabla 2.6, Figura 2.9). Es importante

mencionar que, con excepción de SCR1 y CAM, las diferencias entre todos los pares de haplotipos ensayados pudieron ser detectadas utilizando una única condición de temperatura-gradiente. Para las mezclas SCR1 Haplotipo 1/SCR1 Haplotipo 2 y SCR1 Haplotipo 1/SCR1 Haplotipo 3 la condición óptima de detección de los heterodúplex se consiguió a una temperatura de 59°C, mientras que para resolver la mezcla SCR1 Haplotipo 2/SCR1 Haplotipo 3 fue necesario elevar la temperatura a 61°C y ajustar el gradiente convenientemente (Tabla 2.4). Análogamente, dos de los siete haplotipos identificados por secuenciación para CAM pudieron distinguirse entre sí sólo cuando la temperatura se elevó a 60°C. La falta de resolución para la primera condición evaluada (54°C) podría adjudicarse a que la única diferencia que presentan estos dos haplotipos es un indel de ocho pb.

A pesar de que el tamaño del fragmento es una de las condiciones limitantes para dHPLC, otras consideraciones deben tenerse en cuenta para tener éxito con esta técnica. En primer lugar, debe conocerse la secuencia exacta y completa de la región en estudio, para poder predecir correctamente la temperatura y gradiente a la que resolverán las moléculas heterodúplex. Tal fue el caso del gen 1-ACCO, el cual no pudo evaluarse mediante esta técnica por el desconocimiento de algunas bases dentro del producto de amplificación. Esta condición no se aplica al método CEL1CH, ya que en ese caso, sólo es necesario conocer el tamaño y las posiciones de los SNPs. Otra consideración fundamental a tener en cuenta en la aplicabilidad del método es el número de haplotipos potenciales producidos por un número dado de SNPs, aunque este número de SNPs suele ser menos restrictivo que en el caso del método CEL1CH. Por ejemplo, para CAM, se encontraron siete haplotipos distintos (combinación de 20 SNPs y seis indels), por lo tanto, la caracterización de esta región implica someter a la línea de genotipo incógnita a la mezcla con siete líneas control. A pesar de que esta cantidad de SNPs dificultó la genotipificación de CAM sobre los 22 accesos de girasol del banco de germoplasma, este *locus* pudo ser caracterizado en las 90 RILs provenientes del cruzamiento de RHA266 X PAC2, dado que sólo dos de los haplotipos posibles estaban presentes en esos individuos. Sin embargo, esto no hubiera sido viable mediante CEL1CH, ya que el número implicado de SNPs y/o indels entre estos dos haplotipos (17 SNPs y un indel) es aún demasiado grande

como para obtener una digestión adecuada por parte de la enzima CEL1. Finalmente, es necesario aclarar que, si bien todos los parámetros pueden ser conocidos y todas las combinaciones de temperatura y gradiente probadas, para algunos *loci* pueden no alcanzarse las condiciones que permitan resolver las moléculas homo- y heterodúplex, tal como ocurrió con el *locus* PGIP1. La falta de resolución entre los dos tipos de moléculas pudo deberse, en este caso, a que éstas poseen una temperatura de disociación similar asociada a una variante T/A como única diferencia entre ellas.

Es interesante destacar que se encontraron nuevos haplotipos dentro de los genes evaluados mediante dHPLC. Esto se hizo evidente cuando ninguna de las mezclas línea-genotipo-desconocido/línea-control mostró un perfil de tipo homodúplex en el análisis, mientras que las muestras homodúplex se comportaron de acuerdo a lo esperado para individuos homocigotas (un único pico). La secuenciación de los productos correspondientes a estos individuos reveló nuevos sitios polimórficos para los genes CAB y SCR1 (Figura 2.8), mientras que el *locus* LIM evidenció que el nuevo haplotipo correspondía a una combinación diferente de los alelos encontrados previamente (Tabla 2.6). Si bien la técnica de dHPLC se aplica generalmente a SNPs conocidos, el hecho de haber detectado nuevos haplotipos ofrece una perspectiva promisoriosa en la utilización de la técnica para la identificación de SNPs de *novus* (Han y col. 2004). Al igual que con CEL1CH, se detectaron individuos heterocigotas para los genes SCR1 y PGIC.

Con respecto al costo general de la técnica, dHPLC alcanza un valor de 0,8 USD por corrida. Esto es muy bajo comparado con cualquier otra técnica de alto rendimiento disponible para la genotipificación de SNPs a la fecha (Bagge y Lübberstedt 2008), lo que la convierte en una técnica ideal para el mapeo de genes en poblaciones biparentales. En efecto, Kota y col. (2001) evaluaron el rendimiento de la técnica con cuatro ESTs que habían sido previamente ubicados en un mapa de cebada como marcadores RFLPs, obteniendo la misma ubicación y, posteriormente, caracterizaron los genotipos de tres poblaciones de mapeo para SNPs derivados de 216 ESTs y cuatro indels, añadiendo más marcadores funcionales a estos mapas genéticos (Kota y col. 2008). Incluso para girasol, dHPLC fue utilizado para el mapeo de 11 genes y 243 ESTs, evidenciando las posibilidades de la técnica (Lai y col. 2005, Lexer y col.

2003). En este capítulo, cuatro marcadores que eran polimórficos para las líneas parentales de la población de mapeo se genotificaron en la población de RILs y fueron localizados en distintos grupos de ligamiento (GL). El gen CAM, por ejemplo, se localizó en el GL13, entre dos marcadores muy distantes entre sí (13 cM). Además, PGIC se ubicó junto con un pequeño grupo de marcadores localizados en la región distal del GL9 en el mapa obtenido por Tang y col. (2003) y con un subgrupo de marcadores en el GL9 obtenido por Poormohammad Kiani y col. (2007a) (Figura 2.9). El *locus* LIM se localizó en el GL10, donde se habían detectado previamente seis QTL que explicarían en conjunto un 13% de la resistencia al patógeno fúngico *S. sclerotiorum* (Maringolo 2007). LIM codifica para un factor de transcripción que podría estar involucrado en mecanismos de defensa, aunque también estaría implicado en la formación del citoesqueleto (Arnaud y col. 2007). Si bien algunos GL dentro del mapa RHA266 X PAC2 necesitan mayor saturación a través de la incorporación de nuevos marcadores neutros y funcionales, el análisis de correlación con otros mapas de Compuestas puede realizarse con los datos de los marcadores públicos ubicados en esos mapas (ORSx), adyacentes a los marcadores funcionales mapeados en este trabajo, utilizando las herramientas disponibles en la página del Proyecto Genoma de Compuestas (*Compositae Genome Project*, http://www.sunflower.uga.edu/cgi-bin/cmap/feature_search).

Recapitulando, CEL1CH y dHPLC son metodologías complementarias que sirven para identificar SNPs e indels en girasol, ya que tres genes candidatos pudieron optimizarse para las dos metodologías, mientras que los demás genes pudieron ser resueltos por una u otra técnica (siete genes cada una). Por ejemplo, dHPLC complementa a la técnica CEL1CH en el análisis de aquellas regiones o genes que contienen múltiples polimorfismos, para los cuales la enzima CEL1 pierde eficiencia (CAM, LIM y AALP). Por otro lado, genes o regiones con pocos SNPs o indels son fácilmente puestos a punto con la técnica CEL1CH, como los casos de los genes GIA, MBTF, PGIP3 y RS16. Sin embargo, no todas las regiones con haplotipos complejos pudieron ser resueltas mediante dHPLC (PGIP1, NsLTP), por ello, CEL1CH se presenta como una buena alternativa para la caracterización de al menos uno de los sitios polimórficos presentes en la región estudiada.

En relación a los costos de los reactivos consumibles ninguna de las dos técnicas requiere una gran inversión, dado que la enzima CEL1 puede obtenerse de manera relativamente sencilla en el laboratorio y la técnica dHPLC sólo requiere la amplificación por PCR de las regiones y la mezcla de los fragmentos obtenidos. La inversión en equipos tampoco resulta ser un factor determinante en el caso de CEL1CH, si bien la detección de los fragmentos puede realizarse mediante un servicio externo, existen otras técnicas de detección menos costosas, pero más laboriosas, como la electroforesis en geles de poliacrilamida y de agarosa. Esto convierte a CEL1CH en una técnica rápida y de relativamente bajo costo para detectar diversidad natural en genes candidatos. Con respecto a dHPLC, el proceso de optimización en la detección de las moléculas heterodúplex requiere la evaluación de entre dos y cuatro combinaciones de temperatura y gradiente de acuerdo a lo estimado *in silico* por el programa, haciendo más dificultosa su implementación si no se cuenta con el equipo en el laboratorio. No obstante, no es necesario contar con los equipos específicos destinados a tal fin, y si se dispone con un sistema tradicional de HPLC, la adquisición de una columna específica para ADN puede ser suficiente para llevar adelante la técnica con resultados exitosos, tal como se realizó en este trabajo, y de acuerdo a lo reportado anteriormente por Randall y col. (2005).

En conclusión, desde el surgimiento de los marcadores moleculares de ADN en los años 80, varias técnicas se han desarrollado y optimizado de acuerdo al presupuesto, equipo y personal disponible en los laboratorios (Collard y col. 2008). En este contexto, CEL1CH y dHPLC se proponen como dos técnicas alternativas que deben ser elegidas cuidadosamente, dependiendo de los genes y los polimorfismos que deseen estudiarse. Sin embargo, como se ha demostrado en este capítulo, se trata más bien de dos técnicas complementarias que presentan una buena relación costo-beneficio. Más aún, establecer ambas metodologías simultáneamente permite incrementar el número de marcadores e individuos a analizar en los laboratorios dedicados al mejoramiento de cultivos.

CAPÍTULO 3

SELECCIÓN DE GENES CANDIDATOS PARA LA RESISTENCIA A LA PODREDUMBRE HÚMEDA DEL CAPÍTULO EN GIRASOL

Introducción

El desarrollo de un estudio de mapeo por asociación involucra la evaluación fenotípica del carácter de interés y la genotipificación de un conjunto apropiado de marcadores moleculares. El número y naturaleza de los polimorfismos a caracterizar dependerá de la estrategia de análisis que se haya seleccionado. La estrategia conocida como *genome-wide association* (GWA) consiste en establecer asociaciones con marcadores homogéneamente distribuidos en todo el genoma, lo cual conlleva la genotipificación de cientos o miles de *loci*, generalmente de tipo SNP. Por su parte, la estrategia de genes candidatos permite acotar el número de marcadores a utilizar a través de la evaluación de los polimorfismos contenidos en un conjunto reducido de genes (Haseneyer y col. 2010). Si abordamos esta última, la selección precisa y cuidadosa de los genes candidatos, potencialmente involucrados en el control del fenotipo, es fundamental para la identificación de marcadores útiles asociados al carácter en estudio.

Dentro de los ejemplos de mapeo por asociación en especies vegetales, la estrategia de genes candidatos ha sido aplicada por primera vez por Thornsberry y col. (2001) en maíz, detectando asociación entre el gen *dwarf8* y las variaciones en el tiempo de floración. Este primer antecedente sentó las bases para el mapeo por asociación en este cultivo para otros caracteres, como morfología general de la planta y tolerancia al aluminio del suelo (Krill y col. 2010, Remington y col. 2001). Asimismo, el mapeo por asociación utilizando genes candidatos ha sido aplicado en especies tales como *A. thaliana*, pino, papa, raigrás perenne, cebada y álamo (Ehrenreich y col. 2009, Gonzalez-Martinez y col. 2007, Haseneyer y col. 2010, Ingvarsson y col. 2008, Li y col. 2008, Pajeroska-Mukhtar y col. 2009, Skot y col. 2007). La mayoría de estos trabajos evalúa caracteres de herencia simple, cuyo nivel de variación es ampliamente conocido en las especies estudiadas, de modo que la elección de unos pocos genes candidatos, en algunos casos un único gen (Simko y col. 2004b), es suficiente para lograr el éxito en la detección de asociación.

En girasol, aún no se han reportado estudios de mapeo por asociación. Uno de los caracteres agronómicos buscados en los programas de mejoramiento de girasol es la

resistencia a la Podredumbre Húmeda del Capítulo (PHC), enfermedad producida por el hongo *S. sclerotiorum*. Como se mencionó anteriormente, la mayoría de los trabajos publicados sobre la base genética de PHC realizan la determinación de QTL que, en general, explican individualmente, un bajo porcentaje de la resistencia de girasol frente a *S. sclerotiorum* (Bert y col. 2002, Bert y col. 2004, Gentzbittel y col. 1998, Mestries y col. 1998, Micic y col. 2004, Micic y col. 2005a, Micic y col. 2005b). Otros progresos en la determinación de los factores genéticos involucrados en la resistencia o susceptibilidad a PHC comprenden: (1) un análisis de las diferencias en los cambios metabólicos producidos en líneas tolerantes y sensibles infectadas con el hongo (Peluffo y col. 2010b) y (2) un análisis de los patrones de expresión de los factores de transcripción tipo WRKY (Giacomelli y col. 2010). Adicionalmente, los mecanismos de defensa frente a *S. sclerotiorum* han sido estudiados en otros huéspedes como *B. napus* y *A. thaliana* (*Arabidopsis*) (Carter y Thornburg 2000, Dumas y col. 1993, Hu y col. 2003). A pesar de estos avances, aún no se ha podido identificar cuántos y cuáles son los genes responsables de la tolerancia a PHC presente en algunas líneas y/o cultivares de girasol. En este contexto, la elección de genes candidatos para realizar mapeo por asociación para este carácter no resulta sencilla.

Una de las metodologías más utilizadas para la detección de genes involucrados en procesos biológicos es la de los microarreglos de ADN. Éstos permiten observar de forma casi instantánea la dinámica de expresión de todos los genes del genoma de un organismo. En el trabajo de Zhao y col. (2007a), se analizaron los perfiles de ARNm de dos genotipos de *B. napus* (RV289, parcialmente tolerante y Stellar, altamente susceptible a *S. sclerotiorum*) utilizando un microarreglo de oligonucleótidos sintéticos representativos de 26.090 genes de *Arabidopsis*. Estudios de genómica comparativa entre *Arabidopsis* y *B. napus* avalan la utilización de este tipo de ensayos, dada la proximidad filogenética de ambos géneros (Cavell y col. 1998, Parkin y col. 2005). Los genes identificados en este trabajo, expresados diferencialmente en el genotipo resistente RV289 a 24 horas post infección (hpi), constituyen candidatos potenciales para el estudio de asociación en otras especies vegetales afectadas por el hongo. Sin embargo, esta información podrá ser utilizada convenientemente sólo en los casos en los que puedan identificarse los correspondientes genes ortólogos en la

especie de interés. Los ortólogos se definen como genes que evolucionaron a partir de un ancestro común, a través del proceso de especiación (Fitch 1970). En contraposición, se clasifican como parálogos aquellos que divergieron después de un evento de duplicación génica, ya sea antes de un evento de especiación (*out-paralogs*) o posteriormente (*in-paralogs* o co-ortólogos) (Remm y col. 2001). De este modo, contar con herramientas confiables a la hora de determinar ortología se convierte en un paso de vital importancia en el proceso de selección de genes candidatos. La comparación de pares de secuencias mediante búsquedas en bases de datos usando algoritmos tales como BLAST (*Basic Local Alignment Search Tool*) (Altschul y col. 1990), COG (*Clusters of Orthologous Groups*) (Tatusov y col. 1997, Tatusov y col. 2000), INPARANOID (O'Brien y col. 2005, Remm y col. 2001), RBH (*Reciprocal Blast Hits*) (Hirsh y Fraser 2001, Jordan y col. 2002) y RSD (*Reciprocal Smallest Distance Algorithm*), (Wall y col. 2003) es la estrategia más comúnmente usada para determinar ortología a escala genómica. A pesar de que las búsquedas por similitud de secuencia pueden ser bastante eficientes, han demostrado tener falencias en la identificación de genes relacionados evolutivamente (Koski y Golding 2001). Más aún, en los casos en los que logran hallar ortólogos estrictos, la información que proveen puede resultar insuficiente. La búsqueda de ortólogos mediante métodos filogenéticos se presenta como una alternativa más completa, ya que la construcción de árboles permite explorar la historia evolutiva de las familias génicas que los componen (Chiu y col. 2006). Si bien esta estrategia puede resultar más laboriosa, diversos antecedentes demuestran las ventajas de su uso para determinar los verdaderos ortólogos entre especies (Alexeyenko y col. 2006, Chiu y col. 2006, Fulton y col. 2006, Koski y Golding 2001).

Otra alternativa para identificar genes expresados diferencialmente durante un determinado proceso son las colecciones de ADNc que han sido normalizadas para ecualizar la abundancia de clones que corresponden a los diferentes transcritos, estrategia que se conoce como *Suppression Subtractive Hybridization* (SSH) (Diatchenko y col. 1996). La secuenciación y comparación de los ESTs resultantes permite identificar transcritos que se expresan en una colección dada y están completamente ausentes en otra (Pessino y Martelotto 2004). Si bien existen varias

metodologías para la construcción de colecciones de ADNc diferenciales, la mayoría consiste en dos pasos: (1) la hibridación de una población blanco (*tester*) con otra población que contiene el ARNm en exceso, llamada de referencia (*driver*); y (2) una separación de la fracción no hibridada obteniéndose los transcriptos diferenciales de interés (Diatchenko y col. 1996).

Esta técnica ha sido aplicada exitosamente por el grupo Genómica de Girasol (GG-IB INTA Castelar), permitiendo la obtención de numerosas colecciones de ADNc diferenciales órgano-específicas (Fernandez y col. 2003). Posteriormente, abordando la problemática de las bases genéticas de la resistencia a *S. sclerotiorum*, se construyeron colecciones de ADNc diferenciales para la línea RHA801 (moderadamente resistente) a dos y cuatro días post infección (dpi) (Peluffo y col. 2010a). De esta manera, se arrestaron los ADNc de la planta inoculada con el ADNc de la planta no inoculada. Es decir, se obtuvieron clonotecas que contienen, en gran proporción, los transcriptos (ESTs) que se expresan diferencialmente durante el proceso de infección de *S. sclerotiorum*, en una línea de girasol moderadamente resistente.

En este capítulo se describe el proceso de selección y caracterización de los genes candidatos para resistencia a PHC, que fueron incluidos en el estudio de mapeo por asociación que se presenta en el Capítulo 4. Como punto de partida, se utilizó la información provista por: (1) los perfiles transcripcionales de *B. napus* obtenidos sobre microarreglos de *Arabidopsis* (Zhao y col. 2007a), (2) una de las colecciones de ADNc diferenciales obtenida por el GG-IB INTA Castelar y (3) genes descritos en la literatura como partícipes en el proceso de defensa frente a PHC.

Materiales y Métodos

Material vegetal

La evaluación inicial de SNPs e indels en los genes candidatos se realizó sobre un grupo de 10 líneas endocriadas (PAC2, RHA266, HA89, RHA801, RK416, RHA275, RHA276, HA853, RHA439, RHA440). Estas líneas se seleccionaron porque son parentales de poblaciones segregantes generadas por el grupo de Fitopatología de Girasol (GFPG EEA INTA Balcarce) para el mapeo de QTL de distintos caracteres, entre ellos PHC (Maringolo 2007). Estas 10 líneas se denominaron Grupo Núcleo (GN) de líneas endocriadas.

La extracción de ADN y la cuantificación se realizó de acuerdo a lo descripto previamente (ver Cap. 1).

Selección de genes candidatos

Los genes candidatos se seleccionaron a partir de tres conjuntos de datos: (1) transcriptos que mostraron un aumento de su expresión en plantas resistentes de *B. napus* inoculadas con *S. sclerotiorum* (Zhao y col. 2007a), (2) transcriptos detectados en una colección de ADNc diferencial obtenida a partir de capítulos de girasol de la línea RHA801 (moderadamente resistente) infectados con *S. sclerotiorum* (Peluffo y col. 2010a), (3) genes seleccionados por su rol empírico en los procesos de defensa frente a PHC.

(1) Selección de genes candidatos derivados de *Arabidopsis*: identificación de ortólogos mediante reconstrucción filogenética

A partir del análisis de perfiles de expresión de genes de *B. napus* en respuesta a la infección con *S. sclerotiorum* (Zhao y col. 2007a), se seleccionaron 19 *loci* cuyos ARNm se incrementaron por lo menos al doble en la línea resistente a 24 hpi (Tabla 3.1). Se utilizó como punto de partida para la búsqueda de ortólogos, la secuencia de aminoácidos del gen identificado en el microarreglo de *Arabidopsis* (Zhao y col. 2007a) (Figura 3.1a, *query*). Como primer paso para poder encontrar la secuencia ortóloga de *H. annuus* en un contexto filogenético, se obtuvieron secuencias con similitud al transcripto de *Arabidopsis* mediante dos búsquedas BLAST (tBLASTn)

(Figura 3.1a, salida) (Altschul y col. 1990). El tipo de búsqueda (tBLASTn) se ajustó a la información con que se cuenta en girasol. Debido a que su genoma no está secuenciado, la mayoría de los datos de secuencia provienen de proyectos de secuenciación de transcritos o ESTs. En la primera búsqueda BLAST se determinaron los siguientes parámetros: (1) el rango de posiciones donde alineaban la mayor parte de las secuencias recuperadas, parámetro que dependió de cada gen en particular; (2) la extensión del alineamiento, que se estableció en mayor o igual a 150 aa, (3) la calidad del alineamiento, que implicó no seleccionar las secuencias de ESTs que alineaban en dos regiones distintas con marco de lectura diferentes; y por último (4) la cantidad de secuencias a recuperar por especie identificada (una a cuatro), ya que algunas se encontraban sobre-representadas. Este último criterio no fue aplicado a los géneros *Helianthus* y *Lactuca* (modelo de estudio dentro de las Asteráceas), dado que todas las secuencias que cumplían con las tres primeras condiciones fueron seleccionadas, a fin de tener más posibilidades de encontrar la secuencia ortóloga a la de *Arabidopsis*. En la segunda búsqueda, simplemente se limitó la recuperación de secuencias a los parámetros establecidos (Figura 3.1a; 1 y 2). Una vez seleccionadas las secuencias, se tradujeron en el marco de lectura correspondiente y se alinearon a través de la rutina implementada en el programa MAFFT utilizando la opción L-INS-i (alineamiento muy lento, recomendado para menos de 200 secuencias con un dominio conservado y grandes porciones de baja similitud) (Kato y col. 2002) (Figura 3.1a; 3).

Obtenido el alineamiento de proteínas, se establecieron las relaciones filogenéticas a través del método de máxima parsimonia (MP), utilizando el programa TNT (Goloboff y col. 2008) (Figura 3.1a; 4). Los caracteres fueron analizados como no ordenados o no aditivos, bajo pesos y costos de transformación iguales. La búsqueda de árboles óptimos se realizó mediante métodos heurísticos, partiendo de 20 árboles de Wagner con secuencias de adición al azar y utilizando TBR (*Tree Bisection Reconnection*) como método de permutación de ramas. Cuando todas las réplicas obtuvieron árboles de longitud mínima, no fue necesario adoptar estrategias de búsqueda más rigurosas. En los casos en los que sólo un número reducido de réplicas encontró los árboles más cortos se incrementó el número de réplicas y se realizó una

ronda adicional de permutación de ramas sobre los árboles encontrados a fin de lograr una mejor representación de los cladogramas presentes en las islas halladas.

Los valores de soporte de rama se obtuvieron mediante *jackknife* (Farris y col. 1996), utilizando una probabilidad de 33, para 100 réplicas.

Paralelamente, con el fin de explorar la sensibilidad de los resultados frente a la elección de distintos modelos de evolución molecular, los árboles filogenéticos también fueron construidos mediante el método de Máxima Verosimilitud (*Maximum Likelihood*, ML) (Felsenstein 1981), ajustando los modelos con el programa ProtTest (Abascal y col. 2005). Los árboles obtenidos en ambos casos (MP y ML) fueron coincidentes y sólo se muestran los árboles de MP.

Finalmente, sobre la base de las reconstrucciones halladas se determinaron secuencias ortólogas y parálogas correspondientes a girasol para cada uno de los 19 *loci* evaluados. En un contexto ideal, se espera que el “árbol de genes” reproduzca las relaciones sistemáticas observadas en el “árbol de las especies”. Adicionalmente, si las relaciones sistemáticas se mantienen pero se encuentran representadas más de una vez dentro del árbol de genes, es factible que la historia evolutiva del gen implique la conformación de una familia multigénica. Dado que la información de secuencias de la especie *H. annuus* es limitada, todas las secuencias del género *Helianthus*, *Lactuca* y otros géneros de Asteráceas que se encontraron dentro del mismo clado que la secuencia de *Arabidopsis* fueron consideradas “ortólogos” a la de *Arabidopsis*. En contraposición, se consideraron “parálogos” las secuencias de Asteráceas que quedaron ubicadas por fuera del clado donde se ubicó la secuencia de *Arabidopsis*. Se verificó que cada clado reprodujera, al menos parcialmente, las relaciones sistemáticas para definir familias multigénicas. Las secuencias seleccionadas sobre la base del árbol filogenético obtenido se utilizaron como molde para el diseño de iniciadores (ver diseño de iniciadores, este capítulo) (Figura 3.1a; 5 y 6).

Tabla 3.1. Genes candidatos seleccionados a partir del análisis de expresión en plantas resistentes de *B. napus* a las 24 hpi con *S. sclerotiorum* (Zhao y col. 2007a).

ID	Clase Función probable	Locus	Nivel de expresión ^a	p- valor ^b
Vía de trasducción de señales				
Factores de transcripción				
bHLHP	Proteína de la familia <i>basic helix-loop-helix</i> (bHLH)	At4g09180	22,67	7,71E-06
ZFHDP	Proteína de la familia <i>zinc finger homeobox</i> (ZF-HD)	At1g14687	20,65	2,78E-05
Fosfatasa y kinasa				
PP2C	Fosfatasa 2C (PP2C)	At1g48040	31,47	8,49E-07
PK	Proteína de la familia de kinasa	At3g25250	29,32	2,66E-07
Genes relacionadas con respuestas hormonales				
Vía del ácido jasmónico				
EXLP	Lipasa extracelular de la familia II	At5g42170	37,34	1,63E-06
Maquinaria oxidativa				
CytP450	Citocromo P450 71 ^a 22	At3g48310	61,20	9,53E-07
Modificación de la pared celular				
GRP	Proteína rica en glicinas	At3g27660	46,22	5,11E-06
Metabolismo secundario				
CPE	Proteína de la familia calcineurinas fosfoesterasas	At1g56360	70,10	2,61E-05
NAT	Adenililpolinucleótido transferasa	At1g28090	23,49	1,18E-03
Degradación de proteínas				
FBoxP	Proteína de la familia <i>F-box</i>	At1g13200	21,67	4,83E-05
Transportadores				
MT	Transportador de monosacáridos	At1g34580	103,36	2,08E-06
EXS-FP	Proteína de la familia EXS/ERD1/XPR1/SYG1	At2g03250	27,47	3,11E-06
ABC-TP	Proteína de la familia de transportadores ABC	At1g15520	19,79	3,95E-07
Sin clasificar				
PRP	Proteína de la familia con respuesta a fosfatasa 1	At5g51550	93,31	1,06E-05
HP	Proteína con dominio DTW	At1g03687	26,97	3,57E-06
NPH3-PhotoR	Proteína de la familia NPH3 con respuesta fototrópica	At5g13600	20,12	1,57E-07
Desconocida				
EP	Proteína que se expresa	At2g16650	38,67	1,02E-07
RGEF	Proteína que se expresa Pfam PF03759 (DUF315)	At5g05940	35,50	1,03E-07
RhoBP	Proteína con dominio <i>p21-rho-binding</i>	At1g04450	28,44	1,07E-04

a Número de veces que aumentó la expresión de los transcritos en las plantas inoculadas con el patógeno versus plantas inoculadas con agua.

b p-valor de la probabilidad de cambio en el nivel de expresión (corrección FDR con un α global del 5%).

(2) Selección de genes candidatos derivados de una colección de ADNc de expresión diferencial de girasol

Recientemente, el grupo GG-IB INTA obtuvo una colección de ADNc de expresión diferencial a partir de ARNm de la línea de girasol RHA801 (moderadamente resistente) infectada con *S. sclerotiorum* a los dos y cuatro dpi (Peluffo y col. 2010a).

De acuerdo a lo reportado por Zhao y col. (2007a), la diferencia de los genotipos tolerantes respecto de los susceptibles radica en la respuesta temprana frente a la infección, por lo tanto, se seleccionaron aquellos genes candidatos que: (1) mostraron aumento de expresión en la etapa más temprana de la inoculación (dos dpi) y (2) su

función estuvo descrita y relacionada con la respuesta a estrés (p. ej. factores de transcripción, proteínas de estrés biótico o abiótico, proteínas relacionadas con la membrana celular) (Figura 3.1b).

(3) Selección de genes candidatos descritos en literatura con un rol empírico en los procesos de defensa frente a PHC

Por último, también se incluyeron como genes candidatos dos factores de transcripción WRKY (HaWRKY5 y HaWRKY7), cuyas secuencias fueron cedidas por la Dra. Raquel Chan (FBCB, UNL; Santa Fe), y cinco proteínas tipo-germina de girasol (HaGLPs), seleccionadas del trabajo de Fernández y col. (2003) y del índice de genes girasol (HaGI) del DFCI (Dana Farber Cancer Institute). Evidencias experimentales avalan la elección de estos genes, ya que podrían estar implicados en la respuesta a varios estreses ambientales (Figura 3.1b). Los transcritos seleccionados a partir de la colección de ADNc y a partir de evidencias empíricas de su rol en defensa se muestran en la Tabla 3.2.

Tabla 3.2. Genes candidatos seleccionados a partir de una colección de ADNc de expresión diferencial y por antecedentes experimentales de participación en los procesos de defensa frente a PHC.

ID ^a	Clase Función probable	Origen de la secuencia ^b	Longitud (pb)
Respuesta a estrés			
Hidrolasa			
HaKIV	Kitinasa clase IV	HeAn_C_497 ^c	715
Ha26SP	Subunidad 4 del proteasoma 26S	HeAn_C_1190 ^c	1060
Transferasa			
HaTF	Desconocida	77D2I-NI_Placa1_G05 ^d	439
HaHTFP	Proteína hipotética	82D2I-NI_Placa1_G10 ^d	499
Proteínas tipo Germina			
HaGLP1	Proteína tipo germina 1	EU112647 ^e	878
HaGLP2	Proteína tipo germina 2	AJ540203 ^e	794
HaGLP3	Proteína tipo germina 3	TC17527 ^f	621
HaGLP4	Proteína tipo germina 4	TC18217 ^f	761
HaGLP5	Proteína tipo germina 5	TC17648 ^f	1710
Transportador de electrones			
HaDT-HP	Similar a AT5g54770, relacionada con la biosíntesis de tiamina y la tolerancia al daño del ADN mitocondrial	HeAn_C_11920 ^c	1382

ID ^a	Clase Función probable	Origen de la secuencia ^b	Longitud (pb)
Trasducción de señales			
Factores de transcripción			
HaWRKY5	Factor de transcripción WRKY5	HAWRKY5-HuCF088675 ^g	667
HaWRKY7	Factor de transcripción WRKY7	HAWRKY7-HuCL02697C002 ^g	1721
Kinasas			
HaPK	Proteína quinasa	3D2I-NI_Placa1_A03 ^d	199
Estrés abiótico			
HaDRP	Proteína RD22 con respuesta a deshidratación	HeAn_C_3434 ^c	843
Metabolismo secundario			
Síntesis de fenilpropanoides			
HaPAL	Fenilalanina-amonio-liasa	115D2I-NI_Placa2_H03 ^d	296
Síntesis de ácidos grasos			
HaThE	Proteína de la familia tioesterasas	HeAn_C_11074 ^c	841
Modificación de la pared celular			
HaWP	Proteína de pared (similar a <i>A. thaliana</i> cw _x _arath)	96D2I-NI_Placa1_H12 ^d	425
HaTRP	Proteína relacionada con la formación de tumores	38D2I-NI_Placa1_D02 ^d	266
Proceso metabólico			
Hidrolasas			
Ha_HP	Proteína hipotética	HeAn_C_10084 ^c	1571
Peptidasas			
HaCP	Cisteín-proteinasa	HeAn_S_32288 ^c	578
HaC2SP	Subunidad 2 del complejo peptidasa señal	75D2I-NI_Placa1_G03 ^d	340
Transportadores			
Transporte de lípidos			
HaLTP1	Isoforma 1 de la proteína de transferencia de lípidos	93D2I-NI_Placa1_H09 ^d	230
Transporte de proteínas			
HaGDP-DI	Inhibidor de la disociación de GDP	29D2I-NI_Placa1_C05 ^d	393
Unión a ácidos nucleicos			
Unión a RNA			
HaRNAHel	Helicasa de RNA dependiente de ATP (<i>drh1</i>) <i>dead box ma helicase 1</i>	TC28373 ^f	1561

a Acrónimo designado para denominar el gen candidato.

b Las secuencias seleccionadas de la colección de ADNc se contrastaron contra las bases de datos de ESTs de girasol del HaGI-DFCI (v2010), o contra la base de datos de Unigenes/Girasol (<http://190.103.0.207/>) para hacer la búsqueda con el BLAST2GO.

c Secuencia de la base de datos Unigenes Girasol del IB-INTA (HeAn_C_#, número de UniGen correspondiente a un contig de ESTs; HeAn_S_#, número de UniGen correspondiente a un EST único o *singleton*).

d El identificador se corresponde con el número de secuencia, placa, clonoteca y posición del clon secuenciado en la placa.

e Número de acceso de la base de datos del GenBank NCBI.

f El identificador se corresponde con el número del *Tentative Consensus* del HaGI-DFCI (v 2010).

g Identificadores de las secuencias de WRKYs en la base de datos de UniGenes Girasol del IB (<http://190.103.0.207/>).

Diseño de iniciadores

La amplificación por PCR a partir de ADN genómico de girasol de los genes candidatos derivados de *Arabidopsis* involucró distintas estrategias de diseño de iniciadores (Figura 3.1a; 5 y 6). En aquellos casos en los que se identificaron como posibles ortólogos o co-ortólogos una o más secuencias nucleotídicas de *H. annuus* se diseñaron iniciadores mediante el programa Primer3 (Rozen y Skaletsky 2000). Para los alineamientos nucleotídicos del género *Helianthus* en donde las secuencias recuperadas no eran 100% idénticas entre sí, y no había zonas conservadas, se degeneraron las bases correspondientes de los iniciadores diseñados por el programa Primer3 (Rozen y Skaletsky 2000). Cuando se dispuso de secuencias de otros géneros de Asteráceas distintos de *Helianthus*, se utilizó el programa iCODEHOP (Boyce y col. 2009). Este programa permite diseñar iniciadores sobre CODEHOPs (*C*onsensus-*D*egenerate *H*ybrid *O*ligonucleotide *P*rimers), los cuales son capaces de amplificar por PCR una secuencia blanco desconocida, relacionada con otra, mediante un alineamiento proteico múltiple. Cada iniciador consiste en una secuencia “núcleo” 3’ degenerada y una secuencia 5’ consenso “ancla”, más larga.

Luego de la primera ronda de amplificación con iniciadores degenerados, los productos obtenidos por PCR fueron clonados y secuenciados (ver sección diseño de los productos, clonado y secuenciación automática, este capítulo). A partir de las secuencias obtenidas, se rediseñaron iniciadores con el programa Primer3 (Rozen y Skaletsky 2000), para poder utilizarlos en la amplificación y secuenciación directa de las 10 líneas del GN (Figura 3.1b).

Para los genes candidatos seleccionados a partir de la clonoteca de expresión diferencial de girasol se diseñaron iniciadores a partir de la secuencia del transcripto con el programa Primer3 (Rozen y Skaletsky 2000) (Figura 3.1b). Previo a ello, se identificó el gen correspondiente a *Arabidopsis* con mayor similitud a la secuencia de girasol (BLASTn) (Altschul y col. 1990), con el objeto de predecir la presencia de intrones en la región a amplificar (<http://www.ncbi.nlm.nih.gov/spidey/>) y disminuir las probabilidades de ausencia de amplificación debido a la longitud excesiva de los productos de PCR.

Los genes candidatos HaWRKYs y HaGLPs fueron amplificados con iniciadores diseñados mediante el programa Primer3 (Rozen y Skaletsky 2000) (Figura 3.1b). Las secuencias utilizadas fueron tomadas del *Tentative Consensus* del HaGI-DFCI y de la base de datos UniGenes Girasol de INTA (<http://190.103.0.207/>).

Los iniciadores utilizados para amplificar los genes candidatos y secuenciar las líneas endocriadas del GN se especifican en la Tabla A3.1 del Apéndice.

Reacciones de PCR

Las reacciones de PCR se llevaron a cabo en un volumen final de 50 μ l de acuerdo a lo especificado en el Cap. 1.

Se ensayaron distintos programas variando la temperatura de hibridación de los iniciadores usando como molde ADN genómico de las líneas HA89 y RHA801. Los iniciadores degenerados se evaluaron, en primera instancia, a una temperatura de hibridación de 55°C, dejando las demás condiciones como se describió en el Cap. 1. Aquellos pares de iniciadores que no produjeron amplificación, se evaluaron posteriormente a una temperatura de hibridación menor (50°C). Los pares de iniciadores que dieron múltiples productos de amplificación se ensayaron a una temperatura de hibridación de 60°C, para descartar amplificaciones inespecíficas. Los iniciadores no degenerados se evaluaron mediante los programas *touchdown* descritos en los Cap. 1 y 2 (65°C a 58°C ó 60°C a 55°C) (Figura 3.1a y b).

Purificación de los productos de PCR, clonado y secuenciación automática

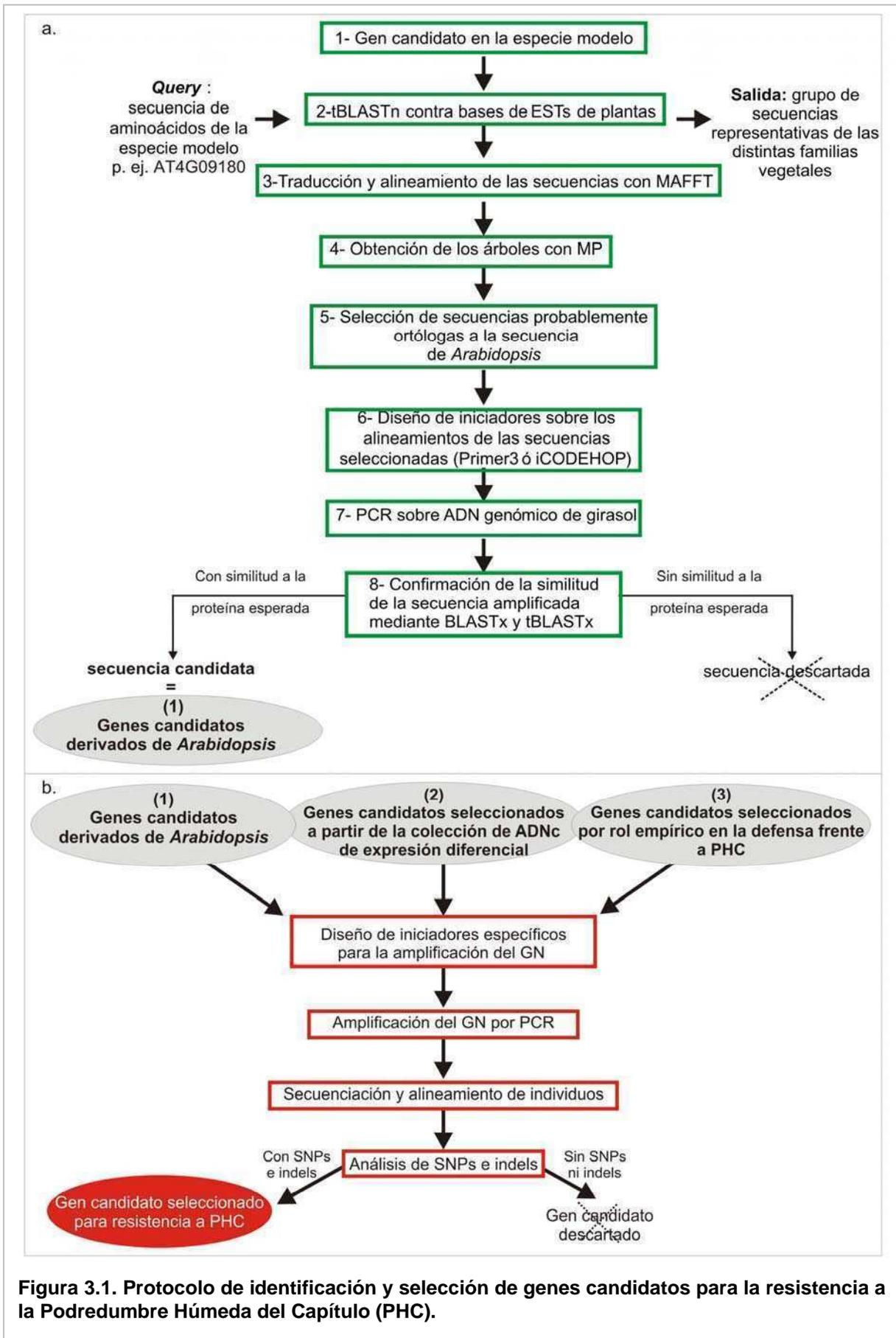
Los productos de PCR obtenidos mediante iniciadores degenerados fueron purificados a partir de tacos de agarosa mediante el sistema QIAEX II (QIAGEN, Alemania), de acuerdo a las especificaciones del fabricante y se clonaron en el vector pGEMT-easy (Promega, EEUU). Se secuenciaron con los iniciadores universales T7 y SP6 entre cinco y diez clones, provenientes de cada una de las reacciones de transformación llevadas a cabo (Sambrook y col. 1989). Los productos de PCR obtenidos mediante iniciadores no degenerados se purificaron y secuenciaron directamente como se describió en el Cap. 1. Las secuencias se obtuvieron de acuerdo a lo detallado en el Cap. 1.

Confirmación de la identidad de los genes candidatos derivados de Arabidopsis

La validación de las secuencias obtenidas en girasol a partir de la amplificación con iniciadores degenerados y no degenerados se realizó por similitud mediante búsquedas BLAST (Altschul y col. 1990). En primer lugar, se realizó una búsqueda BLASTx contra proteínas de *Arabidopsis* para evidenciar si la región obtenida se correspondía con la secuencia de aminoácidos de la proteína candidata de partida. Luego, se realizó una búsqueda tBLASTx contra ESTs de la familia Asteraceae para determinar si la región amplificada se correspondía con las secuencias de partida utilizadas como molde en el diseño de iniciadores (Figura 3.1a; 8).

Identificación y análisis de SNPs y estimación del desequilibrio de ligamiento

La identificación de SNP, el análisis de diversidad nucleotídica y la estimación de DL se realizaron de acuerdo a lo descripto en el Cap. 1.



Resultados

Selección de genes candidatos

Las tres fuentes de información utilizadas resultaron exitosas para la identificación de genes candidatos para resistencia a PHC, dando lugar a un conjunto de 30 genes (17 derivados de *Arabidopsis*, seis provenientes de la colección de ADNc diferencial de girasol y siete seleccionados por su rol empírico en la defensa frente a PHC).

Genes candidatos derivados de *Arabidopsis*: identificación y obtención de secuencias

A partir de la secuencia proteica de 19 transcritos de *Arabidopsis*, cuya expresión se vio aumentada en el genotipo resistente de *B. napus* a 24 hpi con *S. sclerotiorum* (Zhao y col. 2007a), se realizaron búsquedas tBLASTn para llevar a cabo la identificación de ortólogos en girasol mediante reconstrucción filogenética como se detalló en Materiales y Métodos.

El número de secuencias recuperadas para cada uno de los 19 genes descritos en la Tabla 3.1 fue variable, desde 26 secuencias para el gen AT1G03687 (HP), hasta 89 para el gen AT3G25250 (PK) (Tabla 3.3). Estos genes, a su vez, fueron los que presentaron los valores extremos en la cantidad de aminoácidos (aa) que pudieron alinearse, siendo de 148 aa y 273 aa para HP y PK, respectivamente. Si bien se trató de representar la mayor cantidad de géneros posibles, este número dependió en gran medida de la cantidad de secuencias recuperadas, con un máximo 35 géneros para el locus AT5G13600 (NPH3-PhotoR) y un mínimo de 17 géneros para los loci AT4G09180 (bHLHP), AT1G03687 (HP) y AT3G27660 (GRP) (Tabla 3.3).

Tabla 3.3. Identificación de ortólogos mediante reconstrucción filogenética. Datos generales de las búsquedas.

ID	Locus	Nº Secuencias (Nº Géneros) ^a	Nº Árboles de MP	Secuencias utilizadas para el diseño de iniciadores ^b
MT	AT1G34580	41 (19)	3	ehan_19917, ehan_7140, DY948536Hp
PRP	AT5G51550	87 (27)	624	EE645143He, EL412786Hc, EE637816He, DY933824Hp, DY971290Ls, DW124687Ls, DY983551Ls, DY972776Ls, DW155782Lv, DW062617Lg
CPE	AT1G56360	69 (28)	456	BQ849859Ls, ehan_5615, BU027083Ha, DW097601Lp, DW139421Ls, DY918549Ha, EE613110Hg, EH690595Ci, EH725219Cm, ehan_5615, EL425114Hc, EL455341Ht, EL480413Hx
CytP450	AT3G48310	44 (23)	1	EL471301Ht (A) EE606236Hg (B) EL470967Ht (C)
GRP	AT3G27660	51 (17)	30	BU025627Ha, ehan_848, BU027455Ha, DW155978Lv, ehan_3358, DY835993To, DY839004To
EP	AT2G16650	42 (22)	2	AT2G16650, AV543758At, EY161536Tp, CO502033St, CJ762495In, CF513646Vv, ES837663Gh, DV140098Ee, CX288448Cc, EY923572Rr, FK015711Gm
EXLP	AT5G42170	79 (22)	280	DY941701Hp, EL484454Hx
RGEF	AT5G05940	59 (29)	14	EL449780Ht, EL394197Ct (A) EL446390Ht, EL442743Ht (B)
PP2C	AT1G48040	68 (30)	9	DW158280Lv, GE612586Cy, EY101049Aa
PK	AT3G25250	89 (33)	152	BQ990749Lr, EE874594Ib
RhoBP	AT1G04450	47 (20)	12	DY921586Ha (A) DY923043Ha (B)
EXS FP	AT2G03250	55 (27)	4	GE511361Ha
HP	AT1G03687	26 (17)	36	GE513783Ha, GE514181Ha, CD857633Ha
NAT	AT1G28090	50 (25)	1	EE611563Hg
bHLHP	AT4G09180	42 (17)	42	GE508209Ha, EE640456He, DY944559Hp (A) DY915158Ha (B)
FBoxP	AT1G13200	65 (30)	6	GE500048Ha (A) GE499408Ha (B)
ZFHDP	AT1G14687	42 (25)	7	EL422840Hc, GE518153Ha
NPH3-PhotoR	AT5G13600	72 (35)	51	DY966958Ls
ABC-TP	AT1G15520	87 (28)	600	FG482665An, DY966951Ls, EH716520Cm, DY959491Ls, BQ870277Ls

a Número de secuencias recuperadas mediante las búsquedas tBLASTn (número de géneros representados).

b Números de acceso de GenBank o KEGG (*Kyoto Encyclopedia of Genes and Genomes*) junto con la identificación (ID) de género y especie (Tabla A3.2, Apéndice).

Las búsquedas basadas en MP permitieron obtener las relaciones filogenéticas entre las secuencias recuperadas para todos los genes analizados. En la Figura 3.2 se presentan los árboles obtenidos para los *loci* AT1G56360 (CPE), AT3G48310 (CytP450), AT5G42170 (EXLP) y AT5G05940 (RGEF). Estos cuatro casos representan las distintas situaciones encontradas para el total de genes (Figura A3.1, Apéndice). Por ejemplo, para los *loci* EXLP y RGEF las soluciones encontradas fueron las más claras en términos de identificación de ortólogos, ya que una o más secuencias del género *Helianthus* quedaron dentro del mismo clado que la secuencia de *Arabidopsis*

de partida (DY941701Hp, EL484454Hx, EL427700Hc, EL426986Hc y EL449780Ht, EL446390Ht, EL442743Ht, Figura 3.2a y 3.2b, respectivamente), mientras que otras secuencias de *Helianthus* quedaron por fuera de este grupo, posición más consistente con la hipótesis de que se trata de parálogos del gen de *Arabidopsis*. Por el contrario, para algunos *loci* la información contenida en las secuencias analizadas no fue suficiente para obtener árboles con alto nivel de resolución, debiendo elegir al menos una secuencia de cada uno de los grupos involucrados en las distintas politomías (p. ej. CPE, Figura 3.2c). Para otros casos, si bien el/los árboles encontrados alcanzaron un alto grado de resolución, la topología del árbol parece indicar la existencia de un evento de duplicación génica en el grupo de representantes de la familia Asteraceae, haciendo difícil la elección de un único conjunto de secuencias para el diseño de los iniciadores (p. ej. CytP450, Figura 3.2d). En contraposición, para el *locus* AT2G16650 (EP), ninguna de las secuencias de la familia Asteraceae se relacionó como probable ortólogo a la secuencia de *Arabidopsis*, de modo que para esta región, las secuencias ortólogas candidatas correspondieron a géneros de otras familias que quedaron incluidas en el mismo clado que *Arabidopsis* (Figura A3.1d, Apéndice).

Los genes identificados como posibles ortólogos o parálogos se seleccionaron para el diseño de iniciadores (Tabla 3.3, Tabla A3.1) y posterior amplificación a partir de ADN genómico de girasol de las líneas HA89 y RHA801.

Figura 3.2 (pág. sig). Árboles de MP obtenidos para los genes de *Arabidopsis* seleccionados.

a. Consenso de 280 árboles de MP para el locus AT5G42170 (EXLP), longitud: 2427. b. Consenso de 14 árboles de MP para el locus AT5G05940 (RGEF), longitud: 1421. c. Consenso de 456 árboles de MP para el locus AT1G56360 (CPE), longitud: 1172. d. Árbol de máxima parsimonia para el locus AT3G48310 (CytP450), longitud: 2493.

b. Se indican las secuencias de *Arabidopsis* (rojo) y las secuencias ortólogas o parálogas, seleccionadas para el diseño de iniciadores (verde, celeste y azul). Valores de soporte de las ramas mayores a 50% obtenidos usando "Jackknife".

Taxones identificados con el número de acceso del GenBank, seguido por un identificador de género y especie (detallados en la Tabla A3.2 del Apéndice).



La estrategia de amplificación de los genes candidatos AT5G51550 (PRP), AT1G56360 (CPE), AT3G48310 (CytP450), AT3G27660 (GRP), AT2G16650 (EP) y AT5G42170 (EXLP) mediante iniciadores degenerados diseñados con iCODEHOP, no fue eficiente. Sólo dos de los seis *loci* pudieron amplificarse bajo las condiciones iniciales de prueba (55°C) (CytP450 y GRP). No obstante, la amplificación de estos dos *loci* fue inespecífica, debiendo ensayarse una temperatura mayor (65°C) para tratar de reducir la cantidad de productos generados con estos iniciadores. Se obtuvieron productos de PCR para los cuatro *loci* restantes cuando se ensayó una temperatura menor (50°C), utilizando como molde 1 µl de la reacción de PCR a 55°C (Figura 3.3a). La amplificación de *loci* mediante los iniciadores degenerados manualmente resultó, en comparación con los iniciadores iCODEHOP, mucho más eficiente, dado que se obtuvieron productos de PCR en la primera prueba de amplificación a 55°C (Figura 3.3b). Por su parte, los iniciadores específicos amplificaron eficientemente los *loci* AT1G34580 (MT), AT1G04450 (RhoBP), AT1G03687 (HP), AT1G28090 (NAT), AT4G09180 (bHLHP), AT1G13200 (FBoxP), fallando solamente en la amplificación de los *loci* AT2G03250 (EXS-FP) y AT5G13600 (NPH3-PhotoR). De esta manera, se obtuvieron productos de amplificación para 15 de los 19 *loci* descritos en la Tabla 3.3, sin poder amplificar los *loci* AT3G25250 (PK), EXS-FP, NPH3-PhotoR y AT1G15520 (ABC-TP).

Dado que se habían diseñado iniciadores para amplificar distintos grupos de secuencias, seleccionadas de acuerdo a lo observado en los árboles filogenéticos (Figura 3.2 y Figura A3.1), los 15 *loci* de *Arabidopsis* se correspondieron con 21 genes candidatos en girasol. Estos 21 genes se denominaron con una letra adicional, A o B, de acuerdo a la secuencia que se utilizó para diseñar los iniciadores que dieron lugar a cada producto de amplificación (Tabla 3.3). Para los genes PRP, la nomenclatura corresponde a los clones que habían sido secuenciados (Tabla 3.4). Se confirmó la similitud con la secuencia de *Arabidopsis* de partida y/o con la secuencia utilizada para el diseño de iniciadores para 17 de los 21 genes amplificados, siendo los positivos MT, PRP_1, PRP_2, CytP450_A, CytP450_B, CytP450_C, RGEF_A1, RGEF_A2, RGEF_B, PP2C, RhoBP_A, RhoBP_B, HP, bHLHP_B, FBoxP_A, FBoxP_B y ZFHDP. Los resultados de este análisis se muestran en la Tabla 3.4. Las búsquedas

mediante BLASTx arrojaron como primer resultado (*best hit*) la proteína de *Arabidopsis* de partida para cuatro genes (PRP_1, PRP_2, CytP450_B y HP). Adicionalmente, las proteínas de *Arabidopsis* fueron el segundo *hit* de la búsqueda para 3/17 genes (RGEF_A1, RGEF_A2, RGEF_B). Para ocho genes, las proteínas de *Arabidopsis* de partida cayeron entre el cuarto y el vigésimo cuarto lugar (MT, FBoxP_A, CytP450_A, CytP450_C, PP2C, RhoBP_A, RhoBP_B y ZFHDP). Finalmente, dos genes no presentaron similitud con la proteína de *Arabidopsis* de partida (bHLHP y FBoxP_B) pero sí con proteínas de la misma familia (Tabla 3.4, sólo se muestra el resultado contra la proteína de *Arabidopsis* de partida).

Las búsquedas mediante tBLASTx arrojaron como primer *hit* la secuencia utilizada para diseñar los iniciadores en 7/17 genes (PRP_2, RGEF_A1, RGEF_A2, CytP450_A, RhoBP_A, HP y ZFHDP). Tres genes arrojaron como segundo *hit* de la búsqueda tBLASTx la secuencia usada para el diseño de iniciadores (CytP450_B, FBoxP_A y FBoxP_B). Finalmente, seis genes dieron como resultado la secuencia de partida de las Asteráceas entre el cuarto y el décimo lugar (MT, PRP_1, CytP450_C, RGEF_B, PP2C, RhoBP_B, bHLHP).

Debido a que los genes habían sido amplificados a partir de ADN genómico, la mayoría presentó regiones no codificantes, correspondientes a intrones de dichos genes, hecho que contribuyó al aumento del E-valor de los *hits* de las búsquedas mediante BLAST (Tabla 3.4). A pesar de estos resultados dispares, los 17 genes se utilizaron como molde para el diseño de iniciadores específicos para amplificar el GN, con el objetivo de detectar SNPs e indels para ser incorporados al análisis de mapeo por asociación (Figura 3.1).

Tabla 3.4. Resultados de las búsquedas BLAST de los genes candidatos derivados de *Arabidopsis*.

Locus candidato	ID	Longitud (pb)	Resultado del BLASTx contra <i>A. thaliana</i>^{a,c}	E-valor	Resultado del tBLASTx contra Asteraceae^{b,c}	E-valor
AT1G34580 Transporte de monosacáridos	MT	1005	NP_174718 Transportador de monosacáridos AAO64833, locus At1g34580	8 e-07	DY948536 <i>Helianthus petiolaris</i> , cDNA clon CHPY4261	2 e-39
AT5G51550 Proteína con respuesta a fosfatasa	PRP_1	431	AAM65358 AT5g51550/K17N15_10	2 e-67	EE637816 <i>Helianthus exilis</i>	5 e-83
	PRP_2	437	AAM65358 AT5g51550/K17N15_10	3 e-71	EE637816 <i>H. exilis</i>	7 e-96

Locus candidato	ID	Longitud (pb)	Resultado del BLASTx contra <i>A. thaliana</i>^{a,c}	E-valor	Resultado del tBLASTx contra Asteraceae^{b,c}	E-valor
AT3G48310 Citocromo P450 71A22	CytP450_A	785	NP_680110 CYP71A22; transportador de electrones, Citocromo P450 71A22	2 e-06	EL471301 <i>Helianthus tuberosus</i> cDNA clon CHTS9948	4 e-62
	CytP450_B	310	NP_680110 CYP71A22; transportador de electrones, Citocromo P450 71A22	2 e-34	EE606236 <i>Helianthus argophyllus</i> cDNA clon CHWL1183	9 e-65
	CytP450_C	400	NP_680110 CYP71A22; transportador de electrones, Citocromo P450 71A22	2 e-24	EL470967 <i>H. tuberosus</i> cDNA clon CHTS9606	2 e-32
AT5G05940 Proteína que se expresa PFAM PF03759	RGEF_A1	431	ROPGEF5 Farctor 5 de intercambio Rho guanil-nucleótido (<i>locus</i> AT5G05940)	4 e-19	EL449780 <i>H. tuberosus</i> cDNA clon CHTM28362	6 e-36
	RGEF_A2	414	ROPGEF5 Farctor 5 de intercambio Rho guanil-nucleótido (<i>locus</i> AT5G05940)	4 e-03	EL449780 <i>H. tuberosus</i> cDNA clon CHTM28362	7 e-08
	RGEF_B	801	ROPGEF5 Farctor 5 de intercambio Rho guanil-nucleótido (<i>locus</i> AT5G05940)	1 e-12	EL442743 <i>H. tuberosus</i> cDNA clon CHTM19399	1 e-27
AT1G48040 Fosfatasa 2C (PP2C)	PP2C	514	NP_175238 Proteína catalítica serina/treonina fosfatasa GEN: 841222 AT1G48040	3 e-25	GE612586 <i>Cynara scolymus</i> cDNA clon CCPX9338	1 e-31
AT1G04450 Proteína con dominio p21-rho-binding	RhoBP_A	941	NP_171940 RIC3 (Proteína 3 con motivo ROP-INTERACTIVE CRIB; proteína de union (<i>locus</i> AT1G04450))	5 e-09	DY921586 <i>H. annuus</i> cDNA clon CHAY4894	6 e-99
	RhoBP_B	1423	NP_171940 RIC3 (Proteína 3 con motivo ROP-INTERACTIVE CRIB; proteína de union (<i>locus</i> AT1G04450))	6 e-11	DY923043 <i>H. annuus</i> cDNA clon CHAY625	6 e-77
AT1G03687 Proteína con dominio DTW	HP	198	NP_973753 GEN: 839052 AT1G03687, proteína hipotética	1 e-07	GE514181 <i>H. annuus</i> cDNA clon CCFT9312	3 e-11
AT4G09180 Proteína de la familia basic-helix-loop-heli	bHLHP_B	309	-	-	DY915158 <i>H. annuus</i> cDNA clon CHAY11053	3 e-66
AT1G13200 Proteína de la familia F-box	FBoxP_A	265	NP_172779 Proteína de la familia F-box (<i>locus</i> At1g13200)	6 e-02	GE500048 <i>H. annuus</i> cDNA clon CCFT2047	7 e-42
	FBoxP_B	263	-	-	GE499408 <i>H. annuus</i> cDNA clon CCFT1715	4 e-46
AT1G14687 Proteína de la familia zinc finger homeobox	ZFHDP	365	NP_563956 AtHB32 (Proteína HOMEBOX 32) (<i>locus</i> AT1G14687)	3 e-06	EL422840. <i>Helianthus ciliaris</i> cDNA clon CHCM2418	7 e-78

a Búsquedas BLASTx contra la base de datos de proteínas de *Arabidopsis*.

b Búsquedas tBLASTx contra la base de datos ESTs de la familia Asteraceae.

c Números de acceso del GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>) del resultado (*hit*) y la definición, junto con su E-valor. Sólo se muestran los resultados obtenidos contra el *loci* de partida.

Genes candidatos derivados de una colección de ADNc de expresión diferencial

Se identificaron 141 transcritos expresados diferencialmente en la colección de ADNc obtenida para la línea moderadamente resistente RHA801 a los dos dpi

(Peluffo y col. 2010a). Del total, 65 no presentaron similitud con ninguna secuencia de función conocida. Dentro de los 76 genes con anotación según *Gene Ontology* (Ashburner y col. 2000), se seleccionaron aquellos cuya función parecía estar más relacionada con la defensa frente a PHC, incluyendo proteínas de pared (HaWP) y kitinasas (HaKIV), directamente involucradas en los procesos de contención y ataque de hongos fitopatógenos. Asimismo se seleccionaron proteínas relacionadas con estrés abiótico (HaDRP), ya que es posible que la entrada de *S. sclerotiorum* dispare, en girasol, señales relacionadas con estrés en general. Adicionalmente se seleccionó una enzima del metabolismo de compuestos fenólicos (HaPAL), debido a que se detectaron alteraciones en los intermediarios de síntesis y degradación de esta vía metabólica (Peluffo y col. 2010b) (Tabla 3.2).

Se logró la amplificación de un producto del tamaño esperado (entre 300 y 1000 pb) para 6/17 genes (HaDRP, HaCP, HaGDP-DI, HaWP, HaTRP y HaPAL, Tabla 3.2 y Figura 3.3c). Los productos de amplificación de cinco genes fueron mayor a las 2000 pb, a pesar de la búsqueda realizada para contemplar la presencia de intrones (HaRNAHel, HaHP, HaKIV, HaCIISP, HaTF, Figura 3c). Por otro lado, amplificaron inespecíficamente los iniciadores diseñados para los genes Ha26SP, HaHTFP y HaLTP1. El producto de PCR obtenido en la amplificación de HaDT-HP fue muy escaso y, finalmente, dos genes candidatos no pudieron ser amplificados (HaThE, HaPK, Figura 3.3c). Se intentaron secuenciar los productos de amplificación mayores a 2000 pb, correspondientes a los genes HaRNAHel, HaHP, HaKIV, HaCIISP y HaTF, a partir de iniciadores universales, posterior ligación y clonado. Sin embargo, en todos los casos, las reacciones de transformación fracasaron, probablemente debido a la escasa cantidad de producto generado en la amplificación y, consecuentemente, la insuficiente cantidad de moléculas de fragmento de PCR disponibles para su ligación al vector de transformación.

Los genes candidatos HaDRP, HaCP, HaGDP-DI, HaWP, HaTRP y HaPAL se seleccionaron para su posterior amplificación en el GN y búsqueda de SNP e indels.

Genes candidatos descritos en literatura con un rol empírico en los procesos de defensa frente a PHC

A través de la búsqueda en el HaGI-DFCI (Dana Farber Cancer Institute) y en colecciones de ADNc órgano-específicas (Fernandez y col. 2003), se ha podido determinar la presencia de al menos cinco proteínas tipo-germina en girasol (HaGLPs). Debido a sus roles hipotéticos en defensa frente a PHC, se amplificaron estos genes candidatos en el GN a partir de ADN genómico. Las amplificaciones resultaron exitosas, sin embargo, para HaGLP4 y HaGLP5 se precisaron dos y tres pares de iniciadores para lograr la amplificación completa de dichos genes (Tabla A3.1, Apéndice).

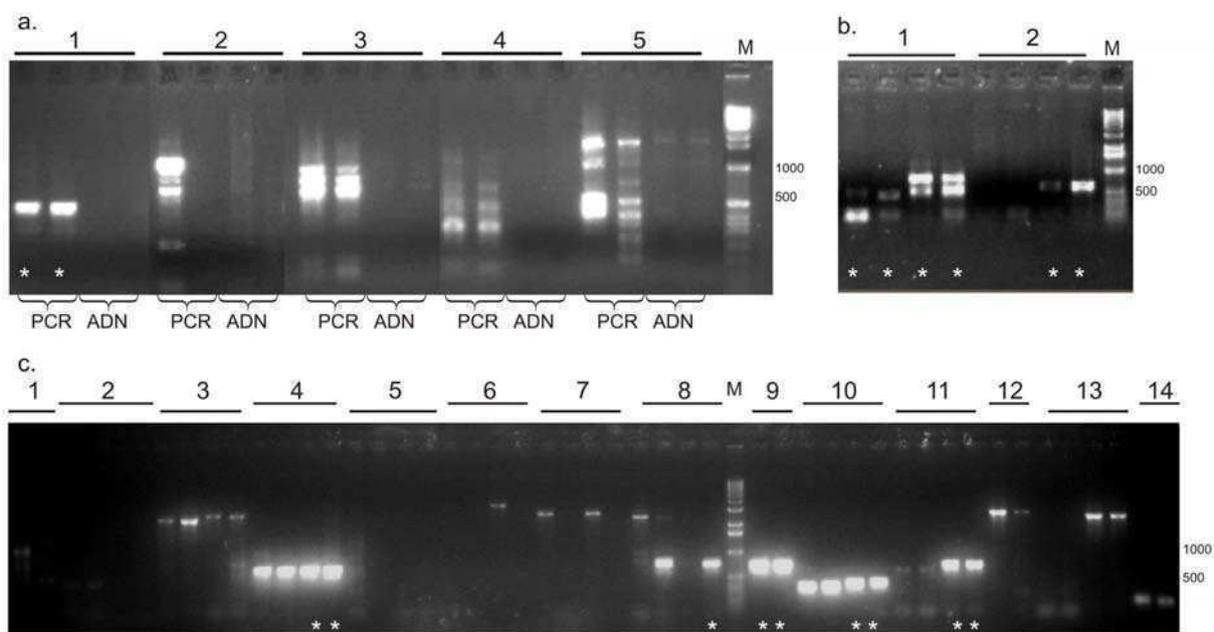


Figura 3.3. Amplificación por PCR de genes candidatos para resistencia a PHC.

La primera y tercera calle de cada barra asociada a un número corresponde a la PCR de la línea HA89. La segunda y cuarta calle de cada referencia corresponde a la PCR de la línea RHA801.

a. Amplificación de genes ortólogos a *Arabidopsis* utilizando iniciadores degenerados. Se utilizó como molde 100 ng de ADN genómico ó 1 µl de PCR de una reacción anterior (ADN y PCR). Temperatura de hibridación: 50 °C. Se muestra la PCR para los *loci* PRP (1), CPE (2), GRP (3), EP (4) y EXLP (5).

b. Amplificación a 55°C de los genes ortólogos a *Arabidopsis* utilizando iniciadores degenerados para los *loci* RGEF (1) y PP2C (2).

c. Amplificación de los genes a partir de la clonoteca de expresión diferencial mediante iniciadores específicos. Se utilizó un programa *touchdown* 65°C-58°C. Se muestran las amplificaciones para los *loci* Ha26SP (1), HaDT-HP (2), HaRNAHel (3), HaDRP (4), HaThE (5), Ha_HP (6), HaKIV (7), HaCP (8), HaGDP-DI (9), HaWP (10), HaTRP (11), HaCIISP (12), HaHTF (13) y HaPK (14).

Los asteriscos (*) indican las regiones seleccionadas para la búsqueda de SNPs e indels en 10 líneas de girasol. M: Marcador de peso molecular.

Con respecto a las proteínas WRKY, inicialmente sólo se seleccionaron las secuencias de HaWRKY5 y HaWRKY7, debido a que los patrones de expresión para estos dos genes se vieron alterados en líneas de girasol susceptible y tolerantes a PHC (Giacomelli y col. 2010) (<http://190.103.0.207/>). Dado que la secuencia de estos transcriptos correspondía a regiones de ~800 y 1700 pb, respectivamente, la amplificación sobre ADN genómico del gen HaWRKY7 se planificó por regiones solapadas de aproximadamente 600 pb. A diferencia de la amplificación del gen HaWRKY5, para el cual se obtuvo un producto de 1200 pb utilizando ADN genómico como molde; la amplificación completa del gen HaWRKY7 no fue posible. Los pares de iniciadores diseñados hacia el extremo 5' del transcripto (Tabla A3.1, HaWRKY_3 y HAWRKY_4) fueron exitosos en la amplificación de fragmentos, aunque el primero produjo amplificaciones inespecíficas. Este hecho, probablemente se deba a la similitud de secuencia con otros genes de la familia, y la especificidad de esta PCR podría incrementarse al diseñar nuevos iniciadores en zonas no conservadas de la familia WRKY y específicas de HaWRKY7. No obstante, los fragmentos correspondientes a estas dos regiones pudieron secuenciarse en el GN.

Caracterización de los genes candidatos en el Grupo Núcleo de líneas de girasol

Se secuenciaron 30 genes candidatos en el GN, correspondiendo a un total de 18.373 pb secuenciadas por individuo. La longitud de los genes varió entre 166 y 1710 pb, excluyendo los indels, con un promedio de 612 pb por región (Tabla 3.5).

Con respecto a los genes candidatos derivados de *Arabidopsis*, 15/17 genes secuenciados presentaron SNPs y/o indels potenciales para la genotipificación de la Población para Mapeo por Asociación (PMA). Sin embargo, solo 9/15 se seleccionaron para su posterior genotipificación en dicha población (MT, PRP_2, CytP450_C, RGEF_A2, PP2C, RhoBP_A, RhoBP_B, bHLHP y ZFHP) (Tabla 3.5). Los genes candidatos CytP450_A, CytP450_B y HP se descartaron ya que presentaban sitios polimórficos en baja frecuencia (un SNP, dos SNPs y un indel, respectivamente). Por otro lado, los genes RGEF_A1, FBoxP_A y FBoxP_B tampoco se seleccionaron para su posterior genotipificación en la PMA, debido a las dificultades encontradas en la secuenciación de los mismos, y por presentar

demasiados sitios polimórficos (Tabla 3.5).

Con respecto a los restantes genes candidatos seleccionados, ya sea a partir de la colección de ADNc de expresión diferencial, o seleccionados por su rol empírico en la defensa frente a PHC, los 13 genes que pudieron ser secuenciados presentaron polimorfismos entre las 10 líneas evaluadas (HaDRP, HaCP, HaGDP-DI, HaWP, HaTRP, HaPAL, HaWRKY5, HaWRKY7, HaGLP1, HaGLP2, HaGLP3, HaGLP4 y HaGLP5). Sin embargo, HaGLP1 presentó un indel de tres pb en una región repetitiva del nucleótido "T" y se descartó para su genotipificación en la PMA, debido a que este tipo de polimorfismos es proclive a generar artefactos de amplificación, resultando inestable.

Se inspeccionaron los índices de polimorfismo, diversidad nucleotídica y número de haplotipos de los 30 genes candidatos para poder seleccionar la estrategia de genotipificación más adecuada sobre la PMA. Considerando el panel completo de datos, los polimorfismos encontrados fueron abundantes, observándose 221 SNPs y 44 indels (Tabla 3.5). En promedio, se detectó un SNP cada 83 pb y un indel cada 417 pb. Por gen, la cantidad de SNPs encontrada varió desde cero (0), para los genes PRP_1, RGEF_B, HP y HaGLP5 hasta 18 SNPs en el *locus* HaGDP-DI, evidenciando una media de 7,4 SNPs por fragmento secuenciado. Los indels también fueron variables, detectando inserciones de un nucleótido (p. ej. HP), hasta una inserción de 46 pb en el gen RGEF_A1. Se contabilizaron en total 183 pb correspondientes a indels. Adicionalmente, la mayor cantidad de haplotipos (seis) se detectó en dos genes (una región de HaWRKY5 y HaGLP3). Finalmente, los valores de diversidad y polimorfismo promedio fueron moderados ($\pi=0,0082$, $\theta=0,0075$) (Tabla 3.5).

Se realizó además, la prospección del desequilibrio de ligamiento (DL) entre los pares de sitios polimórficos para cada uno de los 23 genes candidatos que presentaron dos o más SNPs (MT, PRP_2, CytP450_B, CytP450_C, RGEF_A1, RGEF_A2, PP2C, RhoBP_A, RhoBP_B, FBoxP_A, FBoxBP_B, ZFHP, HaDRP, HaCP, HaGDP-DI, HaWP, HaTRP, HaPAL, HaWRKY5, HaGLP2, HaGLP3, HaGLP4 y HaGLP5). Estas estimas constituyeron sólo una aproximación a nivel descriptivo, debido al reducido número de líneas analizadas. La ocurrencia de DL significativo se

observó para todos los pares de sitios en 9/23 genes analizados (CytP450_B, CytP450_C, RGEF_A2, PP2C, RhoBP_A, FBoxP_A, HaDRP, HaWP y HaGLP2). Adicionalmente, para los genes MT y RhoBP_B se observó DL entre todos los pares de comparaciones, excepto las comparaciones que involucraban el primer sitio SNP. Para los 12 genes restantes se observó DL significativo sólo entre algunas de las comparaciones.

Tabla 3.5. Características de los genes candidatos para resistencia a PHC.

ID Región ^a	Longitud ^b	SNPs ^c	Indels ^d	Nº Haplotipos ^e	Diversidad haplotípica	π_s	θ_w
MT	580	8 (8)	3 (11)	3	0,714	0,0074	0,0056
PRP_1	431	0	0	1	0	0	0
PRP_2	437	11 (11)	0	3	0,762	0,0127	0,0105
CytP450_A	786	1 (0)	0	2	0,222	0,0003	0,0005
CytP450_B	310	2 (0)	0	2	0,222	0,0017	0,0028
CytP450_C	400	2 (2)	1 (1)	2	0,356	0,0021	0,0021
RGEF_A1	447	12 (11)	3 (50)	4	0,778	0,0176	0,0127
RGEF_A2	403	2 (2)	2 (11)	2	0,356	0,0019	0,0019
RGEF_B	801	0	0	1	0	0	0
PP2C	514	4 (4)	0	2	0,200	0,0016	0,0028
RhoBP_A	940	8 (8)	4 (6)	3	0,556	0,0044	0,0033
RhoBP_B	1424	12 (0)	2 (15)	3	0,524	0,0029	0,0042
HP	198	0	1(1)	1	0	0	0
bHLHP	309	1 (1)	0	2	0,386	0,0013	0,0012
FBoxP_A	265	11 (11)	0	3	0,556	0,0219	0,0169
FBoxP_B	262	8 (5)	1 (1)	5	0,905	0,0150	0,0136
ZFHDP	166	9 (7)	0	4	0,733	0,0222	0,0218
HaDRP	587	14 (12)	1 (1)	2	0,556	0,0145	0,0096
HaCP	542	14 (10)	0	3	0,833	0,0166	0,0147
HaGDP-DI	620	18 (14)	2 (2)	3	0,607	0,0137	0,0132
HaWP	351	4 (4)	3 (10)	2	0,533	0,0063	0,0042
HaTRP	658	14 (9)	3 (4)	5	0,786	0,0095	0,0104
HaPAL	317	2 (2)	0	2	0,467	0,0030	0,0023
HaWRKY5_1^f	356	14 (14)	4 (10)	4	0,821	0,0221	0,0192
HaWRKY5_2	203	9 (9)	3 (20)	3	0,762	0,0257	0,0211
HaWRKY5_3	176	10 (2)	1 (4)	6	0,929	0,0200	0,0268
HaWRKY7_3 ^f	939	1 (1)	1 (10)	2	0,500	0,0005	0,0004
HaWRKY7_4	189	0	1 (12)	2	0,600	0	0
HaGLP1	876	0	3 (3)	3	0,433	0	0
HaGLP2	794	6 (6)	0	2	0,303	0,0026	0,0028
HaGLP3	621	13 (9)	0	6	0,758	0,0083	0,0075
HaGLP4	761	9 (2)	3 (11)	4	0,561	0,0026	0,0042
HaGLP5	1710	2 (1)	2 (3)	3	0,618	0,004	0,004
Total	18.373	221 (165)	44 (186)				
Promedio/gen	612,4	7,4	1,4	2,9	0,5254	0,0079	0,0073
Frecuencia		1/83 (1/111)	1/417				

a Acrónimos de los genes candidatos secuenciados y genes seleccionados para la genotipificación de la PMA (subrayados).

b Longitud en pb excluyendo indels.

c Sitios totales (sitios informativos).

d Número de indels (número de pb totales de los indels).

e El número de haplotipos fue determinado por la cantidad de sitios informativos y/o indels.

f Los genes HaWRKY5 y HaWRKY7 se analizaron por regiones donde alineaban al menos cuatro individuos.

Discusión

El mapeo por asociación a través de la estrategia de genes candidatos puede ser una tarea difícil, dependiendo del carácter fenotípico estudiado. En este caso, la resistencia a PHC es un carácter complejo, de herencia poligénica, cuyo estudio ha sido abordado, casi exclusivamente, a través del mapeo de QTL en poblaciones biparentales, para tratar de identificar las regiones genómicas involucradas en la resistencia (Bert y col. 2002, Bert y col. 2004, Gentzbittel y col. 1998, Micic y col. 2004, Micic y col. 2005a, Micic y col. 2005b, Yue y col. 2008).

Al momento de la elección de los genes candidatos a incluir en este trabajo, la información más concreta con la que se contaba acerca de genes involucrados en la respuesta de la planta frente a *S. sclerotiorum* era un análisis transcripcional de *B. napus* basado en un microarreglo de *Arabidopsis* (Zhao y col. 2007a). Para poder encontrar en girasol los genes que poseen la misma función que los de *Arabidopsis*, es necesario entender cómo han evolucionado estos genes. Dos genes en dos especies diferentes que evolucionaron a partir de un ancestro común tienen altas probabilidades de compartir la misma función biológica. Estos genes se denominan ortólogos por definición (Fitch 1970). Por lo tanto, el éxito en la selección de los genes, dependerá, en parte, de la precisión con que se hayan determinado los ortólogos y/o parálogos en girasol de los genes correspondientes a *Arabidopsis*. A pesar de que los métodos basados en comparaciones BLAST se han convertido en los más populares a la hora de determinar homología, se ha demostrado que muchas veces el mejor resultado del BLAST no es el más cercano filogenéticamente (Koski y Golding 2001). En concordancia con lo que postulan algunos trabajos (Chiu y col. 2006, Fulton y col. 2006), se realizó la determinación de ortología a través de la reconstrucción filogenética, ya que resulta más precisa y explicativa que las búsquedas mediante BLAST.

La rutina de identificación de ortólogos implicó la reconstrucción de árboles de genes con la información disponible para el género *Helianthus* y *Lactuca* (especies cuyo genoma aún no está completamente secuenciado), en un contexto general de secuencias de distintas especies y géneros vegetales con similitud a la secuencia de

Arabidopsis de partida. Se construyeron 19 árboles de MP y ML correspondientes a los 19 *loci* candidatos correspondientes, sin observarse diferencias entre los árboles encontrados por uno y otro método. Si bien el objetivo de este análisis no comprendió la optimización exhaustiva de cada uno de los parámetros involucrados en la construcción de los árboles, la aproximación utilizada permitió identificar secuencias ortólogas y/o parálogas al gen de *Arabidopsis*, pertenecientes a géneros y especies de la familia Asteraceae. En detalle, para nueve *loci* las secuencias utilizadas para el diseño de iniciadores pertenecieron a *H. annuus* (MT, CPE, GRP, RhoBP, EXS-FP, HP, bHLHP, FBoxP, ZFHDP), para cinco *loci* se seleccionaron secuencias de otras especies del género *Helianthus* (PRP, CytP450, EXLP, RGEF, NAT) y para otros cuatro *loci* se seleccionaron secuencias de otras especies de la familia Asteraceae (PP2C, PK, NPH3-PhotoR, ABC-TP) (Tabla 3.3). Un caso destacable es el del *locus* AT2G16650 (EP), para el cual se habían detectado *hits* significativos de *Helianthus*, *Lactuca* y otros géneros de la familia Asteraceae mediante BLAST, pero ninguna de estas secuencias resultó ortóloga a EP, según el árbol MP obtenido (Figura A3.1d). Esta situación hubiera pasado inadvertida al utilizar únicamente un método de comparación de pares de secuencias (p. ej. BLAST o RBH), ya que en estos métodos sólo se tiene en cuenta el mejor *hit* de la especie de interés (girasol u otra especie de la familia), sin reparar en que es posible que la secuencia del verdadero ortólogo no esté aún depositada en las bases de datos (Fulton y col. 2006). Sin duda, la estrategia filogenética aquí establecida cobra completo significado a partir de estos resultados.

Adicionalmente, este análisis enriqueció el conocimiento acerca de la naturaleza evolutiva de estos genes, ya que, a juzgar por los árboles obtenidos, dichos *loci* conforman familias multigénicas (Figura 3.2 y Figura A3.1). De hecho, se conocen en detalle la naturaleza y diversidad de las familias para algunos de ellos (p. ej. CytP450, bHLHP, ZFHDP) (Ehrling y col. 2008, Hu y col. 2008, Mukherjee y col. 2009, Pires y Dolan 2010).

Dentro de la diversidad de soluciones encontradas para cada gen, pueden mencionarse, por ejemplo, los casos de los genes AT3G48310 (CytP450) y AT1G28090 (NAT), para los cuales se encontró un único árbol más parsimonioso, poniendo de manifiesto la ausencia de contradicción en la matriz básica de datos. Para otros seis

loci el número de árboles de longitud mínima fue menor a 10 (MT, EP, PP2C, EXS-FP, FBoxP y ZFHDP). Finalmente, en cinco *loci* la cantidad de árboles igualmente parsimoniosos superó la centena, conformando las soluciones más conflictivas (PRP, CPE, EXLP, PK, ABC-TP). Esto último podría deberse a la insuficiente información con que se contaba para definir los grupos, es decir, alineamientos muy cortos, demasiadas indeterminaciones en las secuencias, junto con la sobre-representación de los géneros *Helianthus* y *Lactuca* respecto de los demás. Sin embargo, a pesar de estas distintas situaciones, el hecho de haber recurrido a la reconstrucción filogenética permitió explorar alternativas que finalmente condujeron a identificar secuencias potencialmente ortólogas a la secuencia de *Arabidopsis*.

Con respecto al diseño de iniciadores para amplificar las secuencias seleccionadas a partir de ADN genómico de girasol, la utilización del programa iCODEHOP (Boyce y col. 2009) no resultó ser una buena opción, ya que la mayoría de las secuencias amplificadas con estos iniciadores presentaron similitud a transposones de otras especies de *Helianthus* (datos no mostrados). Existe evidencia que avala la eficiencia de amplificación de iniciadores diseñados usando CODEHOPs (*C*onsensus-*D*egenerate *H*ybrid *O*ligonucleotide *P*rimers) (Navaud y col. 2007, Rose 2005). No obstante, la *performance* del programa iCODEHOP no ha sido descripta aún. Por su parte, los iniciadores diseñados con degeneraciones de forma manual fueron más exitosos, lo que sugiere que la mejor elección para amplificar secuencias en girasol a partir de secuencias de otras especies, es diseñar los iniciadores de esta última forma, sobre todo, si no se dispone de regiones conservadas entre las secuencias utilizadas como molde.

A pesar de que cuatro *loci* de *Arabidopsis* no pudieron ser amplificados (PK, EXS-FP, NPH3-PhotoR y ABC-TP), el hecho de haber diseñado iniciadores sobre distintos grupos de secuencias de *Helianthus* o géneros relacionados, de acuerdo a la información de los árboles, permitió que los 15 *loci* de *Arabidopsis* se correspondieran con 21 genes amplificados en girasol. No obstante, sólo 17/21 genes secuenciados mostraron similitud con las correspondientes secuencias de *Arabidopsis* de partida o con la secuencia usada para diseñar los iniciadores (Tabla 3.4). Este resultado, que en principio podría parecer desalentador, demuestra, por el contrario, que la riqueza de

información obtenida con la reconstrucción de los árboles filogenéticos, y la minuciosidad con la que se seleccionaron las secuencias y se diseñaron los iniciadores, se vieron finalmente reflejadas en la posibilidad de haber contado con más de un gen en girasol a partir de la información de un único *locus* de *Arabidopsis*.

La colección de ADNc diferencial de girasol puede parecer *a priori* una fuente de información más precisa para la selección de genes candidatos derivados de *Arabidopsis*, sin embargo, la selección propiamente dicha y la amplificación de estas regiones a partir de ADN genómico de girasol son dos cuestiones que merecen especial consideración.

El gran número de transcritos expresados diferencialmente (141) que fueron identificados en esta clonoteca se presentó como la primera dificultad. La selección se restringió, por lo tanto, a transcritos con similitud a secuencias ya depositadas, lo que implicó descartar el 46% de las secuencias identificadas (65). Si bien la falta de similitud con otras secuencias pudo ser consecuencia de la longitud del fragmento clonado, también podría deberse a que los genes implicados en la defensa a PHC en girasol son escasamente conocidos y, por tanto, estas secuencias inéditas sean realmente específicas para el proceso infeccioso e impliquen la determinación de la tolerancia en la línea RHA801. De cualquier modo, encontrar asociaciones fenotípicas con genes de función aún desconocida sería un desafío mayor.

La amplificación de estos transcritos a partir de ADN genómico de girasol, fue, inesperadamente, poco exitosa, ya que sólo 6/17 genes pudieron ser evaluados en el GN. La escasa o nula amplificación de los genes HaDT-HP, HaThE y HaPK pudo deberse a que las secuencias de girasol sobre las que se diseñaron los iniciadores provinieron de una sola lectura del clon correspondiente, lo que pudo haber llevado a la generación de iniciadores con errores que no hibridaron correctamente con el ADN molde. Otra posibilidad es que al utilizar secuencias provenientes de consensos de transcritos, como en el caso de HaDT-HP y HaRThE, los iniciadores hayan sido diseñados justo en la región de *splicing* del ARNm, provocando que éstos no hibriden con el gen correspondiente. Por otro lado, la amplificación inespecífica de los genes Ha26SP, HaHTFP y HALTP1, podría ser consecuencia de la presencia de más de una

región del genoma de girasol con similitud a los iniciadores diseñados, probablemente debido a que pertenecen a familias multigénicas (p. ej. LTPs; (Kader 1996)). Finalmente, la observación de productos de amplificación mayores a 2000 pb para los genes HaRNAHel, HaHP, HaKIV, HaC2SP y HaTF, a instancias de la presencia de probables intrones, permite inferir que la estructura de estos genes en girasol es diferente a la estructura de los *loci* en *Arabidopsis*. De acuerdo a lo expuesto por Deutsh y Long (1999), una explicación de esta diferencia, aunque no la única, podría ser que la longitud de los intrones correlaciona con el tamaño del genoma, con lo cual los intrones del genoma de girasol serían de mayor tamaño que los respectivos en el genoma de *Arabidopsis*.

Con respecto a los índices de polimorfismos encontrados en los 30 genes secuenciados, es destacable que 28 genes hayan presentado SNPs y/o indels potenciales para la genotipificación de la PMA. Sin embargo, debido a las características de algunos polimorfismos (p. ej. sitios no informativos en el caso de CytP450_A, CytP450_B y HP, indels inestables en HaGLP1 y dificultades en la secuenciación en RGEF_A1, FboxP_A y FBoxP_B), sólo 21 genes se seleccionaron como candidatos finales. La ocurrencia de SNPs (1/83 pb) fue menor a la encontrada para el panel de datos del Cap. 1 (1/61 pb), al igual que la cantidad de indels (44 vs. 61 en el Cap. 1), correlacionando, este último dato, con una menor cantidad de pb insertadas totales (186 pb vs. 302 pb, respectivamente). Es probable que estas diferencias se deban a la cantidad de individuos analizados en uno u otro caso, es decir el tamaño de la muestra. Mientras que en el Cap. 1 este análisis se realizó con alineamientos de al menos 15 líneas por gen, el análisis para los genes aquí evaluados se realizó con 10 líneas como máximo, contando en algunos casos con las secuencias para sólo cuatro líneas (p. ej. HaWRKY7). Asimismo, una inserción de 425 pb para el gen HaWRKY5 no fue tomada en cuenta, debido a que los datos de diversidad debieron analizarse en tres regiones distintas que no la contemplaban por la falta de datos en el contexto de este indel.

Por otra parte, la evidencia experimental sugiere que el DL en líneas endocriadas de girasol se hace despreciable más allá de las 5500 pb ($r^2=0,32$), es decir, supera los tamaños esperados para cualquier gen (Cap. 1 y (Kolkman y col. 2007)). Sin embargo,

la ocurrencia de pares de comparaciones no significativas entre los sitios SNPs dentro de un gen pudo deberse a la baja frecuencia alélica de uno de los sitios respecto de los otros, haciendo que el análisis de este parámetro en un tamaño de muestra tan reducido pueda no ser representativo de lo que ocurre en la PMA a estudiar (Flint-García y col. 2003). De hecho, en el caso de los genes MT y RhoBP_B se observa DL entre todos los sitios, excepto las comparaciones contra el primer SNP, que por observación de los alineamientos resultó ser el de menor frecuencia respecto a los demás.

Por otro lado, la ubicación de los 21 genes candidatos en el mapa genético saturado de girasol cultivado aún se desconoce, y deberán ser ubicados en los mapas disponibles para girasol cultivado para evaluar su co-localización con los QTL identificados para PHC (Bert y col. 2002, Gentzbittel y col. 1998, Micic y col. 2004, Micic y col. 2005a, Micic y col. 2005b, Yue y col. 2008). De hecho, 14/21 genes resultaron polimórficos entre las líneas parentales de la población de mapeo de referencia RHA266 X PAC2 (MT, PRP, RGEF_A1, RhoBP_B, ZFHDP, HaDRP, HaDGP-DI, HaWP, HaPAL, HaGLP2, HaGLP3, HaGLP4, HaGLP5, HaWRKY7). Dependiendo del método de genotipificación seleccionado para la PMA, será posible incorporarlos y correlacionar las posiciones con otros mapas de ligamiento de girasol (ver detalles en Cap. 4).

La mayoría de los trabajos de mapeo por asociación con genes candidatos realizados hasta el momento involucran la genotipificación de uno, dos y hasta cuatro genes en su mayoría (Haseneyer y col. 2010, Skot y col. 2007, Sulpice y col. 2009, Thornsberry y col. 2001). En estos trabajos, los genes seleccionados contaban con evidencia indiscutible para su elección. Por ejemplo en el trabajo de Skot y col. (2007) se investigaron dos caracteres (tiempo de floración y contenido de carbohidratos solubles) utilizando un gen candidato en cada caso, para el primero un ortólogo de *CONSTANS* en *Lolium* y para el segundo una enzima que co-localizaba con un QTL para el contenido de glucosa y fructosa en el cromosoma 6. Asimismo, en el trabajo de Sulpice y col. (2009) los dos genes elegidos habían sido correlacionados con la biomasa de la roseta a través de un estudio metabólico. Entre los trabajos realizados con más de 10 genes candidatos se puede citar el trabajo de Erenreich y col. (2009),

en el cual evalúan 51 *loci* en *Arabidopsis* asociados al tiempo de floración. Sin duda, dado que el genoma de *Arabidopsis* está completamente secuenciado, la amplificación de estos *loci* no presentó mayores obstáculos. Por otro lado, Pajerowska-Mukhtar y col. (2009) seleccionaron 21 genes para mapeo por asociación del tizón tardío en papa. En este caso, la selección se basó en: (1) genes que co-localizaban con QTL detectados para la enfermedad, (2) genes asociados a la resistencia al tizón e identificados en el trabajo de Malosetti y col. (2007) y (3) otros genes relacionados con defensa, previamente mapeados y que habían sido identificados en procesos de defensa para ésta y otras enfermedades en papa. Si bien el número de genes seleccionados es idéntico al de este trabajo de tesis, los antecedentes del comportamiento de estos genes en papa y de los fenotipos asociados, prácticamente aseguran el resultado final. En el trabajo de Krill y col. (2010) se seleccionaron 21 genes candidatos para asociación con la tolerancia al aluminio basados en: (1) un análisis de expresión y (2) por similitud con genes responsables de la tolerancia al aluminio en otras especies. Seis genes candidatos mostraron asociación significativa con el carácter en dicho estudio, apoyando la validez de las fuentes de selección utilizadas en este trabajo de tesis. Por otro lado, en el trabajo de Ingvarsson y col. (2008) se evaluaron 13 regiones, pero todas ellas físicamente cercanas al gen *phyB2*, de manera de evaluar si la región cromosómica estaba asociada al tiempo de brote en *Populus*.

En este capítulo se expuso el proceso de selección, amplificación y caracterización de genes candidatos para mapeo por asociación, y, debido a que PHC es un carácter complejo, la cantidad final de 21 genes seleccionados, resulta adecuada. A su vez, las fuentes de selección (microarreglo de *Arabidopsis*, colección de ADNc diferencial y genes seleccionados por su rol empírico en los procesos de defensa) resultaron válidas dada la escasa información con que se cuenta sobre los genes implicados en la determinación de resistencia a PHC en girasol. Finalmente, además de la función que estos genes puedan tener sobre el carácter, la caracterización de estos 30 genes permitió aumentar la información sobre la variabilidad intraespecífica y la estructura génica en girasol. El proceso de genotipificación de la PMA, la asociación de estos genes con el carácter y las implicancias finales se presentan en el Capítulo 4.

CAPÍTULO 4

MAPEO POR ASOCIACIÓN PARA LA RESISTENCIA A LA PODREDUMBRE HÚMEDA DEL CAPÍTULO EN GIRASOL

Introducción

El mapeo por asociación es una estrategia poderosa que utiliza los eventos de recombinación históricos y la diversidad nucleotídica en una población determinada para encontrar genes causales de caracteres complejos con efectos moderados en el fenotipo (Hirschhorn y Daly 2005). Como se mencionó anteriormente, existen dos formas de abordar los estudios de asociación: (1) a través de la genotipificación de marcadores homogéneamente distribuidos en el genoma (*genome-wide association, GWA*), y (2) a través de la selección y genotipificación de genes candidatos para un determinado carácter. Si bien el abordaje GWA permite establecer asociaciones para múltiples caracteres fenotípicos utilizando siempre el mismo conjunto de marcadores, la estrategia de genes candidatos se adapta mejor al estudio de uno o dos caracteres en especies que poseen información genómica limitada, como girasol.

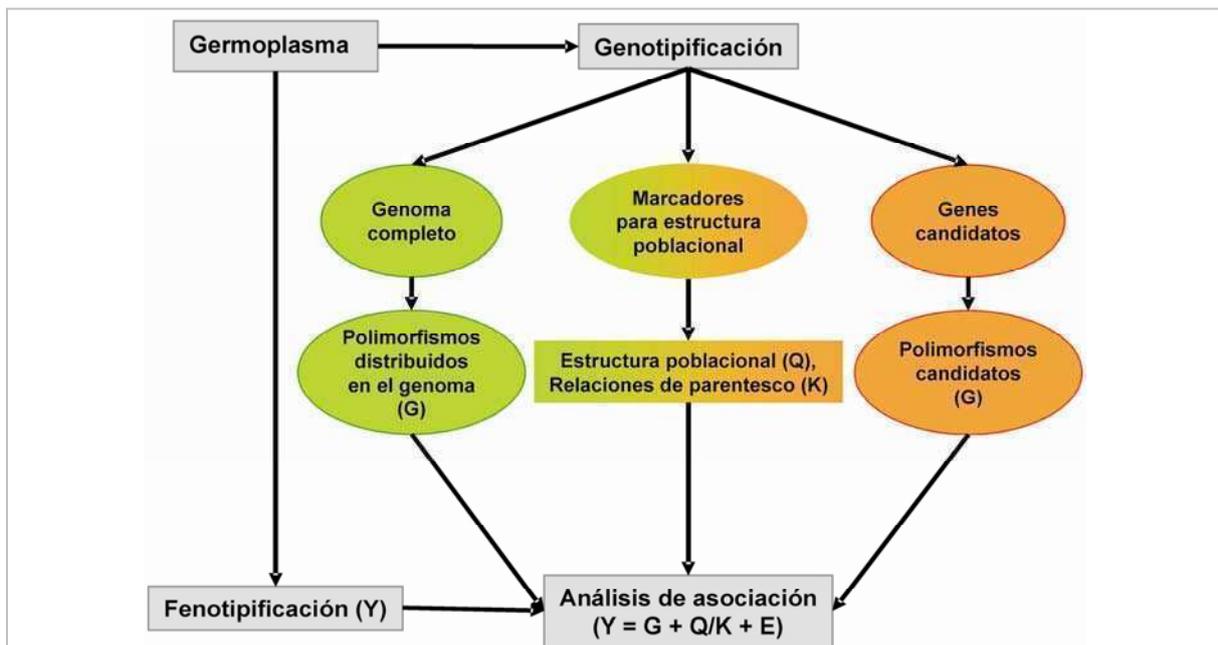


Figura 4.1. Posibles abordajes para el mapeo por asociación.

Componentes del mapeo por asociación: el germoplasma, la fenotipificación, la genotipificación y el análisis de asociación. Tanto para la estrategia de **genoma completo** (verde) como para la de **genes candidatos** (naranja) se obtiene un conjunto de polimorfismos (G). En ambos casos se realiza la inferencia de estructura poblacional (Q) y de las relaciones de parentesco (K) utilizando un conjunto independiente de marcadores moleculares.

El modelo estadístico del análisis de asociación intenta explicar la variación fenotípica (Y) en términos de la variación alélica (G), teniendo en cuenta la estructura y/o relaciones de parentesco de la población en estudio (Q y/o K). Un factor de error (E) completa el modelo.

Para encarar un estudio de mapeo por asociación es necesario determinar con precisión cada uno de los aspectos involucrados: (1) la **selección del germoplasma** que constituirá la población de mapeo por asociación (PMA), (2) la **fenotipificación** de la PMA para el carácter a evaluar (3) la **genotipificación** de la PMA a través de GWA o genes candidatos, (4) la comprobación de la existencia (o no) de **estructura poblacional** en la muestra, y (5) el **análisis estadístico** y los modelos de comparación de los genotipos y fenotipos para detectar asociaciones (Ersoz y col. 2007) (Figura 4.1).

La correcta selección del **germoplasma** a conformar la PMA es crítica para asegurar el éxito del mapeo por asociación (Zhu y col. 2008). Si bien es ideal trabajar con individuos no emparentados, esto a veces no es factible, por lo tanto, la población experimental debe ser lo más representativa posible de la población para la cual se realizará la inferencia de asociación (Ersoz y col. 2007). De acuerdo a la fuente de germoplasma, las PMA pueden clasificarse en: (1) colecciones núcleo de los banco de germoplasma, (2) poblaciones sintéticas o (3) líneas elite representativas del germoplasma utilizado en los programas de mejoramiento (Breseghello y Sorrells 2006a). Cualquiera sea la población de elección, debido a la historia de adaptación, selección, recombinación y endocria, la mayoría de las poblaciones disponibles presentan estructura y relaciones de parentesco que deben ser tenidas en cuenta para no sesgar los resultados de asociación (Yu y Buckler 2006). De todas maneras, de acuerdo a la PMA elegida o disponible, el enfoque del mapeo por asociación puede variar, según se detalla a continuación.

Las colecciones núcleo están conformadas por un número limitado de accesos que representan, en gran medida, la variabilidad genética presente en la especie. Sin embargo, la heterogeneidad genética que puede existir dentro de estas colecciones (razas nativas, poblaciones naturales), puede dificultar los distintos pasos del experimento (esquema de fenotipificación, extracción del ADN, genotipificación). Si bien estos materiales resultan interesantes para la asociación de caracteres a través de la estrategia de genes candidatos por la corta extensión del desequilibrio de ligamiento (DL), es probable que algunos no estén adaptados a las condiciones de crecimiento a campo y a las enfermedades más comunes del cultivo, lo cual

repercutirá en una medida fenotípica de baja calidad (Breseghello y Sorrells 2006a).

Las poblaciones sintéticas parecieran ser la mejor aproximación a una PMA ideal, dado que se diseñan y mantienen para minimizar los efectos de endocría. Si bien la diversidad alélica de estas poblaciones depende del número y divergencia de los fundadores utilizados para crear la población, la resolución de los análisis de asociación se maximiza debido a la existencia de panmixia y se minimizan los riesgos de encontrar falsos positivos (Breseghello y Sorrells 2006a). Sin embargo, si la población se encuentra sometida a selección de forma continua, los análisis de asociación deben contemplar la genotipificación de los individuos en las sucesivas generaciones para poder reflejar correctamente la constitución genética de la población en estudio, siendo ésta su principal desventaja de utilización.

Las líneas elite son una buena elección cuando se pretende estudiar caracteres complejos de baja heredabilidad, ya que son estables y están adaptadas a las condiciones de crecimiento normales del cultivo (Breseghello y Sorrells 2006a). En los programas de mejoramiento, las líneas emparentadas tienden a ser admitidas simultáneamente en las sucesivas etapas de selección de materiales, haciendo que la estructura poblacional se convierta inevitablemente en un factor a corregir, ya sea mediante información del *pedigree*, o utilizando marcadores moleculares. Para este tipo de PMA, la estrategia de GWA es la más apropiada, ya que el DL se mantiene alto con la distancia, reduciendo el número de marcadores necesarios para lograr una cobertura completa del genoma (Rafalski 2010). No obstante, aún para este tipo de PMA, en especies como maíz, trigo o girasol se necesitan demasiados marcadores para cubrir homogéneamente el genoma (más de 10.000 SNPs), haciendo que la estrategia de elección siga siendo la de genes candidatos. La gran variabilidad presente en este tipo de PMA puede conducir a la detección de alelos favorables, con posibilidad de incorporación inmediata en nuevas variedades, a través de cruzamientos dirigidos en los programas de mejoramiento.

Los **ensayos de fenotipificación** son otro aspecto importante del mapeo por asociación. Mientras que la precisión y eficiencia de la genotipificación ha aumentado notablemente, la obtención de medidas fenotípicas robustas parece ser un obstáculo

para los proyectos de mapeo por asociación a gran escala (Zhu y col. 2008). Por lo tanto, es importante planificar un diseño apropiado del ensayo (p. ej. bloques completos o incompletos con alfa látice), un método estadístico adecuado para el tratamiento de los datos (p. ej. utilizando modelos espaciales) y considerar las interacciones genotipo-ambiente (p. ej. mediante la repetición del ensayo en condiciones controladas o en distintos ambientes) (Eskridge 2003).

A pesar de que la estrategia GWA resulta más apropiada para la asociación de caracteres donde no se conoce en detalle la fisiología y genética que los gobiernan, la **estrategia de genotipificación** a través de genes candidatos es, en gran medida, la elegida debido a su menor costo de implementación (Rafalski 2002). De todas maneras, aún esta estrategia requiere contar con metodologías simples y que permitan la genotipificación de cientos de individuos de manera eficiente. Si bien la secuenciación directa de los genes candidatos es una metodología ampliamente utilizada por su rapidez y eficiencia, la genotipificación de indels (*insertion/deletion*) mediante electroforesis capilar fluorescente (ECF) y la detección de SNPs y/o indels mediante el análisis de las moléculas heterodúplex (CEL1CH y dHPLC, Cap. 2) son más económicas e igualmente eficientes y precisas (Comai y col. 2004, Giordano y col. 1999).

Los **niveles de estructuración y las relaciones de parentesco** en una PMA pueden determinarse a través de la información del *pedigree* o a partir de la selección de un conjunto de marcadores no ligados que permitan caracterizar la composición genética de los individuos. Los marcadores microsatélites (SSRs, *Simple Sequence Repeats*) son ideales para establecer estructura, relaciones de parentesco y presencia de endocria, debido a su polialelismo, codominancia y posibilidad de genotipificación automática. Los SNPs (*Single Nucleotide Polymorphisms*) también son utilizados para estimar y corregir estructura, sin embargo, al ser bialélicos se necesita de un mayor número de marcadores comparado con los SSRs (Zhu y col. 2008). Los métodos empleados para determinar estructura y relaciones de parentesco dependen del modelo estadístico de asociación a utilizar. Existen dos aproximaciones utilizadas en mapeo por asociación para detectar estructura poblacional: (1) a través del modelo bayesiano implementado en el programa STRUCTURE, en cuyo caso se obtiene la

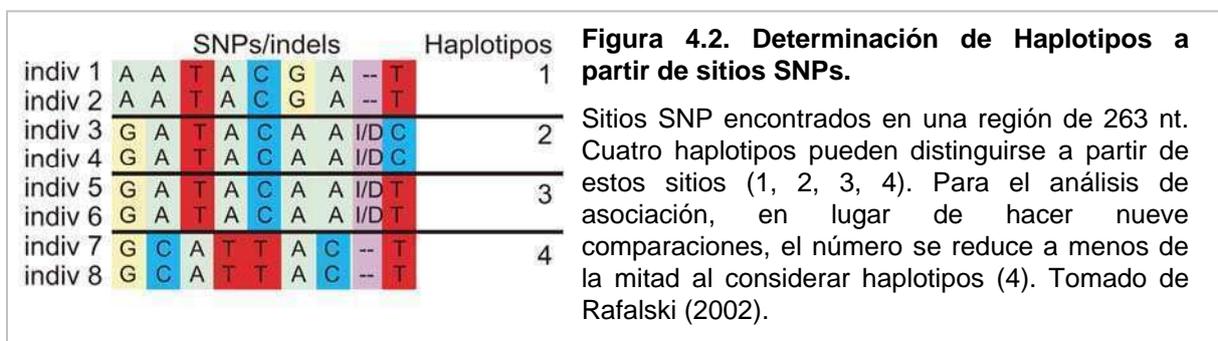
matriz de estructura poblacional, generalmente denominada Q (Figura 4.1) (Pritchard y col. 2000) y (2) a través de un análisis de coordenadas principales (ACP), para el cual la matriz obtenida se denomina P (Price y col. 2006). Asimismo, pueden inferirse las relaciones de parentesco entre los individuos a través de: (1) coeficientes de co-ancestralidad derivados de la información del *pedigree* o (2) coeficientes de parentesco derivados de la información provista por marcadores moleculares neutros (matrices denominadas K generalmente, Figura 4.1) (Bernardo 1993, Lynch y Ritland 1999).

El **modelo de asociación** a utilizar para el análisis estadístico de las asociaciones genotipo-fenotipo dependerá de la PMA seleccionada y de sus características. Si bien los modelos de regresión lineal, comparación de varianzas (ANOVA), pruebas-t y pruebas X^2 (chi-cuadrado) pueden resultar adecuados; en el caso de las especies vegetales la información de estructura poblacional no puede obviarse, y por ello se han desarrollado varios modelos que la contemplan (Stich y Melchinger 2009). Estos modelos comprenden: (1) modelos lineales (Brescaglio y Sorrells 2006b), (2) modelos de regresión logística (Thornsberry y col. 2001) y (3) modelos lineales mixtos (MLM), que contemplan tanto la estructuración poblacional (Q o P) como las relaciones de parentesco (K), denominándose QK o PK en cada caso (Stich y col. 2008, Yu y col. 2006).

A pesar de que muchos de los MLM utilizan STRUCTURE (Pritchard y col. 2000) para inferir estructura poblacional y eliminar eficientemente los errores tipo I del análisis, el uso de este programa presenta algunas desventajas. Entre las más importantes pueden mencionarse: (1) el gran requerimiento computacional, (2) las dificultades que generalmente implica establecer el verdadero número de poblaciones ideales que conforman la muestra y (3) la inestabilidad frente al supuesto de equilibrio de Hardy-Weinberg dentro de las poblaciones. Por lo tanto, los resultados obtenidos de este modo deben tratarse con cautela (Stich y col. 2008, Yu y col. 2006). Debido a esto, diversos autores han aplicado los MLM utilizando ACP para corregir las asociaciones espurias, demostrando que esta aproximación tiene el mismo poder resolutivo, con la consecuente ganancia de tiempo (Stich y col. 2008, Zhao y col. 2007b). Adicionalmente, Stich y col. (2008) modificaron además la

matriz de parentesco (K), de acuerdo a lo propuesto por Bernardo (1993), en donde se corrige el coeficiente de parentesco para los alelos idénticos en estado pero no idénticos por descendencia, disminuyendo aún más los errores de asociación tipo I y aumentando el poder de detección de asociación (Stich y col. 2008).

Los análisis de asociación pueden aplicarse teniendo en cuenta sitios SNP individuales, o bien, haplotipos. Se denominan haplotipos a los bloques polimórficos en desequilibrio de ligamiento (Figura 4.2). Debido a las correcciones que deben hacerse por comparaciones múltiples, los análisis de asociación basados en haplotipos suelen ser más poderosos en detectar asociaciones que al utilizar los sitios por separado (Rafalski 2010).



Las correcciones por comparaciones múltiples que generalmente se utilizan son: (1) la prueba de Bonferroni, que generalmente conlleva a la pérdida de poder de la prueba de asociación (Ersoz y col. 2007), (2) el método FDR (*false discovery rate*), que es el más utilizado y el que mejor se adapta a comparaciones múltiples o cuando se comparan marcadores que no estén en fuerte desequilibrio de ligamiento (Benjamini y Hochberg 1995, Storey 2002, Storey y Tibshirani 2003), (3) las pruebas de permutación que controlan la tasa de error para GWA (Churchill y Doerge 1994) y (4) las estimaciones no paramétricas a través de la comparación de las distribuciones de los p-valores con valores fenotípicos simulados (Ersoz y col. 2007). No obstante, es necesario evaluar en qué casos aplicarlas, debido a que estas correcciones aumentan la tasa de error tipo II, pudiendo llevar a la pérdida total del poder de detección de asociación.

En cuanto a la metodología de aplicación de los estudios de mapeo por asociación en especies vegetales, los modelos estadísticos utilizados varían en cada caso, siendo los

MLM los más populares actualmente, ya sea a través de las aproximaciones QK o PK (p. ej. en (Ehrenreich y col. 2009, Ingvarsson y col. 2008, Skot y col. 2007, Zhao y col. 2007b)). Excepto para los estudios realizados en pino, eucalipto y raigrás, que utilizan poblaciones naturales o germoplasma diverso (Ingvarsson y col. 2008, Skot y col. 2007, Thumma y col. 2005), la mayoría de las PMA utilizadas poseen estructura, siendo las líneas endocriadas las más utilizadas en los estudios de asociación para cultivos agronómicos (Casa y col. 2008, Gonzalez-Martinez y col. 2006, Stich y Melchinger 2009, Thornsberry y col. 2001). Erzos y col. (2007) sugieren que la cantidad de individuos necesarios para establecer asociaciones significativas es de 300-500, sin embargo, la mayoría de los trabajos realizados hasta el momento involucran alrededor de 100 individuos (Bressegello y Sorrells 2006b, Palaisa y col. 2004, Thornsberry y col. 2001, Zhao y col. 2007b).

En girasol, a pesar de que la información sobre marcadores moleculares funcionales (SNPs) ha aumentado notablemente en los últimos cinco años (Kolkman y col. 2007, Lai y col. 2005, Liu y Burke 2006), la estrategia de mapeo por asociación aún no ha sido implementada en la especie. El INTA cuenta con todas las herramientas necesarias para establecer un estudio de asociación en girasol para las enfermedades más importantes del cultivo. Por ejemplo, las líneas elite disponibles en el Banco Activo de Girasol-EEA INTA Manfredi (BAG-IM) conforman un material de partida diverso y propicio para la realización de este tipo de estudios. Estas líneas, al formar parte de los programas de mejoramiento llevados a cabo en INTA, se evalúan anualmente para caracteres agronómicos de importancia, incluyendo la Podredumbre Húmeda del Capítulo (PHC). El Grupo de Fitopatología de Girasol de la EEA Balcarce (GFPG EEA Balcarce) posee amplia experiencia e idoneidad en el diseño experimental y la evaluación de las medidas fenotípicas para PHC utilizando un método de infección asistido que reproduce la infección causada por *S. sclerotiorum* y evita la variación en el grado de infección entre plantas causada por diferencia en la cantidad de ascosporas recibidas (Maringolo 2007). Debido a esto, y dado que PHC constituye un problema para la producción del cultivo, éste carácter es uno de los blancos más interesantes para estudios de asociación.

La estrategia de genes candidatos, a pesar de la extensión del DL en girasol (~100

kb), pareciera ser más apropiada que GWA, reduciendo los costos y la cantidad de marcadores a genotipificar. Sin embargo, al utilizar esta aproximación debe recurrirse a la determinación de estructura poblacional y relaciones de parentesco con un conjunto adicional de marcadores. Los marcadores microsatélites (SSRs) constituyen la fuente más importante de variación neutra en el genoma de girasol (Paniego y col. 2002, Tang y col. 2002, Tang y col. 2003), por lo que se cuenta con una herramienta efectiva para la detección de diversidad que permite inferir estructura poblacional y relaciones de parentesco, y, de esta manera, evitar asociaciones espurias.

Por todo lo expuesto, el objetivo de este capítulo implicó la realización de un estudio de mapeo por asociación para el carácter “resistencia a PHC”, incorporando en el modelo de asociación la información sobre estructura y parentesco a través de marcadores SSRs y utilizando SNPs detectados en genes candidatos identificados como partícipes de los procesos de defensa en girasol.

Materiales y Métodos

Material vegetal

Se trabajó con una población de 134 líneas puras de girasol denominada Población de Mapeo por Asociación (PMA, Tabla A4.1, Apéndice), compuesta por 124 líneas correspondientes al programa de mejoramiento para PHC del BAG-IM, más 10 líneas de comportamiento contrastante al ataque de *S. sclerotiorum* seleccionadas por el GFPG-EEA INTA Balcarce y utilizadas en los capítulos anteriores de este trabajo de tesis (HA89, PAC2, RHA266 y RHA801, Cap. 1; HA853, RHA275, RK416, HA441, RHA440 y RHA439, Cap. 2). Se eligieron líneas endocriadas o de endocría avanzada para evitar el problema de la heterocigosis y disminuir la cantidad de individuos a analizar. Estos accesos contaban con datos sobre incidencia de infección frente a *S. sclerotiorum* en distintas campañas (2002-03 hasta 2007-08), demostrando una amplia diversidad en la respuesta fenotípica observada y permitiendo su uso para este tipo de estudios. Durante el ensayo a campo 2008-09 (ver detalles en evaluación fenotípica, este capítulo), se recolectaron hojas de plantas jóvenes (tres semanas de edad), para la extracción de ADN y el análisis genotípico. Se siguió el protocolo de recolección y extracción de ADN descrito previamente en el Cap. 1.

Preparación del inóculo

Se utilizó una población de *S. sclerotiorum* de Balcarce (Buenos Aires, Argentina). Los esclerocios fueron recolectados en el campo y almacenados en bolsas de papel a 13°C por tres meses. La producción de apotecios se realizó mediante el procedimiento descrito en el trabajo de Escande y col. (2002). Brevemente, los esclerocios previamente expuestos a $-18^{\circ}\text{C} \pm 2^{\circ}\text{C}$ durante siete días se enterraron a 1 cm de profundidad en suelo húmedo pasteurizado hasta su emergencia. Los cultivos se incubaron a 16°C y a 2500 lux de luz continua, aproximadamente. Los apotecios maduros se colectaron hacia arriba en una caja de Petri y se incubaron durante 4 h para favorecer la generación de ascosporas. Finalmente, las ascosporas se reservaron a -18°C hasta su uso. Para la inoculación de los capítulos, las ascosporas se resuspendieron en agua bidestilada estéril a una concentración de 2500 ascosporas/ml.

Evaluación fenotípica

Para evitar variación entre las medidas fenotípicas y determinar con mayor precisión el comportamiento de los cultivares frente a *S. sclerotiorum*, el GFPG-EEA INTA Balcarce utiliza exitosamente un método de inoculación asistida con esporas del hongo. Los ensayos a campo se realizaron en la EEA INTA Balcarce (Buenos Aires, Argentina) en dos campañas sucesivas: 2008-09 y 2009-10, en condiciones de secano. Las semillas se plantaron en el campo experimental de acuerdo a un diseño de dos bloques completos aleatorizados, en un suelo Arguidol típico conteniendo 5% de materia orgánica a pH 6,2. Los surcos (6 m de longitud) conteniendo 30 plantas por acceso de la PMA, se espaciaron entre sí 70 cm y la distancia entre los bloques fue de 1 m. Los capítulos de girasol se rociaron con la suspensión de ascosporas de *S. sclerotiorum* cuando se encontraron en el estadio R5.2. Este estadio se corresponde al momento en que el 20% de las flores externas se encuentran en antesis (Schneiter y Miller 1981). Inmediatamente, los capítulos se cubrieron con bolsas de papel manila (desde la inoculación hasta 15 dpi). La medida fenotípica relacionada con la infección por *S. sclerotiorum* registrada fue la **Incidencia de PHC** que se obtiene como el número de plantas enfermas sobre el número de plantas totales de cada línea. Si bien este parámetro se determina cada 10 días a partir de la inoculación, se tomó el valor de incidencia al final del ensayo a campo.

Evaluación genotípica

Genotipificación de SSRs para análisis de la estructura poblacional

Para determinar la estructura de la población se genotipificaron ocho SSRs en los 134 accesos que conforman la PMA. Estos SSRs se seleccionaron por su ubicación en diferentes grupos de ligamiento (GL) del mapa de referencia de girasol y por su reproducibilidad y facilidad de interpretación en un estudio de diversidad realizado en aproximadamente 300 accesos del BAG-IM (Poormohammad Kiani y col. 2007a) (MSc. María V. Moreno, com. pers.). Las características de los SSRs y los iniciadores utilizados para su amplificación se muestran en la Tabla 4.1. Se genotipificaron tres individuos para 130 líneas y dos individuos para las cuatro líneas restantes (RK416, GP365, 2021 y 2067), haciendo un total de 398 individuos genotipificados mediante

electroforesis capilar fluorescente (ECF). Los detalles metodológicos de las corridas de ECF se detallan en el Cap. 2. Los genotipos de cada línea se establecieron por el consenso de los alelos observados en por lo menos dos de los tres individuos genotificados. Es decir: (1) si dos individuos poseían el mismo genotipo y el tercer individuo poseía al menos un alelo idéntico al de los otros individuos, el genotipo de la línea se consideró homocigota para el alelo mayoritario; (2) si dos individuos eran heterocigotas y el tercer individuo era homocigota para alguno de los alelos observados en los otros dos individuos, el genotipo de la línea se consideró heterocigota para esos alelos, (3) si 2/3 individuos tenían genotipo homocigota para dos alelos diferentes, la línea se consideró heterocigota para esos dos alelos y (4) si los individuos genotificados presentaban tres o más alelos en total, el genotipo de la línea se consideró dato faltante.

Tabla 4.1. Características de los SSRs utilizados para la evaluación de la estructura poblacional e iniciadores utilizados para su genotipificación.

Nombre SSR	Nº de Acceso ^a	Tipo	Motivo	Fluoróforo	Tamaño (pb)	Iniciadores	GL
HA77	BV727857	Dinucleótido	GA	HEX	249	F: 5' TGTAATCTGTATCACTTCCACC 3' R: 5' GTTGTTCTGTTAGGTCGTTCCG 3'	1
HA293	G67519	Dinucleótido	GA	HEX	120	F: 5' GGGGACATCTCCCGTCCACC 3' R: 5' CCTCATCCATCTCCATCCAATC 3'	14
HA928	BV728013	di-trinucleótido	GT-ATT	HEX	220	F: 5' CCTTTGTAGTCCCTTACTGG 3' R: 5' GGTCGATCATGTATGCGTGTG 3'	10
HA991	G67520	Dinucleótido	GA	FAM	146	F: 5' GCCCCCTTGATGCCCTTTTC 3' R: 5' GAATCGCCATTTGAATCGCCAG 3'	4
HA2063	BV728113	Dinucleótido	GA-GT	FAM	171	F: 5' CACCGCAGATGGCCAATTCAAC 3' R: 5' GGTTCTGCTAACCTGCAATGCG 3'	9
HA2920	BV728085	Trinucleótido	ATG	FAM	177	F: 5' AACGTGTTTCATTCCGGATGC 3' R: 5' GAAGATTGGTATGATGGGGC 3'	3
HA3239	BV728208	Trinucleótido	ATG	NED	142	F: 5' CTTAATTCCGACATAAGCAGG 3' R: 5' CTTTCATTATTCACAACCCCC 3'	2
HA4103	BV728338	Dinucleótido	GA	FAM	190	F: 5' TCACTTATCACCAATCTCATCCA 3' R: 5' CTGTCTCAAATCGGGTGGTT 3'	6

^a Número de acceso del GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>)

Genotipificación de genes candidatos para mapeo por asociación

A partir del análisis de los SSRs, para las líneas que presentaron los tres individuos con genotipo idéntico, los genes candidatos detallados en el Cap. 3 se genotipificaron sobre uno o dos individuos solamente. Para el resto de las líneas, se genotipificaron los tres individuos para la totalidad de genes candidatos. Para ello se utilizaron uno de los siguientes métodos: secuenciación directa, detección de heterodúplex mediante dHPLC o diferencia del tamaño del fragmento mediante ECF.

El genotipo de los individuos, obtenido por cualquiera de estos métodos, fue asociado a un determinado haplotipo (Figura 4.2). Dado que se genotificaron hasta tres individuos por línea, el haplotipo de la línea se determinó como el consenso de los haplotipos observados en los individuos analizados. De este modo, el criterio para la asignación del haplotipo de la línea fue que por lo menos dos de los tres individuos genotificados posean el mismo haplotipo para ese gen candidato. Los individuos heterocigotas, debido a la dificultad en la determinación de la fase en que se hallan los alelos para dos *loci*, se consideraron datos faltantes. Es decir, si un *locus* presenta un polimorfismo A/G y el *locus* adyacente un polimorfismo C/T, no es inmediatamente evidente si los haplotipos del individuo heterocigota son **A**(*locus* 1)-**C**(*locus* 2) y **G**(*locus* 1)-**T**(*locus* 2), o **A**(*locus*1)-**T**(*locus* 2) y **G**(*locus*1)-**C**(*locus* 2).

La selección de los métodos de genotificación a utilizar en la caracterización de la PMA se estableció de acuerdo al costo general del ensayo y a las posibilidades técnicas disponibles. De este modo, cuando los genes presentaron únicamente dos haplotipos en el Grupo Núcleo (GN), o cuando el gen pudo acotarse a una región donde se establecieran dos haplotipos únicos, se realizó la genotificación de la PMA mediante la detección de moléculas heterodúplex con dHPLC. Los genes candidatos que cumplieron esta condición fueron: PRP_2, CytP450_C, PP2C, bHLHP, HaDRP, HaCP, HaPAL, HaGLP2, HaGLP3 y HaGLP5. Para cada individuo se realizaron tres inyecciones de muestra en el HPLC, dos inyecciones correspondientes a las mezclas heterodúplex con cada uno de los productos de PCR de las líneas control representantes de los dos haplotipos encontrados, y una inyección del control homodúplex correspondiente al producto de PCR del individuo de genotipo incógnita. Los detalles de la asignación del haplotipo de cada individuo se encuentran en el Cap. 2.

Cuando las regiones presentaron indels o microsatélites, se realizó la genotificación de la PMA por diferencia en el tamaño del fragmento amplificado mediante ECF (detalles de la técnica en el Cap. 2). Los genes seleccionados para este método de genotificación fueron MT, RGEF_A2, RhoBP_A, RhoBP_B, HaWP, HaTRP, HaWRKY5, HaWRKY7 y HaGLP4.

Para aquellos casos en los cuales el tamaño del fragmento del producto de PCR del individuo no se correspondió con ninguno de los identificados en las líneas del GN, los genes candidatos se amplificaron con iniciadores no marcados y se secuenciaron directamente para determinar los haplotipos nuevos (protocolo de secuenciación en Cap. 1).

Finalmente, se seleccionaron para secuenciación directa de los productos de PCR a aquellos genes que presentaron varios SNPs y más de dos haplotipos entre las 10 líneas del GN, siendo ZFHDP y HaGDP-DI los que cumplieron esta condición. La secuenciación directa de productos de PCR se realizó con un único iniciador, que fue seleccionado previamente para obtener el mejor rendimiento en cuanto a cantidad de pb secuenciadas y calidad de la secuencia. Para esta técnica se procedió de acuerdo a lo descrito en el Cap. 1.

Todos los iniciadores diseñados y utilizados para la genotipificación de las 134 líneas de la PMA se detallan en la Tabla A4.2 (Apéndice).

Análisis estadístico

Para realizar el análisis estadístico se siguieron los lineamientos propuestos en Stich y col. (2008).

Para ello, las medias ajustadas de la incidencia de PHC por línea se ajustaron al modelo según la ecuación E4.1. El modelo ajustado contempló a las “líneas” como efecto fijo y al “año de realización de los experimentos” y los “bloques dentro de años” como efectos aleatorios.

$$Y_{ijk} = \mu + Linea_i + Año_j + Bloque(Año)_{jk} + \varepsilon_{ijk} \quad (E4.1)$$

En E4.1, Y_{ijk} representa la incidencia de PHC, μ una media general común a todas las observaciones, $Linea_i$ el efecto (fijo) de la i -ésima línea, $Año_j$ el efecto aleatorio del j -ésimo año en que se realizó el experimento (años 2008-09 y 2009-10), $Bloque(Año)_{jk}$ el efecto aleatorio del k -ésimo bloque dentro del j -ésimo año y ε_{ijk} el error aleatorio de cada observación. El modelo se ajustó para todas las líneas evaluadas. Las medias

ajustadas se obtuvieron mediante la siguiente expresión:

$$M_i = \hat{\mu} + \hat{L}ínea_i$$

Para establecer asociaciones, se utilizó un modelo lineal mixto (MLM). Las medias ajustadas para el j -ésimo gen candidato fueron modeladas según la siguiente expresión:

$$\mathbf{M} = \boldsymbol{\mu} + \mathbf{m}_j + (\mathbf{P} \text{ ó } \mathbf{Q}) \boldsymbol{\beta} + \mathbf{U}\mathbf{b} + \boldsymbol{\varepsilon} \quad (\text{E 4.2})$$

Donde \mathbf{M} es el vector de incidencia estimada de PHC de longitud igual a L (número de líneas), \mathbf{m}_j es el vector de estados (haplotipos de las líneas) del gen candidato j -ésimo, \mathbf{P} (ó \mathbf{Q}) son matrices $L \times P$ (ó $L \times Q$) de regresoras que se incluyen para explicar la variación entre líneas debida a la estructura poblacional y $\boldsymbol{\beta}$ el vector de dimensión P (ó Q) que contienen los coeficientes de regresión que acompañan a \mathbf{P} (ó \mathbf{Q}). \mathbf{U} es una matriz de pesos para los efectos aleatorios de dimensiones $L \times L$ y \mathbf{b} el vector de dimensión L de efectos aleatorios de las líneas con esperanza cero y matriz de covarianzas $\sigma_b^2 \mathbf{K}$, siendo \mathbf{K} la matriz de parentesco entre la líneas. Finalmente, $\boldsymbol{\varepsilon}$ es un vector de dimensión L conteniendo los términos de error que se suponen normales con esperanza cero y matriz de covarianzas $\sigma_\varepsilon^2 \mathbf{I}$.

La matriz \mathbf{P} para determinar estructura poblacional se obtuvo a través del análisis de coordenadas principales (ACP) a partir de la distancia de Rogers (1972) obtenida con los datos de la matriz de frecuencias alélicas \mathbf{F} , utilizando el programa Infostat (Di Rienzo y col. 2009). La matriz \mathbf{F} , a su vez, se calculó a partir de los marcadores SSR (Tabla A4.3, Apéndice). Con la inclusión de la matriz \mathbf{P} se pretendió capturar los niveles superiores de estructuración poblacional (Zhao y col. 2007b).

La matriz \mathbf{Q} de estructura poblacional se obtuvo a través del método bayesiano implementado en el programa STRUCTURE (Pritchard y col. 2000), a partir de la matriz de SSRs (Tabla A4.3, Apéndice). El número de grupos evaluados osciló entre uno y 12, utilizando cinco réplicas por cada valor K , un largo del período de *burn-in* de 100.000 iteraciones, y un largo de la corrida de 5×10^5 iteraciones. Las frecuencias alélicas se mantuvieron correlacionadas. No se utilizó ninguna información adicional sobre el origen de los individuos para definir los grupos. Para cada valor K , se tuvo

en cuenta la corrida que mostró la mayor probabilidad *a posteriori* de los datos.

La estructura de parentesco se incorporó mediante la matriz de pesos \mathbf{U} . Para calcular \mathbf{U} se define inicialmente la matriz $\mathbf{S} = \{S_{ij}\}$, donde S_{ij} es la proporción de marcadores (SSRs) con variantes coincidentes entre las líneas i y j . A partir de esta matriz, y utilizando la expresión $K_{ij}^T = \frac{S_{ij} - 1}{1 - T} + 1$, se obtiene lo que se denomina matriz de parentesco (Bernardo 1993), donde T es la probabilidad de que dos marcadores sean iguales en estado pero distintos por descendencia. Finalmente, la matriz \mathbf{U} es tal que $\mathbf{K} = \mathbf{U}'\mathbf{U}$. La probabilidad T es desconocida, pero se puede estimar maximizando la verosimilitud para el modelo E4.2. Se eligieron valores de $T=0,20$; $0,30$; $0,70$ y $0,80$, pero sólo se muestran los valores $T=0,20$ y $T=0,80$. Este modelo fue estimado mediante la rutina **lme** del paquete **R** (Ihaka y Gentleman 1996).

No se realizaron pruebas de comparaciones múltiples debido al bajo número de marcadores evaluados simultáneamente, con el fin de evitar un incremento del error tipo II en el análisis.

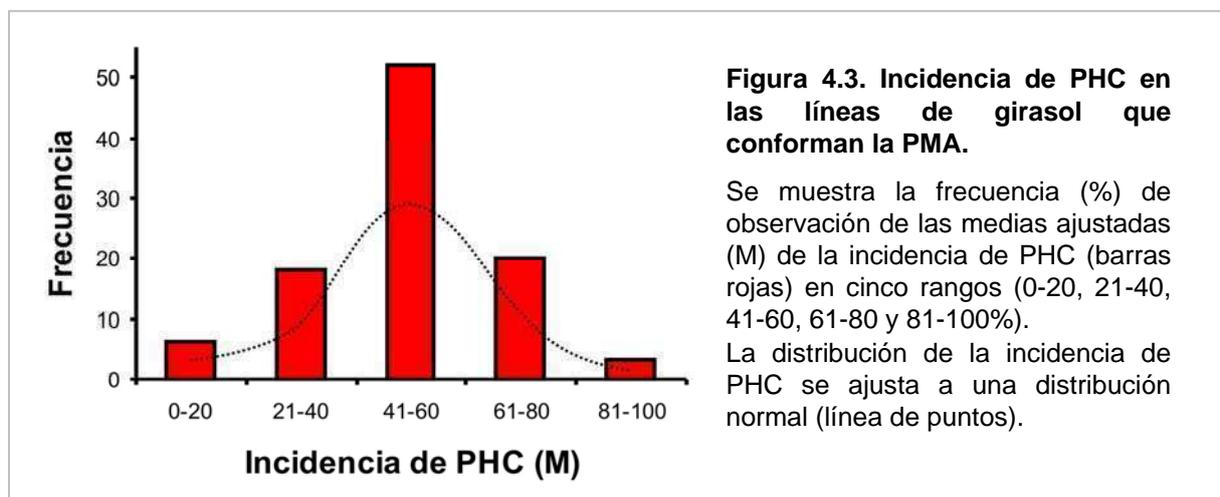
Construcción del mapa genético

La genotipificación de la población RIL mediante ECF y la construcción del mapa de ligamiento se realizaron de acuerdo a lo descrito en el Cap. 2.

Resultados

Evaluación fenotípica de las líneas analizadas

Se evaluaron fenotípicamente 134 líneas endocriadas provenientes del BAG-IM en dos campañas sucesivas: 2008-09 y 2009-10 y en las instalaciones de la EEA Balcarce. Las condiciones climáticas fueron muy diferentes entre ensayos: la campaña 2008-09 presentó condiciones de baja humedad y escasas precipitaciones en comparación con la campaña 2009-10, que se comportó de manera contraria y en concordancia con la media de precipitaciones esperada para la zona de Balcarce. La sequía que se presentó en la campaña 2008-09 afectó en el número de líneas que pudieron ser evaluadas durante el ensayo. La Figura 4.3 muestra la diversidad de la respuesta frente a PHC en el conjunto de líneas endocriadas ensayadas, presentando valores desde 0% (cero) hasta el 100% de incidencia de PHC (valores correspondientes a las medias ajustadas finales). De acuerdo a estos resultados, el 52% de las líneas se comportó de manera intermedia, alcanzando valores de incidencia de PHC de entre un 40 y un 60% (Figura 4.3).



Genotipificación de SSRs para análisis de estructura poblacional

La amplificación por PCR de los ocho SSRs seleccionados para el cálculo de estructura poblacional presentó algunas dificultades, fallando para el 38% de los individuos en la primera ronda de amplificación. Posteriormente, se ajustaron las concentraciones de iniciadores en la reacción de PCR, logrando que el porcentaje de datos faltantes totales se redujera al 5% (154 genotipos). En particular, HA77 y

HA3239 conformaron los *loci* con mayor porcentaje de datos faltantes debido a problemas en la PCR (9% y 11%, respectivamente). La amplificación de los restantes *loci* fue exitosa, evidenciando entre el 0,8% y el 5% de datos faltantes (*loci* HA991 y HA4103, respectivamente) por fallas en la PCR. Con respecto al número de individuos genotificados para cada *locus*, éste varió entre 395 para HA991 y 355 para HA3239.

La caracterización de la PMA mediante SSRs evidenció una gran diversidad molecular entre las líneas seleccionadas para este estudio (Tabla 4.2). El número de alelos detectados para cada uno de los ocho *loci* genotificados fue variable, siendo cinco el número mínimo y 12 el número máximo de alelos encontrados (*loci* HA4103 y HA991, respectivamente). Los tamaños alélicos encontrados se mantuvieron dentro del rango de tamaños esperado para siete de los ocho SSRs analizados, de acuerdo a la información de la secuencia depositada en la bases de datos UniSTS de GenBank (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=unists>). Sin embargo, para HA4103 se detectaron alelos cuyo tamaño difirió del esperado (152-138, Tabla 4.2 vs. 190, Tabla 4.1).

Tabla 4.2. Datos de los SSRs genotificados en la PMA.

Nombre SSR	Nº de alelos	Rango (pb)	Nº Individuos genotificados ^a	Nº Líneas genotificadas ^b	% Líneas genotificadas
HA77	6	262-239	362	123	91,8%
HA293	7	133-113	377	117	87,3%
HA928	7	227-209	385	130	97,0%
HA991	12	157-127	395	128	95,5%
HA2063	9	180-158	388	121	90,3%
HA2920	9	187-161	384	127	94,8%
HA3239	6	144-126	355	119	88,8%
HA4103	5	152-138	380	130	97,0%

a Total de individuos: 398

b Total de líneas: 134

Una de las primeras dificultades detectadas al realizar la genotipificación mediante SSRs fue que, a diferencia de la homocigosis y homogeneidad esperada en la genotipificación de líneas endocriadas, se observó una gran variabilidad entre los individuos pertenecientes a una misma línea. Entre los casos más significativos se encuentran: (1) líneas en donde los tres individuos dieron genotipos distintos entre sí

(p. ej. para el *locus* HA2920, los genotipos identificados para la línea C432-2 fueron 173/173, 170/170 y 175/175), (2) líneas donde dos de los individuos genotipificados fueron homocigotas para dos alelos distintos y el tercer individuo heterocigota para un tercer alelo (p. ej. para el *locus* HA293 los individuos de la línea 2125 fueron 119/119, 121/121 y 115/121), y (3) líneas donde dos individuos resultaron heterocigotas para los mismos alelos y el tercer individuo homocigota para otro alelo (p. ej. para el *locus* HA2063, los individuos de la línea C173 presentaron los genotipos 174/180, 174/180 y 178/178). Sumado a estos inconvenientes en la asignación de un genotipo consenso para la línea, se encontraron otros problemas menores, donde los genotipos de dos individuos fueron homocigotas para un alelo y el tercer individuo resultó homocigota para otro alelo (p. ej. para el *locus* HA77, la línea C618-1M presentó dos individuos con genotipo 247/247 y un individuo con genotipo 259/259). Finalmente, las líneas menos conflictivas fueron las que presentaron dos individuos con genotipo homocigota idéntico y el individuo restante con genotipo heterocigota, presentando el alelo de los individuos anteriores (p. ej. para el *locus* HA3239, los individuos de la línea RHA439 resultaron 144/144, 144/144 y 137/144). Para 47 líneas los tres individuos tuvieron genotipo idéntico.

Debido a esta variabilidad intra-línea, se reconstruyó el genotipo consenso de las mismas de acuerdo al criterio establecido en Materiales y Métodos. De este modo, 83 líneas contaron con información para los ocho *loci* SSRs genotipificados. El número de líneas para las cuales no se pudieron reconstruir los genotipos, varió de *locus* en *locus*, por ejemplo, mientras que con 362 individuos genotipificados fue posible reconstruir el genotipo de 123 líneas para el *locus* HA77, con 377 individuos genotipificados para el *locus* HA293 sólo fue posible reconstruir el genotipo de 117 líneas (Tabla 4.2). Una situación similar se observa al comparar los datos de los *loci* HA928 y HA991, en donde los individuos genotipificados para uno y otro *locus* fueron 385 y 395, mientras que los genotipos reconstruidos a partir de esa información fueron 130 y 128, respectivamente (Tabla 4.2). Debido a que la cantidad de SSRs genotipificados fue limitada, se consideró aceptable que la línea posea información para al menos 7/8 *loci* SSRs.

Genotipificación de genes candidatos para el análisis de asociación

Considerando los resultados del análisis de los SSRs, nueve de las 47 líneas cuyos individuos presentaron genotipo idéntico fueron representadas por uno o dos individuos en la caracterización de los genes candidatos, a fin de ajustar el total de muestras a sistemas de placas de 384 pocillos.

Como se describió en el Cap. 3, se seleccionaron 21 genes candidatos para la resistencia a PHC, de los cuales 16 pudieron ser genotipificados satisfactoriamente. La genotipificación mediante dHPLC de los genes HaDRP, HaCP, HaGLP2, HaGLP3 y HaGLP5 no pudo optimizarse adecuadamente, siendo necesarias evaluaciones adicionales de las condiciones de gradiente y temperatura para detectar diferencias entre los tiempos de elución de las moléculas hetero- y homodúplex.

Si bien el método de genotipificación de cada gen candidato se eligió *a priori*, por observación de la cantidad de haplotipos y sitios presentes en el GN de líneas secuenciadas, no siempre fue posible adoptar el método pre-seleccionado. Por ejemplo, los genes bHLHP y PP2C que poseían características adecuadas para ser genotipificados por dHPLC, debieron ser genotipificados mediante secuenciación directa en la PMA (Tabla 4.3). En el caso del gen bHLHP, que poseía un sitio SNP, ninguna de las condiciones preestablecidas por el programa *dHPLC MeltProgram* permitió la separación de las moléculas heterodúplex de las homodúplex. Por otra parte, para el gen PP2C se encontraron las condiciones óptimas de genotipificación mediante dHPLC, sin embargo, al amplificar el gen mediante PCR sobre la PMA, se detectó un indel que no había sido identificado en el GN. Este hecho implicó un replanteo del método y los gastos implicados en ello. Debido a que la genotipificación mediante dHPLC requeriría nuevas evaluaciones para optimizar las condiciones de detección de este nuevo haplotipo, y esto, a su vez, implicaría una corrida extra en dHPLC por cada individuo de genotipo incógnita, se decidió secuenciar directamente el producto de PCR de PP2C con el iniciador sentido utilizado en la amplificación.

Finalmente, tres genes se genotipificaron por dHPLC (PRP_2, CytP450_C y HaPAL), nueve genes por ECF (MT, RGEF_A2, RhoBP_A, RhoBP_B, HaWP, HaTRP,

HaWRKY5, HaWRKY y HaGLP4) y cuatro genes por secuenciación directa del fragmento de PCR (PP2C, bHLHP, ZFHDP y HaGDP-DI) (Tabla 4.3).

Tabla 4.3. Resultado de la genotipificación de los genes candidatos en la PMA.

Gen	Método ^a	Nº Haplotipos ^b	Nº Individuos genotipificados	Nº Individuos heterocigotas	Nº Líneas genotipificadas ^c
MT	ECF	2	375	23	122
PRP_2	dHPLC	2	384	28	123
CytP450_C	dHPLC	6	384	31	123
RGEF_A2	ECF	2	378	29	116
PP2C	Secuenciación	4	374	28	120
RhoBP_A	ECF	2	381	15	125
RhoBP_B	ECF	3	384	9	130
bHLHP	Secuenciación	5	377	30	122
ZFHDP	Secuenciación	9	377	42	112
HaGDP-DI	Secuenciación	10	375	47	107
HaWP	ECF	2	381	16	128
HaTRP	ECF	2	369	9	126
HaPAL	dHPLC	2	377	26	120
HaWRKY5	ECF	6	374	13	128
HaWRKY7	ECF	2	375	32	121
HaGLP4	ECF	4	369	69	98

a Método de genotipificación final utilizado. ECF: electroforesis capilar fluorescente, dHPLC: cromatografía líquida de alto rendimiento en condiciones desnaturizantes, Secuenciación directa del producto de PCR.

b Número de haplotipos detectados en la genotipificación de 384 individuos. Los genotipos finales de las líneas mostraron todos los haplotipos excepto para HaGDP-DI, para el cual sólo se observaron seis haplotipos en lugar de 10.

c Total de líneas 134.

La caracterización de los 16 genes candidatos en la PMA fue exitosa, registrando un 98% de los individuos genotipificados para el total de los genes. Se identificaron 10 haplotipos como máximo para el gen HaGDP-DI, y ocho genes presentaron únicamente dos haplotipos en la PMA (MT, PRP_2, RGEF_A2, RhoBP_A, HaWP, HaTRP, HaPAL y HaWRKY7) (Tabla 4.3). La cantidad de individuos genotipificados fue variable, sin mostrar relación con el método utilizado en la genotipificación. Por ejemplo, 384 individuos fueron genotipificados para PRP_2, CytP450_C (dHPLC) y RhoBP_B (ECF). Los genes cuyo haplotipo se definió por secuenciación presentaron valores intermedios de individuos genotipificados (374 a 377), siendo ZFHDP y HaGDP-DI los genes secuenciados que más individuos heterocigotas evidenciaron (42 y 47, respectivamente), después de HaGLP4 (69), genotipificado mediante ECF.

Como se menciona en Materiales y Métodos, se definió un criterio de “reconstrucción” del haplotipo de la línea a partir de la información de los haplotipos individuales. Este criterio fue que al menos dos de los individuos genotipificados dieran haplotipos idénticos entre sí. A diferencia del criterio asumido para los SSRs, en el caso de genes candidatos se consideró únicamente el alelo mayoritario y los individuos heterocigotas no se tuvieron en cuenta para consensuar el haplotipo de la línea. En este aspecto, RhoBP_B, HaWP y HaWRKY5 fueron los genes con menos individuos heterocigotas y mayor cantidad de líneas para las cuales pudo determinarse un genotipo consenso, siendo 130 para RhoBP_B, y 128 para HaWP y HaWRKY5 (Tabla 4.3).

Estructura poblacional y análisis de asociación

A partir de los datos obtenidos sobre los SSRs y genes candidatos, el análisis de asociación se realizó sobre un conjunto reducido de 94 líneas, para el cual se contó con la información completa sobre SSRs. Las matrices de microsatélites y genes candidatos utilizadas en el análisis se muestran en las Tablas A4.3 y A4.4 del Apéndice, respectivamente.

La estructura de la PMA se determinó a través de un análisis de coordenadas principales (ACP). Para ello, se tuvieron en cuenta cinco coordenadas que capturaron el 43% de la varianza observada entre las 94 líneas (Tabla A4.5, Apéndice). En la Figura 4.4 se muestra el resultado del ACP para las dos primeras coordenadas (CP1 y CP2), que explican el 12,6% y el 9,8% de la varianza, respectivamente. Sin embargo, no se observa un agrupamiento claro de las líneas a partir de la información de estas dos coordenadas principales.

Para poder determinar si el análisis de asociación presenta diferencias cuando la estructura poblacional es incorporada mediante distintas aproximaciones, además de la inferencia de estructuración a partir de ACP (matriz P), se utilizó el método bayesiano implementado en el programa STRUCTURE (matriz Q) (Pritchard y col. 2000).

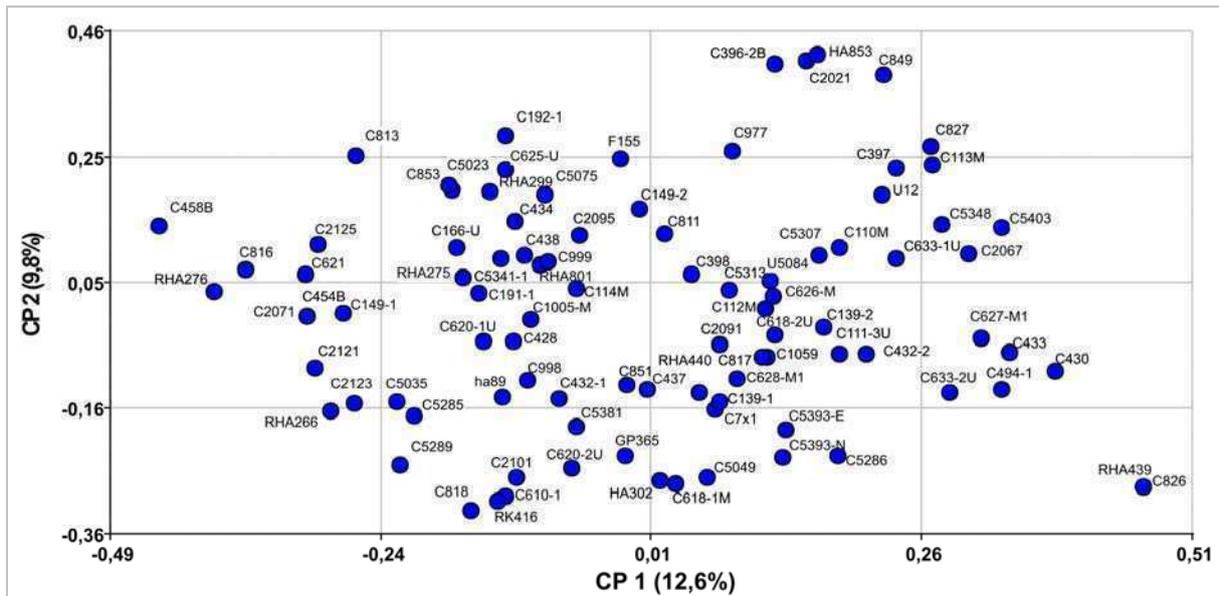
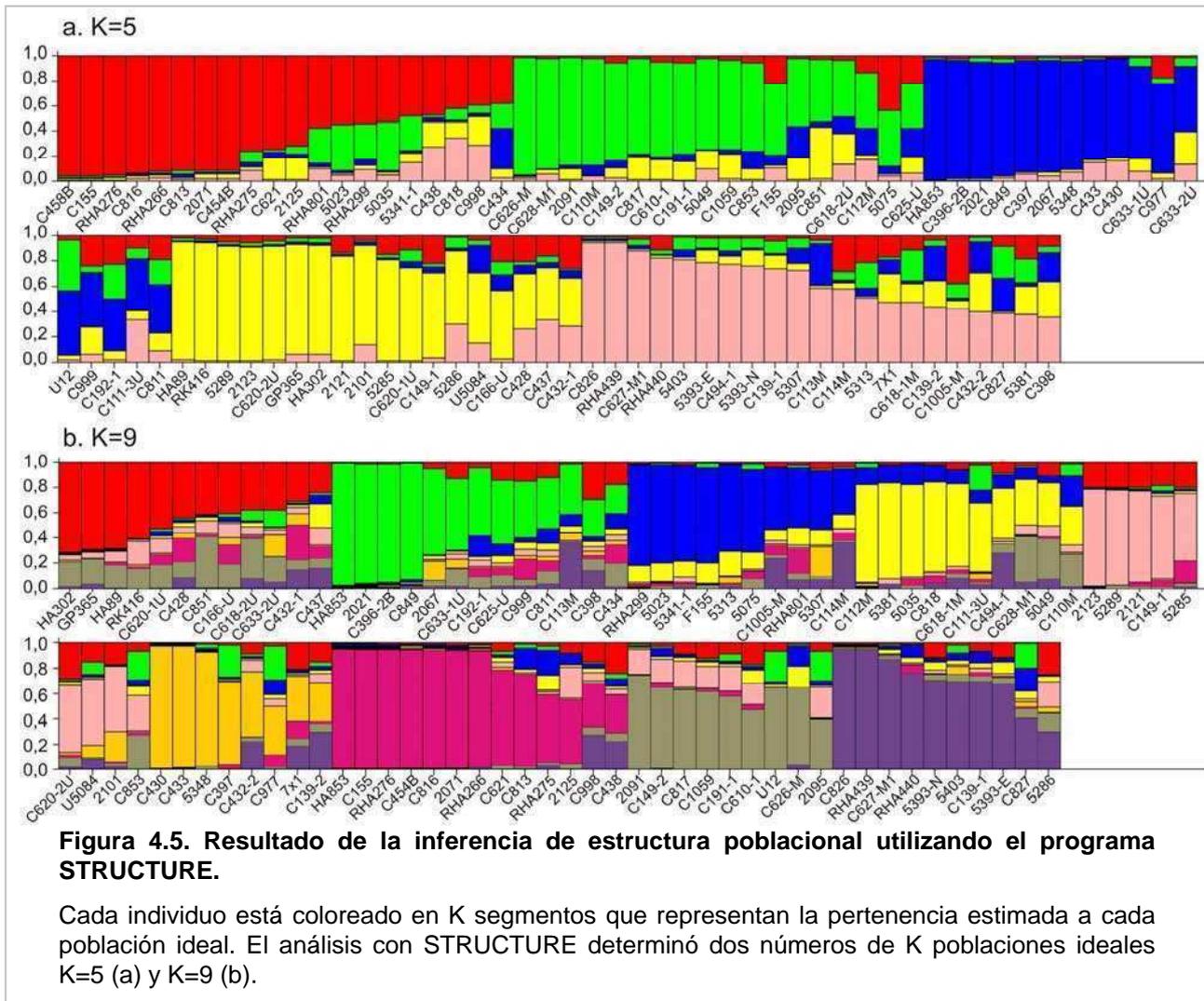


Figura 4.4. Análisis de Coordenadas Principales para las 94 líneas de girasol que conforman la PMA.

El análisis se realizó a partir de la matriz de frecuencias alélicas de SSRs, calculando la distancia de Rogers. Los porcentajes entre paréntesis refieren a la proporción de la varianza que explica cada coordenada.

El análisis mediante STRUCTURE, variando K de uno a 12, mostró que la probabilidad *a posteriori* de los datos aumentaba hasta estabilizarse en K=9. Sin embargo, al graficar el estadístico ΔK , basado en la tasa de cambio de la probabilidad (Evanno y col. 2005), se encontró también un número de K poblaciones ideales igual a cinco (K=5) (Figura A4.1). Por lo tanto, se utilizaron las asignaciones de las líneas considerando independientemente ambos valores de K para el análisis de asociación (K=5 y K=9). Los gráficos correspondientes a las corridas que mostraron la mayor probabilidad *a posteriori* de los datos para cada valor K se observan en la Figura 4.5. Si bien la mayor parte de los accesos (el 62%) muestran una constitución genética mayoritaria (> 60%) correspondiente a una de las cinco o nueve poblaciones ideales, el 30% restante mostró mezcla en su constitución genética. Las matrices Q utilizadas (K=5 y K=9) se muestran en la Tabla A4.6 del Apéndice.



Dado que las relaciones de parentesco se obtuvieron de acuerdo a lo expuesto en Bernardo (1993), el modelo lineal mixto de asociación (MLM) requirió que se estime el valor T desconocido, maximizando la verosimilitud del mismo. Al considerarse varios valores de T (0,20; 0,30; 0,50; 0,70 y 0,80) utilizando la matriz P (ACP) o las matrices Q (con K=5 ó K=9), los resultados no cambiaron sustancialmente como consecuencia de los valores T tomados. En la Tabla 4.4 se muestran los p-valores para T=0,2 y T=0,8, observando una asociación significativa con una menor incidencia de PHC para el gen RhoBP_B ($p < 0,05$). Por otro lado, el gen HaWRKY7 muestra una asociación marginalmente significativa ($p < 0,1$), pero sólo cuando se tomaron los valores T=0,2 y la matriz Q para K=9 (Tabla 4.4).

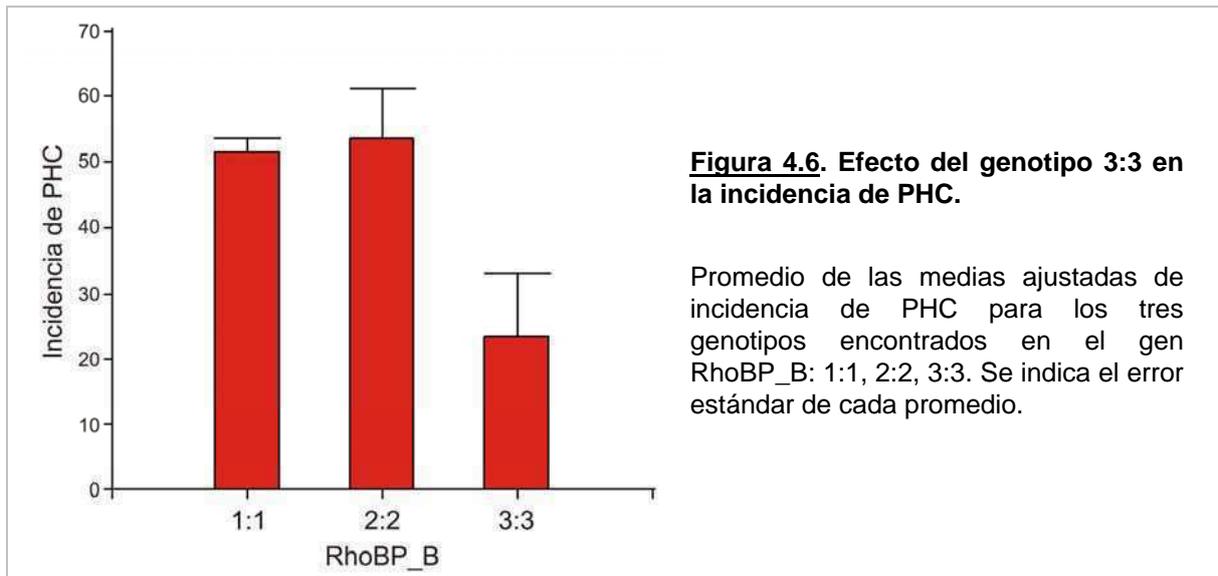
Tabla 4.4. Análisis estadístico. Datos de los p-valores obtenidos para la prueba de asociación entre cada gen candidato y la incidencia de PHC, tomando en el modelo distintos valores de T.

Gen	Matriz P (ACP)		Matriz Q (K=5)		Matriz Q (K=9)	
	p-valor (T= 0,2)	p-valor (T= 0,8)	p-valor (T= 0,2)	p-valor (T= 0,8)	p-valor (T= 0,2)	p-valor (T= 0,8)
MT	0,8362	0,8872	0,8362	0,8876	0,8192	0,8192
PRP_2	0,9227	0,7572	0,9227	0,7477	0,9932	0,9136
CytP450_C	0,7868	0,8240	0,7828	0,7653	0,7514	0,7852
RGEF_A2	0,9134	0,9489	0,9131	0,9488	0,9072	0,9454
PP2C	0,7203	0,7354	0,7179	0,7341	0,6725	0,6918
RhoBP_A	0,2124	0,2404	0,2155	0,4025	0,2064	0,3124
RhoBP_B ^a	0,0093	0,0093	0,0102	0,0101	0,0066	0,0066
bHLHP	0,5582	0,5321	0,5668	0,5011	0,4789	0,5113
ZFHDP	0,4571	0,4584	0,4454	0,4457	0,3461	0,3370
HaGDP-DI	0,9779	0,9807	0,9791	0,9817	0,9747	0,9775
HaWP	0,4697	0,4397	0,4736	0,4439	0,4531	0,4232
HaTRP	0,3813	0,3690	0,3832	0,3711	0,3567	0,3451
HaPAL	0,1669	0,1669	0,1708	0,1708	0,1501	0,1359
HaWRKY5	0,2164	0,2316	0,2070	0,2226	0,1394	0,1545
HaWRKY7 ^b	0,1023	0,2481	0,1019	0,2246	0,0959	0,2219
HaGLP4	0,3466	0,2009	0,3616	0,2112	0,3186	0,1914

a Asociación significativa ($p < 0,05$) entre el gen RhoBP_B y una menor incidencia de PHC (negrita).

b Asociación marginalmente significativa ($p < 0,1$) entre el gen HaWRKY7 y una menor incidencia de PHC (negrita).

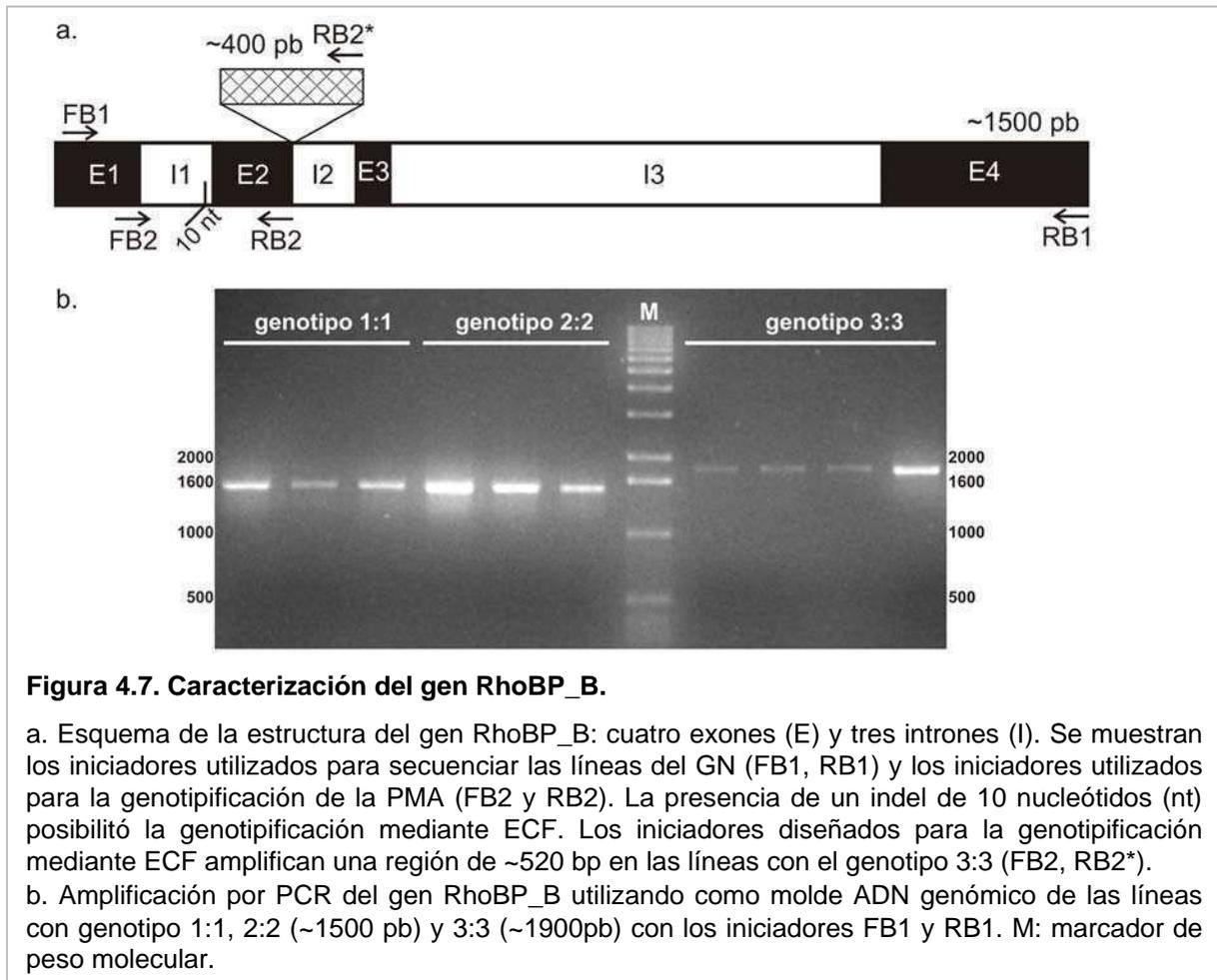
Para observar el efecto de los alelos detectados para el gen RhoBP_B sobre la incidencia de PHC, se graficaron los promedios de las medias ajustadas de incidencia de PHC, en función de los haplotipos encontrados (Figura 4.6). Se puede observar que el genotipo 3:3 (homocigota para el haplotipo 3) está asociado con los niveles más bajos de incidencia. Este genotipo estuvo presente en cuatro líneas GP365, C192-1, 5381, y 5393-E, y no había sido detectado en el GN de líneas de girasol secuenciadas. Para las líneas GP365, C192-1 y 5381 todos los individuos genotipificados presentaron el haplotipo 3, mientras que para 5393-E, dos individuos presentaron el haplotipo 3 y un individuo evidenció el haplotipo 1. Es interesante mencionar que para esta última línea, la media ajustada de incidencia de PHC fue del 46,25% mientras que las otras líneas presentaron medias ajustadas de 0,00; 22,5 y 25,5% (GP365, C192-1 y 5381, respectivamente).



Caracterización del gen RhoBP_B

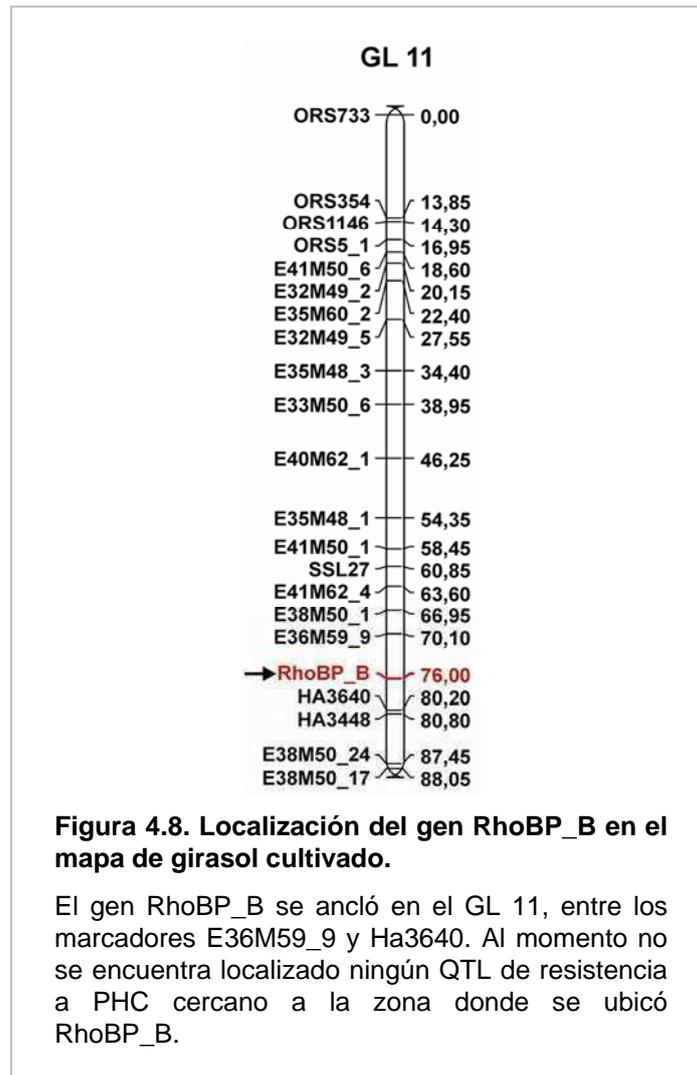
La secuencia genómica de RhoBP_B se contrastó con el índice de genes de girasol (HaGI) del DFCI (Dana Farber Cancer Institute), para poder conocer su estructura y organización genómica en girasol. Este gen constaría de cuatro exones y tres intrones (Figura 4.7a). Los iniciadores utilizados para genotipificar dicho gen (FB2 y RB2, Figura 4.7a) amplifican un fragmento de 236 y 226 pb en los genotipos 1:1 y 2:2, respectivamente. Sin embargo en las líneas con el genotipo 3:3 amplifican un fragmento de ~520 pb. La secuenciación directa de los productos de PCR del fragmento de ~520 pb evidenció la presencia de una inserción de aproximadamente 400 pb en la región contigua al segundo exón en las líneas con este último genotipo (Figura 4.7a). Sin embargo, debido a una región repetitiva del nucleótido “T”, esta región no pudo ser secuenciada por completo.

La amplificación mediante PCR del gen RhoBP_B con los iniciadores utilizados para secuenciar las líneas del GN usando como molde ADN genómico de las líneas portadoras de cada uno de los tres genotipos reveló que, en los individuos portadores del genotipo 3:3, el gen posee una longitud de aproximadamente 1900 pb, en contraposición a las ~1500 pb que poseen los individuos portadores del genotipo 1:1 y 2:2 (FB1 y RB1, Figura 4.7b).



Dado que aún no se posee la secuencia completa del alelo 3 del gen RhoBP_B, y que uno de los exones quedaría dentro de esta región, no es posible inferir cuál es la variante de este alelo respecto de los otros dos a nivel de ARN mensajero.

Debido a que las líneas RHA266 y PAC2, parentales de la población RIL, resultaron polimórficas para el gen RhoBP_B presentando únicamente los haplotipos 2 y 1, respectivamente, se determinó la posición de RhoBP_B en el mapa de referencia de girasol cultivado. RhoBP_B se localizó en el grupo de ligamiento 11 (GL 11) entre los marcadores E36M59_9 y HA3640, a 5,9 y 4,2 cM, respectivamente (Figura 4.8). Este GL no se encuentra demasiado saturado y sólo existe un QTL relacionado con la enfermedad PHC descrito sobre una población F2 (Micic y col. 2004) y alejado de la zona donde se ubicó este marcador (Figura A4.1, Apéndice).



Discusión

Los estudios de asociación constituyen una herramienta alternativa al análisis de ligamiento tradicional (QTL) que se ha probado con éxito en la disección de diversos caracteres complejos en especies vegetales (p. ej. maíz: tiempo de floración, papa: tizón tardío, trigo: calidad molinera y tamaño de semilla) (Breseghello y Sorrells 2006b, Pajeroska-Mukhtar y col. 2009, Thornsberry y col. 2001).

Este trabajo constituye el primer antecedente de aplicación de mapeo por asociación en girasol para un carácter de importancia como la resistencia a PHC, enfermedad provocada por *S. sclerotiorum*. El panel de datos experimentales para establecer la asociación estuvo formado por dos ensayos a campo con infección controlada durante las campañas 2008-09 y 2009-10, ocho SSRs para establecer estructura poblacional y 16 genes candidatos a estar involucrados en la defensa del cultivo frente a PHC, para evaluar asociaciones fenotipo-genotipo.

A pesar de que fueron 134 las líneas evaluadas a campo y genotipificadas para los SSRs y los genes candidatos, el grupo que se utilizó como población de mapeo final (PMA), estuvo compuesto por 94 líneas. Esto se debió, en parte, a la complicación encontrada al genotipificar tres individuos por línea y hallar líneas heterocigotas o individuos de una misma línea con alelos distintos (en mayor medida para los SSRs que para los genes candidatos). Sin embargo, la dificultad en la asignación de un genotipo consenso para las líneas endocriadas puede explicarse por varias causas. En principio, se conoció *a posteriori* que 20 accesos no eran líneas endocriadas, sino individuos provenientes de poblaciones segregantes, hecho que determinó su exclusión del análisis de asociación (Ing. Agr. Daniel Álvarez, com. pers.). Otro factor determinante de la diferencia entre los genotipos de los individuos de una línea dada es la presencia de heterocigosis residual, esperable en un porcentaje que varía entre el 6% al 12% (Haldane y Waddington 1931), ya que algunos accesos se encontraban en generación F4 o F5 del proceso de endocría. La heterogeneidad encontrada en este trabajo también ha sido descrita para líneas endocriadas de maíz (Heckenberger y col. 2003). Sin embargo, a pesar de la variación encontrada en los *loci* SSR en dicho trabajo, los individuos asignados a una misma línea sobre la base de características

morfológicas, continuaban mostrando mayor similitud genética entre sí que con los de otras líneas. Finalmente se seleccionó un conjunto de 94 líneas, incluyendo a aquellas que mostraron la menor cantidad de datos faltantes y menos ambigüedades en la genotipificación de SSRs y genes candidatos. A pesar de ser un número de líneas inferior a lo sugerido por algunos autores (p. ej. (Ersoz y col. 2007, Rafalski 2010)), esta cantidad de individuos concuerda con los números analizados para otras especies como maíz (92) (Thornsberry y col. 2001), *Arabidopsis* (95) (Olsen y col. 2004), trigo (95) (Brescaglio y Sorrells 2006b) y raigrás perenne (96) (Skot y col. 2007).

Con respecto a los métodos de genotipificación utilizados, los tres métodos fueron exitosos, alcanzando la genotipificación del 98% de los individuos para el panel completo de genes. La técnica de dHPLC fue una de las más exitosas, siendo el método que pudo caracterizar el conjunto de 384 individuos para dos de los tres genes genotipificados (PRP_2, CytP450_C). Sin embargo, dHPLC no pudo utilizarse para la genotipificación de bHLHP, ya que no se encontraron las condiciones óptimas para poder resolver las moléculas heterodúplex formadas por mezcla de los dos haplotipos encontrados en el GN de líneas control secuenciadas. Este hecho, lejos de haber sido una dificultad, redundó en una ganancia de variabilidad entre los individuos (y líneas, finalmente), habiendo detectado cinco haplotipos en lugar de dos (Tabla 4.3). Otro gen en el que se encontraron más de dos haplotipos inicialmente identificados en el GN, fue CytP450_C. Sin embargo, estos nuevos haplotipos sí pudieron ser detectados mediante dHPLC, dado que no se observó un patrón de tipo homodúplex para ninguna de las mezclas con los haplotipos de las líneas control, pero el control homodúplex se comportó como tal. Por lo tanto, en este caso, se secuenciaron los individuos correspondientes para no perder esta variabilidad. Si bien la genotipificación mediante dHPLC tiene una etapa laboriosa que implica mezclar los productos de PCR en forma equimolecular (manualmente), la misma se ve compensada por la automatización posterior, pudiendo contar con la información completa para un gen al cabo de una semana. Además, como se mencionó en el Cap. 2, el costo del experimento por muestra es mucho menor que el de los otros dos sistemas (dHPLC: 0,8 USD, ECF: 2,5 USD, Secuenciación: 3,5 USD), convirtiéndola en una técnica excepcional para el escrutinio de al menos dos haplotipos en un conjunto

de individuos, para luego restringir la secuenciación sólo a los individuos portadores de los nuevos haplotipos detectados.

La técnica ECF fue la más empleada, utilizándose en la genotipificación de nueve genes candidatos y alcanzando un 98% de individuos genotipificados para el conjunto de los genes evaluados (3386/3456 genotipos) (MT, RGEF_A2, RhoBP_A, RhoBP_B, HaWP, HaTRP, HaWRKY5, HaWRKY7 y HaGLP4). Esta técnica tiene grandes ventajas en cuanto a la facilidad en el procesamiento de grandes números de muestras, una alta precisión en la adjudicación de los alelos, la posibilidad de automatización del proceso y de uso de programas de análisis de datos sencillos (Chen y Sullivan 2003). Sin embargo, el gen HaGLP4, genotipificado con ECF, exhibió un comportamiento excepcional, dado que se identificaron fragmentos del mismo tamaño pero cuyas secuencias eran distintas, resultando en una gran cantidad de individuos para los cuales no pudo determinarse el genotipo exacto.

Muchos de los trabajos de mapeo por asociación utilizan la secuenciación como método de genotipificación por excelencia (Haseneyer y col. 2010, Ingvarsson y col. 2008, Krill y col. 2010). La mayor ventaja de este método, respecto de ECF y dHPLC, radica en que los haplotipos no detectados en el GN utilizado para el desarrollo de los SNPs son identificados directamente. Por el contrario, dHPLC y ECF, deberán pasar por una instancia adicional de secuenciación si se desea determinar la secuencia de los polimorfismos determinantes de los nuevos haplotipos hallados. Sin embargo, el método de secuenciación no siempre resulta eficiente, como se demuestra en este trabajo, en donde los genes aquí secuenciados mostraron resultados intermedios en la cantidad de individuos genotipificados (PP2C, bHLHP, ZFHDP y HaGDP-DI, entre 374 y 377).

Por otro lado, a pesar de haber detectado una gran diversidad en los genes ZFHDP y HaGDP-GI (nueve y 10 haplotipos, respectivamente), esto fue en detrimento de la cantidad de líneas que pudieron finalmente genotipificarse (112 y 107 para ZFHDP y HaGDP-DI, respectivamente). Esto se debió al conflicto en la asignación de haplotipos a las líneas a partir de la información de los tres individuos analizados, ya que muchos resultaron heterocigotas (42 para ZFHDP y 47 para HaGDP-DI).

Con respecto al análisis de asociación propiamente dicho, a pesar de la existencia de distintos modelos estadísticos propuestos para evaluar asociaciones (ANOVA, Modelos Lineales Simples y Modelos Lineales Mixtos), los modelos que tienen en cuenta la estructura poblacional, y además, las relaciones de parentesco entre los individuos estudiados (Modelos Lineales Mixtos), son los que han logrado disminuir el error tipo I (Stich y col. 2008).

El programa STRUCTURE (Pritchard y col. 2000) que implementa un método bayesiano para inferir estructura poblacional es ampliamente utilizado en los análisis de asociación con MLM (p. ej. (Casa y col. 2008, Gonzalez-Martinez y col. 2007, Malosetti y col. 2007, Thornsberry y col. 2001)). Si bien uno de los argumentos para desestimar el uso de este programa es la demanda de tiempo computacional que requiere (Stich y col. 2008), este factor no fue limitante en el análisis de los datos de este capítulo. Además, para poder comparar resultados se infirió la estructura poblacional utilizando tanto este programa, como realizando un análisis de coordenadas principales (ACP) (Figura 4.4 y Figura 4.5).

En el análisis de ACP las líneas mostraron una distribución dispersa, sin ningún patrón de agrupamiento específico al graficarlas en función de las dos primeras coordenadas (CP1 12,6% y CP2 9,8%, respectivamente), lo que coincide con lo observado por Stich y col. (2008) para 303 accesos de trigo. Sin embargo, Yan y col. (2009b) pudieron detectar tres grupos distintivos para una colección de 1794 accesos de arroz analizados mediante ACP sobre la base de 14 caracteres fenotípicos diferentes. Es probable que las diferencias encontradas entre los resultados de ACP aquí presentados y el trabajo de Yan y col. (2009b), radiquen en la cantidad de individuos analizados y en la naturaleza de los caracteres utilizados para la diversidad pre-estimada de éstos.

El resultado del análisis mediante STRUCTURE mostró que existían dos números de poblaciones hipotéticas ideales ($K=5$ y $K=9$, Figura 4.5, Figura A4.1). Sin embargo, el haber utilizado una u otra matriz de pertenencia no implicó una variación en el resultado del análisis de asociación (Tabla 4.4). Más aún, el componente que más parece afectar el resultado es el valor de T utilizado para la construcción de la matriz

de parentesco. Si bien es mínima la variación de los p-valores obtenidos con los distintos valores de T evaluados, el único gen que mostró resultados dispares fue HaWRKY7, mostrando una asociación marginalmente significativa con $T=0,2$ y no significativa con $T=0,8$ (Tabla 4.4). Debido a que este gen era uno de los candidatos más promisorios para resistencia a PHC, sería importante ampliar el panel de líneas genotipificadas y/o genotipificar otra región de HaWRKY7. Esto aumentaría las posibilidades de genotipificar haplotipos diferentes a los dos evaluados en este capítulo y permitiría confirmar o desestimar la asociación marginalmente significativa aquí detectada (Tabla 4.4).

Si la estructura poblacional está controlada, el poder de la prueba de asociación (error tipo II) dependerá en gran medida del carácter estudiado y de las frecuencias de los haplotipos encontrados, perdiendo poder la detección de asociación cuando los alelos se encuentran en baja frecuencia, a menos que su efecto sea importante (Rafalski 2010). Si bien la resistencia a *S. sclerotiorum* es un carácter complejo, haber detectado asociación entre una menor incidencia de PHC y el haplotipo 3 del gen RhoBP_B constituye un excelente antecedente para la implementación de esta estrategia en el estudio de otros caracteres de importancia económica en girasol.

El MLM utilizado demostró robustez en la asociación de RhoBP_B con la incidencia de PHC, ya que los p-valores fueron menores a 0,05 utilizando tanto la aproximación QK como PK. Más aún, los p-valores fueron significativos para cualquier valor de T tomado en la matriz de parentesco (Tabla 4.4). El uso de correcciones por comparaciones múltiples es un tema recurrente en los estudios de asociación y no se ha llegado aún a un consenso sobre cuál utilizar. No obstante, es necesario tener en cuenta que en los casos en los cuales las comparaciones que se realizan son pocas (menores a 50), la corrección por comparaciones múltiples pierde sentido dado que aumenta la tasa de error tipo II, pudiendo incluso conducir a la pérdida total del poder de detección de la asociación (Saville 1990) (MSc. Julio di Rienzo, com. pers.). Por estos motivos, en este trabajo de tesis no se consideró necesaria la utilización de este tipo de correcciones.

Con respecto al gen RhoBP-B, éste fue seleccionado a partir de los datos del

microarreglo de *Arabidopsis* (Cap. 3), evidenciando la idoneidad de la estrategia de detección de ortólogos utilizada en este trabajo, y la importancia de la selección de genes candidatos a partir de distintas fuentes de información.

El gen RhoBP_B deriva del *locus* AT1G04450 (RIC3) de *Arabidopsis*, una proteína de unión a pequeñas GTPasas denominadas ROP. Existen 11 proteínas RIC blanco directo de las proteínas ROP en *Arabidopsis*, y sus funciones están involucradas en la formación y funcionamiento del tubo polínico (Wu y col. 2001). RIC3 interactúa con ROP (ROP1) y pareciera estar involucrado en la elevación de la concentración de Ca^{2+} citosólico y en la regulación de: (1) la depolimerización de los filamentos de actina, (2) la exocitosis y (3) las señales mediadas por Ca^{2+} (Gu y col. 2005, Lee y col. 2008).

Por otro lado, en un estudio reciente han sido caracterizados dos factores de necrosis y de inducción de respuestas a etileno en *S. sclerotiorum* (SsNep1 y SsNep2) que han evidenciado una expresión fuertemente dependiente de la concentración de Ca^{2+} circundante (Bashi y col. 2010). En dicho trabajo se describe que la incorporación de compuestos que activan las señales de Ca^{2+} en la planta (p. ej. cafeína) provoca la inhibición completa de la expresión de SsNep2 y, consecuentemente, impide el desarrollo de necrosis en las hojas (Bashi y col. 2010).

A partir de estos antecedentes, es posible hipotetizar que la expresión de RhoBP_B esté participando en la defensa contra el ataque del *S. sclerotiorum*, reprimiendo la expresión de estos factores de necrosis, fundamentales para la patogenicidad y virulencia del hongo (Bashi y col. 2010).

Dado que tanto RIC3 como otras proteínas RIC son efectores directos de las proteínas ROP, también es posible que la elevada expresión de estas proteínas (observado por Zhao y col. (2007a), para *Brassica*) sea directa consecuencia de la activación de las proteínas ROP durante el proceso de infección. Las GTPasas ROP cumplen diversas funciones y abarcan una gran cantidad de procesos, incluyendo el desarrollo embrionario, el desarrollo de la semilla, la regulación de la acumulación de hormonas, la activación de la producción de H_2O_2 y las respuestas de defensa (Kawasaki y col. 1999, Ono y col. 2001, Wu y col. 2001).

Si bien no se conoce aún la función que tiene RhoBP_B en girasol, la estructura entre

los haplotipos es diferente, siendo el haplotipo 3 el más distinto al presentar una inserción de ~400 pb. Esta inserción podría estar afectando la expresión del ARNm o los sitios de *splicing* de los exones 2 y 3, (Figura 4.7). De cualquier modo, es necesario evaluar la funcionalidad y el patrón de expresión de este gen en girasol en condiciones normales y de infección para poder determinar si efectivamente es este gen el que determina la menor incidencia o existe otro gen en DL que lo está haciendo.

Si bien la co-localización de RhoBP_B con algún QTL de resistencia a PHC hubiera servido como validación inmediata de la asociación encontrada, este resultado no resulta extraño, ya que se conoce (excepto para HA280) que los parentales de las poblaciones utilizadas para investigar este carácter (RHA266, PAC2, RHA801, HA89) no poseen el alelo que muestra menor incidencia de PHC. Por lo tanto, no existiría segregación posible de los marcadores que puedan correlacionar con la segregación del carácter. Por otra parte, si bien esto acarrea una ardua tarea, la posibilidad de contar con una población de RILs generada a partir de un parental portador del haplotipo 3 sería invaluable para la validación mediante análisis de ligamiento tradicional.

En conclusión, la demanda de nuevas estrategias biotecnológicas para el mejoramiento asistido a través de marcadores moleculares ha posibilitado la exploración del mapeo por asociación como alternativa al mapeo tradicional en poblaciones biparentales. El esfuerzo de un grupo multidisciplinario ha contribuido a establecer los modelos estadísticos apropiados para eliminar las asociaciones espurias debido a la estructura poblacional. De esta manera, la estrategia de mapeo por asociación pudo ser establecida en girasol para el carácter resistencia a PHC, detectando una correlación entre la menor incidencia de PHC y el gen RhoBP_B. La genotipificación de RhoBP_B en otra población de mapeo, la inclusión de datos fenotípicos de otras campañas sobre esta PMA y la introgresión del gen en distintos fondos genéticos posibilitarán la validación de este resultado. Finalmente, experimentos de genómica funcional serán necesarios para desentrañar el rol que pueda tener RhoBP_B en la respuesta a la infección de *S. sclerotiorum*.

**DISCUSIÓN GENERAL Y
CONCLUSIONES FINALES**

Discusión general

El mapeo por asociación enriquece el repertorio de herramientas disponibles para la disección de caracteres complejos, aplicable tanto en especies modelos como en cultivos de importancia económica (Rafalski 2010). Este enfoque comenzó a desarrollarse en especies vegetales a comienzos del Siglo XXI, gracias al salto de tecnología producido en la biología molecular en términos de información genómica disponible en bases de datos y de técnicas de genotipificación de alto rendimiento (*high-throughput*).

Este trabajo de tesis se propuso demostrar la aplicabilidad de estudios de genética de asociación en girasol, como enfoque alternativo al tradicional mapeo de QTL en poblaciones biparentales; y como herramienta para dilucidar los componentes genéticos involucrados en caracteres agronómicos complejos, como la resistencia a la Podredumbre Húmeda del Capítulo (PHC).

La resistencia a PHC es un carácter que involucra numerosos genes, cada uno con efectos menores sobre el fenotipo (Jan y Seiler 2007). El análisis de ligamiento en poblaciones biparentales ha sido la herramienta más utilizada para la identificación de fuentes de resistencia contra PHC (p. ej. (Maringolo 2007, Micic y col. 2004, Micic y col. 2005a, Micic y col. 2005b, Yue y col. 2008)). A pesar de su amplia utilización, este tipo de análisis presenta algunas desventajas en comparación con el mapeo por asociación. Por ejemplo, la cantidad de variación genética segregante es limitada en las poblaciones biparentales, debido a que sólo dos alelos pueden ser evaluados a la vez para un marcador dado. Adicionalmente, si no existe polimorfismo entre los parentales seleccionados para el cruzamiento respecto al marcador ligado al carácter, el QTL nunca podrá ser identificado. El mapeo por asociación, en cambio, al utilizar poblaciones naturales o colecciones de individuos disponibles en los bancos de germoplasma, permite explotar la mayor diversidad alélica presente en ellos y evita el gran insumo de tiempo que conlleva la creación de poblaciones RILs creadas a partir del cruzamiento de parentales de *pedigree* conocido.

Debido a la ubicuidad de los SNPs e indels y a su fácil adaptación a técnicas de genotipificación a gran escala, éstos se han convertido en los marcadores de elección

para los estudios de asociación (Gupta y col. 2001). Por ello, en este trabajo de tesis se caracterizaron las variantes alélicas de un conjunto de 31 regiones genómicas sobre 19 accesos de girasol representativos de los materiales de mejoramiento del cultivo en Argentina. Esta parte del trabajo, expuesto en el Cap. 1, permitió determinar la frecuencia SNPs e indels, estimar la diversidad y el polimorfismo nucleotídico y evaluar el alcance del desequilibrio de ligamiento (DL) en el girasol cultivado.

Se encontró una frecuencia de 1 SNP/61pb y 1 indel/288 pb. De acuerdo a lo esperado para materiales endocriados, las estimas de diversidad y polimorfismo nucleotídico exhibieron valores moderados ($\pi_T = 0,0074$ y $\theta_W = 0,0064$). La frecuencia de SNPs y estimas de diversidad se analizaron también para el panel de genes candidatos y accesos evaluados en el Cap. 3, siendo los valores obtenidos para este conjunto de datos levemente diferentes a los encontrados en el Cap. 1 (1 SNP/83 pb, 1 indel/417 pb, $\pi_T = 0,0079$ y $\theta_W = 0,0073$). Sin embargo, es probable que estas diferencias radiquen tanto en la cantidad de accesos de girasol analizados (19 vs. 10), como en la naturaleza de los genes examinados (regiones genómicas de origen y funciones diversas respecto de genes candidatos para resistencia a PHC).

El DL alcanzó valores de $r^2=0,48$ a las 643 pb, con una tendencia a valores despreciables de DL ($r^2<0,1$) hacia las 100 kb (Cap. 1). Esta extensión, más típica de especies autógamas que de alógamas, sugiere que es posible realizar mapeo por asociación en girasol con una densidad de marcadores menor a lo reportado para otras especies como maíz (Yan y col. 2009a). Sin embargo, debido al tamaño del genoma de girasol (~3500 Mb), la cantidad de marcadores requeridos para implementar una estrategia de asociación sobre el genoma completo (*genome-wide association*, GWA), resulta demasiado onerosa (> 30.000 SNPs). Más aún, para llevar a cabo GWA, muchas veces la tendencia del DL evaluado en un panel de genes al azar no siempre es representativa de lo que ocurre en la totalidad del genoma, tal como se describe en un estudio de *Arabidopsis* (Aranzana y col. 2005). Por lo tanto, se consideró más apropiada utilizar una estrategia de mapeo a través de genes candidatos para el carácter resistencia a PHC.

Debido a la poca información existente acerca de los genes involucrados en la

defensa frente a *S. sclerotiorum*, la selección de genes candidatos aquí realizada trató de compilar todas las fuentes de información disponible acerca del patógeno. (Cap. 3). La aproximación filogenética utilizada para hallar en girasol los ortólogos a los genes de *Arabidopsis* descritos por Zhao y col. (2007a) enriqueció la nómina de genes candidatos para tolerancia a PHC, dando lugar a la identificación de 17 genes de girasol con similitud a las secuencias originales de partida. En el caso de los genes PRP, CytP450, RGEF, RhoBP y FBoxP la estrategia empleada dio como resultado la amplificación de más de un *locus* parálogo en girasol a partir de un único *locus* de *Arabidopsis*. Por otro lado, tanto la selección de genes a partir de la colección de ADNc de expresión diferencial, como la incorporación de secuencias de girasol evaluadas experimentalmente como partícipes de los procesos de defensa frente a PHC, ampliaron el número de genes a 30, siendo 21 los genes polimórficos plausibles de genotipificación en la población de mapeo por asociación (PMA). Además de poner de manifiesto la gran variabilidad encontrada (70% de los genes polimórficos), el número de genes candidatos seleccionados es comparable al de otros trabajos que estudian caracteres de base compleja (p. ej. maíz, (Krill y col. 2010)) y resulta mucho mayor que la cantidad de genes candidatos elegidos para el mapeo de caracteres de herencia simple, como por ejemplo la resistencia a *Verticillium* en papa (Simko y col. 2004a).

Contar con técnicas de genotipificación de SNPs e indels eficientes y aplicables a un gran número de individuos es fundamental para realizar experimentos de mapeo por asociación. Si bien existen distintas plataformas de genotipificación masiva *high-throughput* (Gupta y col. 2008), muchas veces este tipo de tecnología es excesivamente costosa para laboratorios académicos o instituciones gubernamentales. Por lo tanto, la posibilidad de disponer de metodologías simples y de gran capacidad de procesamiento, es fundamental para llevar a cabo con éxito este tipo de experimentos. La evaluación del desempeño y la escalabilidad de las técnicas CEL1CH y dHPLC permitió establecerlas como técnicas de genotipificación a mediana escala, eficientes y de bajo costo, constituyendo este trabajo de tesis uno de los primeros antecedentes de la aplicación de CEL1CH en girasol (Cap. 2). Por otro lado, dHPLC pudo utilizarse en la genotipificación de la población de RILs (90

individuos) proveniente del cruzamiento de RHA266 X PAC2 para incorporar marcadores funcionales al mapa saturado de girasol cultivado (Poormohammad Kiani y col. 2007a), ampliando las evidencias de la factibilidad de su empleo en este tipo de experimentos (Kota y col. 2001, Kota y col. 2008, Lai y col. 2005, Lexer y col. 2003). Asimismo, en este trabajo de tesis se utilizó la técnica dHPLC para la genotipificación de la PMA (= 384 individuos, Cap. 4), demostrando la gran eficiencia y versatilidad de este método para escalar la genotipificación de marcadores funcionales a un mayor número de individuos.

En el presente trabajo de tesis, el establecimiento de la estrategia de mapeo por asociación para la resistencia a PHC ha sido consecuencia de un esfuerzo multidisciplinario, que incluyó biólogos moleculares, genetistas, fitopatólogos, mejoradores y estadísticos. Este abordaje resultó exitoso en la identificación de un haplotipo del gen RhoBP_B asociado a una menor incidencia de PHC (Cap. 4, genotipo 3:3, $p < 0,05$). Este resultado, además de aportar una herramienta para el mejoramiento asistido por marcadores y de incrementar la información acerca de los factores genéticos en girasol involucrados en la defensa contra PHC, permitió establecer las bases para el estudio de otros caracteres agronómicos.

Es importante mencionar que el haplotipo 3 del RhoBP_B no se encontraba presente en los parentales de la población del mapa de referencia de girasol cultivado (Poormohammad Kiani y col. 2007a), ni en las líneas de comportamiento contrastante utilizadas para la identificación de QTL (Maringolo 2007). Por lo tanto, la estrategia de mapeo por asociación realizada en este trabajo de tesis refleja las ventajas de su implementación, identificando nuevas fuentes de resistencia a PHC inéditas, y complementando el mapeo de QTL en poblaciones biparentales.

La asociación encontrada no puede considerarse espuria, dado que el modelo utilizado para inferir asociación tuvo en cuenta la presencia de estructura poblacional y las relaciones de parentesco entre los individuos analizados. Por el contrario, uno de los genes con mayores evidencias empíricas de estar involucrado en el proceso de defensa de la planta, como HaWRKY7, no mostró estar asociado con menores niveles de incidencia de PHC (Cap. 4). Sin embargo, HaWRKY7 fue el único gen para el cual

los p-valores para HaWRKY7 variaron de acuerdo al valor de T ensayado.

Con respecto a RhoBP_B, resulta interesante que la expresión de este gen en *Arabidopsis* esté circunscripta a las inflorescencias, en relación a la formación del tubo polínico, y que además, regule las concentraciones de Ca^{2+} en las células (Wu y col. 2001). De este modo, la función de RhoBP_B en la defensa frente a *S. sclerotiorum* podría deberse a la inhibición de la expresión de un factor de necrosis del hongo (SsNep2), cuya expresión está regulada por las señales de Ca^{2+} (Bashi y col. 2010). Sin embargo, serán necesarios estudios de genómica funcional para establecer una correlación entre la función del haplotipo y la menor incidencia de PHC en los accesos de girasol que lo portan.

Conclusiones finales y perspectivas

Los resultados de este trabajo de tesis demuestran que es posible establecer estudios de genética de asociación en girasol para dilucidar los componentes genéticos involucrados en caracteres agronómicos complejos como la resistencia a hongos.

- I. Los niveles de polimorfismo y diversidad nucleotídica encontrados en el conjunto de accesos de girasol, representativo de los materiales de cultivo en Argentina, fueron moderados y no se modificaron sustancialmente con la naturaleza de los genes evaluados.
- II. La extensión del desequilibrio de ligamiento en girasol cultivado alcanza distancias comparables a las encontradas en especies autógamas. De acuerdo a ello, la densidad de marcadores necesarios para un estudio de *genome-wide association* sería mucho menor que lo reportado para otras especies (entre 30.000 y 40.000 SNPs). Sin embargo, este valor deberá ajustarse de acuerdo a la presencia de regiones genómicas con un desequilibrio de ligamiento de menor alcance.
- III. Las técnicas de genotipificación a mediana escala optimizadas en este trabajo de tesis (CEL1CH y dHPLC) resultaron eficientes, simples y versátiles para: (1) aumentar la cantidad de individuos y genes a caracterizar en los estudios de mapeo por asociación y (2) incorporar marcadores funcionales al mapa genético de girasol cultivado.
- IV. La función del gen RhoBP_B descrita para *Arabidopsis* (RIC3) es consistente con la respuesta de resistencia a la infección de *S. sclerotiorum* observada en girasol, y, consecuentemente, con una menor incidencia de PHC. De este modo, RhoBP_B se convierte en un gen candidato de alta prioridad para estudios fisiológicos y funcionales de la resistencia a esta enfermedad en girasol.
- V. La identificación de un alelo del gen RhoBP_B asociado a una baja incidencia de PHC acredita la utilización del mapeo por asociación como estrategia alternativa para la generación de marcadores moleculares para

mejoramiento asistido. La estructura genómica de RhoBP_B permite el desarrollo de un marcador molecular basado en PCR. Éste puede ser transferido y utilizado inmediatamente en los programas de mejoramiento de girasol para la obtención de líneas con características mejoradas para la resistencia a esta enfermedad.

Este trabajo de tesis constituye el primer antecedente de aplicación de mapeo por asociación en un cultivo de importancia económica en nuestro país, como lo es el girasol. Si bien el resultado aquí presentado deberá ser validado, es realmente promisorio la aplicación de este nuevo enfoque para la disección de factores genéticos involucrados en caracteres agronómicos de importancia en girasol.

Dr. Verónica V. Lia
Director

Dr. Norma B. Paniego
Director Asistente

Lic. Corina M. Fusari
Doctorando

BIBLIOGRAFÍA

Aardse A, Argyris J, Bervillé A, Bradford K, Burke J, Caldwell D, Cervantes J, Chapman M, Dahal P, Wit Jd, Dong N, Endow J, Heinz R, Hewitt J, Hills M, Hiscock S, Hoefft E, Hu J, Jenni S, Kane N, Kesseli R, Knapp S, Kondo K, Lai Z, Lavelle D, Lawton-Rauh A, Matsuda S, McHale L, Michelmore R, Michiels A, Oberprieler C, Ochoa O, Osborn T, Paniago N, Remington D, Rieseberg L, Ro D-K, Roberts RP, Serquen F, Shintani D, Shirmohamadali A, Simko I, Still D, Tang S, Truco MJ, Tulsieram L, Ulloa M, Urbatsch L, Wal Mvd, Deynze AV, Eijk Mv, Vick B, Vincourt P, Whalen M, Whitton J, Yang J. Priorities for research, education and extension in genomics, genetics, and breeding of the Compositae. (2006) White Paper Meeting, USA.

Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. (2005) *Bioinformatics* 21:2104-2105.

Abbas A, Lepelley M, Lechevrel M, Sichel F. Assessment of DHPLC usefulness in the genotyping of GSTP1 exon 5 SNP: comparison to the PCR-RFLP method. (2004) *J Biochem Biophys Methods* 59:121-126.

Agrama HA, Eizenga GC, Yan W. Association mapping of yield and its components in rice cultivars. (2007) *Mol Breed* 19:341-356.

Agüero M, Pereyra V, Escande A. Efecto de la pudrición húmeda del capítulo del girasol (*Sclerotinia sclerotiorum* (lib) de Bary) sobre el contenido de impurezas del producto cosechado, y el contenido y acidez del aceite. (2001) *J of Agric Univ of Puerto Rico* 85:177-186.

Al-Chaarani GR, Gentzbittel L, Huang XQ, Sarrafi A. Genotypic variation and identification of QTLs for agronomic traits, using AFLP and SSR markers in RILs of sunflower (*Helianthus annuus* L.). (2004) *Theor Appl Genet* 109:1353-1360.

Al-Chaarani GR, Roustaei A, Gentzbittel L, Mokrani G, Barrault G, et al. A QTL analysis of sunflower partial resistance to downy mildew (*Plasmopara halstedii*) and black stem (*Phoma macdonaldii*) by the use of recombinant inbred lines (RILs). (2002) *Theor Appl Genet* 104:490-496.

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. (2006) *Bioinformatics* 22:e9-15.

Alfadil AT, Poormohammad Kiani G, Dechamp-Guillaume L, Gentzbittel L, Sarrafi A. QTL mapping of partial resistance to Phoma basal stem and root necrosis in sunflower (*Helianthus annuus* L.). (2007) *Plant Sci* 172:

Alignan M, Hewezi T, Petitprez M, Dechamp-Guillaume G, Gentzbittel L. A cDNA microarray approach to decipher sunflower (*Helianthus annuus*) responses to the necrotrophic fungus *Phoma macdonaldii*. (2006) *New Phytol* 170:523-536.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. (1990) *J Mol Biol* 215:403-410.

Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, Nordborg M. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. (2005) *PLoS Genet* 1:e60.

Arnaud D, Dejardin A, Leple JC, Lesage-Descauses MC, Pilate G. Genome-wide analysis of LIM gene family in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*. (2007) *DNA Res* 14:103-116.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. (2000) *Nat Genet* 25:25-29.

- Baack EJ, Whitney KD, Rieseberg LH.** Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. (2005) *New Phytol* 167:623-630.
- Bagge M, Lübberstedt T.** Functional markers in wheat: technical and economic aspects. (2008) *Mol Breed* 22:319-328.
- Barker G, Batley J, O' SH, Edwards KJ, Edwards D.** Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. (2003) *Bioinformatics* 19:421-422.
- Barkley NA, Wang ML.** Application of TILLING and EcoTILLING as Reverse Genetic Approaches to Elucidate the Function of Genes in Plants and Animals. (2008) *Curr Genomics* 9:212-226.
- Barreiro LB, Henriques R, Mhlanga MM.** High-throughput SNP genotyping: combining tag SNPs and molecular beacons. (2009) *Methods Mol Biol* 578:255-276.
- Bashi ZD, Hegedus DD, Buchwaldt L, Rimmer R, Borhan MH.** Expression and regulation of *Sclerotinia sclerotiorum* necrosis and ethylene-inducing peptides (NEPs). (2010) *Molec Plant Pathol* 11:43-53.
- Bateman DF, Beer SV.** Simultaneous production and synergistic action of oxalic acid and polygalacturonase during pathogenesis by *Sclerotium rolfsii*. (1965) *Phytopathol* 55:204-211.
- Benjamini Y, Hochberg Y.** Controlling the false discovery rate: a practical and powerful approach to multiple testing. (1995) *J R Statist Soc (B)* 57:289-300.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL.** GenBank. (2008) *Nucl Acids Res* 36:D25-30.
- Bernardo R.** Estimation of coefficient of coancestry using molecular markers in maize. (1993) *Theor Appl Genet* 85:
- Bert PF, Dechamp-Guillaume G, Serre F, Jouan I, de Labrouhe DT, Nicolas P, Vear F.** Comparative genetic analysis of quantitative traits in sunflower (*Helianthus annuus* L.) 3. Characterisation of QTL involved in resistance to *Sclerotinia sclerotiorum* and *Phoma macdonaldi*. (2004) *Theor Appl Genet* 109:865-874.
- Bert PF, Jouan I, De Labrouhe DT, Serre F, Nicolas P, Vear F.** Comparative genetic analysis of quantitative traits in sunflower (*Helianthus annuus* L.) 1. QTL involved in resistance to *Sclerotinia sclerotiorum* and *Diaporthe helianthi*. (2002) *Theor Appl Genet* 105:985-993.
- Bishop JG.** Directed mutagenesis confirms the functional importance of positively selected sites in polygalacturonase inhibitor protein. (2005) *Mol Biol Evol* 22:1531-1534.
- Bouzidi MF, Franchel J, Tao Q, Stormo K, Mraz A, Nicolas P, Mouzeyar S.** A sunflower BAC library suitable for PCR screening and physical mapping of targeted genomic regions. (2006) *Theor Appl Genet* 113:81-89.
- Boyce R, Chilana P, Rose TM.** iCODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. (2009) *Nucl Acids Res* 37:W222-228.
- Breeze E, Harrison E, Page T, Warner N, Shen C, Zhang C, Buchanan-Wollaston V.** Transcriptional regulation of plant senescence: from functional genomics to systems biology. (2008) *Plant Biol (Stuttg)* 10 Suppl 1:99-109.
- Brenner EA, Zein I, Chen Y, Andersen JR, Wenzel G, Ouzunova M, Eder J, Darnhofer B, Frei U, Barriere Y, Lübberstedt T.** Polymorphisms in O-methyltransferase genes are associated with stover cell wall digestibility in European maize (*Zea mays* L.). (2010) *BMC Plant Biol* 10:27.

- Bresegghello F, Sorrells ME.** Association Analysis as a Strategy for Improvement of Quantitative Traits in Plants. (2006a) *Crop Sci* 46:1323-1330.
- Bresegghello F, Sorrells ME.** Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. (2006b) *Genetics* 172:1165-1177.
- Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB.** Nucleotide diversity and linkage disequilibrium in loblolly pine. (2004) *Proc Natl Acad Sci U S A* 101:15255-15260.
- Buckler ESt, Thornsberry JM.** Plant molecular diversity and applications to genomics. (2002) *Curr Opin Plant Biol* 5:107-111.
- Burke J, Davison D, Hide W.** d2_cluster: a validated method for clustering EST and full-length cDNA sequences. (1999) *Genome Res* 9:1135-1142.
- Carter C, Thornburg RW.** Tobacco nectarin I. Purification and characterization as a germin-like, manganese superoxide dismutase implicated in the defense of floral reproductive tissues. (2000) *J Biol Chem* 275:36726-36733.
- Casa AM, Pressoira G, Brown PJ, Mitchell SE, Rooney WL, M.R.Tuinstrac, Franks CD, Kresovicha. S.** Community resources and strategies for association mapping in sorghum. (2008) *Crop Sci* 48:30-40.
- Cavell AC, Lydiate DJ, Parkin IA, Dean C, Trick M.** Collinearity between a 30-centimorgan segment of *Arabidopsis thaliana* chromosome 4 and duplicated regions within the *Brassica napus* genome. (1998) *Genome* 41:62-69.
- Cellier F, Conejero G, Breitler JC, Casse F.** Molecular and physiological responses to water deficit in drought-tolerant and drought-sensitive lines of sunflower. Accumulation of dehydrin transcripts correlates with tolerance. (1998) *Plant Physiol* 116:319-328.
- Cellier F, Conejero G, Casse F.** Dehydrin transcript fluctuations during a day/night cycle in drought-stressed sunflower. (2000) *J Exp Bot* 51:299-304.
- Coberly LC, Rausher MD.** Analysis of a chalcone synthase mutant in *Ipomoea purpurea* reveals a novel function for flavonoids: amelioration of heat stress. (2003) *Mol Ecol* 12:1113-1124.
- Collard BCY, Cruz CMV, McNally KL, Virk PS, Mackill DJ.** Rice Molecular Breeding Laboratories in the Genomics Era: Current Status and Future Considerations. (2008) *Int J Plant Genomics* 2008:524847.
- Comai L, Young K, Till BJ, Reynolds SH, Greene EA, Codomo CA, Enns LC, Johnson JE, Burtner C, Odden AR, Henikoff S.** Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. (2004) *Plant J* 37:778-786.
- Cooper JL, Till BJ, Laport RG, Darlow MC, Kleffner JM, Jamai A, El-Mellouki T, Liu S, Ritchie R, Nielsen N, Bilyeu KD, Meksem K, Comai L, Henikoff S.** TILLING to detect induced mutations in soybean. (2008) *BMC Plant Biol* 8:9.
- Cordeiro GM, Elliott F, McIntyre CL, Casu RE, Henry RJ.** Characterisation of single nucleotide polymorphisms in sugarcane ESTs. (2006) *Theor Appl Genet* 113:331-343.
- Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC, Jr., Rimmler JB, Locke PA, Conneally PM, Schmechel KE, et al.** Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. (1994) *Nat Genet* 7:180-184.
- Costabile M, Quach A, Ferrante A.** Molecular approaches in the diagnosis of primary immunodeficiency diseases. (2006) *Hum Mutat* 27:1163-1173.

- Chen X, Sullivan PF.** Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. (2003) *Pharmacogenomics J* 3:
- Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ.** SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. (2002) *BMC Genetics* 3:19.
- Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R.** OrthologID: automation of genome-scale ortholog identification within a parsimony framework. (2006) *Bioinformatics* 22:699-707.
- Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, Hubbell E, Au M, Chung EY, Lashkari D, Lemieux B, Dean C, Lipshutz RJ, Ausubel FM, Davis RW, Oefner PJ.** Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. (1999) *Nat Genet* 23:203-207.
- Chu Y, Su X, Huang Q, Zhang X.** Patterns of DNA sequence variation at candidate gene loci in black poplar (*Populus nigra* L.) as revealed by single nucleotide polymorphisms. (2009) *Genetica* 137:141-150.
- Churchill GA, Doerge RW.** Empirical threshold values for quantitative trait mapping. (1994) *Genetics* 138:963-971.
- Dana Farber Cancer Institute TGIP.** The *Helianthus annuus* Gene Index (HaGI) [<http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gireport.pl?gudb=Sunflower>].
- Deutsch M, Long M.** Intron-exon structures of eukaryotic model organisms. (1999) *Nucl Acids Res* 27:3219-3228.
- Dezar CA, Tioni MF, Gonzalez DH, Chan RL.** Identification of three MADS-box genes expressed in sunflower capitulum. (2003) *J Exp Bot* 54:1637-1639.
- Di Rienzo JA, Casanoves F, Balzarini MG, Gonzalez L, Tablada M, Robledo CW.** InfoStat versión 2009. (2009) *Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina*
- Diatchenko L, Lau Y, Campbell A, Chenchick A, Moqaddam F, Huang B, Lukyanov S, Konstantin L, Gurskaya N, Sverdlov E, Siebert P.** Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. (1996) *Proc Natl Acad Sci U S A* 93:6025-6030.
- Díaz-Zorita M, Duarte GA, Plante E, Díaz-Zorita, Asociados D.** N°7 El cultivo del girasol. (2003) *Cuadernillo Informativo* Buenos Aires, Argentina. 10pp.
- Dorrell DG, Vick BA.** Properties and processing of oilseed sunflower en: Sunflower Technology and Production. (1997) Series (ed). ASA, CSSA, and ASSA. Madison, WI. . 35:709-745pp.
- Dracatos PM, Cogan NO, Dobrowolski MP, Sawbridge TI, Spangenberg GC, Smith KF, Forster JW.** Discovery and genetic mapping of single nucleotide polymorphisms in candidate genes for pathogen defence response in perennial ryegrass (*Lolium perenne* L.). (2008) *Theor Appl Genet* 117:203-219.
- Dumas B, Sailland A, Cheviet JP, Freyssinet G, Pallett K.** Identification of barley oxalate oxidase as a germin-like protein. (1993) *C R Acad Sci III* 316:793-798.
- Eason JR, Ryan DJ, Watson LM, Hedderley D, Christey MC, Braun RH, Coupe SA.** Suppression of the cysteine protease, aleurain, delays floret senescence in *Brassica oleracea*. (2005) *Plant Mol Biol* 57:645-657.
- Ehrling J, Sauveplane V, Olry A, Ginglinger JF, Provart NJ, Werck-Reichhart D.** An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in *Arabidopsis thaliana*. (2008) *BMC Plant Biol* 8:47.

- Ehrenreich I, Hanzawa Y, Chou L, Roe J, Kover P, Purugganan M.** Candidate gene association mapping of Arabidopsis flowering time. (2009) *Genetics*
- Elias R, Till BJ, Mba C, Al-Safadi B.** Optimizing TILLING and Ecotilling techniques for potato (*Solanum tuberosum* L). (2009) *BMC Res Notes* 2:141.
- Ersoz E, Yu J, Buckler ES.** Applications of Linkage Disequilibrium and Association Mapping in Crop Plants en: Genomics-Assisted Crop Improvement. (2007) Varshney, Tuberosa (ed). Springer Netherlands. Vol 1: Genomics Approaches and Platforms:97-119pp.
- Escande AR, Laich FS, Pedraza MV.** Field testing of honeybee-dispersed *Trichoderma* spp. to manage sunflower head rot (*Sclerotinia sclerotiorum*). (2002) *Plant Pathol* 51:346-351.
- Eskridge KM.** Field design and the search for quantitative trait loci in plants. (2003) Graybill Conference on Microarray bioinformatics and related topics, Fort Collins, CO.
- Evanno G, Regnaut S, Goudet J.** Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. (2005) *Mol Ecol* 14:2611-2620.
- Falush D, Stephens M, Pritchard JK.** Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. (2003) *Genetics* 164:1567-1587.
- Farris JS, Albert VA, Källersjö M, Lipscomb D, Kluge AG.** PARSIMONY JACKKNIFING OUTPERFORMS NEIGHBOR-JOINING. (1996) *Cladistics* 12:99-124.
- Felsenstein J.** Evolutionary trees from DNA sequences: a maximum likelihood approach. (1981) *J Mol Evol* 17:368-376.
- Felsenstein J.** Phylip (Phylogeny Inference Package). Washington, University of Washington. (1991-2004)
- Feltus FA, Singh HP, Lohithaswa HC, Schulze SR, Silva TD, Paterson AH.** A comparative genomics strategy for targeted discovery of single-nucleotide polymorphisms and conserved-noncoding sequences in orphan crops. (2006) *Plant Physiol.* 140:1183-1191.
- Feng J, Vick BA, Lee MK, Zhang HB, Jan CC.** Construction of BAC and BIBAC libraries from sunflower and identification of linkage group-specific clones by overgo hybridization. (2006) *Theor Appl Genet* 113:23-32.
- Fernandez P, Di Rienzo J, Fernandez L, Hopp HE, Paniego N, Heinz RA.** Transcriptomic identification of candidate genes involved in sunflower responses to chilling and salt stresses based on cDNA microarray analysis. (2008) *BMC Plant Biol* 8:11.
- Fernandez P, Paniego N, Lew S, Hopp HE, Heinz RA.** Differential representation of sunflower ESTs in enriched organ-specific cDNA libraries in a small scale sequencing project. (2003) *BMC Genomics* 4:40.
- Fick GN, Zimmer DE.** Yield stability of sunflower hybrids and open pollinated varieties. (1976) Proc. 7TH Int. Sunflower Conference, Paris, France.
- Fitch WM.** Distinguishing homologous from analogous proteins. (1970) *Syst Zool* 19:99-113.
- Flint-Garcia SA, Thornsberry JM, Buckler ES.** Structure of linkage disequilibrium in plants. (2003) *Annu Rev Plant Biol* 54:357-374.
- Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES.** Maize association population: a high-resolution platform for quantitative trait locus dissection. (2005) *Plant J* 44:1054-1064.

- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM.** International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. (2007) *Nature* 449:851–861.
- Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, Brinkman FS.** Improving the specificity of high-throughput ortholog prediction. (2006) *BMC Bioinformatics* 7:270.
- Galeano CH, Gomez M, Rodriguez LM, Blair MW.** CEL I Nuclease Digestion for SNP Discovery and Marker Development in Common Bean (*Phaseolus vulgaris* L.). (2009) *Crop Sci* 49:1-14.
- Ganal MW, Altmann T, Roder MS.** SNP identification in crop plants. (2009) *Curr Opin Plant Biol* 12:211-217.
- Garris AJ, McCouch SR, Kresovich S.** Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). (2003) *Genetics* 165:759-769.
- Gaut BS, Long AD.** The lowdown on linkage disequilibrium. (2003) *Plant Cell* 15:1502-1506.
- Gentzbittel L, Abbott A, Galaud JP, Georgi L, Fabre F, Liboz T, Alibert G.** A bacterial artificial chromosome (BAC) library for sunflower, and identification of clones containing genes for putative transmembrane receptors. (2002) *Mol Genet Genom* 266:979-987.
- Gentzbittel L, Mestries E, Mouzeyar S, Mazeyrat F, Badaoui S, Vear F, De Labrouhe DT, Nicolas P.** A composite map of expressed sequences and phenotypic traits of the sunflower *Helianthus annuus* L. genome (1999) *Theor Appl Genet* 99:218-234.
- Gentzbittel L, Mouzeyar S, Badaoui S, Mestries E, Vear F, D. TDL, Nicolas P.** Cloning of molecular markers for disease resistance in sunflower, *Helianthus annuus* L. (1998) *Theor Appl Genet* 96:519-525.
- Gentzbittel L, Vear F, Zhang YX, Berville A, Nicolas P.** Development of a consensus linkage RFLP map of cultivated sunflower (*Helianthus annuus* L.). (1995) *Theor Appl Genet* 90:1079-1086.
- Gepstein S, Sabehi G, Carp MJ, Hajouj T, Neshar MF, Yariv I, Dor C, Bassani M.** Large-scale identification of leaf senescence-associated genes. (2003) *Plant J* 36:629-642.
- Giacomelli C, Ribichich K, Dezar C, Chan RL.** Caracterización funcional y participación en los mecanismos de respuesta a estreses bióticos y abióticos de genes de girasol pertenecientes a la familia WRKY. (2010) Taller ASAGIR 2010, Mar del Plata, Buenos Aires.
- Giancola S, McKhann HI, Berard A, Camilleri C, Durand S, Libeau P, Roux F, Reboud X, Gut IG, Brunel D.** Utilization of the three high-throughput SNP genotyping methods, the GOOD assay, Amplifluor and TaqMan, in diploid and polyploid plants. (2006) *Theor Appl Genet* 112:1115-1124.
- Gilchrist EJ, Haughn GW, Ying CC, Otto SP, Zhuang J, Cheung D, Hamberger B, Aboutorabi F, Kalynyak T, Johnson L, Bohlmann J, Ellis BE, Douglas CJ, Cronk QCB.** Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. (2006) *Mol Ecol* 15:1367–1378.
- Gilles PN, Wu DJ, Foster CB, Dillon PJ, Chanock SJ.** Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips. (1999) *Nat Biotechnol* 17:365-370.
- Giordano M, Oefner PJ, Underhill PA, Cavalli Sforza LL, Tosi R, Richiardi PM.** Identification by denaturing high-performance liquid chromatography of numerous polymorphisms in a candidate region for multiple sclerosis susceptibility. (1999) *Genomics* 56:247-253.
- Goloboff PA, Farris JS, Nixon KC.** TNT, a free program for phylogenetic analysis. (2008) *Cladistics* 24:774-786.

Gonzalez-Martinez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB. DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. (2006) *Genetics* 172:1915-1926.

Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB. Association genetics in *Pinus taeda* L. I. Wood property traits. (2007) *Genetics* 175:399-409.

Gu Y, Fu Y, Dowd P, Li S, Vernoud V, Gilroy S, Yang Z. A Rho family GTPase controls actin dynamics and tip growth via two counteracting downstream pathways in pollen tubes. (2005) *The Journal of cell biology* 169:127.

Gupta PK, Roy JK, Prasad M. Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. (2001) *Curr Sci* 80:524-535.

Gupta PK, Rustgi S, Mir RR. Array-based high-throughput DNA markers for crop improvement. (2008) *Heredity* 101:5-18.

Hagenblad J, Tang C, Molitor J, Werner J, Zhao K, Zheng H, Marjoram P, Weigel D, Nordborg M. Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. (2004) *Genetics* 168:1627-1638.

Haldane JB, Waddington CH. Inbreeding and Linkage. (1931) *Genetics* 16:357-374.

Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. . (1999) *Nucl Acids Symp Ser* 41:95-98.

Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA, Paterson AH, Kresovich S. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of sorghum bicolor. (2004) *Genetics* 167:471-483.

Han W, Yip SP, Wang J, Yap MK. Using denaturing HPLC for SNP discovery and genotyping, and establishing the linkage disequilibrium pattern for the all-trans-retinol dehydrogenase (RDH8) gene. (2004) *J Hum Genet* 49:16-23.

Haseneyer G, Stracke S, Piepho HP, Sauer S, Geiger HH, Graner A. DNA polymorphisms and haplotype patterns of transcription factors involved in barley endosperm development are associated with key agronomic traits. (2010) *BMC Plant Biol* 10:5.

Heckenberger M, van der Voort JR, Peleman J, Bohn M. Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties: II. Genetic and technical sources of variation in AFLP data and comparison with SSR data. (2003) *Molecular Breeding* 12:97-106.

Hegedus DD, Rimmer SR. *Sclerotinia sclerotiorum*: When “to be or not to be” a pathogen? (2005) *FEMS Microbiol Let* 251:177–184.

Heiser CB, Smith DM. New chromosome numbers in *Helianthus* and related genera. (1955) *Proc Indiana Acad Sci* 64:250–253.

Hewezi T, Petitprez M, Gentzbittel L. Primary metabolic pathways and signal transduction in sunflower (*Helianthus annuus* L.): comparison of transcriptional profiling in leaves and immature embryos using cDNA microarrays. (2006) *Planta* 223:948-964.

Higuchi R, Fockler C, Dollinger G, Watson R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. (1993) *Biotechnology (N Y)* 11:1026-1030.

Hill WG, Weir BS. Maximum-likelihood estimation of gene location by linkage disequilibrium. (1994) *Am J Hum Genet* 54:705-714.

- Hirschhorn JN, Daly MJ.** Genome-wide association studies for common diseases and complex traits. (2005) *Nat Rev Genet* 6:95-108.
- Hirsh AE, Fraser HB.** Protein dispensability and rate of evolution. (2001) *Nature* 411:1046-1049.
- Holland JB.** Genetic architecture of complex traits in plants. (2007) *Curr Opin Plant Biol* 10:156-161.
- Hu W, dePamphilis CW, Ma H.** Phylogenetic analysis of the plant-specific zinc finger-homeobox and mini zinc finger gene families. (2008) *J Integr Plant Biol* 50:1031-1045.
- Hu X, Bidney DL, Yalpani N, Duvick JP, Crasta O, Folkerts O, Lu G.** Overexpression of a gene encoding hydrogen peroxide-generating oxalate oxidase evokes defense responses in sunflower. (2003) *Plant Physiol* 133:170-181.
- Huang X, Madan A.** CAP3: A DNA sequence assembly program. (1999) *Genome Res* 9:868-877.
- Hung CC, Su YN, Lin CY, Chang YF, Chang CH, Cheng WF, Chen CA, Lee CN, Lin WL.** Comparison of the mismatch-specific endonuclease method and denaturing high-performance liquid chromatography for the identification of HBB gene mutations. (2008) *BMC Biotechnol* 8:62.
- Ihaka R, Gentleman R.** R: A language for data analysis and graphics. (1996) *J Comput Graph Stat* 5:299-314.
- Ingaramo J.** N° 11 Plan Estratégico ASAGIR 2006-2015. (2005) *Cuadernillo Informativo*. Buenos Aires, Argentina. 24pp.
- Ingvarsson PK.** Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). (2005) *Genetics* 169:945-953.
- Ingvarsson PK, Garcia MV, Luquez V, Hall D, Jansson S.** Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 Locus in European aspen (*Populus tremula*, Salicaceae). (2008) *Genetics* 178:2217-2226.
- Jan C-C, Seiler GJ.** Sunflower en: Genetic resources, chromosome engineering, and crop improvement: oilseed crops. (2007) Singh (ed). Taylor & Francis Group. Boca Raton, FL 33487-2742. 4:166pp.
- Jan CC, Vick BA, Miller JF, Kahler AL, Butler ET.** Construction of an RFLP linkage map for cultivated sunflower. (1998) *Theor Appl Genet* 96:15-22.
- Jones AC, Austin J, Hansen N, Hoogendoorn B, Oefner PJ, Cheadle JP, O'Donovan MC.** Optimal temperature selection for mutation detection by denaturing HPLC and comparison to single-stranded conformation polymorphism and heteroduplex analysis. (1999) *Clin Chem* 45:1133-1140.
- Jordan IK, Rogozin IB, Wolf YI, Koonin EV.** Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. (2002) *Genome Res* 12:962-968.
- Kader JC.** Lipid-Transfer Proteins in Plants. (1996) *Annu Rev Plant Physiol Plant Mol Biol* 47:627-654.
- Kashuk C, SenGupta S, Eichler E, Chakravarti A.** ViewGene: a graphical tool for polymorphism visualization and characterization. (2002) *Genome Res* 12:333-338.
- Katinas L, Gutiérrez DG, Grossi MA, Crisci JV.** Panorama de la familia Asteraceae (= Compositae) en la Republica Argentina. (2007) *Bol Soc Argent Bot* 42:113-129.
- Katoh K, Misawa K, Kuma K, Miyata T.** MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. (2002) *Nucl Acids Res* 30:3059-3066.

- Kawasaki T, Henmi K, Ono E, Hatakeyama S, Iwano M, Satoh H, Shimamoto K.** The small GTP-binding protein rac is a regulator of cell death in plants. (1999) *Proc Natl Acad Sci U S A* 96:10922–10926.
- Kearsey MJ, Farquhar AG.** QTL analysis in plants; where are we now? (1998) *Heredity* 80 (Pt 2):137-142.
- Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC.** Identification of the cystic fibrosis gene: genetic analysis. (1989) *Sci* 245:1073-1080.
- Kim MJ, Yoo YA, Kim HJ, Kang S, Kim YG, Kim JS, Yoo YD.** Mitochondrial ribosomal protein L41 mediates serum starvation-induced cell-cycle arrest through an increase of p21(WAF1/CIP1). (2005) *Biochem Biophys Res Commun* 338:1179-1184.
- Kolkman JM, Berry ST, Leon A, Slabaugh MB, Tang S, Gao W, Shintani DK, Burke JM, Knapp SJ.** Single Nucleotide Polymorphism and Linkage Disequilibrium in Sunflower. (2007) *Genetics* 177:457-468.
- Kosambi DD.** The estimation of a map distance from recombination values. (1944) *Ann Eugen* 12:172–175.
- Koski LB, Golding GB.** The closest BLAST hit is often not the nearest neighbor. (2001) *J Mol Evol* 52:540-542.
- Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A.** Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). (2003) *Mol Genet Genom* 270:24-33.
- Kota R, Varshney RK, Prasad M, Zhang H, Stein N, Graner A.** EST-derived single nucleotide polymorphism markers for assembling genetic and physical maps of the barley genome. (2008) *Funct Integr Genom* 8:223-233.
- Kota R, Wolf M, Michalek W, Graner A.** Application of denaturing high-performance liquid chromatography for mapping of single nucleotide polymorphisms in barley (*Hordeum vulgare* L.). (2001) *Genome* 44:523-528.
- Kraakman AT, Niks RE, Van den Berg PM, Stam P, Van Eeuwijk FA.** Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. (2006) *Genetics* 168:435-446.
- Krill AM, Kirst M, Kochian LV, Buckler ES, Hoekenga OA.** Association and linkage analysis of aluminum tolerance genes in maize. (2010) *PLoS One* 5:e9958.
- Laemmli UK.** Cleavage of structural proteins during the assembly of the head of bacteriophage T4. (1970) *Nature* 227:680-685.
- Lai Z, Gross BL, Zou Y, Andrews J, Rieseberg LH.** Microarray analysis reveals differential gene expression in hybrid sunflower species. (2006) *Mol Ecol* 15:1213-1227.
- Lai Z, Livingstone K, Zou Y, Church SA, Knapp SJ, Andrews J, Rieseberg LH.** Identification and mapping of SNPs from ESTs in sunflower. (2005) *Theor Appl Genet* 111:1532-1544.
- Lam E, Kato N, Lawton M.** Programmed cell death, mitochondria and the plant hypersensitive response. (2001) *Nature* 411:848-853.
- Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newberg LA.** MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. (1987) *Genomics* 1:174-181.

- Lee HH, Kim DJ, Ahn HJ, Ha JY, Suh SW.** Crystal structure of T-protein of the glycine cleavage system. Cofactor binding, insights into H-protein recognition, and molecular basis for understanding nonketotic hyperglycinemia. (2004) *J Biol Chem* 279:50514-50523.
- Lee YJ, Szumlanski A, Nielsen E, Yang Z.** Rho-GTPase-dependent filamentous actin dynamics coordinate vesicle targeting and exocytosis during tip growth. (2008) *The Journal of cell biology* 181:1155.
- Lewontin RC.** The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. (1964) *Genetics* 49:49-67.
- Lexer C, Lai Z, Rieseberg LH.** Candidate gene polymorphisms associated with salt tolerance in wild sunflower hybrids: implications for the origin of *Helianthus paradoxus*, a diploid hybrid species. (2003) *New Phytol* 161:225-233.
- Li L, Paulo MJ, Strahwald J, Lubeck J, Hofferbert HR, Tacke E, Junghans H, Wunder J, Draffehn A, van Eeuwijk F, Gebhardt C.** Natural DNA variation at candidate loci is associated with potato chip color, tuber starch content, yield and starch yield. (2008) *Theor Appl Genet* 116:1167-1181.
- Lijavetzky D, Cabezas JA, Ibanez A, Rodriguez V, Martinez-Zapater JM.** High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. (2007) *BMC Genomics* 8:424.
- Lin CH, Yeakley JM, McDaniel TK, Shen R.** Medium- to High-Throughput SNP Genotyping Using VeraCode Microbeads en: DNA and RNA Profiling in Human Blood: Methods and Protocols. (2009) Press (ed). 496:
- Liu A, Burke JM.** Patterns of nucleotide diversity in wild and cultivated sunflower. (2006) *Genetics* 173:321-330.
- Liu JH, Lee-Tamon SH, Reid DM.** Differential and wound-inducible expression of 1-aminocyclopropane-1-carboxylate oxidase genes in sunflower seedlings. (1997) *Plant Mol Biol* 34:923-933.
- Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J.** Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. (2003) *Genetics* 165:2117-2128.
- Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K.** Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. (1995) *PCR Methods Appl* 4:357-362.
- Lynch M, Ritland K.** Estimation of pairwise relatedness with molecular markers. (1999) *Genetics* 152:1753-1766.
- Makridakis NM, Reichardt JK.** Multiplex automated primer extension analysis: simultaneous genotyping of several polymorphisms. (2001) *Biotechniques* 31:1374-1380.
- Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA.** A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. (2007) *Genetics* 175:879-889.
- Mallappa C, Yadav V, Negi P, Chattopadhyay S.** A basic leucine zipper transcription factor, G-box-binding factor 1, regulates blue light-mediated photomorphogenic growth in *Arabidopsis*. (2006) *J Biol Chem* 281:22190-22199.
- Maringolo CA.** Regiones Cromosómicas Asociadas a Resistencia a Podredumbre Húmeda del Capítulo de Girasol (*Sclerotinia sclerotiorum* (Lib.) de Bary). (2007) Posgrado en Producción Vegetal, FCA, Mar del Plata, UNMdP.

- Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD.** The extent of linkage disequilibrium in rice (*Oryza sativa* L.). (2007) *Genetics* 177:2223-2232.
- Mestries E, Gentzbittel L, de Labrouhe DT, Nicolas P, Vear F.** Analyses of quantitative trait loci associated with resistance to *Sclerotinia sclerotiorum* in sunflower (*Helianthus annuus* L.). (1998) *Mol Breed* 4:215-226.
- Micic Z, Hahn V, Bauer E, Melchinger AE, Knapp SJ, Tang S, Schon CC.** Identification and validation of QTL for *Sclerotinia* midstalk rot resistance in sunflower by selective genotyping. (2005b) *Theor Appl Genet* 111:233-242.
- Micic Z, Hahn V, Bauer E, Schon CC, Knapp SJ, Tang S, Melchinger AE.** QTL mapping of *Sclerotinia* midstalk-rot resistance in sunflower. (2004) *Theor Appl Genet* 109:1474-1484.
- Micic Z, Hahn V, Bauer E, Schon CC, Melchinger AE.** QTL mapping of resistance to *Sclerotinia* midstalk rot in RIL of sunflower population NDBLOSSel x CM625. (2005a) *Theor Appl Genet* 110:1490-1498.
- Minch E, Linares AR, Goldstein D, Feldman M, Cavalli-Sforza LL.** Microsat (version 1.5b): a computer program for calculating various statistics on microsatellite allele data. (1995-1996)
- Monna L, Ohta R, Masuda H, Koike A, Minobe Y.** Genome-wide searching of single-nucleotide polymorphisms among eight distantly and closely related rice cultivars (*Oryza sativa* L.) and a wild accession (*Oryza rufipogon* Griff.). (2006) *DNA Res* 13:43-51.
- Morales M, Roig E, Monforte AJ, Arus P, Garcia-Mas J.** Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). (2004) *Genome* 47:352-360.
- Mukherjee K, Brocchieri L, Burglin TR.** A comprehensive classification and evolutionary analysis of plant homeobox genes. (2009) *Mol Biol Evol* 26:2775-2794.
- Navaud O, Dabos P, Carnus E, Tremousaygue D, Herve C.** TCP transcription factors predate the emergence of land plants. (2007) *J Mol Evol* 65:23-33.
- Nieto C, Piron F, Dalmais M, Marco CF, Moriones E, Gómez-Guillamón ML, Truniger V, Gómez P, Garcia-Mas J, Aranda MA, Bendahmane A.** EcoTILLING for the identification of allelic variants of melon eIF4E, a factor that controls virus susceptibility. (2007) *BMC Plant Biol* 7:34.
- Ning Z, Cox AJ, Mullikin JC.** SSAHA: a fast search method for large DNA databases. (2001) *Genome Res* 11:1725-1729.
- Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D.** The extent of linkage disequilibrium in *Arabidopsis thaliana*. (2002) *Nat Genet* 30:190-193.
- Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J.** The pattern of polymorphism in *Arabidopsis thaliana*. (2005) *PLoS Biol* 3:e196.
- Nordborg M, Tavare S.** Linkage disequilibrium: what history has to tell us. (2002) *Trends Genet* 18:83-90.
- Noyes RD, Hancock JG.** Role of oxalic acid in the *Sclerotinia* wilt of sunflower. (1981) *Physiol Plant Pathol* 18:123-132.
- O'Brien KP, Remm M, Sonnhammer EL.** Inparanoid: a comprehensive database of eukaryotic orthologs. (2005) *Nucl Acids Res* 33:D476-480.

- Oefner P, Underhill P.** DNA Mutation Detection Using Denaturing High-Performance Liquid Chromatography (DHPLC) en: *Current Protocols in Human Genetics*. (1998) John Wiley & Sons (ed). 7.10.11-17.10.12pp.
- Oleykowski CA, Bronson Mullins CR, Godwin AK, Yeung AT.** Mutation detection using a novel plant endonuclease. (1998) *Nucl Acids Res* 26:4597-4602.
- Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD.** Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. (2004) *Genetics* 167:1361-1369.
- Ono E, Wong HL, Kawasaki T, Hasegawa M, Kodama O, Shimamoto K.** Essential role of the small GTPase Rac in disease resistance of rice. (2001) *Proceedings of the National Academy of Sciences* 98:759.
- Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN.** Association Mapping in Plants. (2007) Oraguzie, Rikkerink, Gardiner, De Silva. Springer Science+Business Media. New York, USA.
- Osterberg MK, Shavorskaya O, Lascoux M, Lagercrantz U.** Naturally occurring indel variation in the Brassica nigra COL1 gene is associated with variation in flowering time. (2002) *Genetics* 161:299-306.
- Özdemir N, Horn R, Friedt W.** Construction and characterization of a BAC library for sunflower (*Helianthus annuus* L.). (2004) *Euphytica* 138:177-183.
- Pajerowska-Mukhtar K, Stich B, Achenbach U, Ballvora A, Lubeck J, Strahwald J, Tacke E, Hofferbert HR, Ilarionova E, Bellin D, Walkemeier B, Baskow R, Kersten B, Gebhardt C.** Single nucleotide polymorphisms in the allene oxide synthase 2 gene are associated with field resistance to late blight in populations of tetraploid potato cultivars. (2009) *Genetics* 181:1115-1127.
- Palaisa K, Morgante M, Tingey S, Rafalski A.** Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. (2004) *Proc Natl Acad Sci U S A* 101:9885-9890.
- Paniego N, Echaide M, Munoz M, Fernandez L, Torales S, Faccio P, Fuxan I, Carrera M, Zandomeni R, Suarez EY, Hopp HE.** Microsatellite isolation and characterization in sunflower (*Helianthus annuus* L.). (2002) *Genome* 45:34-43.
- Paniego N, Heinz R, Fernandez P, Talia P, Nishinakamasu V, Hopp HE.** Sunflower en: *Genome Mapping and Molecular Breeding in Plants*. (2007) Kole. (ed). Springer-Verlag. Berlin. 2:153-177pp.
- Parkin IA, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ.** Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. (2005) *Genetics* 171:765-781.
- Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP.** Light-generated oligonucleotide arrays for rapid DNA sequence analysis. (1994) *Proc Natl Acad Sci U S A* 91:5022-5026.
- Peluffo L, Lia V, Hopp HE, Paniego N, Carrari F, Heinz R.** Identificación de genes candidatos para la resistencia a la podredumbre húmeda del capítulo causada por *Sclerotinia sclerotiorum* en girasol. (2010a) Taller ASAGIR 2010, Mar del Plata, Buenos Aires.
- Peluffo L, Lia V, Troglia C, Maringolo C, Norma P, Escande A, Esteban Hopp H, Lytovchenko A, Fernie AR, Heinz R, Carrari F.** Metabolic profiles of sunflower genotypes with contrasting response to *Sclerotinia sclerotiorum* infection. (2010b) *Phytochem* 71:70-80.
- Pereyra V, Bazzalo ME.** Podredumbre del capítulo de girasol. (1988) *Jornada de actualización técnica para profesionales*. Balcarce, Buenos Aires, Argentina. sppp.

- Pereyra V, Escande A.** Enfermedades del girasol en la Argentina. Manual de reconocimiento. (1994) INTA. Buenos Aires.
- Perry JA, Wang TL, Welham TJ, Gardner S, Pike JM, Yoshida S, Parniske M.** A TILLING reverse genetics tool and a web-accessible collection of mutants of the legume *Lotus japonicus*. (2003) *Plant Physiol* 131:866-871.
- Pessino S, Martelotto LG.** Métodos para el estudio de la expresión de genes en: Biotecnología y Mejoramiento Vegetal. (2004) Echenique, Rubinstein, Mroginski (ed). Ediciones INTA. Buenos Aires. VII-2:424pp.
- Pires N, Dolan L.** Early evolution of bHLH proteins in plants. (2010) *Plant Signal Behav* 5:
- Ponting RC, Drayton MC, Cogan NOI, Dobrowolski MP, Spangenberg GC, Smith KF, Forster JW.** SNP discovery, validation, haplotype structure and linkage disequilibrium in full-length herbage nutritive quality genes of perennial ryegrass (*Lolium perenne* L.). (2007) *Molecular Genetics and Genomics* 278:585-597.
- Poormohammad Kiani G, Maury P, Sarrafi A, Grieu P.** QTL analysis of chlorophyll fluorescence parameters in sunflower (*Helianthus annuus* L.) under well-watered and water-stressed conditions. (2008) *Plant Sci* 175:565-573.
- Poormohammad Kiani S, Grieu P, Maury P, Hewezi T, Gentzbittel L, Sarrafi A.** Genetic variability for physiological traits under drought conditions and differential expression of water stress-associated genes in sunflower (*Helianthus annuus* L.). (2007b) *Theor Appl Genet* 114:193-207.
- Poormohammad Kiani S, Talia P, Grieu P, Maury P, Hewezi T, Gentzbittel L, Sarrafi A.** Genetic analysis of plant water status and osmotic adjustment in recombinant inbred lines of sunflower under two water treatments. (2007a) *Plant Sci* 178:773-787.
- Prats E, Galindo JC, Bazzalo ME, Leon A, Macias FA, Rubiales D, Jorrin JV.** Antifungal activity of a new phenolic compound from capitulum of a head rot-resistant sunflower genotype. (2007) *J Chem Ecol* 33:2245-2253.
- Price AH.** Believe it or not, QTLs are accurate! (2006) *Trends Plant Sci* 11:213-216.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D.** Principal components analysis corrects for stratification in genome-wide association studies. (2006) *Nat Genet* 38:904-909.
- Pritchard JK, Przeworski M.** Linkage disequilibrium in humans: models and data. (2001) *Am J Hum Genet* 69:1-14.
- Pritchard JK, Stephens M, Donnelly P.** Inference of population structure using multilocus genotype data [<http://www.pritch.bsd.uchicago.edu>] (2000) *Genetics* 155:945-959.
- Rafalski A.** Applications of single nucleotide polymorphisms in crop genetics. (2002) *Curr Opin Plant Biol* 5:94-100.
- Rafalski JA.** Association genetics in crop improvement. (2010) *Curr Opin Plant Biol* 13:174-180.
- Randall LP, Coldham NG, Woodward MJ.** Detection of mutations in *Salmonella enterica* gyrA, gyrB, parC and parE genes by denaturing high performance liquid chromatography (DHPLC) using standard HPLC instrumentation. (2005) *J Antimicrob Chemother* 56:619-623.
- Ravel C, Praud S, Murigneux A, Canaguier A, Sapet F, Samson D, Balfourier F, Dufour P, Chalhou B, Brunel D, Beckert M, Charmet G.** Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). (2006) *Genome* 49:1131-1139.
- Raven P, Axelrod D.** Angiosperm biogeography and past continental movements. (1974) *Ann Missouri Bot Gard* 61:539-673.

- Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ESt.** Structure of linkage disequilibrium and phenotypic associations in the maize genome. (2001) *Proc Natl Acad Sci U S A* 98:11479-11484.
- Remm M, Storm CE, Sonnhammer EL.** Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. (2001) *J Mol Biol* 314:1041-1052.
- Ren B, Zhou J-M, Komiyama M.** Straightforward detection of SNPs in doublestranded DNA by using exonuclease III/nuclease S1/PNA system. (2004) *Nucl Acids Res* 34:e42.
- Rieseberg LH, Seiler GJ.** Molecular evidence and the origin and development of the domesticated sunflower (*Helianthus annuus*, Asteraceae). (1990) *Economical Botany* 44:79-91.
- Rodriguez MA, Venedikian N, Bazzalo ME, Godeas A.** Histopathology of *Sclerotinia sclerotiorum* attack on flower parts of *Helianthus annuus* heads in tolerant and susceptible varieties. (2004) *Mycopathologia* 157:291-302.
- Rogers JS.** Measures of genetic similarity and distance. (1972) *Studies Genetics VII. University Texas Publ.*(7213), Texas 145-153.
- Rose TM.** CODEHOP-mediated PCR - a powerful technique for the identification and characterization of viral genomes. (2005) *Virology* 2:20.
- Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MA, Hillel J, Maki-Tanila A, Tixier-Boichard M, Vignal A, Wimmers K, Weigend S.** Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. (2001) *Genetics* 159:699-713.
- Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A, Svensson JT, Wanamaker SI, Walia H, Rodriguez EM, Hedley PE, Liu H, Morris J, Close TJ, Marshall DF, Waugh R.** Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. (2005) *Mol Genet Genomics* 274:515-527.
- Rozas J, Rozas R.** DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. (1999) *Bioinformatics* 15:174-175.
- Rozen S, Skaletsky H.** Primer3 on the WWW for general users and for biologist programmers [<http://frodo.wi.mit.edu/>] en: *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. (2000) Krawetz, Misener (ed). Humana Press. Totowa, NJ. 365-386pp.
- Russell J, Booth A, Fuller J, Harrower B, Hedley P, Machray G, Powell W.** A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. (2004) *Genome* 47:389-398.
- Sambrook J, Fritsch E, Maniatis T.** *Molecular cloning: a laboratory manual*. (1989) Laboratory. Cold Spring Harbor. New York.
- Sato Y, Fukuda Y, Hirano HY.** Mutations that cause amino acid substitutions at the invariant positions in homeodomain of OSH3 KNOX protein suggest artificial selection during rice domestication. (2001) *Genes Genet Syst* 76:381-392.
- Saville DJ.** Multiple comparison procedures: the practical solution. (1990) *American Statistician* 44:174-180.
- Schiex T, Gaspin C.** CARTHAGENE: constructing and joining maximum likelihood genetic maps. (1997) *Proc Int Conf Intell Syst Mol Biol* 5:258-267.
- Schilling EE, Heiser CB.** Infrageneric classification of *Helianthus* (Compositae). (1981) *Taxon* 30:393-403.

- Schneiter AA, Miller JF.** Description of sunflower growth stages. . (1981) *Crop Sci* 21:901-903.
- Seiler JG, Rieseberg LH.** Systematics, origin, resources of the wild and domesticated sunflower en: Sunflower Technology and Production. (1997) Agronomy series Agronomy (*ed*). Madison, Wisconsin. 21-65 pp.
- Shirasawa K, Maeda H, Monna L, Kishitani S, Nishio T.** The number of genes having diVerent alleles between rice cultivars estimated by SNP analysis. (2007) *Theor Appl Genet* 115:1067–1074.
- Simko I, Costanzo S, Haynes KG, Christ BJ, Jones RW.** Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. (2004a) *Theor Appl Genet* 108:217-224.
- Simko I, Haynes KG, Ewing EE, Costanzo S, Christ BJ, Jones RW.** Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. (2004b) *Mol Genet Genom* 271:522-531.
- Simko I, Haynes KG, Jones RW.** Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. (2006) *Genetics* 173:2237-2245.
- Skot L, Humphreys J, Humphreys MO, Thorogood D, Gallagher J, Sanderson R, Armstead IP, Thomas ID.** Association of candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.). (2007) *Genetics* 177:535-547.
- Skøt L, Humphreys MO, Armstead I, Heywood S, Skøt KP, Sanderson R, Thomas ID, Chorlton KH, Hamilton NRS.** An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.). (2005) *Mol Breed* 15:233-245.
- Smalle J, Vierstra RD.** The ubiquitin 26S proteasome proteolytic pathway. (2004) *Annu Rev Plant Biol* 55:555-590.
- Somers DJ, Kirkpatrick R, Moniwa M, Walsh A.** Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. (2003) *Genome* 46:431-437.
- Song BH, Windsor AJ, Schmid KJ, Ramos-Onsins S, Schranz ME, Heide AJ, Mitchell-Olds T.** Multilocus patterns of nucleotide diversity, population structure and linkage disequilibrium in *Boechera stricta*, a wild relative of *Arabidopsis*. (2009) *Genetics* 181:1021-1033.
- Southern EM.** High-density gridding: techniques and applications. (1996) *Curr Opin Biotechnol* 7:85-88.
- Spiegelman JI, Mindrinos MN, Oefner PJ.** High-accuracy DNA sequence variation screening by DHPLC. (2000) *Biotechniques* 29:1084-1090, 1092.
- Stich B, Melchinger AE.** Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and *Arabidopsis*. (2009) *BMC Genomics* 10:94.
- Stich B, Moring J, Piepho HP, Heckenberger M, Buckler ES, Melchinger AE.** Comparison of mixed-model approaches for association mapping. (2008) *Genetics* 178:1745-1754.
- Storey JD.** A direct approach to false discovery rates. (2002) *J R Statis Soc (B)* 64:479-498.
- Storey JD, Tibshirani R.** Statistical significance for genomewide studies. (2003) *Proc Natl Acad Sci U S A* 100:9440.
- Sulpice R, Pyl ET, Ishihara H, Trenkamp S, Steinfath M, Witucka-Wall H, Gibon Y, Usadel B, Poree F, Piques MC, Von Korff M, Steinhauser MC, Keurentjes JJ, Guenther M, Hoehne M, Selbig J, Fernie AR, Altmann T, Stitt M.** Starch as a major integrator in the regulation of plant growth. (2009) *Proc Natl Acad Sci U S A* 106:10348-10353.

- Syvanen AC.** Accessing genetic variation: genotyping single nucleotide polymorphisms. (2001) *Nat Rev Genet* 2:930-942.
- Tajima F.** Evolutionary relationship of DNA sequences in finite populations. (1983) *Genetics* 105:437-460.
- Talia P.** Desarrollo de un mapa genético y físico integrado para girasol cultivado y su aplicación en la caracterización de regiones genómicas involucradas en la resistencia a enfermedades. (2008) Departamento de Ecología, Genética y Evolución. Facultad de Ciencias Exactas y Naturales, Buenos Aires, UBA.
- Talia P, Greizerstein E, Quijano CD, Peluffo L, Fernandez L, Fernandez P, Hopp HE, Paniego N, Heinz RA, Poggio L.** Cytological characterization of sunflower by in situ hybridization using homologous rDNA sequences and a BAC clone containing highly represented repetitive retrotransposon-like sequences. (2010) *Genome* 53:172-179.
- Tang S, Kishore VK, Knapp SJ.** PCR-multiplexes for a genome-wide framework of simple sequence repeat marker loci in cultivated sunflower. (2003) *Theor Appl Genet* 107:6-19.
- Tang S, Yu JK, Slabaugh B, Shintani K, Knapp J.** Simple sequence repeat map of the sunflower genome. (2002) *Theor Appl Genet* 105:1124-1136.
- Tatusov RL, Galperin MY, Natale DA, Koonin EV.** The COG database: a tool for genome-scale analysis of protein functions and evolution. (2000) *Nucl Acids Res* 28:33-36.
- Tatusov RL, Koonin EV, Lipman DJ.** A genomic perspective on protein families. (1997) *Sci* 278:631-637.
- Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS.** Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). (2001) *Proc Natl Acad Sci U S A* 98:9161-9166.
- Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ESt.** Dwarf8 polymorphisms associate with variation in flowering time. (2001) *Nat Genet* 28:286-289.
- Thumma BR, Nolan MF, Evans R, Moran GF.** Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in *Eucalyptus* spp. (2005) *Genetics* 171:1257-1265.
- Till BJ, Burtner C, Comai L, Henikoff S.** Mismatch cleavage by single-strand specific nucleases. (2004) *Nucl Acids Res* 32:2632-2641.
- Till BJ, Colbert T, Tompa R, Enns LC, Codomo CA, Johnson JE, Reynolds SH, Henikoff JG, Greene EA, Steine MN, Comai L, Henikoff S.** High-throughput TILLING for Functional Genomics en: *Plant Functional Genomics: Methods and Protocols*. (2003) Grotewold (ed). Humana Press. Totowa, NJ. 236:205-220pp.
- Till BJ, Cooper J, Tai TH, Colowit P, Greene EA, Henikoff S, Comai L.** Discovery of chemically induced mutations in rice by TILLING. (2007) *BMC Plant Biol* 7:19.
- Troglia C.** Factores ambientales que afectan la supervivencia y producción de inóculo de *Sclerotinia sclerotiorum* y umbral mínimo de inóculo para la podredumbre húmeda del capítulo de girasol. (2003) Facultad de Ciencias Agrarias, Unidad Integrada Balcarce, Mar del Plata, Buenos Aires, Argentina, UNMdP.
- Tsuchihashi Z, Dracopoli NC.** Progress in high throughput SNP genotyping methods. (2002) *Pharmacogenomics J* 2:103-110.

- Uauy C, Paraiso F, Colasuonno P, Tran RK, Tsai H, Berardi S, Comai L, Dubcovsky J.** A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. (2009) *BMC Plant Biol* 9:115.
- Valluru R, Van den Ende W.** Plant fructans in stress environments: emerging concepts and future prospects. (2008) *J Exp Bot* 59:2905-2916.
- Van K, Hwang EY, Kim MY, Kim IH, Cho YI, Cregan PB, Lee SH.** Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs. (2004) *Euphytica* 139:147-157.
- Van K, Hwang EY, Kim MY, Park HJ, Lee SH, Cregan PB.** Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. (2005) *J Hered* 96:529-535.
- Voorrips RE.** MapChart: software for the graphical presentation of linkage maps and QTLs. (2002) *J Hered* 93:77-78.
- Wagner T, Stoppa-Lyonnet D, Fleischmann E, Muhr D, Pages S, Sandberg T, Caux V, Moeslinger R, Langbauer G, Borg A, Oefner P.** Denaturing high-performance liquid chromatography detects reliably BRCA1 and BRCA2 mutations. (1999) *Genomics* 62:369-376.
- Wall DP, Fraser HB, Hirsh AE.** Detecting putative orthologs. (2003) *Bioinformatics* 19:1710-1711.
- Wang H-Y, Luo M, Tereshchenko IV, Frikker DM, Cui X, Li JY, Hu G, Chu Y, Azaro MA, Lin Y, Shen L, Yang Q, Kambouris ME, Gao R, Shih W, Li H.** A genotyping system capable of simultaneously analyzing > 1000 single nucleotide polymorphisms in a haploid genome. (2005) *Genome Res* 15:276-283.
- Wang N, Wang Y, Tian F, King GJ, Zhang C, Long Y, Shi L, Meng J.** A functional genomics resource for *Brassica napus*: development of an EMS mutagenized population and discovery of FAE1 point mutations by TILLING. (2008) *New Phytol* 180:751-765.
- Watterson GA.** On the number of segregating sites in genetical models without recombination. (1975) *Theor Popul Biol* 7:256-276.
- Wei X, Jackson PA, McIntyre CL, Aitken KS, Croft B.** Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. (2006) *Theor Appl Genet* 114:155-164.
- Wi SJ, Kim WT, Park KY.** Overexpression of carnation S-adenosylmethionine decarboxylase gene generates a broad-spectrum tolerance to abiotic stresses in transgenic tobacco plants. (2006) *Plant Cell Rep* 25:1111-1121.
- Wilson LM, Whitt SR, Ibanez AM, Rocheford TR, Goodman MM, Buckler Est.** Dissection of maize kernel composition and starch production by candidate gene association. (2004) *Plant Cell* 16:2719-2733.
- Wood AJ, Joel Duff R, Oliver MJ.** The translational apparatus of *Tortula ruralis*: polysomal retention of transcripts encoding the ribosomal proteins RPS14, RPS16 and RPL23 in desiccated and rehydrated gametophytes. (2000) *J Exp Bot* 51:1655-1662.
- Wright SI, Gaut BS.** Molecular population genetics and the search for adaptive evolution in plants. (2005) *Mol Biol Evol* 22:506-519.
- Wu G, Gu Y, Li S, Yang Z.** Genome-Wide Analysis of Arabidopsis Rop-Interactive CRIB Motif-Containing Proteins That Act as Rop GTPase Targets. (2001) *Plant Cell* 13:2841-2856.
- Xiao W, Oefner PJ.** Denaturing high-performance liquid chromatography: A review. (2001) *Hum Mutat* 17:439-474.

- Xin Z, Wang ML, Barkley NA, Burow G, Franks C, Pederson G, Burke J.** Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. (2008) *BMC Plant Biol* 8:103.
- Xing Y, Frei U, Schejbel B, Asp T, Lubberstedt T.** Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*. (2007) *BMC Plant Biol* 7:43.
- Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J.** Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. (2009a) *PLoS One* 4:e8451.
- Yan WG, Li Y, Agrama HA, Luo D, Gao F, Lu X, Ren G.** Association mapping of stigma and spikelet characteristics in rice (*Oryza sativa* L.). (2009b) *Mol Breed* 24:277–292.
- Yang W, Bai X, Kabelka E, Eaton C, Kamoun S, van der Knaap E, Francis D.** Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. (2004) *Mol Breed* 14:21–34.
- Yu, J. K., Tang S, Slabaugh MB, Heesacker A, Cole G, Herring M, Soper J, Han F, Chu W-C, Webb DM, Thompson L, Edwards KJ, Berry S, Leon AJ, Grondona M, Olungu C, Maes N, Knapp SJ.** Towards a saturated molecular genetic linkage map for cultivated sunflower. (2003) *Crop Sci* 43:367–387.
- Yu B, Sawyer NA, Caramins M, Yuan ZG, Saunderson RB, Pamphlett R, Richmond DR, Jeremy RW, Trent RJ.** Denaturing high performance liquid chromatography: high throughput mutation screening in familial hypertrophic cardiomyopathy and SNP genotyping in motor neurone disease. (2005) *J Clin Pathol* 58:479-485.
- Yu J, Buckler ES.** Genetic association mapping and genome organization of maize. (2006) *Curr Opin Biotechnol* 17:155-160.
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES.** A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. (2006) *Nat Genet* 38:203-208.
- Yue B, Radi SA, Vick BA, Cai X, Tang S, Knapp SJ, Gulya TJ, Miller JF, Hu J.** Identifying quantitative trait loci for resistance to *Sclerotinia* head rot in two USDA sunflower germplasms. (2008) *Phytopathol* 98:926-931.
- Zhao J, Wang J, An L, Doerge RW, Chen ZJ, Grau CR, Meng J, Osborn TC.** Analysis of gene expression profiles in response to *Sclerotinia sclerotiorum* in *Brassica napus*. (2007a) *Planta* 227:13-24.
- Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M.** An *Arabidopsis* example of association mapping in structured samples. (2007b) *PLoS Genet* 3:e4.
- Zhu C, Gore M, Buckler ES, Yu J.** Status and Prospects of Association Mapping in Plants. (2008) *The Plant Genome* 1:5–20.
- Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB.** Single-nucleotide polymorphisms in soybean. (2003) *Genetics* 163:1123-1134.
- Zimmermann G, Baumlein H, Mock HP, Himmelbach A, Schweizer P.** The multigene family encoding germin-like proteins of barley. Regulation and function in Basal host resistance. (2006) *Plant Physiol* 142:181-192.

APÉNDICE

Tabla A1.1. Funciones descritas para genes con similitud a las regiones genómicas amplificadas para la búsqueda de SNPs e indels en girasol.

Estrategia de selección	ID	Similitud (búsqueda por BLASTx)	Descripción	Cita
Análisis de las colecciones de ADNc del GG-IB INTA Castelar	GO	Glicolato oxidasa	Transporta de electrones en la maquinaria de las especies reactivas del oxígeno (ROS)	(Fernandez y col. 2003)
	PGIP1	Proteína inhibidora de poligalacturonasa	Defensa de la planta contra diversos patógenos que usan la poligalacturonasa para penetrar la pared celular	(Bishop 2005)
	PGIP3	Proteína inhibidora de poligalacturonasa		(Bishop 2005)
	LZP	Probable proteína de cierre de leucinas	Factor de transcripción involucrado en el desarrollo vegetal, la fotomorfogénesis y las respuestas a estrés	(Mallappa y col. 2006)
	GLP	Proteína tipo germina	Proteína apoplástica y glicosilada involucrada en defensa	(Zimmermann y col. 2006)
Búsqueda bibliográfica	MADSB-TF3	Factor de Transcripción MADS-box	Factor de transcripción que regula varios aspectos del desarrollo de la planta	(Dezar y col. 2003)
	AALP	Proteasa tipo Aleuraina	Enzima involucrada en la degradación de macromoléculas, su expresión aumenta durante el proceso de envejecimiento y de senescencia inducida por cosecha	(Breeze y col. 2008, Eason y col. 2005)
	LIM	Proteína con dominio LIM, PLIM1b	Factor de transcripción involucrado en la construcción del citoesqueleto y la traducción de señales	(Arnaud y col. 2007, Gepstein y col. 2003)
	1-ACCO	Oxidasa del ácido 1-aminociclopropano-1-carboxílico	Expresión relacionada con células en alta actividad de división, expresión y actividad en concordancia con el aumento de etileno en plantas heridas o tratadas con iones de plata	(Liu y col. 1997)
	ZFP	Protina dedos de zinc	Factor de transcripción involucrado en la regulación de la senescencia y la protección frente a estrés hídrico	(Hu y col. 2008)
	CAM	Calmodulina	Rol central en la señalización mediada por calcio	(Liu y Burke 2006)
	CHS	Chalcona sintasa	Rol central en la biosíntesis de fenilpropanoides y en las respuestas a estrés abiótico	(Coberly y Rausher 2003, Liu y Burke 2006)
	GAPDH	Glicerol 3-fosfato deshidrogenasa	Proteína tetramérica de unión a NAD1, involucrada en la glicólisis y la gluconeogénesis	(Liu y Burke 2006)

Estrategia de selección	ID	Similitud (búsqueda por BLASTx)	Descripción	Cita
Búsqueda bibliográfica (cont.)	GIA	Proteína insensible a ácido giberélico	Modulador de la respuesta a giberelinas	(Liu y Burke 2006)
	GPX	Probable glutation peroxidasa	Enzimas antioxidantes indicadas como importantes factores en los mecanismos de protección contra el daño oxidativo	(Liu y Burke 2006)
	GST	Glutation S-transferasa		
	PGIC	Fosfoglucosa isomerasa citosólica	Cataliza de forma reversible la isomerización de la 6-fosfoglucosa, reacción esencial que precede a la biosíntesis de sacarosa	(Liu y Burke 2006)
	SCR1 SCR2	Factor de transcripción "Scarecrow" tipo 1 y tipo 2	Genes tipo "SCARECROW" conocidos por su acción en la división asimétrica de las células vegetales	(Liu y Burke 2006)
Análisis <i>in silico</i> con el programa AutoSNP	RL41	Proteína ribosomal L41 de la subunidad 60S	Proteína importante en el aparato de traducción relacionada con apoptosis y ciclo celular	(Kim y col. 2005)
	ANT	Precursor mitocondrial del translocador de adenin-nucleótido	Transportador ubicado en la membrana interna mitocondrial, rol en la integración del estrés celular y la muerte celular programada	(Lam y col. 2001)
	RS16	Proteína ribosomal S16 de la subunidad 40S	Componente ribosomal retenido durante los procesos de deshidratación en plantas tolerantes a estrés hídrico	(Wood y col. 2000)
	NsLTP	Precursor de la proteína de transferencia de lípidos no-específica	Participa en la formación de cutina, embriogénesis, defensa contra patógenos, simbiosis y adaptación a distintas condiciones ambientales	(Kader 1996)
	SEM	Probable subunidad sem 1-2 del complejo proteasoma	Complejo involucrado en el recambio de proteínas ayudando a remover proteínas mal plegadas, degradadas o dañadas debido a estrés ambiental o radicales libres	(Smalle y Vierstra 2004)
	SAMC	S-adenosil-metionina decarboxilasa	Enzima clave e la biosíntesis de poliaminas (PAs) La síntesis de PA se induce por alta presión osmótica, baja temperatura, bajo pH y estrés oxidativo Se supone que las PAs cumplen un rol en las interacciones planta-microorganismo	(Wi y col. 2006)
	GCvT	Proteína del sistema T de corte de Glicina	El complejo cataliza la decarboxilación oxidativa de glicina en mitocondrias de plantas	(Lee y col. 2004)

Estrategia de selección	ID	Similitud (búsqueda por BLASTx)	Descripción	Cita
Análisis <i>in silico</i> con el programa AutoSNP (cont.)	SBP	Seudo-heptulosa-1,7-bisfosfatasa chloroplastídica	Enzima del ciclo de Calvin: punto de bifurcación entre la regeneración de ribulosa 1,5 fosfato y la biosíntesis de almidón Su sobre-expresión incrementa la fijación de carbono fotosintético y la biomasa en plantas	(Fernandez y col. 2008)
	LHCP	Precursor de la proteína de unión a clorofila a/b receptora de luz	Genes que codifican para componentes involucra dos en la fotosíntesis, mostraron patrones diferenciales de expresión en condiciones de estrés salino y por frío en girasol	(Fernandez y col. 2008)
	CPSI	Subunidad V de Fotosistema I, precursor		
	PSI-III-CAB	Proteína tipo III de unión a clorofila a/b del Fotosistema I		
	CAB	Proteína de unión a clorofila a/b		

Tabla A3.1. Iniciadores utilizados para la amplificación y secuenciación de los genes candidatos para resistencia a la Podredumbre Húmeda del Capítulo (PHC).

Gen candidato^a	Iniciadores diseñados para amplificar la región genómica en girasol	Iniciadores diseñados para secuenciar 10 líneas de girasol
AT1G34580, MT Transportador de monosacáridos	F: 5' CGTTGTTGTCATCGTTGGTC 3' R: 5'AGCCTCCATCCCAAATTACC 3'	F: 5' CGTTGTTGTCATCGTTGGTC 3' R: 5'AGCCTCCATCCCAAATTACC 3'
AT5G51550, PRP Proteína con respuesta a fosfatasa	F: 5' CGVATYATYCGBGAVTTYATYAACTC 3' R: 5' GGRTCHTSRCCDCRTACCARSC 3'	F1: 5' CGAAGCCATCTCAGCAATCT 3' R1:5' AGCTGTACACCGACCAAACC 3" F2: 5' AGGCGTTCACTAGCGGATTA 3' R2: 5' GACCGTTTGAACATCTCTC 3'
AT1G56360, CPE Calcineurin fosfoesterasa	F: 5' CTCCAACATCACCCCTGWSNCAYTAYGA 3' R: 5' GGAGGCCCGCTTGWNSWRTACCA 3'	-
AT3G48310, CytP450 Citocromo P450 71 ²²	FA: 5' TAGGGTGACGTTTGGGAAGA 3' RA: 5' CGCCCTTTGAAATCAATGTT 3' FB: 5' GTTTTAGAGGGTGTCTAG 3' RB: 5' CCACTCGTCTCGCTCATATCC 3' FC: 5' GTTCTTGGATGGTGTGGTGG 3' RC: 5' GGACACCCTCTTCGTCCAGAAC 3'	FA: 5' ATGGGTGGATGAAGACGAAG 3' RA:5' GCTCGGTGGGCTTAACATAG 3' FB: 5' TCTCCATCACGACCAAATCA 3' RB: 5' ACCGGTTGATGTAAATGC 3' FC: 5' AGTGGGATTCAATGGCAAAC 3' RC:5' CGCCCTTTGAAATCAATGTT 3'
AT3G27660, GRP Proteína rica en glicinas	F: 5' CCCWBCAKBTHCAAGTHMAMRMVYT 3' R: 5' CATGAMCYKSACCCTCYGGTCC 3'	-
AT2G16650, EP Proteína que se expresa	F: 5' CATCGGCCTGTACCARCARAAYTTYG 3' R: 5' TCGCAGGCCACGGGNAMRTGCCA 3'	-
AT5G42170, EXLP Lipasa extracelular II	F: 5' TCTGGWGTAGCAAMCTTCAAGA 3' R: 5' ACTCATAACAACCACCCTTCGT 3'	-
AT5G05940, RGEF Proteína que se expresa PFAM PF03759	FA1: 5' ATATGTCYGGTTGTGGGAAT 3' RA1: 5' TACGAAGAGCCGGRAGRRTTAC 3' FA2: 5' TAAYCTYCCGGCTCTTCGTA 3' RA2: 5' CRGGRACATCMATATCGGCTA 3' FB: 5' CTGCATTTGGTCAGTTGTGG 3' RB: 5' AGGCTGGACTATCGGGTTCT 3'	FA1: 5' CAGCCTTGGCTATCTCGAAC 3' RA1: 5' AAGAGCCGGGAGATTACCAT 3' FA2: 5' AATTGATCGGGGAACATCAA 3' RA2: 5'CCGCCACGTGTAATCCTAGT 3' FB: 5' CTGCATTTGGTCAGTTGTGG 3' RB: 5' AGGCTGGACTATCGGGTTCT 3'
AT1G48040, PP2C Fosfatasa 2C	F: 5' GAYGGYCATGGMGGSTCTGAAGC 3' R: 5' CAAATCCCRTCACAYCCRATDATC 3'	F: 5' AAGCGGCTGCCTATGTTAAG 3' R: 5' TGCTGAATTTCTGGCTCTGA 3'
AT3G25250, PK Proteína Kinasa	F: 5' GAGATTTAAGCCAGAMAATG 3' R: 5' CCTYTRAACGGCGTCKYYCCRTAC 3'	-

Gen candidato ^a	Iniciadores diseñados para amplificar la región genómica en girasol	Iniciadores diseñados para secuenciar 10 líneas de girasol
AT1G04450, RhoBP Proteína con dominio <i>p21-rho-binding</i>	FA: 5' GGCCTACGTTACATCTCTGAA 3' RA: 5' GAATGAACCGAAATGCTACAA 3' FB: 5' TTGAGGGATTCTAATTGTTATAGTTGA 3' RB: 5' TTCGGGTGTTTCGTCCTTTT 3'	FA: 5' GGCCTACGTTACATCTCTGAA 3' RA: 5' GAATGAACCGAAATGCTACAA 3' FB1: 5' TTGAGGGATTCTAATTGTTATAGTTGA3' RB1: 5' TTCGGGTGTTTCGTCCTTTT 3'
AT2G03250, EXS-FP Transportador de la familia EXS/ERD1/XPR1/SYG1	F: 5' GGCGACTGTTTGAGGAAAAA 3' R: 5' TTTCCCGACGTTGTTCAAAT 3'	-
AT1G03687, HP Proteína con dominio DTW	F: 5' CAAAGTGATGCCAGATGAA 3' R: 5' CGTATCCGATGCCAAAGATT 3'	F: 5' CAGAGCAAAGTACAAAGCGTTAGA 3' R: 5' CGTATCCGATGCCAAAGATT 3'
AT1G28090, ANT Adenililtransferasa polinucleótido	F: 5' TCGTCTCAATTCTCTCTCTTCC 3' R: 5' CCGCTTCCCAACAATCTCACACC 3'	-
AT4G09180, bHLHP Proteína de la familia <i>Basic-helix-loop-helic</i>	FA: 5' ACACACGGTGGCACTAGMGG 3' RA: 5' AGGATGAGTCGCACACCCG 3' FB: 5' CTGCTATCCAGGCTCATTCA 3' RB: 5' AGAATGGCAGGGCGACCAAG 3'	FB: 5' CTGCTATCCAGGCTCATTCA 3' RB: 5' AGAATGGCAGGGCGACCAAG 3'
AT1G13200, FboxP Proteína de la familia <i>F-box</i>	FA: 5' ACCGAGATTCTTCTCCGCCTC 3' RA: 5' CTTCCATTCTTTGTGTGTCAGAC 3' FB: 5' TAACCGAGATCCTTCTCCGCCT 3' RB: 5' CTGTTTGATGATATGCCAGACCC 3'	FA: 5' CGCTGTCTTGATTCTACTT 3' RA: 5' ATACGAATCCACATCAGGC 3' FB: 5' AGCGACAGACATTTACCCGA 3' RB: 5' TGTGGATTGGTGAGATGAGTGCC 3'
AT1G14687, ZFHDP Proteína de la familia zinc finger homeobox	F: 5' GCCACCATCGGCAGMCACGCCT 3' R: 5' ATCATCRCAYCKYTGCATCT 3'	F: 5' TCATGCCCTCACTAACATGC 3' R: 5' TTTGTCCGGAATCTTTTTCG 3'
AT5G13600, NPH3-PhotoR Proteína de la familia NPH3	F: 5' GCATCCAACCCGAAAAGTTA 3' R: 5' GCACAAGCATCAATGGAAAGT 3'	-
AT1G15520, ABC-TP Transportador tipo ABC	F: 5' CCRAARAAVCAAGAAACDTT 3' R: 5' CATCCAAGTTGCBGGATTRTAACCATCT 3'	-
Ha26SP Subunidad 4 del proteasoma 26S	F: 5' TTCTTGCGTGTGTTGGAAG 3' R: 5' GGCAACTAACGGGTCTGAATA 3'	-
HaDT-HP Biosíntesis de tiamina y la tolerancia al daño mitocondrial	F: 5' CCACCACCTTCACCTCATCT 3' R: 5' ACGACATTCGTGTTGGTTCA 3'	-

Gen candidato ^a	Iniciadores diseñados para amplificar la región genómica en girasol	Iniciadores diseñados para secuenciar 10 líneas de girasol
HaRNAHel Helicasa de RNA	F: 5' ACAAGCCGAAGCAGTCAAGT 3' R: 5' TCTGAAGCGTGTTCATC 3'	-
HaDRP Proteína RD22 con respuesta a deshidratación	F: 5' CAAAGGAAGCAACAACACCA 3' R: 5' TCGCTGATTCGATACCCTTT 3'	F: 5' CAAAGGAAGCAACAACACCA 3' R: 5' TCGCTGATTCGATACCCTTT 3'
HaThE Tioesterasa	F: 5' AAGTACGAAGGCGAGGGTTT 3' R: 5' TGAGCTTGACCCCTCTACCA 3'	-
Ha_HP Proteína hipotética	F: 5' ATACCCGGCCTTTTCATTTT 3' R: 5' CGGAAACAACCTTCGTCCAAT 3'	-
HaKIV Kitinasa clase IV	F: 5' TCGTAACAACACACGCACTTT 3' R: 5' GGGTTGGTTCCGAAGAAGAT 3'	-
HaCP Cisteína proteinasa	F: 5' TGCTATTGATGCTGGCAGTT 3' R: 5' AATCATGTTTCACATACCAAATCTT 3'	F: 5' TGCTATTGATGCTGGCAGTT 3' R: 5' AATCATGTTTCACATACCAAATCTT 3'
HaGDP-DI Inhibidor de la disociación de GDP	F: 5' CAGAAACTGATCAACCCGAAA 3' R: 5' TGCATGCATCTTGGAAAATAG 3'	F: 5' CAGAAACTGATCAACCCGAAA 3' R: 5' TGCATGCATCTTGGAAAATAG 3'
HaWP Proteína de pared	F: 5' CCAAACCCGATGATGATAAA 3' R: 5' AACAAACAAAACAAGCCACATT 3'	F: 5' CCAAACCCGATGATGATAAA 3' R: 5' AACAAACAAAACAAGCCACATT 3'
HaTRP Proteína relacionada con la formación de tumores	F: 5' TTCTTTAGGCCAACCCCTCAC 3' R: 5' CCCTTAATCATAATTCACGAATGTC 3'	F: 5' TTCTTTAGGCCAACCCCTCAC 3' R: 5' CCCTTAATCATAATTCACGAATGTC 3'
HaC2SP Subunidad 2 del complejo peptidasa señal	F: 5' CGGAAAGAATCGGAGATGAC 3' R: 5' AAACCACAACCATTCTTTTCAA 3'	-
HaTF Transferasa	F: 5' ACCCAGGTTTCATCAAGTTGC 3' R: 5' CAGCTTCTGACGCCATGTAA 3'	-
HaPK Proteína quinasa	F: 5' CGAAAAGCCAAAGTCTGAGG 3' R: 5' TTCACATCGCGGTGTACAAT 3'	-
HaHTFP Transferasa hipotética	F: 5' AAATGGCGAATGAGGTGAAG 3' R: 5' TGCTGCTTATTATCGCATCG 3'	-
HaLTP1 Proteína de transferencia de lípidos, isoforma I	F: 5' AGTTGTGAAGATGGCCTTGG 3' R: 5' AACACGCGATAACAAACACG 3'	-

Gen candidato ^a	Iniciadores diseñados para amplificar la región genómica en girasol	Iniciadores diseñados para secuenciar 10 líneas de girasol
HaPAL Fenilalanina amonio-liasa	F: 5' TGTGGTCTTCAAATTCATTAATAACC 3' R: 5' GGCCATTCCTAACAGGATCA 3'	F: 5' TGTGGTCTTCAAATTCATTAATAACC 3' R: 5' GGCCATTCCTAACAGGATCA 3'
HaWRKY5 Factor de transcripción		F2: 5' ACACCTCTCAACTTGCACCAAA 3' R2: 5' GCCCATGGATCTGAAGACAAA 3'
HaWRKY7 Factor de transcripción	F1: 5' GCTTCATCAGGTGGCAGTTT 3' R1: 5' TGTTGGACAATTCGGGTATG 3' F2: 5' CACCAACTGCTTTGAACCAA 3' R2: 5' CCAAATGATGCCGAAGAATC 3' F3: 5' CAAGCCGCTAGAGATGGAAC 3' R3: 5' TCAACCCTGTGGTGTTTTGA 3' F4: 5' TGATTCTTCGGCATCATTTG 3' R4: 5' GGCCAAAACCTGATTCATGGT 3' ^c	- - F3: 5' CAAGCCGCTAGAGATGGAAC 3' R3: 5' TCAACCCTGTGGTGTTTTGA 3' F4: 5' TGATTCTTCGGCATCATTTG 3' R4: 5' GGCCAAAACCTGATTCATGGT 3'
HaGLP1 Oxalato oxidasa	-	F: 5' CATGGCGACACCTTTTACCT 3' R: 5' GGGACAACGTCTTGTGGTTG 3'
HaGLP2 Proteína tipo germina (AJ540203)	-	F: 5' AAACAAGTTTTTCGCCTCTAGC 3' R: 5' AAACAAATGTTACAAATACAAGC 3'
HaGLP3 Proteína tipo germina (TC17527)	-	F: 5' TCACATTCTTTCATCCTATGCTC 3' R: 5' CCAGTTCCACCAAGAACAGC 3'
HaGLP4 Proteína tipo germina (TC18217)	-	1F: 5' TGCTTGTAACCTCCTCCTCCT 3' 1R: 5' TGTGGGAAAAGCATGATGTC 3' 2F: 5' GCTCCGAACCTTTGGTGGT 3' 2R: 5' AGCTAAGCTTCAATCCAGAAACA 3'
HaGLP5 Proteína tipo germina (TC17648)	-	1F: 5' TCAAGTGCTATGGCTACTAATGATTT 3' 1R: 5' TCGATGTATAAATGTATGGTAAAAGAA 3' 2F: 5' TTTTCTCCCAATTTTAGTTCCAA 3' 2R: 5' CACCTTTCTTGATGGTCTTGG 3' 3F: 5' GCGAATTGGATGTTGGTTTT 3' 3R: 5' TTTTCTCCCAATTTTAGTTCCAA 3'

^a Número de acceso del *locus* de *A. thaliana* en el NCBI o TAIR (p. ej. AT1G34580) y acrónimo de identificación. Entre paréntesis se muestran los *Tentative Consensus* (TC) del índice de genes de girasol del Dana Farber Cancer Institute (Ha-DFCI) o los números de acceso del NCBI para las proteínas tipo germina (HaGLPs).

Tabla A3.2. Identificación utilizada (ID) junto con el número de acceso de GenBank para reconocer el género y la especie de la secuencia utilizada en la construcción de los árboles de Máxima Parsimonia y Máxima Verosimilitud.

Género-especie	ID	Género-especie	ID
<i>Actinidia chinensis</i>	Ac	<i>Cynodon dactylon</i>	Cd
<i>Actinidia deliciosa</i>	Ad	<i>Dunaliella salina</i>	Ds
<i>Actinidia eriantha</i>	An	<i>Elaeis guineensis</i>	EI
<i>Actinidia hemsleyana</i>	Ah	<i>Eragrostis curvula</i>	Ec
<i>Adonis aestivalis</i>	Ae	<i>Eschscholzia californica</i>	Es
<i>Agrostis alba</i>	Al	<i>Eucalyptus globulus</i>	Eg
<i>Agrostis capillaris</i>	Ag	<i>Euphorbia esula</i>	Ee
<i>Agrostis stolonifera</i>	As	<i>Euphorbia tirucalli</i>	Et
<i>Allium cepa</i>	Ap	<i>Festuca arundinacea</i>	Fa
<i>Amborella trichopoda</i>	Ab	<i>Fragaria vesca</i>	Fv
<i>Antirrhinum majus</i>	Am	<i>Ginkgo biloba</i>	Gi
<i>Arabidopsis thaliana</i>	At	<i>Glycine max</i>	Gm
<i>Arachis hypogaea</i>	Ai	<i>Gnetum gnemon</i>	Gg
<i>Aristolochia fimbriata</i>	Af	<i>Gossypium arboreum</i>	Ga
<i>Artemisia annua</i>	Aa	<i>Gossypium barbadense</i>	Gb
<i>Barnadesia spinosa</i>	Be	<i>Gossypium hirsutum</i>	Gh
<i>Beta vulgaris</i>	Bv	<i>Gossypium raimondii</i>	Gr
<i>Brachypodium distachyon</i>	Bd	<i>Guizotia abyssinica</i>	Gy
<i>Brassica napus</i>	Bn	<i>Helianthus annuus</i>	Ha
<i>Brassica rapa</i>	Br	<i>Helianthus anomalous</i>	Ho
<i>Cannabis sativa</i>	Cv	<i>Helianthus argophyllus</i>	Hg
<i>Capsicum annuum</i>	Ca	<i>Helianthus atrorubens</i>	HZ
<i>Carica papaya</i>	Cp	<i>Helianthus bolanderi</i>	HB
<i>Carthamus tinctorius</i>	Ct	<i>Helianthus californicus</i>	HC
<i>Catharanthus roseus</i>	Cq	<i>Helianthus carnosus</i>	HN
<i>Cenchrus ciliaris</i>	Cx	<i>Helianthus ciliaris</i>	Hc
<i>Centaurea maculosa</i>	Cm	<i>Helianthus cusickii</i>	Hk
<i>Centaurea solstitialis</i>	Co	<i>Helianthus debilis</i>	Hd
<i>Ceratopteris richardii</i>	Ch	<i>Helianthus decapetalus</i>	HE
<i>Chamaecyparis formosensis</i>	Cz	<i>Helianthus deserticola</i>	HS
<i>Chlamydomonas reinhardtii</i>	CH	<i>Helianthus divaricatus</i>	HV
<i>Cichorium endivia</i>	Ce	<i>Helianthus eggertii</i>	HK
<i>Cichorium intybus</i>	Ci	<i>Helianthus exilis</i>	He
<i>Citrullus lanatus</i>	Ck	<i>Helianthus floridanus</i>	FI
<i>Citrus aurantiifolia</i>	cA	<i>Helianthus giganteus</i>	Gu
<i>Citrus aurantium</i>	CA	<i>Helianthus gracilentus</i>	Gn
<i>Citrus clementina</i>	Cc	<i>Helianthus grandiflorus</i>	Gd
<i>Citrus latifolia</i>	Cf	<i>Helianthus grosseserratus</i>	Gr
<i>Citrus reshni</i>	CE	<i>Helianthus heterophyllus</i>	He
<i>Citrus reticulata</i>	Cr	<i>Helianthus hirsutus</i>	HZ
<i>Citrus sinensis</i>	Cs	<i>Helianthus laciniatus</i>	Hq
<i>Citrus uinshu</i>	CU	<i>Helianthus laevigatus</i>	HQ
<i>Coffea canephora</i>	Cn	<i>Helianthus longifolius</i>	HW
<i>Cryptomeria japonica</i>	Cj	<i>Helianthus maximiliani</i>	HX
<i>Cucumis melo</i>	Cl	<i>Helianthus microcephalus</i>	HM
<i>Curcuma longa</i>	Cg	<i>Helianthus mollis</i>	Mo
<i>Cyamopsis tetragonoloba</i>	Cb	<i>Helianthus neglectus</i>	Ne
<i>Cycas rumphii</i>	Cu	<i>Helianthus niveus</i>	Hi
<i>Cynara scolymus</i>	Cy	<i>Helianthus occidentalis</i>	HO

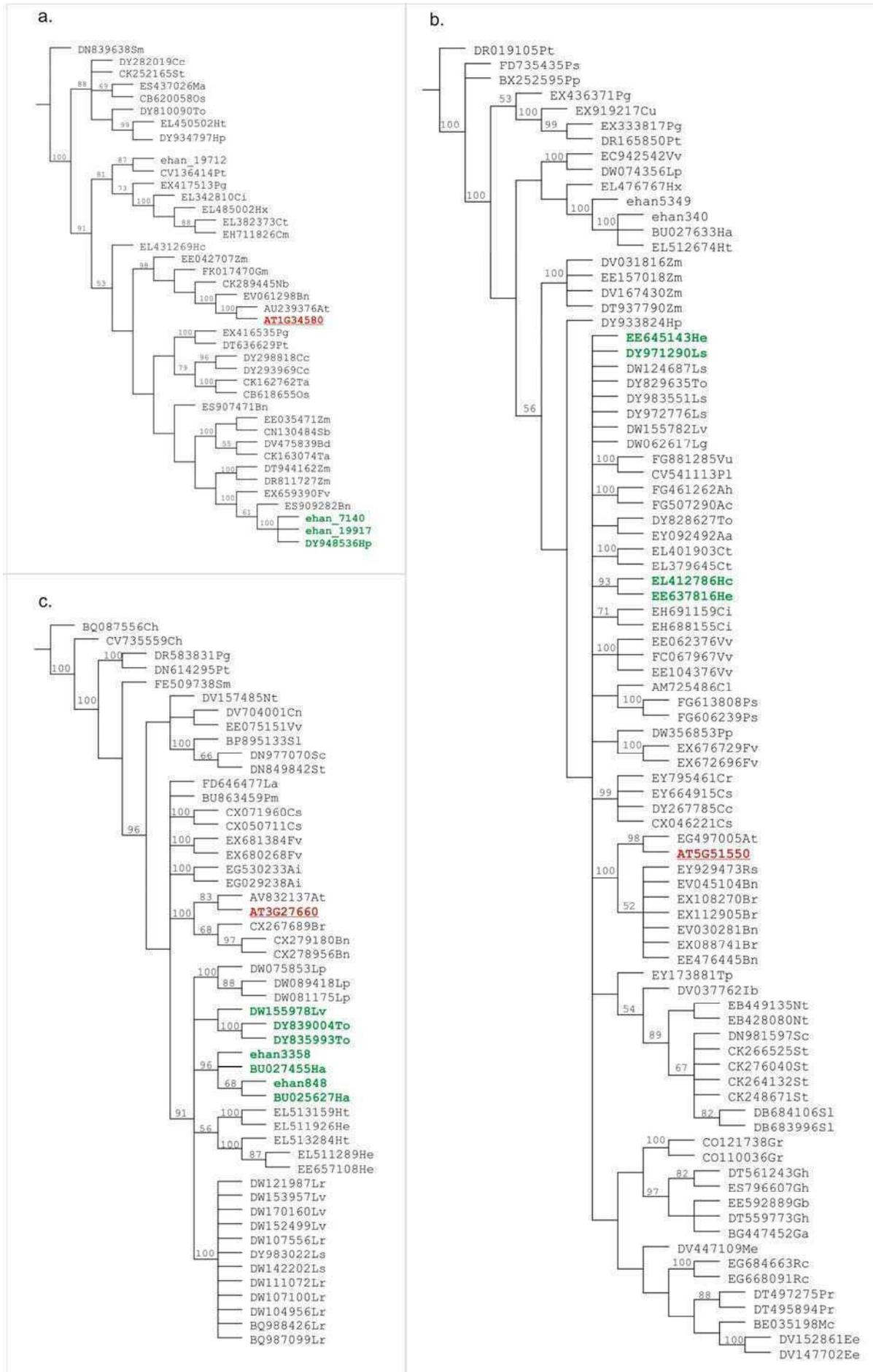
Género-especie	ID
<i>Helianthus paradoxus</i>	Hx
<i>Helianthus petiolaris</i>	Hp
<i>Helianthus porteri</i>	HP
<i>Helianthus praecox</i>	Hr
<i>Helianthus pumilus</i>	Hu
<i>Helianthus radula</i>	HD
<i>Helianthus resinosus</i>	Hs
<i>Helianthus salicifolius</i>	Hf
<i>Helianthus schweinitzii</i>	Hw
<i>Helianthus silphoides</i>	HU
<i>Helianthus simulans</i>	Hm
<i>Helianthus smithii</i>	Hh
<i>Helianthus strumosus</i>	HT
<i>Helianthus tuberosus</i>	Ht
<i>Helianthus verticillatus</i>	HL
<i>Hevea brasiliensis</i>	Hb
<i>Hordeum vulgare</i>	Hv
<i>Humulus lupulus</i>	HI
<i>Ipomea batata</i>	Ib
<i>Lactuca aculeata</i>	Lc
<i>Lactuca altaica</i>	LI
<i>Lactuca canadensis</i>	Ln
<i>Lactuca dissecta</i>	Ld
<i>Lactuca dregeana</i>	Le
<i>Lactuca indica</i>	Li
<i>Lactuca lessertiana</i>	LL
<i>Lactuca perennis</i>	Lp
<i>Lactuca saligna</i>	Lg
<i>Lactuca sativa</i>	Ls
<i>Lactuca serviola</i>	Lr
<i>Lactuca sibirica</i>	Lb
<i>Lactuca tatarica</i>	LT
<i>Lactuca tenerrima</i>	Lm
<i>Lactuca viminea</i>	LV
<i>Lactuca virosa</i>	Lv
<i>Lathyrus odoratus</i>	Lo
<i>Lemna gibba</i>	LB
<i>Limnanthes alba</i>	La
<i>Liriodendron tulipifera</i>	Lt
<i>Lotus japonicus</i>	Lj
<i>Lupinus albus</i>	LA
<i>Malus sieboldii</i>	Mi
<i>Manihot esculenta</i>	Me
<i>Medicago sativa</i>	Ms
<i>Medicago truncatula</i>	Mt
<i>Mesembryanthemum crystallinum</i>	Mc
<i>Mimulus guttatus</i>	Mu
<i>Musa acuminata</i>	Ma
<i>Nicotiana benthamiana</i>	Nb
<i>Nicotiana tabacum</i>	Nt
<i>Nuphar advena</i>	Na
<i>Ocimum basilicum</i>	Ob
<i>Oryza sativa</i>	Os

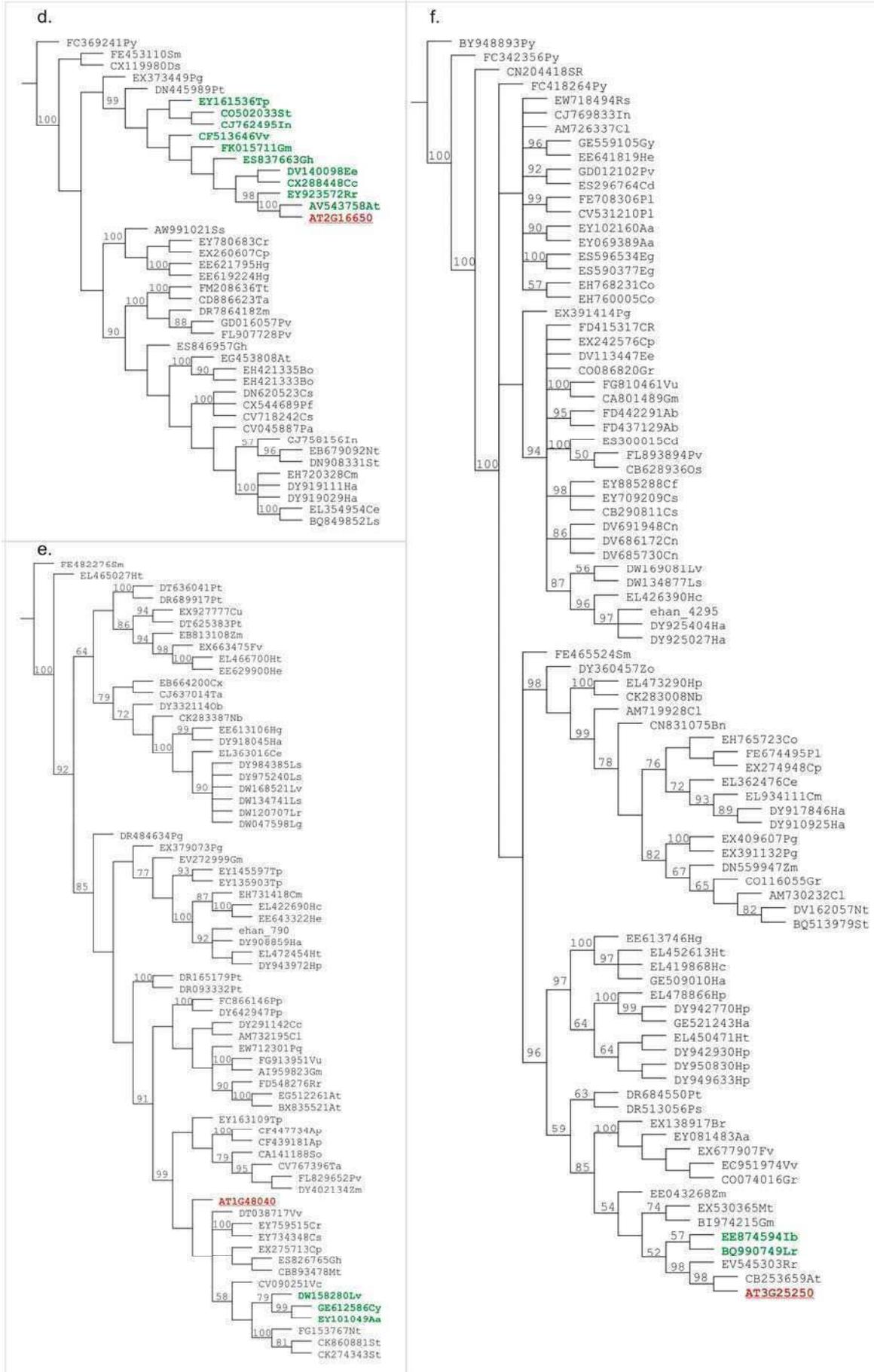
Género-especie	ID
<i>Panax ginseng</i>	Px
<i>Panax quinquefolius</i>	Pq
<i>Panicum virgatum</i>	Pv
<i>Papaver somniferum</i>	Pn
<i>Phaseolus vulgaris</i>	PI
<i>Physcomitrella patens</i>	Py
<i>Picea glauca</i>	Pg
<i>Picea sitchensis</i>	Ps
<i>Pinus taeda</i>	Pt
<i>Poncirus trifoliata</i>	Pf
<i>Populus alba</i>	Pa
<i>Populus nigra</i>	Po
<i>Populus tomentosa</i>	Pm
<i>Populus tremula</i>	Pe
<i>Populus tremuloides</i>	Pi
<i>Populus trichocarpa</i>	Pr
<i>Prunus armeniaca</i>	PA
<i>Prunus persica</i>	Pp
<i>Pseudoroegneria spicata</i>	Pu
<i>Raphanus sativus</i>	Rs
<i>Ricinus communis</i>	Rc
<i>Saccharum officinarum</i>	So
<i>Salvia fruticosa</i>	Sf
<i>Saruma henryi</i>	Sh
<i>Selaginella moellendorffii</i>	Sm
<i>Senecio cambrensis</i>	SB
<i>Solanum chacoense</i>	Sc
<i>Solanum lycopersicum</i>	SI
<i>Solanum pennelli</i>	Sp
<i>Solanum purheja</i>	Sj
<i>Solanum tuberosum</i>	St
<i>Sorghum bicolor</i>	Sb
<i>Stevia rebaudiana</i>	Sr
<i>Suaeda salsa</i>	Ss
<i>Syntrichia ruralis</i>	SR
<i>Taraxacum officinale</i>	To
<i>Thellungiella halophila</i>	Th
<i>Theobroma cacao</i>	Tc
<i>Trifolium pratense</i>	Tr
<i>Triphysaria pusilla</i>	Tp
<i>Triphysaria versicolor</i>	Tv
<i>Triticum aestivum</i>	Ta
<i>Triticum monococcum</i>	Tm
<i>Triticum turgidum</i>	Tt
<i>Tropaeolum majus</i>	Tj
<i>Vaccinium corymbosum</i>	Vc
<i>Vigna unguiculata</i>	Vu
<i>Vitis shuttleworthii</i>	Vs
<i>Vitis vinifera</i>	Vv
<i>Zantedeschia aethiopica</i>	Za
<i>Zea mays</i>	Zm
<i>Zingiber officinale</i>	Zo
<i>Zinnia violacea</i>	Zv

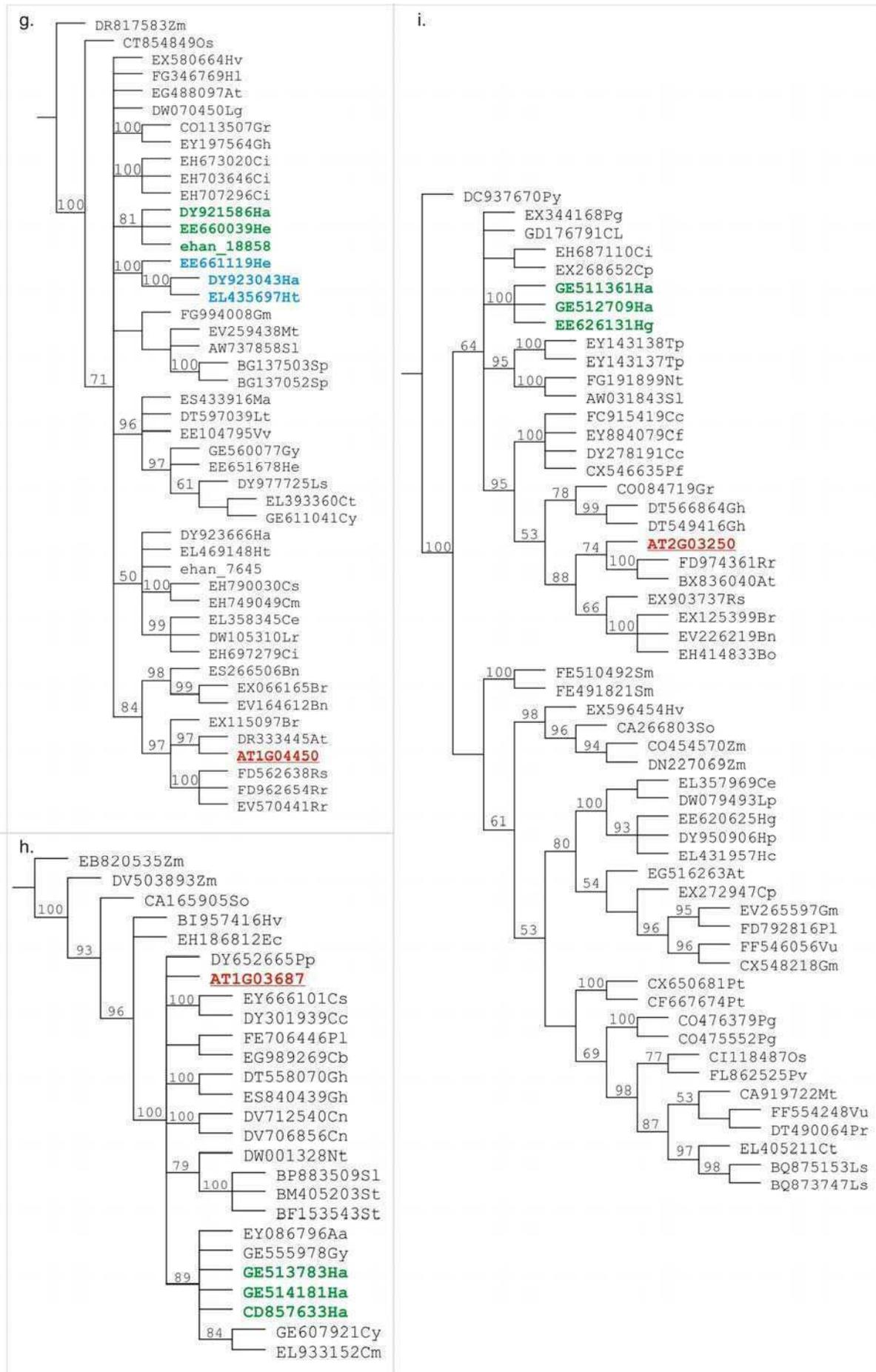
Figura A3.1. Árboles de Máxima Parsimonia (MP) obtenidos para encontrar los genes de girasol ortólogos a los genes de *Arabidopsis* seleccionados a partir de los datos de expresión del microarreglo realizado por Zhao y col. (2007a).

Se indican las secuencias de *Arabidopsis* (rojo) y las secuencias ortólogas o parálogas, seleccionadas para el diseño de iniciadores (verde, celeste y azul). Valores de soporte de las ramas mayores a 50% obtenidos usando "Jackknife".

- a. AT1G34580: consenso de tres árboles MP, longitud 1712.
- b. AT5G51550: consenso de 624 árboles MP, longitud 1766.
- c. AT3G27660: consenso de 30 árboles MP, longitud 811.
- d. AT2G16650: consenso de dos árboles de MP, longitud 1049.
- e. AT1G48040: consenso de nueve árboles MP, longitud: 1587.
- f. AT3G25250: consenso de 152 árboles MP, longitud 2484.
- g. AT1G04450: consenso de 12 árboles MP, longitud: 1031.
- h. AT2G03250: consenso de cuatro árboles MP, longitud: 1330.
- i. AT1G03687: consenso de 36 árboles MP, longitud 95.
- j. AT1G28090: árbol de MP, longitud: 1076.
- k. AT4G09180: consenso de 42 árboles MP, longitud: 1268.
- l. AT1G13200: consenso de seis árboles MP, longitud: 4062.
- m. AT1G14687: consenso de siete árboles MP, longitud 2267.
- n. AT5G13600: consenso de 51 árboles MP, longitud 2452.
- o. AT1G15520, consenso de 600 árboles MP, longitud 892.







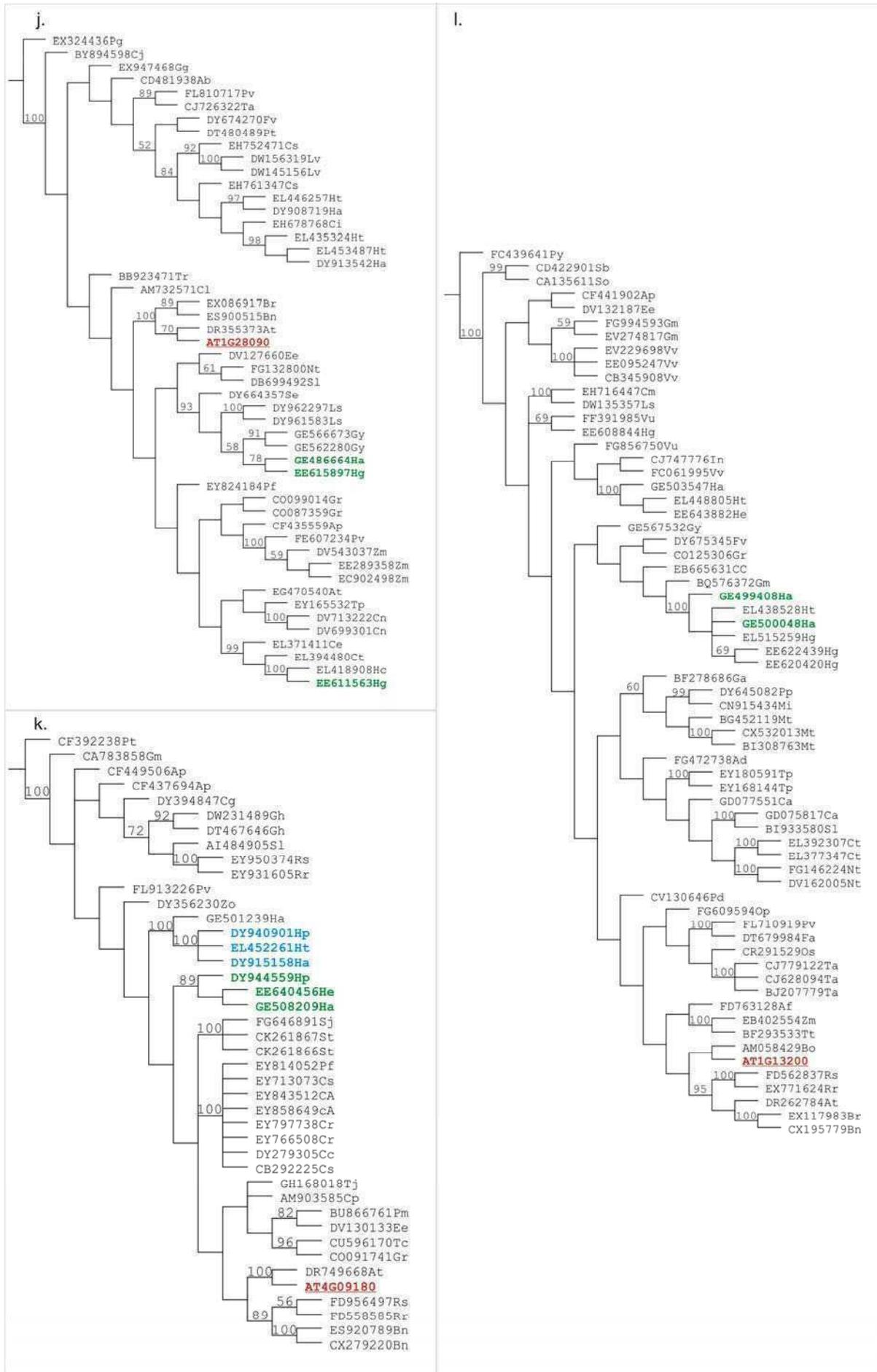


Tabla A4.1. Material utilizado como Población de Mapeo por Asociación (PMA) para la resistencia a la Podredumbre Húmeda del Capítulo (PHC). Nomenclatura de los accesos según el Banco Activo de Girasol de la EEA INTA Manfredi (BAG-IM).

Nº	Acceso	Nº	Acceso	Nº	Acceso
1	2021	51	C202-1	101	C827
2	2067	52	C202-2	102	C849
3	2071	53	C203	103	C851
4	2091	54	C204-1	104	C853
5	2095	55	C204-2	105	C977
6	2101	56	C206-1	106	C998
7	2121	57	C207-F	107	C999
8	2123	58	C208-F	108	F155
9	2125	59	C369B	109	GP365
10	5023	60	C373B	110	HA302
11	5035	61	C374	111	HA441
12	5049	62	C375	112	HA850
13	5075	63	C378-1	113	HA853
14	5285	64	C386M	114	HA89
15	5286	65	C386U	115	PAC2
16	5289	66	C396-2B	116	RHA266
17	5307	67	C397	117	RHA275
18	5313	68	C398	118	RHA276
19	5348	69	C428	119	RHA299
20	5381	70	C430	120	RHA439
21	5403	71	C432-1	121	RHA440
22	5341-1	72	C432-2	122	RHA801
23	5393-E	73	C433	123	RK416
24	5393-N	74	C434	124	u12
25	7_1_1	75	C436	125	u1264
26	7_1_2	76	C437	126	u5045
27	7X1	77	C438	127	u5084
28	C1005-M	78	C454B	128	u703
29	C1059	79	C458B	129	u704
30	C110M	80	C494-1	130	u797
31	C111-3U	81	C610-1	131	v22603
32	C112M	82	C618-1M	132	v2601
33	C113M	83	C618-1U	133	v2602
34	C114M	84	C618-2U	134	v2652
35	C139-1	85	C620-1U		
36	C139-2	86	C620-2U		
37	C149-1	87	C621		
38	C149-2	88	C625-U		
39	C154B	89	C626-M		
40	C155	90	C627-M1		
41	C166-U	91	C628-M1		
42	C173	92	C632-1		
43	C174	93	C633-1U		
44	C177-1	94	C633-2U		
45	C177-2	95	C811		
46	C177M	96	C813		
47	C181	97	C816		
48	C188	98	C817		
49	C191-1	99	C818		
50	C192-1	100	C826		

Tabla A4.2. Iniciadores utilizados para la amplificación por PCR y la genotipificación de la Población de Mapeo por Asociación (PMA).

Gen candidato	Iniciadores diseñados para genotipificar 134 líneas endocriadas de girasol
AT1G34580 (MT) Transportador de monosacáridos	F*:5' GAAAGGACATGCTACTTTATGG 3' R:5' CTTTACTTGAATTAAGTTACT 3'
AT5G51550 (PRP_2) Proteína con respuesta a fosfatasa	F2: 5' GGACGGGAACGTAAAATAATG 3' R2: 5' CCGTCTGTCCGTACAATCG 3'
AT3G48310 (CytP50_C) Citocromo P450 71 ^a 22	FC: 5' AAGTGACTTTAGCAACGTCC 3' RC: 5' GAGTTGGTATGGGTGGATGAA 3'
AT5G05940 (RGEF_A2) Proteína que se expresa PFAM PF03759	FA2*: 5' TCGTAGTGGTTCTAAAATTGG 3' RA2: 5' CGTCAATCATTACCCCAACC 3'
AT1G48040 (PP2C) Fosfatasa 2C (PP2C)	F: 5' ACTGGGACTACGGCATTGAC 3' R: 5' TGCTGAATTTCTGGCTCTGA 3'
AT1G04450 (RhoBP) Proteína con dominio <i>p21-rho-binding</i>	FA*: 5' GCACGAATAGTGACATTGAAAC 3' RA:5' ACATAAAACAGTTTTCGGTCC 3' FB2*: 5' GGCTTGC GTTACATCTCTGA 3' RB2: 5' CCCAACTAGGAGCATTGGAA 3'
AT4G09180 (bHLHP) Proteína de la familia <i>Basic-helix-loop-helic</i>	FB: 5' CTGCTATCCAGGCTCATTCA 3' RB: 5' AGAATGGCAGGGCGACCAAG 3'
AT1G14687 (ZFHDP) Proteína de la familia zinc finger homeobox	F:5' TCATGCCCTCACTAACATGC 3' R: 5' TTTGTCCGGAATCTTTTTTCG 3'
HaPRD Proteína RD22 con respuesta a deshidratación	F: 5' TCTTCAGGTAAGAAATGGCAAA 3' R: 5' GCATACGCGGTCTTTTATTG 3' ^a
HaCP Cisteína proteinasa	F: 5' TTCGAATCCAACCTTCACTGG 3' R: 5' GGATAAGAGGCTTCCATTGC 3' ^a
HaGDP-DI Inhibidor de la disociación de GDP	F: 5' CAGAAACTGATCAACCCGAAA 3' R: 5' TGCATGCATCTTGAAAATAG 3'
HaWP Proteína de pared	F*: 5' CAGGAATCACGGTCCCTAGT 3' R: 5' TGAACATGAGGGATGAGCA 3'
HaTRP Proteína relacionada con la formación de tumores	F*: 5' TCCAACAGTGTGTGACCTTTG 3' R*: 5' CATTAGTTACGTTACAAAGCTAT 3'
HaPAL Fenilalanina amonio-liasa	F: 5' TGTGGTCTTCAAATTCATTAATAACC 3' R: 5' GGCCATTCTAACAGGATCA 3'
HaWRKY5 Factor de transcripción	F*: 5' CCGATCAAAGGCTCAATCTA 3' R: 5' CACATCCGCTAGTTCACACC 3'
HaWRKY7 Factor de transcripción	F4*: 5' CATTGTTGGTCAACCCCTGTG 3' R4: 5' AGGGAAGCATAACCATGACG 3'
HaGLP2 Proteína tipo germina (AJ540203)	F: 5' TTCGGACCCTAGTCTCTCC 3' R: 5' GGGTTTGATCCAAACACAGC 3' ^a
HaGLP3 Proteína tipo germina (TC17527)	F: 5' GACTTTTGTGTTGCGGACCT 3' R: 5' TAGCCTCCACCAATTTTCGAC 3' ^a
HaGLP4 Proteína tipo germina (TC18217)	F*: 5' TGGCTGCAACAACCTTTCCTT 3' R: 5' TTCAATCCAGAAACAAACTTCTAA 3'
HaGLP5 Proteína tipo germina (TC17648)	F: 5' TGGGTTGATAATTGGCATGAT 3' R: 5' TAAACGGCAACATTTTTTCATGT 3' ^a

a No pudieron optimizarse las condiciones de temperatura y gradiente para la genotipificación mediante dHPLC.

Tabla A4.3. Matriz de datos de los ocho SSRs genotipificados para estimar la estructura poblacional y las relaciones de parentesco entre las líneas de la Población de Mapeo por Asociación (PMA).

Línea	HA77_1	HA77_2	HA991_1	HA991_2	HA293_1	HA293_2	HA2063_1	HA2063_2	HA928_1	HA928_2	HA2920_1	HA2920_2	HA4103_1	HA4103_2	HA3239_1	HA3239_2
U5084	247	247	146	146	121	121	171	171	227	227	173	173	141	141	144	144
U12	253	253	146	146	119	119	178	178	217	217	175	175	150	150	137	137
RK416	259	259	137	137	121	121	170	170	217	217	173	173	150	150	144	144
RHA801	259	259	146	146	113	113	171	171	217	221	170	173	141	150	134	134
RHA440	259	259	146	146	121	121	171	180	221	221	170	170	150	141	134	134
RHA439	247	247	146	146	121	121	180	180	221	221	170	170	150	150	144	144
RHA299	259	259	146	146	119	119	171	171	221	209	173	173	141	141	137	144
RHA276	259	259	137	137	119	119	171	171	217	217	173	173	141	150	134	134
RHA275	259	259	137	146	119	119	171	180	217	221	173	173	141	150	137	134
RHA266	259	259	137	137	119	119	171	171	217	217	170	170	150	150	134	134
HA89	259	259	137	157	121	121	170	180	217	225	173	173	141	150	146	146
HA853	247	247	146	146	123	123	178	178	217	217	173	173	141	141	137	137
HA302	259	259	146	157	121	121	170	180	217	217	173	175	150	150	144	144
GP365	259	259	146	157	121	121	180	180	217	217	173	173	150	150	144	144
F155	259	259	146	146	113	113	171	171	221	221	173	173	141	141	137	137
C999	259	259	137	146	119	121	168	180	217	217	173	173	141	150	137	137
C998	259	259	137	146	113	121	171	180	217	217	170	170	141	150	134	144
C977	247	247	146	146	119	119	171	171	227	217	173	173	141	150	137	137
C853	259	259	137	137	113	113	178	178	227	227	173	173	141	141	137	137
C851	253	259	137	146	113	113	180	170	217	217	173	173	150	150	144	144
C849	247	247	146	146	123	123	178	178	221	217	173	173	141	141	137	137
C827	247	247	146	146	113	113	178	178	221	221	170	170	141	141	134	134
C826	247	247	146	146	121	121	180	180	221	221	170	170	150	150	144	144
C818	259	259	137	137	121	121	171	171	221	221	173	161	150	150	134	134
C817	253	253	146	146	113	113	171	171	227	227	173	173	150	150	144	144
C816	259	259	137	137	119	119	171	171	217	217	170	170	141	141	126	134
C813	259	259	146	146	119	119	171	171	217	217	173	173	141	141	134	134
C811	247	259	146	146	113	119	171	180	217	217	173	173	150	141	137	144
C633-2U	247	247	146	146	121	121	170	178	217	217	173	175	150	150	144	144
C633-1U	247	247	146	146	121	121	178	178	217	217	173	175	141	150	137	144
C628-M1	262	259	137	146	113	113	178	180	221	221	173	173	150	150	144	140
C627-M1	247	247	146	146	121	121	180	180	221	221	170	173	141	150	140	134

Línea	HA77_1	HA77_2	HA991_1	HA991_2	HA293_1	HA293_2	HA2063_1	HA2063_2	HA928_1	HA928_2	HA2920_1	HA2920_2	HA4103_1	HA4103_2	HA3239_1	HA3239_2
C626-M	259	259	146	146	113	113	178	178	221	221	173	173	150	150	140	140
C625-U	259	259	151	151	113	113	180	180	217	217	173	173	141	141	137	137
C621	259	259	137	137	119	119	176	176	217	217	173	173	141	150	137	134
C620-2U	259	259	137	146	121	121	171	176	227	217	173	173	150	150	144	144
C620-1U	259	259	137	146	119	113	170	176	217	225	173	173	141	150	144	144
C618-2U	259	247	146	146	113	113	180	180	217	217	173	173	150	150	144	144
C618-1M	247	259	137	137	113	121	180	171	221	221	173	175	150	150	134	144
C610-1	259	259	137	137	119	131	178	178	219	219	187	187	150	150	144	144
C494-1	244	259	146	146	121	121	171	180	221	221	175	175	150	150	137	137
C458B	259	259	137	137	119	119	171	171	217	217	173	173	141	141	134	134
C454B	259	259	137	137	119	119	180	180	217	217	173	173	141	150	134	134
C438	259	259	146	146	121	121	180	180	217	217	173	173	141	141	134	134
C437	259	247	137	146	121	121	176	171	217	221	173	175	141	150	134	144
C434	259	259	146	146	113	113	168	170	217	217	173	173	141	150	134	134
C433	247	247	143	143	119	121	170	170	217	227	170	170	150	150	137	137
C432-2	247	259	146	146	121	121	170	180	227	227	170	173	141	150	137	144
C432-1	247	259	137	137	121	119	170	180	217	224	170	173	141	150	144	134
C430	247	247	143	143	119	121	170	170	227	227	170	170	150	150	137	137
C428	247	259	137	137	121	119	170	180	217	224	175	175	141	141	144	134
C398	259	259	146	146	121	121	180	180	217	217	173	173	141	150	137	137
C397	247	247	146	146	113	119	170	170	217	217	170	173	150	141	137	137
C396-2B	247	247	146	146	113	119	168	168	217	217	173	173	141	141	137	137
C192-1	259	259	146	146	119	119	178	178	217	217	173	173	141	141	137	144
C191-1	259	259	137	146	113	113	178	178	217	217	173	173	141	150	144	144
C166-U	253	259	137	146	119	119	170	170	217	217	173	173	141	141	144	144
C155	259	259	137	137	119	119	171	171	217	217	173	173	141	141	134	134
C149-2	259	259	146	146	113	113	178	178	227	227	175	175	141	141	134	144
C149-1	259	259	137	137	121	121	171	171	217	217	173	173	141	141	137	144
C139-2	247	247	137	146	113	121	170	170	221	217	170	170	141	141	144	144
C139-1	259	259	146	146	121	113	171	180	221	224	170	170	141	150	144	144
C114M	259	259	146	146	121	121	171	180	221	221	173	173	141	141	134	134
C113M	247	247	146	146	121	119	180	180	221	221	173	173	141	141	137	137
C112M	247	247	137	137	113	113	171	171	221	221	173	173	150	150	137	137
C111-3U	247	247	137	137	119	121	168	180	221	221	173	173	150	150	137	137
C110M	259	259	146	146	113	113	178	178	221	221	173	173	150	150	137	137

Línea	HA77_1	HA77_2	HA991_1	HA991_2	HA293_1	HA293_2	HA2063_1	HA2063_2	HA928_1	HA928_2	HA2920_1	HA2920_2	HA4103_1	HA4103_2	HA3239_1	HA3239_2
C1059	253	253	146	146	113	113	171	171	217	217	175	175	150	150	144	144
C1005-M	259	259	146	148	119	119	171	180	221	221	170	173	141	141	144	144
7x1	259	259	146	146	121	121	170	170	217	217	170	170	150	150	134	134
5393-N	259	259	146	146	121	121	180	180	221	221	173	173	150	150	144	144
5393-E	259	259	146	146	121	121	180	180	221	221	173	173	150	150	135	144
5341-1	259	259	146	146	119	119	171	171	221	221	173	173	141	141	144	144
5403	253	253	146	146	115	121	180	180	221	221	170	170	141	141	137	137
5381	259	259	137	137	121	121	171	171	221	221	173	173	150	150	137	137
5348	247	247	146	146	119	119	170	170	227	227	173	173	150	150	137	137
5313	259	259	146	146	119	119	171	171	221	221	164	164	150	150	137	137
5307	259	259	146	146	121	121	170	170	221	221	164	164	141	141	137	137
5289	259	259	137	137	121	121	171	171	227	227	173	173	141	150	144	144
5286	244	259	146	146	121	121	180	180	227	227	173	173	150	150	144	144
5285	259	259	137	137	121	119	180	171	227	217	173	173	150	141	144	144
5075	259	259	146	146	119	119	178	178	221	221	173	173	148	141	134	134
5049	253	253	137	137	113	113	180	180	221	221	173	173	150	150	144	144
5035	259	259	137	137	113	113	171	171	221	221	173	173	150	150	134	134
5023	259	259	146	146	113	113	171	171	209	221	173	173	141	141	134	134
2125	259	259	137	146	119	119	171	171	227	217	173	173	141	141	144	134
2123	259	259	137	137	121	121	171	171	227	227	173	173	141	141	144	144
2121	259	259	137	137	121	121	171	171	217	217	173	173	141	141	144	144
2101	259	259	137	137	121	121	170	170	227	227	170	170	141	141	144	144
2095	259	259	146	146	113	113	168	168	227	227	173	173	141	141	144	144
2091	259	259	146	146	113	113	178	178	227	227	173	173	150	150	144	144
2071	259	259	137	137	119	119	180	180	217	217	173	173	141	150	134	134
2067	247	247	146	146	121	121	178	178	217	217	173	173	150	150	137	137
2021	247	247	146	146	113	113	168	168	217	217	173	173	141	141	137	137

Tabla A4.4. Matriz de datos de los 16 genes candidatos genotipificados en la Población de Mapeo por Asociación (PMA).

Líneas	Genes Candidatos															
	MT	PRP_2	CytP450_C	RGEF_A2	PP2C	RhoBP_A	RhoBP_B	bHLHP	ZFHDP	HaGDP-DI	HaWP	HaTRP	HaPAL	HaWRKY5	HaWRKY7	HaGLP4
U5084	2:2	1:1	2:2	1:1	3:3	1:1	1:1	1:1	4:4	1:1	1:1	1:1	1:1	2:2	1:1	1:1
U12	2:2	1:1	1:1	2:2	3:3	1:1	1:1	1:1	4:4	7:7	1:1	1:1	2:2	2:2	2:2	
RK416	2:2	1:1	2:2	1:1	1:1	1:1	1:1	1:1	4:4	1:1	1:1	2:2	1:1	1:1	2:2	4:4
RHA801		1:1		1:1	1:1	1:1	1:1		1:1	3:3	1:1	1:1	1:1	1:1	2:2	1:1
RHA440	2:2	2:2	1:1	1:1	1:1	1:1	1:1	4:4	1:1	2:2	1:1		1:1	2:2	2:2	1:1
RHA439	1:1	2:2	1:1	1:1	1:1	2:2	1:1		1:1	1:1	2:2	1:1	1:1	2:2	2:2	4:4
RHA299	2:2	1:1	1:1	1:1	1:1	1:1	1:1	2:2	1:1	3:3	1:1	1:1	1:1	3:3	2:2	3:3
RHA276	2:2	2:2	2:2	1:1	1:1	2:2	1:1	1:1	4:4	1:1	1:1	1:1	1:1	1:1	1:1	3:3
RHA275	2:2	2:2	1:1	1:1	1:1	1:1	1:1	1:1	4:4	1:1	1:1	1:1	1:1	1:1	1:1	4:4
RHA266	2:2	2:2	1:1	1:1	1:1	1:1	1:1	1:1	2:2	3:3	1:1	1:1	1:1	1:1	1:1	4:4
HA89	2:2	1:1	1:1	1:1	1:1		1:1	1:1				1:1	1:1	1:1	1:1	
HA853	1:1	1:1	1:1	2:2	4:4	1:1	1:1	1:1	7:7	3:3	1:1	2:2	2:2	2:2	1:1	3:3
HA302	2:2	1:1	1:1	2:2	1:1	1:1	1:1	3:3	4:4	3:3	1:1	1:1	1:1	1:1	2:2	1:1
GP365	2:2	1:1	1:1	1:1	3:3	1:1	3:3	1:1	4:4	1:1	1:1	1:1	1:1	2:2	2:2	3:3
F155	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	1:1	3:3	1:1	1:1	1:1	3:3	2:2	
C999	2:2	2:2	2:2		1:1			1:1		1:1		1:1	1:1	1:1		
C998	2:2	2:2	1:1	1:1	3:3	1:1	1:1	1:1	5:5		1:1	1:1	1:1	1:1		4:4
C977	2:2	1:1	5:5	1:1	3:3	2:2	1:1	1:1	5:5	1:1	1:1	1:1	1:1	2:2	2:2	3:3
C853	2:2	1:1	1:1	1:1	3:3	1:1	1:1	1:1	5:5	2:2	1:1	1:1	1:1	2:2	2:2	3:3
C851	2:2	1:1	1:1	2:2	1:1	1:1	1:1	3:3	5:5	3:3	1:1	1:1	1:1	1:1	1:1	4:4
C849	1:1	1:1	1:1	2:2	4:4	1:1	1:1	1:1	7:7	3:3	1:1	2:2	2:2	2:2	1:1	3:3
C827	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	5:5	3:3	1:1	1:1	1:1	2:2	2:2	3:3
C826	1:1	2:2	1:1	1:1	1:1	2:2	1:1	2:2	1:1	1:1	2:2	1:1	1:1	2:2	2:2	4:4
C818	2:2	2:2	1:1	1:1	1:1	1:1	1:1	2:2	1:1	1:1	1:1	1:1	1:1	3:3	2:2	3:3
C817	2:2	2:2	1:1	2:2	1:1	1:1	1:1	4:4	5:5	1:1	1:1	1:1	1:1	2:2	2:2	1:1
C816	2:2	2:2	3:3	1:1	1:1	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	2:2	1:1	3:3
C813	1:1	1:1	2:2	1:1	3:3	2:2	1:1	1:1	4:4	1:1	1:1	1:1	1:1	1:1	1:1	3:3
C811	2:2	1:1	1:1	1:1	3:3	1:1	1:1	1:1	5:5	2:2	1:1	1:1	1:1	1:1	2:2	4:4
C633-2U	2:2			1:1	3:3	1:1	2:2	1:1		2:2	1:1	1:1	1:1			
C633-1U				1:1	3:3	1:1	1:1			6:6	1:1	1:1	1:1	2:2	2:2	
C628-M1	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	8:8	6:6	1:1		2:2	1:1	1:1	4:4

Líneas	Genes Candidatos															
	MT	PRP_2	CytP450_C	RGEF_A2	PP2C	RhoBP_A	RhoBP_B	bHLHP	ZFHDP	HaGDP-DI	HaWP	HaTRP	HaPAL	HaWRKY5	HaWRKY7	HaGLP4
C627-M1	1:1	2:2	1:1	1:1	3:3	1:1	2:2	1:1			2:2	1:1	1:1	2:2	2:2	1:1
C626-M	2:2	1:1	1:1	1:1		1:1	1:1	1:1	8:8	1:1	1:1		2:2	1:1	2:2	4:4
C625-U	2:2	2:2	1:1	1:1	3:3	1:1	2:2	5:5	1:1	2:2	1:1	1:1	1:1	2:2	2:2	4:4
C621	1:1	2:2	1:1	1:1	1:1	2:2	1:1	2:2	4:4	1:1	1:1	1:1	1:1	1:1	2:2	3:3
C620-2U	2:2	1:1	1:1	2:2	3:3	2:2	1:1	1:1	4:4	1:1	1:1	1:1	2:2	2:2	2:2	1:1
C620-1U	1:1	1:1	1:1	1:1	3:3	1:1	1:1	1:1	5:5	1:1	1:1	1:1	2:2	5:5	2:2	4:4
C618-2U			1:1	1:1	3:3	1:1	1:1	1:1	1:1	2:2	1:1	1:1		5:5	2:2	4:4
C618-1M		2:2	1:1	1:1	1:1	1:1	1:1	1:1		1:1	1:1	1:1		2:2	2:2	1:1
C610-1	2:2	2:2	6:6	2:2	1:1	1:1	1:1	2:2	7:7	2:2	1:1	1:1	1:1	6:6	2:2	1:1
C494-1	2:2	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	3:3	1:1	1:1	1:1	2:2	2:2	
C458B	2:2	2:2	2:2	1:1	1:1	2:2	1:1	1:1	4:4	1:1	1:1	1:1	1:1	1:1	1:1	
C454B	2:2	2:2	2:2	1:1	1:1	2:2	1:1	1:1	6:6	1:1	1:1	1:1	1:1	2:2	1:1	
C438	2:2	2:2	2:2	1:1	3:3	1:1	1:1	1:1	1:1	2:2	1:1	1:1	1:1	1:1	1:1	3:3
C437	2:2	1:1	1:1	2:2	1:1	1:1	2:2		5:5	2:2	1:1	1:1	2:2	2:2	1:1	
C434		1:1	1:1	1:1		1:1	1:1	3:3	4:4	1:1	1:1	1:1	1:1			1:1
C433	1:1	1:1	1:1	1:1	2:2	2:2	1:1	2:2	2:2	1:1	1:1	1:1	1:1	1:1	2:2	3:3
C432-2	1:1	1:1	1:1	1:1	2:2	2:2	1:1	2:2	2:2	1:1	1:1	1:1	2:2	1:1	2:2	4:4
C432-1	2:2	1:1	1:1	1:1	2:2	1:1	1:1	2:2		3:3	1:1	1:1	2:2	2:2	1:1	
C430	1:1	1:1	4:4	1:1	2:2	2:2	1:1			1:1	1:1	1:1	2:2	1:1	2:2	3:3
C428	2:2	1:1	1:1	1:1	2:2	1:1	1:1	2:2		3:3	1:1	1:1	1:1	1:1	2:2	4:4
C398	2:2	1:1	2:2		1:1	1:1	1:1		5:5		1:1	1:1	1:1	1:1	2:2	
C397	2:2	2:2	2:2			1:1	1:1	2:2	5:5		1:1	1:1	1:1	1:1	2:2	
C396-2B	1:1	2:2	1:1	1:1	3:3	1:1	1:1	1:1	4:4	3:3	1:1	1:1	1:1	2:2	2:2	4:4
C192-1			1:1		3:3	1:1	3:3	1:1		2:2	1:1	1:1		1:1		
C191-1	2:2	1:1	1:1	2:2	1:1		1:1	1:1	4:4		1:1	1:1		2:2	1:1	4:4
C166-U	2:2	1:1		1:1			1:1	1:1	4:4		1:1	1:1		1:1	2:2	
C155	2:2	2:2	2:2	1:1	1:1	2:2	1:1	1:1	4:4	1:1	1:1	1:1	1:1	1:1	2:2	3:3
C149-2	2:2		1:1	1:1		2:2	1:1	1:1	5:5		1:1	1:1	1:1	1:1	2:2	3:3
C149-1		1:1	1:1	1:1		2:2	1:1	1:1				1:1	1:1	1:1	2:2	
C139-2	2:2	2:2	1:1		3:3	1:1	1:1				1:1	1:1	2:2	1:1		
C139-1	2:2		1:1	1:1	3:3	1:1	1:1		2:2		1:1	1:1	1:1	1:1		1:1
C114M	1:1	2:2	1:1	2:2	3:3	1:1	2:2	2:2	2:2	3:3	2:2	2:2	2:2	2:2	2:2	1:1
C113M	2:2	2:2	1:1	1:1		1:1	1:1	1:1			1:1	1:1		2:2	2:2	4:4

Líneas	Genes Candidatos															
	MT	PRP_2	CytP450_C	RGEF_A2	PP2C	RhoBP_A	RhoBP_B	bHLHP	ZFHDP	HaGDP-DI	HaWP	HaTRP	HaPAL	HaWRKY5	HaWRKY7	HaGLP4
C112M	2:2	2:2	1:1	2:2	1:1	1:1	1:1	1:1	1:1	1:1	2:2	1:1	1:1	2:2	2:2	4:4
C111-3U	2:2	2:2	1:1	1:1	1:1	1:1	1:1	1:1				1:1	2:2	1:1	2:2	3:3
C110M	1:1	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	3:3	2:2	1:1	1:1	1:1	1:1	2:2
C1059	2:2	2:2	1:1	2:2	1:1	1:1	2:2	2:2	5:5	3:3	1:1	1:1	2:2	4:4	2:2	3:3
C1005-M		1:1	1:1		1:1	2:2	1:1	1:1	4:4	1:1	2:2	1:1	1:1	2:2	2:2	4:4
7x1	1:1	1:1	1:1		3:3	1:1	1:1	1:1	5:5		1:1	1:1	1:1	2:2		4:4
5393-N	2:2	1:1	2:2	1:1	3:3	1:1	1:1	2:2	4:4	3:3	1:1	1:1	1:1	2:2	2:2	3:3
5393-E	2:2	1:1	2:2	1:1	3:3	1:1	3:3	2:2	4:4	3:3	1:1	1:1	1:1	2:2	2:2	3:3
5341-1	2:2	1:1	1:1	1:1	3:3	1:1	1:1	1:1	5:5	2:2	1:1	1:1	1:1	1:1	2:2	3:3
5403	2:2	1:1	1:1	2:2	1:1	1:1	1:1	1:1	2:2	3:3	2:2	1:1	2:2	2:2	2:2	3:3
5381	2:2	1:1	1:1	1:1	3:3	1:1	3:3	1:1	4:4	1:1	1:1	1:1	1:1	1:1	2:2	4:4
5348	2:2	1:1	1:1	1:1	3:3	1:1	1:1	1:1	9:9	1:1	1:1	1:1	1:1	3:3	2:2	1:1
5313	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	3:3	2:2	1:1	2:2	2:2	3:3	2:2	4:4
5307	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	1:1	3:3	1:1	1:1	1:1	2:2	2:2	4:4
5289	2:2	2:2	1:1	1:1	1:1	1:1	1:1	1:1	4:4	1:1	1:1	1:1	1:1	2:2	1:1	3:3
5286	2:2	2:2	2:2	1:1	3:3	2:2	1:1	1:1	4:4	3:3	1:1	1:1	1:1	2:2	2:2	3:3
5285	2:2	2:2	2:2	1:1	1:1	2:2	1:1	1:1	4:4	3:3	1:1	1:1	1:1	2:2	2:2	
5075	2:2	2:2	2:2	1:1	3:3	1:1	1:1	2:2	1:1	1:1	1:1	1:1	1:1	1:1	2:2	3:3
5049	2:2	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	3:3	2:2	1:1	1:1	2:2	2:2	1:1
5035	1:1	2:2	2:2	1:1	1:1	1:1	1:1	1:1	4:4	2:2	1:1	1:1	1:1	1:1	2:2	3:3
5023	2:2	1:1	1:1	1:1	1:1	1:1	1:1	1:1	1:1	3:3	1:1	1:1	1:1	3:3	2:2	3:3
2125	2:2	1:1	2:2	1:1	1:1	2:2	1:1	1:1	4:4	3:3	1:1	1:1	1:1	1:1	1:1	3:3
2123	2:2	1:1	2:2	1:1	1:1	2:2	1:1	2:2	4:4	3:3	1:1	1:1	1:1	1:1	1:1	3:3
2121	2:2	1:1	2:2	1:1	1:1	2:2	1:1	1:1	4:4	3:3	1:1	1:1	1:1	1:1	1:1	3:3
2101	2:2	2:2	1:1	1:1	3:3	1:1	1:1	3:3	1:1	3:3	1:1	1:1	1:1	2:2	1:1	3:3
2095	2:2	1:1	1:1	1:1	3:3	2:2	1:1	1:1	5:5	3:3	1:1	1:1	1:1	1:1	1:1	4:4
2091	2:2	1:1	1:1	1:1	3:3	2:2	1:1	1:1	5:5	3:3	1:1	1:1	2:2	1:1	1:1	4:4
2071	2:2	2:2	2:2	1:1	1:1	2:2	1:1	1:1	6:6	1:1	1:1	1:1	1:1	2:2	1:1	
2067	2:2	1:1	1:1	2:2	3:3	1:1	1:1	1:1	5:5	1:1	1:1	2:2	2:2	1:1	1:1	4:4
2021	1:1	2:2	1:1	1:1	3:3	1:1	1:1	1:1	4:4	3:3	1:1	1:1	1:1	2:2	2:2	4:4

Las celdas en blanco son datos faltantes.

Tabla A4.5. Autovalores del Análisis de Coordenadas Principales (ACP) determinados a partir de la distancia de Rogers (Rogers 1972), utilizando la matriz de SSRs para las 94 líneas de la Población de Mapeo por Asociación (PMA).

Autovalor	Valor	Proporción	Proporción Acumulada
1	4,04	0,13	0,13
2	3,12	0,10	0,22
3	2,4	0,07	0,30
4	2,17	0,07	0,37
5	2,03	0,06	0,43
6	1,64	0,05	0,48
7	1,58	0,05	0,53
8	1,17	0,04	0,57
9	1,07	0,03	0,60
10	0,92	0,03	0,63
11	0,78	0,02	0,65
12	0,76	0,02	0,68
13	0,6	0,02	0,70
14	0,54	0,02	0,71
15	0,5	0,02	0,73
16	0,41	0,01	0,74
17	0,39	0,01	0,75
18	0,39	0,01	0,77
19	0,34	0,01	0,78
20	0,33	0,01	0,79

Sólo se tomaron las primeras cinco coordenadas de la matriz para el análisis de asociación.

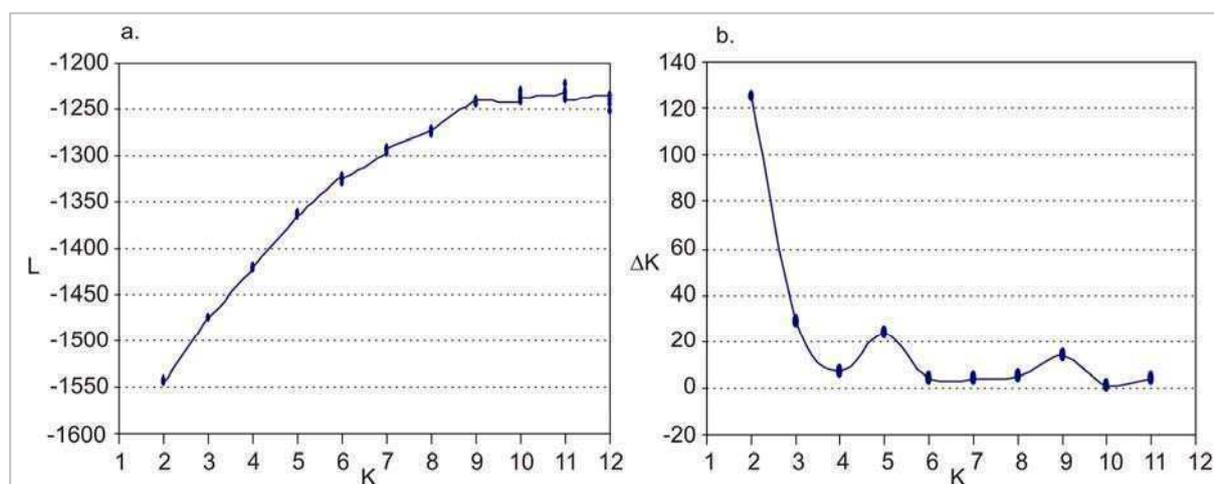


Figura A4.1. Análisis de los resultados del programa STRUCTURE para encontrar las K poblaciones ideales que forman la Población de Mapeo por Asociación (PMA).

a. Probabilidad *a posteriori* de los datos (L) como función del número de poblaciones ideales evaluadas (K=1 a 12). Los distintos puntos representan las corridas independientes para cada valor K.

b. ΔK calculado de acuerdo a Evanno y col. (2005), en función de los números de poblaciones ideales evaluadas (K=1 a 12). El valor modal de esta distribución es considerado como el número real de poblaciones K. Se observan dos valores modales a K=5 y K=9.

Tabla A4.6. Matriz de pertenencia de cada individuo con respecto a cada número de población (K) tomado como ideal.

Líneas	Pertenencia K=5				
U5084	0,029	0,047	0,216	0,556	0,153
U12	0,029	0,406	0,505	0,037	0,023
RK416	0,024	0,015	0,015	0,929	0,017
RHA801	0,574	0,277	0,028	0,021	0,1
RHA440	0,113	0,032	0,011	0,021	0,824
RHA439	0,008	0,014	0,017	0,018	0,943
RHA299	0,539	0,284	0,047	0,036	0,095
RHA276	0,946	0,014	0,011	0,02	0,01
RHA275	0,763	0,08	0,042	0,026	0,089
RHA266	0,916	0,014	0,014	0,023	0,033
ha89	0,021	0,015	0,013	0,929	0,022
HA853	0,012	0,015	0,953	0,01	0,01
HA302	0,016	0,036	0,023	0,858	0,067
GP365	0,021	0,028	0,017	0,864	0,071
F155	0,216	0,572	0,079	0,02	0,113
C999	0,241	0,048	0,428	0,216	0,067
C998	0,392	0,065	0,028	0,227	0,288
C977	0,175	0,041	0,719	0,038	0,029
C853	0,054	0,704	0,136	0,091	0,014
C851	0,027	0,496	0,042	0,408	0,028
C849	0,015	0,041	0,901	0,01	0,032
C827	0,085	0,248	0,266	0,01	0,392
C826	0,008	0,014	0,017	0,018	0,944
C818	0,421	0,091	0,011	0,131	0,346
C817	0,018	0,767	0,02	0,177	0,018
C816	0,929	0,011	0,014	0,021	0,025
C813	0,914	0,025	0,026	0,018	0,018
C811	0,192	0,189	0,386	0,138	0,095
C633-2U	0,011	0,074	0,517	0,257	0,14
C633-1U	0,012	0,068	0,74	0,099	0,081
C628-M1	0,018	0,876	0,013	0,035	0,058
C627-M1	0,022	0,051	0,028	0,016	0,883
C626-M	0,013	0,931	0,016	0,013	0,027

Líneas	Pertenencia K=9								
U5084	0,146	0,104	0,02	0,012	0,525	0,098	0,006	0,017	0,072
U12	0,059	0,245	0,022	0,021	0,113	0,101	0,018	0,415	0,006
RK416	0,602	0,006	0,006	0,013	0,182	0,02	0,015	0,15	0,007
RHA801	0,02	0,02	0,475	0,138	0,02	0,019	0,184	0,052	0,071
RHA440	0,013	0,005	0,094	0,043	0,014	0,008	0,05	0,008	0,765
RHA439	0,008	0,006	0,005	0,007	0,006	0,013	0,004	0,007	0,945
RHA299	0,01	0,011	0,794	0,128	0,017	0,007	0,015	0,009	0,008
RHA276	0,008	0,005	0,009	0,01	0,01	0,005	0,943	0,006	0,005
RHA275	0,022	0,03	0,203	0,116	0,014	0,015	0,548	0,017	0,036
RHA266	0,008	0,004	0,007	0,011	0,007	0,017	0,925	0,006	0,014
ha89	0,679	0,007	0,009	0,011	0,085	0,008	0,014	0,177	0,01
HA853	0,005	0,96	0,005	0,004	0,005	0,006	0,005	0,006	0,005
HA302	0,708	0,007	0,007	0,009	0,045	0,01	0,006	0,191	0,018
GP365	0,687	0,008	0,007	0,009	0,052	0,006	0,009	0,186	0,036
F155	0,007	0,032	0,756	0,159	0,01	0,007	0,007	0,015	0,008
C999	0,148	0,448	0,033	0,062	0,051	0,024	0,161	0,048	0,023
C998	0,144	0,015	0,037	0,035	0,063	0,023	0,343	0,067	0,273
C977	0,023	0,264	0,115	0,037	0,058	0,388	0,087	0,02	0,009
C853	0,06	0,229	0,05	0,073	0,276	0,016	0,019	0,272	0,005
C851	0,409	0,011	0,011	0,03	0,097	0,014	0,016	0,401	0,011
C849	0,005	0,917	0,021	0,016	0,005	0,007	0,004	0,008	0,016
C827	0,006	0,194	0,178	0,063	0,02	0,042	0,015	0,072	0,409
C826	0,009	0,006	0,005	0,007	0,006	0,013	0,004	0,007	0,945
C818	0,02	0,004	0,133	0,708	0,033	0,005	0,052	0,008	0,036
C817	0,081	0,007	0,02	0,033	0,206	0,011	0,008	0,628	0,007
C816	0,008	0,005	0,01	0,007	0,01	0,012	0,931	0,005	0,012
C813	0,013	0,027	0,163	0,032	0,012	0,008	0,72	0,013	0,011
C811	0,119	0,404	0,118	0,053	0,053	0,035	0,077	0,099	0,042
C633-2U	0,375	0,133	0,008	0,013	0,043	0,176	0,005	0,19	0,057
C633-1U	0,128	0,569	0,01	0,019	0,04	0,073	0,005	0,101	0,056
C628-M1	0,027	0,011	0,096	0,36	0,084	0,005	0,009	0,352	0,056
C627-M1	0,011	0,013	0,022	0,036	0,008	0,013	0,01	0,015	0,872
C626-M	0,009	0,017	0,154	0,297	0,087	0,005	0,005	0,394	0,033

Líneas	Pertenenca K=5				
C625-U	0,218	0,359	0,225	0,129	0,069
C621	0,745	0,025	0,042	0,171	0,018
C620-2U	0,04	0,029	0,015	0,892	0,024
C620-1U	0,114	0,092	0,046	0,729	0,02
C618-2U	0,029	0,455	0,134	0,242	0,139
C618-1M	0,109	0,246	0,021	0,152	0,472
C610-1	0,049	0,752	0,024	0,156	0,018
C494-1	0,015	0,106	0,037	0,065	0,776
C458B	0,955	0,011	0,01	0,016	0,008
C454B	0,909	0,016	0,012	0,038	0,024
C438	0,465	0,024	0,034	0,201	0,275
C437	0,16	0,053	0,039	0,409	0,34
C434	0,376	0,197	0,326	0,066	0,034
C433	0,013	0,01	0,8	0,02	0,158
C432-2	0,014	0,035	0,243	0,301	0,406
C432-1	0,26	0,018	0,056	0,377	0,288
C430	0,01	0,011	0,789	0,023	0,167
C428	0,204	0,031	0,063	0,437	0,266
C398	0,081	0,049	0,23	0,28	0,359
C397	0,017	0,016	0,894	0,015	0,057
C396-2B	0,017	0,018	0,943	0,011	0,011
C192-1	0,227	0,268	0,412	0,07	0,023
C191-1	0,057	0,722	0,057	0,15	0,014
C166-U	0,201	0,105	0,127	0,537	0,03
C155	0,954	0,011	0,01	0,016	0,009
C149-2	0,051	0,784	0,055	0,076	0,034
C149-1	0,22	0,028	0,041	0,675	0,036
C139-2	0,032	0,049	0,271	0,212	0,437
C139-1	0,037	0,1	0,016	0,111	0,736
C114M	0,283	0,067	0,016	0,049	0,584
C113M	0,034	0,029	0,329	0,019	0,589
C112M	0,134	0,443	0,214	0,037	0,173
C111-3U	0,096	0,085	0,406	0,074	0,34
C110M	0,016	0,854	0,074	0,013	0,043
C1059	0,03	0,713	0,041	0,188	0,028

Líneas	Pertenenca K=9								
C625-U	0,139	0,51	0,091	0,058	0,029	0,009	0,056	0,085	0,023
C621	0,064	0,027	0,023	0,033	0,042	0,019	0,763	0,021	0,006
C620-2U	0,284	0,009	0,014	0,026	0,531	0,015	0,028	0,074	0,019
C620-1U	0,528	0,018	0,036	0,02	0,115	0,015	0,078	0,18	0,009
C618-2U	0,376	0,087	0,015	0,032	0,063	0,016	0,011	0,316	0,084
C618-1M	0,052	0,007	0,112	0,657	0,025	0,008	0,016	0,039	0,085
C610-1	0,17	0,01	0,031	0,109	0,159	0,008	0,036	0,47	0,006
C494-1	0,081	0,017	0,107	0,401	0,019	0,024	0,006	0,062	0,284
C458B	0,007	0,006	0,009	0,007	0,01	0,004	0,948	0,005	0,004
C454B	0,014	0,006	0,007	0,01	0,008	0,005	0,932	0,009	0,01
C438	0,236	0,047	0,057	0,022	0,039	0,009	0,298	0,068	0,224
C437	0,238	0,023	0,067	0,189	0,136	0,016	0,094	0,077	0,16
C434	0,175	0,226	0,123	0,053	0,041	0,033	0,147	0,181	0,02
C433	0,006	0,007	0,004	0,005	0,005	0,956	0,006	0,005	0,006
C432-2	0,083	0,018	0,017	0,014	0,09	0,515	0,01	0,035	0,219
C432-1	0,305	0,011	0,013	0,023	0,049	0,095	0,282	0,076	0,147
C430	0,005	0,005	0,004	0,005	0,005	0,963	0,004	0,004	0,006
C428	0,437	0,019	0,019	0,035	0,052	0,04	0,187	0,121	0,091
C398	0,292	0,297	0,038	0,041	0,037	0,037	0,027	0,093	0,138
C397	0,017	0,251	0,016	0,01	0,009	0,658	0,012	0,018	0,009
C396-2B	0,01	0,938	0,009	0,006	0,005	0,011	0,009	0,007	0,005
C192-1	0,039	0,542	0,162	0,033	0,049	0,011	0,069	0,085	0,01
C191-1	0,079	0,06	0,02	0,025	0,191	0,006	0,039	0,572	0,007
C166-U	0,407	0,028	0,036	0,013	0,109	0,06	0,16	0,176	0,011
C155	0,007	0,006	0,01	0,007	0,011	0,004	0,947	0,005	0,004
C149-2	0,022	0,02	0,06	0,022	0,184	0,008	0,027	0,643	0,014
C149-1	0,185	0,043	0,019	0,022	0,65	0,014	0,049	0,012	0,007
C139-2	0,149	0,033	0,028	0,03	0,07	0,313	0,017	0,062	0,298
C139-1	0,061	0,008	0,083	0,041	0,037	0,012	0,016	0,048	0,695
C114M	0,032	0,011	0,369	0,112	0,034	0,006	0,055	0,015	0,365
C113M	0,014	0,397	0,093	0,051	0,008	0,052	0,012	0,009	0,364
C112M	0,007	0,035	0,131	0,771	0,009	0,022	0,007	0,009	0,01
C111-3U	0,019	0,202	0,096	0,549	0,011	0,049	0,021	0,011	0,043
C110M	0,01	0,096	0,238	0,31	0,06	0,012	0,005	0,251	0,017
C1059	0,119	0,01	0,02	0,035	0,168	0,007	0,013	0,621	0,007

Líneas	Pertenenca K=5				
C1005-M	0,376	0,113	0,02	0,067	0,424
7x1	0,171	0,019	0,106	0,23	0,474
5393-N	0,018	0,077	0,014	0,127	0,763
5393-E	0,02	0,078	0,015	0,096	0,791
5341-1	0,474	0,28	0,022	0,067	0,158
5403	0,013	0,099	0,054	0,02	0,814
5381	0,18	0,187	0,033	0,221	0,379
5348	0,016	0,02	0,866	0,026	0,072
5313	0,211	0,203	0,059	0,017	0,511
5307	0,02	0,079	0,122	0,055	0,724
5289	0,044	0,024	0,011	0,906	0,015
5286	0,013	0,082	0,023	0,58	0,302
5285	0,148	0,026	0,016	0,79	0,02
5075	0,434	0,44	0,064	0,013	0,048
5049	0,021	0,722	0,011	0,144	0,103
5035	0,524	0,39	0,011	0,024	0,052
5023	0,547	0,364	0,021	0,016	0,052
2125	0,718	0,066	0,027	0,169	0,019
2123	0,057	0,022	0,011	0,895	0,015
2121	0,13	0,016	0,012	0,825	0,016
2101	0,034	0,014	0,026	0,79	0,136
2095	0,021	0,543	0,247	0,168	0,021
2091	0,011	0,859	0,029	0,086	0,015
2071	0,91	0,016	0,012	0,037	0,024
2067	0,011	0,022	0,891	0,032	0,044
2021	0,014	0,032	0,932	0,012	0,011

Líneas	Pertenenca K=9								
C1005-M	0,029	0,009	0,497	0,085	0,029	0,012	0,081	0,017	0,242
7x1	0,202	0,008	0,022	0,018	0,013	0,354	0,14	0,055	0,188
5393-N	0,114	0,007	0,027	0,064	0,025	0,006	0,008	0,047	0,702
5393-E	0,108	0,008	0,034	0,092	0,019	0,007	0,008	0,046	0,677
5341-1	0,019	0,008	0,758	0,122	0,037	0,005	0,018	0,015	0,018
5403	0,021	0,057	0,051	0,036	0,019	0,059	0,006	0,054	0,698
5381	0,019	0,007	0,134	0,756	0,043	0,009	0,01	0,007	0,014
5348	0,009	0,03	0,012	0,009	0,009	0,904	0,009	0,011	0,007
5313	0,009	0,01	0,682	0,196	0,007	0,049	0,014	0,012	0,02
5307	0,045	0,014	0,474	0,119	0,019	0,234	0,006	0,017	0,072
5289	0,197	0,004	0,008	0,012	0,747	0,007	0,009	0,01	0,007
5286	0,243	0,008	0,013	0,037	0,206	0,039	0,006	0,147	0,301
5285	0,199	0,01	0,016	0,018	0,542	0,011	0,157	0,029	0,017
5075	0,008	0,047	0,657	0,133	0,019	0,006	0,048	0,068	0,014
5049	0,101	0,006	0,054	0,34	0,086	0,004	0,01	0,324	0,075
5035	0,009	0,005	0,158	0,735	0,012	0,004	0,055	0,014	0,008
5023	0,007	0,009	0,78	0,141	0,009	0,004	0,023	0,016	0,009
2125	0,072	0,012	0,087	0,02	0,239	0,013	0,512	0,037	0,009
2123	0,197	0,004	0,007	0,006	0,762	0,004	0,008	0,006	0,005
2121	0,2	0,009	0,008	0,007	0,723	0,005	0,033	0,01	0,005
2101	0,165	0,004	0,006	0,006	0,512	0,249	0,012	0,013	0,033
2095	0,059	0,239	0,037	0,014	0,239	0,008	0,007	0,388	0,009
2091	0,014	0,009	0,011	0,015	0,199	0,008	0,005	0,733	0,007
2071	0,015	0,006	0,008	0,009	0,008	0,005	0,931	0,009	0,01
2067	0,044	0,688	0,01	0,025	0,016	0,147	0,006	0,029	0,034
2021	0,01	0,945	0,007	0,006	0,006	0,007	0,005	0,009	0,005

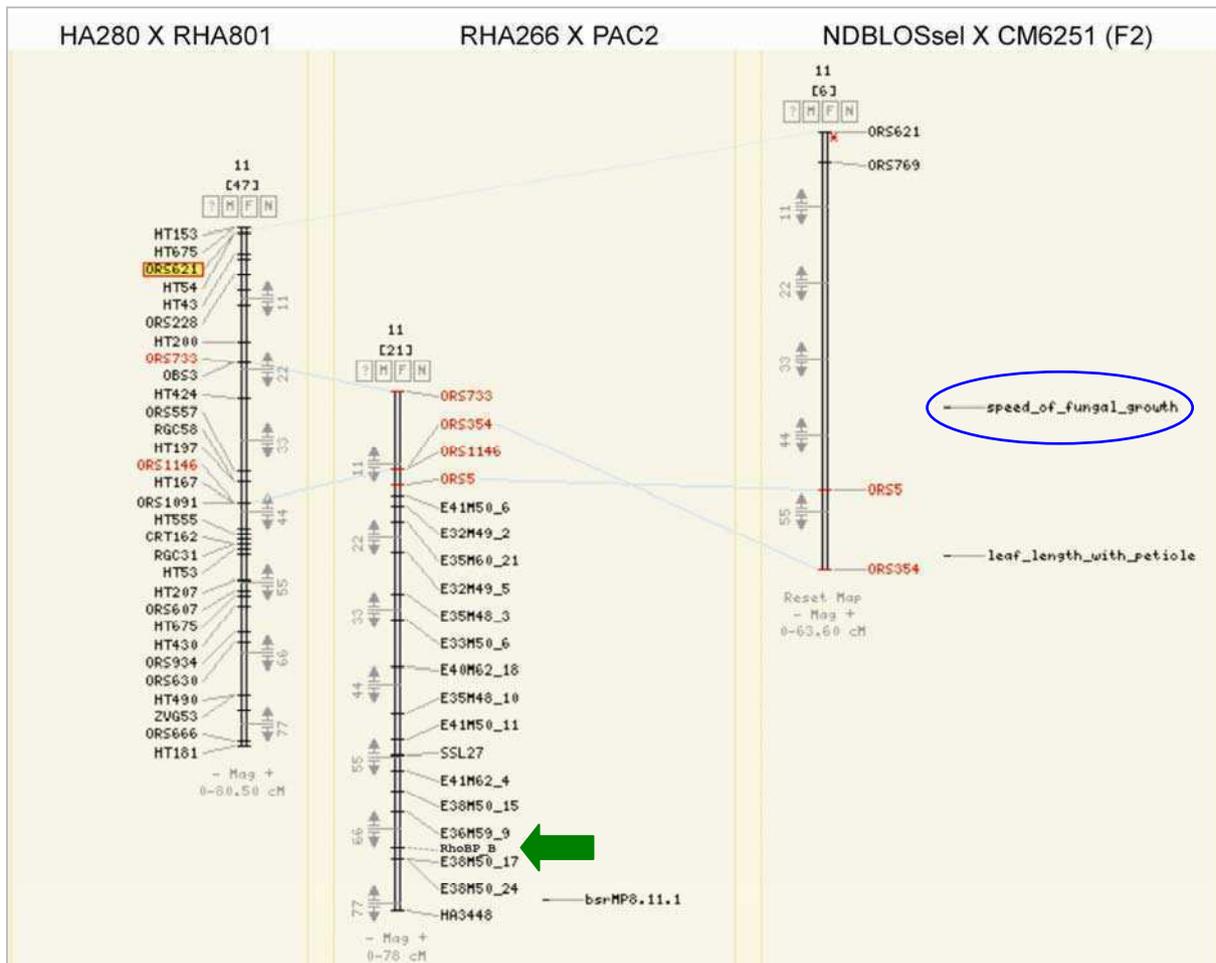


Figura A4.2. Correlación entre los GL11 para distintos mapas de girasol cultivado.

Los marcadores comunes (ORSx, en rojo) permiten correlacionar los GL11 de los distintos mapas de ligamiento: el mapa del GL11 obtenido con la población RIL proveniente del cruzamiento de las líneas HA280 X RHA801 (izquierda, no publicado), el GL11 del mapa de referencia de girasol RHA266 X PAC2 (central, Poormohammad Kiani y col. 2007a), y el GL11 del mapa obtenido con la población F2 proveniente del cruzamiento de NDBLOSsel X CM6251 (derecha, Micic y col. 2004). Se indica la posición de RhoBP_B en el mapa RHA266 X PAC2 (flecha verde). Un QTL relacionado con la tasa de crecimiento de *S. sclerotiorum* (*speed of fungal growth*, óvalo azul) se ubicó por encima del marcador ORS5 en el mapa de Micic y col. (2004), alejado de la posición de RhoBP_B en el mapa RHA266 X PAC2.