



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES

Respuestas a la temperatura en especies del género *Nothofagus* distribuidas en un gradiente altitudinal

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

LIC. MAXIMILIANO ESTRAVIS BARCALA

DIRECTORA DE TESIS: DRA. MARÍA VERÓNICA ARANA

DIRECTOR ASISTENTE: DR. NICOLÁS BELLORA

CONSEJERO DE ESTUDIOS: DR. PABLO Cerdán

Lugar de trabajo: Instituto de Investigaciones Forestales y Agrícolas
(IFAB, INTA-EEA Bariloche – CONICET)

FECHA DE DEFENSA: 7 DE MARZO DE 2019

*Non alia magis est libera et vitio carens
ritusque melius vita quae priscos colat,
quam quae relictis moenibus silvas amat.*

No existe otra vida más libre y carente de vicio
y que mejor cultive los antiguos ritos
que aquella que, abandonadas las murallas, ama los bosques.

Séneca, *Fedra*, vv. 483-485

Índice general

Resumen	1
Palabras clave	2
Abstract	3
Keywords	4
Agradecimientos	5
Publicaciones	11
Convenciones utilizadas en esta Tesis	13
1. Introducción general	15
1.1. Los bosques y el Cambio Climático	15
1.2. El bosque andino-patagónico y el Cambio Climático	19
1.3. El reloj circadiano y los procesos biológicos	24
1.3.1. Abordaje propuesto	24
1.3.2. Ritmos circadianos	28
1.3.3. Compensación por temperatura	30
1.4. Objetivo general	32
1.5. Objetivos específicos	32
2. Secuenciación, ensamblado y anotación de transcriptomas de Nothofagus pumilio	35
2.1. Introducción	35
2.2. Materiales y Métodos	40
2.2.1. Material vegetal y diseño experimental	40

2.2.2.	Extracción de ARN, construcción de <i>libraries</i> y secuenciación . . .	41
2.2.3.	Procesamiento de datos y ensamblado <i>de novo</i>	42
2.2.4.	Control de calidad de los ensamblados	44
2.2.5.	Anotación	46
2.2.6.	Predicción de microsatélites	47
2.3.	Resultados	47
2.3.1.	Depósito público de datos	47
2.3.2.	Control de calidad de la secuenciación	48
2.3.3.	Control de calidad del ensamblado	49
2.3.4.	Control de calidad de la anotación	54
2.3.5.	Predicciones de microsatélites	58
2.3.6.	Código desarrollado	58
2.4.	Discusión	60
2.4.1.	Análisis bioinformático	60
2.4.2.	Aportes a la comunidad bioinformática	65
3.	Análisis filogenómico del orden Fagales: aplicación de recursos y herramientas bioinformáticas	69
3.1.	Introducción	69
3.1.1.	La filogenética y la filogenómica	69
3.1.2.	Filogenómica en organismos vegetales	73
3.1.3.	El orden Fagales	74
3.2.	Materiales y Métodos	78
3.2.1.	Especies y secuencias utilizadas	78
3.2.2.	Obtención de genes conservados en común	78
3.2.3.	Análisis filogenéticos	80
3.2.4.	Análisis de incongruencias	80
3.3.	Resultados	82
3.3.1.	Especies y genes conservados comunes elegidos	82
3.3.2.	Análisis filogenéticos	83
3.4.	Discusión	89
3.4.1.	Genes conservados y construcción de la matriz de caracteres . . .	90
3.4.2.	Aportes a la filogenia del orden Fagales	92

4. Caracterización y respuesta a la temperatura del reloj circadiano de <i>Nothofagus</i> spp.	95
4.1. Introducción	96
4.2. Materiales y Métodos	97
4.2.1. Material vegetal y diseño experimental en condiciones controladas	97
4.2.2. Medición de <i>outputs</i> fisiológicos	98
4.2.3. Caracterización <i>in silico</i> del reloj circadiano	100
4.2.4. Medición de la expresión génica	101
4.2.5. Diseño de primers	103
4.2.6. Análisis de expresión diferencial	104
4.2.7. Muestreos en condiciones naturales	107
4.3. Resultados	109
4.3.1. <i>Outputs</i> fisiológicos del reloj circadiano	109
4.3.2. Caracterización <i>in silico</i> de genes del reloj	109
4.3.3. Puesta a punto de protocolos de laboratorio	112
4.3.4. Caracterización de la expresión de genes del oscilador central de <i>Nothofagus</i> spp. en condiciones controladas	116
4.3.5. Expresión diferencial	116
4.3.6. Efecto de la temperatura en el funcionamiento del reloj y su relación con la distribución natural de <i>Nothofagus</i> spp.	129
4.4. Discusión	135
4.4.1. Puesta a punto de la medición de expresión génica en <i>Nothofagus</i> spp.	137
4.4.2. Caracterización del funcionamiento del reloj circadiano en <i>Notho- fagus</i> spp.	139
4.4.3. Efecto del estrés térmico en la expresión génica global de <i>N. pumilio</i>	142
4.4.4. Respuestas a la temperatura y establecimiento de nichos ecológicos	144
5. Conclusiones generales y perspectivas futuras	151
Bibliografía	155
Glosario	187

Apéndices

A. Genes conservados para filogenómica	191
B. Figuras suplementarias	197
C. Tablas suplementarias	203
D. Descargos de derechos de autor de Figuras	217

Índice de figuras

1.1. Distribución mundial de los bosques	16
1.2. Tendencias globales en temperatura y precipitaciones para el siglo XXI según la proyección RCP6.0	17
1.3. Mapa de isohietas de la Patagonia noroccidental	20
1.4. Tendencias para Argentina en temperatura para el siglo XXI	21
1.5. Tendencias para Argentina en precipitaciones para el siglo XXI	22
1.6. Distribución natural de <i>Nothofagus obliqua</i> (roble pellín) en Argentina . .	25
1.7. Distribución natural de <i>Nothofagus pumilio</i> (lenga) en Patagonia norte argentina	26
1.8. Distribución natural de <i>Nothofagus pumilio</i> (lenga) en Patagonia austral argentina	27
1.9. Esquema molecular del reloj circadiano en plantas	29
2.1. Esquema de distintas técnicas de NGS	36
2.2. Esquema de un grafo de de Bruijn	39
2.3. Diseño experimental y condiciones de muestreo para experimentos de RNA- seq	41
2.4. Resumen esquemático del Capítulo	43
2.5. Correlación entre réplicas biológicas para cada ensamblado	51
2.6. Distribución de largo de contigs para cada ensamblado	52
2.7. Justificación del uso del ORF más largo por marco y por contig para la anotación (ensamblado ZT48_20C)	55
2.8. Distribución del puntaje de anotación para cada ensamblado	56
2.9. Ejemplo de salida del programa de búsqueda de contigs	60
2.10. Relación entre expresión en hojas y raíces de genes BUSCO de plantas .	64

3.1. Distribución geográfica de las familias del orden Fagales	76
3.2. Filogenia aceptada para las familias del orden Fagales	77
3.3. Frecuencia de modelos de evolución molecular que mejor ajustan a cada uno de los 199 genes CEG comunes	84
3.4. Árbol filogenético de máxima verosimilitud obtenido a partir de genes CEGs. Matriz de 7 taxones y 199 genes	85
3.5. Distribución por largo y ΔGLS de los 199 CEGs conservados entre especies	87
3.6. Árbol filogenético de máxima verosimilitud obtenido a partir de genes CEGs. Matriz de 7 taxones y 184 genes	89
3.7. Ejemplo de posible conflicto en alineamientos de genes provenientes de ensamblados <i>de novo</i>	90
4.1. Diseño experimental y condiciones de muestreo para experimentos de reloj circadiano en condiciones controladas	99
4.2. Cuenca del Lago Lácar	107
4.3. Outputs fisiológicos del reloj circadiano	110
4.4. Puesta a punto de métodos de extracción de ARN	112
4.5. Puesta a punto de la purificación del ARN	114
4.6. Puesta a punto de la retrotranscripción y la PCR cuantitativa	115
4.7. Puesta a punto de la PCR cuantitativa	117
4.8. Expresión de <i>NpTOC1</i> en condiciones diurnas para temperatura control .	118
4.9. Expresión de <i>NoTOC1</i> en condiciones diurnas para temperatura control .	118
4.10. Expresión de <i>NpTOC1</i> y <i>NpLHY</i> en condiciones circadianas para temperatura control	119
4.11. Expresión de <i>NoTOC1</i> y <i>NoLHY</i> en condiciones circadianas para temperatura control	119
4.12. Expresión de <i>NpTOC1</i> y <i>NpLHY</i> en condiciones circadianas a 34°C . . .	120
4.13. <i>Heat-map</i> para genes diferencialmente expresados según DESeq en condiciones contrastantes de temperatura y momento del día	122
4.14. Genes regulados ordenados por cambio en expresión	124
4.15. <i>Heat-map</i> para genes diferencialmente expresados según criterio <i>ad hoc</i> de significancia en condiciones contrastantes de temperatura y momento del día	126

4.16. Algunos términos de Gene Ontology sobrerrepresentados en genes de los grupos destacados según el criterio <i>ad hoc</i> de significancia	127
4.17. <i>MA-plot</i> entre 20°C y 34°C	129
4.18. Expresión de <i>NpTOC1</i> <i>NoTOC1</i> en condiciones diurnas para temperaturas control y tratamiento	131
4.19. Comparación de la temperatura media y el índice de sequía a lo largo de 3 años a 680 y 1100 msnm	132
4.20. Temperatura máxima diaria para 680 y 1100 msnm de mayo de 2012 a noviembre de 2013	133
4.21. Expresión de <i>NpTOC1</i> en condiciones naturales	134
4.22. Expresión de <i>NoTOC1</i> en condiciones naturales	134
4.23. Mortalidad de plántulas de roble y lenga en condiciones naturales	135
4.24. Concentración de clorofila de plántulas de roble y lenga en condiciones naturales	136
4.25. Peso seco de plántulas de lenga en condiciones naturales	136
4.26. Área foliar de plántulas de roble y lenga en condiciones naturales	137
4.27. Algunos ejemplos de debilitamiento de ritmos circadianos en condiciones constantes	141
4.28. Algunos ejemplos de disrupción del reloj circadiano por temperatura	146
B.1. Justificación del uso del ORF más largo por marco y por contig para la anotación (ensamblado ZT60_20C)	197
B.2. Justificación del uso del ORF más largo por marco y por contig para la anotación (ensamblado ZT48_34C)	198
B.3. Justificación del uso del ORF más largo por marco y por contig para la anotación (ensamblado ZT60_34C)	198
B.4. Árboles filogenéticos de máxima verosimilitud de genes <i>outliers</i> que apoyan a la hipótesis T1 (valores positivos de ΔGLS)	199
B.5. Árboles filogenéticos de máxima verosimilitud de genes <i>outliers</i> que apoyan a la hipótesis T2 (valores negativos de ΔGLS)	200
B.6. <i>Heat-map</i> para genes diferencialmente expresados según DESeq en condiciones contrastantes de temperatura	201

Índice de tablas

2.1. Secuencias de GenBank utilizadas para analizar la completitud e identidad de secuencias ensambladas de <i>N. pumilio</i>	46
2.2. Depósito público de los datos crudos	48
2.3. Depósito público de los ensamblados	48
2.4. Métricas de calidad de las lecturas antes y después del filtrado y recorte para cada conjunto de datos	49
2.5. Filtrado de lecturas	49
2.6. Comparación entre los ensambladores Trinity y SPAdes	50
2.7. Métricas de los ensamblados <i>de novo</i>	52
2.8. Métricas de mapeo	53
2.9. Número y porcentaje de genes conservados	53
2.10. Comparación de secuencias ensambladas y obtenidas de GenBank	54
2.11. Métricas de la anotación de los ensamblados contra <i>Arabidopsis thaliana</i>	56
2.12. Métricas de la anotación de los ensamblados contra SwissProt	57
2.13. Resumen de SSRs predichos en ensamblados de <i>N. pumilio</i>	58
3.1. Resumen de ventajas y desventajas de los métodos de reconstrucción filogenética más comunes	73
3.2. Descripción de los ensamblados del orden Fagales del proyecto Hardwood Genomics	79
3.3. Número y porcentaje de genes conservados de cada ensamblado	82
3.4. Genes conservados comunes entre ensamblados utilizados para el análisis filogenético	83
3.5. Valor de ΔGLS para cada gen CEG conservado común	86
3.6. Genes CEG <i>outliers</i>	88

4.1. Genes de <i>Arabidopsis thaliana</i> candidatos para referencia de qPCR . . .	104
4.2. Secuencias de <i>primers</i> de PCR de punto final y cuantitativa utilizados en este Capítulo	105
4.3. Genes del reloj circadiano de <i>Arabidopsis thaliana</i> encontrados <i>in silico</i> en <i>N. pumilio</i>	111
4.4. Dominios proteicos de genes del reloj circadiano de <i>A. thaliana</i> encontrados <i>in silico</i> en <i>N. pumilio</i>	113
4.5. Algunos términos de Gene Ontology sobrerrepresentados en genes más expresados a 20°C que a 34°C	123
4.6. Algunos términos de Gene Ontology sobrerrepresentados en genes más expresados a 34°C que a 20°C	123
A.1. Genes CEGs conservados entre especies	191
C.1. Términos de GO sobrerrepresentados en genes más expresados a 20°C que a 34°C	203
C.2. Términos de GO sobrerrepresentados en genes más expresados a 34°C que a 20°C	205
C.3. Términos de GO sobrerrepresentados en genes del grupo A del criterio <i>ad hoc</i> de significancia	206
C.4. Términos de GO sobrerrepresentados en genes del grupo B del criterio <i>ad hoc</i> de significancia	207
C.5. Términos de GO sobrerrepresentados en genes del grupo C del criterio <i>ad hoc</i> de significancia	210
C.6. Términos de GO sobrerrepresentados en genes del grupo D del criterio <i>ad hoc</i> de significancia	211
C.7. Términos de GO sobrerrepresentados en genes del grupo A de la Figura B.6213	
C.8. Términos de GO sobrerrepresentados en genes del grupo B de la Figura B.6214	

Respuestas a la temperatura en especies del género *Nothofagus* distribuidas en un gradiente altitudinal

Resumen

Las predicciones del Cambio Climático Global para el noroeste de la Patagonia indican un aumento de la temperatura media, acompañado de mayor frecuencia de eventos extremos. En este contexto, es fundamental comprender la respuesta de las especies nativas a la temperatura, en vistas de su conservación y uso sustentable. Entre estas especies, los árboles del género *Nothofagus* son de gran importancia ecológica y económica regional. En particular, dos de estas especies se encuentran distribuidas marcadamente en distintos estratos altitudinales en el norte de la Patagonia. Mientras que *N. obliqua* (roble pellín) domina las zonas bajas y cálidas, con un máximo de abundancia en los 650 m sobre el nivel del mar (msnm), *N. pumilio* (lenga) habita las zonas más frías del gradiente, encontrándose a partir de los 1000-1100 msnm y hasta el límite altitudinal del bosque. En este ambiente, el factor que varía más significativamente con la altura es la temperatura, y sugiere que cada especie posee preferencias por distintos nichos térmicos. En esta Tesis se estudió, entonces, la respuesta a la temperatura en especies de *Nothofagus* desde un enfoque novedoso, el cual consiste en el estudio de la influencia de la temperatura sobre el funcionamiento del reloj circadiano. Estas respuestas fueron abordadas integrando aspectos moleculares, bioinformáticos, fisiológicos, y ecológicos.

Demostramos con experimentos de laboratorio que el reloj circadiano de *N. pumilio* pierde la compensación por temperatura entre los 28°C y los 31°C, mientras que *N. obliqua* la pierde entre 31°C y 34°C. A un nivel de expresión génica global (mediante

análisis bioinformáticos de RNA-seq), nuestros resultados sugieren que en *N. pumilio* la temperatura alta reprime vías reguladas a 20°C por el reloj, como la fotosíntesis o el metabolismo primario, y que, además, a 34°C esta especie se encuentra en un estado de severo estrés térmico, independientemente del reloj circadiano, promoviéndose a esa temperatura la expresión de genes de señalización hormonal y estrés abiótico, entre otros. Por otro lado, en condiciones naturales, se observó que fuera de su área natural de distribución, *N. pumilio* posee un patrón inusual de expresión del gen del oscilador central del reloj circadiano, *TOC1*, además de una disminución en la acumulación de materia seca, clorofila a, clorofila b, expansión foliar y supervivencia de plántulas. Este comportamiento no fue observado en *N. obliqua* fuera de su área de distribución natural, es decir en los estratos más fríos del gradiente, aportando fuertes evidencias a favor del efecto disruptivo de la alta temperatura. En otras palabras, patrones inusuales de oscilación del reloj circadiano se encuentran asociados con una disminución en la capacidad de crecimiento y *fitness* de *N. pumilio*.

También se presenta como producto de esta Tesis el primer transcriptoma con alto grado de completitud para *Nothofagus*, secuenciado para *N. pumilio* en temperaturas y momentos del día contrastantes, en condiciones circadianas. Esta base de datos fue utilizada no sólo para estudios de expresión diferencial, sino también para diseño de primers y caracterización *in silico* de genes de reloj circadiano. También a través del uso de este transcriptoma, se llevó a cabo el primer estudio filogenómico que involucra distintas especies dentro del importante orden Fagales.

Palabras clave

Nothofagus — Estrés térmico — Reloj circadiano — Transcriptómica — Cambio climático — Gradientes ambientales — Filogenómica

Responses to temperature in *Nothofagus* species distributed in an altitudinal gradient

Abstract

Global Climate Change predictions for northwestern Patagonia indicate a rise in mean temperature, together with more frequent extreme events. In this context, understanding native species' responses to temperature is of utmost importance, considering their future conservation and sustainable use. Among these species, trees of the *Nothofagus* genus have great ecological and economical regional importance. Particularly, two of these species are markedly distributed in different altitudinal strata in northern Patagonia. While *N. obliqua* (roble pellín) dominates the lower, warmer areas, with highest abundance around 650 m above sea level (masl), *N. pumilio* (lenga) inhabits the colder areas, from 1000-1100 masl up to the altitudinal forest limit (tree-line) of these mountains. Here, the environmental factor that most significantly correlates with altitude is temperature, suggesting that each species has preferences for different thermic niches. Thus, in this Thesis we studied responses to temperature in *Nothofagus* species from a novel approach, namely the study of the influence of temperature in the circadian clock function. These responses were tackled integrating molecular, bioinformatic, physiological and ecological approaches.

We show in laboratory experiments that *N. pumilio*'s circadian clock loses temperature compensation at between 28°C and 31°C, while *N. obliqua* loses it between 31°C and 34°C. Regarding global gene expression (through RNA-seq bioinformatic analysis), our results suggest that in *N. pumilio* high temperature suppresses pathways regulated by the clock at 20°C, like photosynthesis or primary metabolism, and, in addition, that this species is under severe thermic stress at 34°C, regardless of circadian clock function, having promoted expression of genes related to hormone signa-

ling and abiotic stress, among others. Furthermore, in natural conditions, we observed that outside its natural distribution area, *N. pumilio* has an unusual expression pattern in a circadian clock central oscillator gene, *TOC1*, together with less dry weight, chlorophyll a, chlorophyll b, leaf expansion and seedling survival. This behavior was not observed in *N. obliqua* outside its natural distribution area, that is, in the colder strata of the gradient, adding strong evidence in favor of the disruptive effect of high temperature. In other words, unusual patterns of circadian clock oscillation are associated with a decrease in growth capacity and fitness in *N. pumilio*.

We also present as a result of this Thesis the first transcriptome with a high degree of completeness for *Nothofagus*, sequenced for *N. pumilio* in contrasting temperature and circadian conditions. This database was not only used for differential expression analysis, but also for primer design and *in silico* circadian clock genes' characterization. It was also through the use of this transcriptome that we conducted the first phylogenomic study for different species of the important order Fagales.

Keywords

Nothofagus — Thermic stress — Circadian clock — Transcriptomics — Climate change — Environmental gradients

Agradecimientos

En primer lugar, a mi directora, Vero, por confiar en mí para en gran parte iniciar varias líneas de trabajo que hoy vemos como bien encaradas y con un gran futuro. Gracias también por dejarme volar y nunca desanimarme en mis inquietudes e ideas. Desde aquella primera charla por Skype hasta la última discusión para esta Tesis, no cambiaría nada.

A mi codirector, Nico, que apareció como del cielo en el momento y lugar adecuados, y hoy es parte integral del equipo que se está consolidando. Por bancarme cuando no podía ni siquiera tipear una línea de código sin él al lado, y por guiarme y confiar en mis incipientes conocimientos bioinformáticos. Muy especialmente a Male Garavaglia por bancarnos tantas tardes, noches y hasta madrugadas trabajando y disrumpiendo la rutina diaria familiar. Por suerte siempre hubo lugar para largas charlas y picadas...

A mi codirector de beca, Marcelo Yanovsky, por compartir su profundo conocimiento del reloj circadiano, confiar en nuestro proyecto, y abrirnos la cabeza en las reuniones intensivas que tuvimos en estos años.

A Pablo Cerdán y Jorge Muschietti, de mi Comité de Seguimiento de Tesis, por sus agudas observaciones y por preocuparse por nuestro pequeño proyecto patagónico. Realmente a uno lo hacen sentir como un futuro colega, y eso es invaluable.

A mis compañeros becarios de grupo, Jorge Arias, Juli Cagnacci, Virginia Duboscq, Gaby Mattera y Ana Torres, por todas las mañanas, tardes y tardecitas en el laboratorio y las oficinas, haciendo que todo se pase más relajadamente. También por tantas charlas científicas, catárticas, divagantes...en fin, por acompañarme, en el sentido más formativo de la palabra.

A todo el grupo de Genética Ecológica y Mejoramiento Forestal del INTA Bariloche por recibirme tan abiertamente y enseñarme tanto sobre el mundo forestal, el bosque patagónico y la vida institucional y de investigación en general. A María Marta Az-

pilicuenta por tantas charlas sobre cine, literatura, “historia” y sistemática de plantas. A Alejandro Aparicio, mi tío favorito, por todas las charlas después de hora sobre los temas más diversos, desde política hasta estadística, pasando por autos y caminatas recomendadas. A Verónica El Mujtar por compartirme tanto conocimiento sobre genómica de árboles y bioinformática, y por tener un espíritu crítico tan agudo que contagia. A Leonardo Gallo por su perspectiva y por el camino que abrió. A Paula Marchelli por ser casi una codirectora. A Mario Pastorino por sus contribuciones gracias a su profundo conocimiento de las especies y del ámbito forestal. A Carolina Soliani por ser para mí un modelo de trabajo en laboratorio y en genética, cuando yo nunca había hecho una PCR antes de llegar a Bariloche. A nuestra secretaria de área, Inés Bertoldi, por hacer todo lo burocrático mucho más fácil, despejarme tantas dudas y arreglar desprolijidades. Muy especialmente a nuestros queridos auxiliares, Mario Huentú por su manejo y cuidado inmejorable de las plantas en vivero, y Fabián Jaque por todas las salidas de campo que no habría soportado sin su humor y trabajo duro, además de por tantas charlas y enseñanzas sobre arte y naturaleza.

Al INTA Bariloche y su gente por haber sido un lugar de trabajo cómodo, bien equipado, y en cuyo interior se dan discusiones sobre investigación estratégica y otras cuestiones que muchas veces quedan de lado en el mundo puramente academicista. Si bien, como en todos los ámbitos, no siempre estamos todos de acuerdo ni se manejan de la mejor manera algunos temas, estoy profundamente agradecido de haber transcurrido una etapa tan importante como el doctorado en una institución como el INTA, que admiro desde la carrera de grado. A su director Mauro Sarasola por tener siempre la puerta abierta y hacernos sentir y notar que le importan nuestras opiniones y percepciones del día a día, sobre todo en un momento de incertidumbre nacional como la que atravesamos los últimos años. A todos los investigadores, becarios, técnicos y personal de apoyo, que son demasiados para nombrar individualmente, que me ayudaron, apoyaron y escucharon en estos años. En especial a Alejandro Martínez-Meier, por su humor único y su visión integradora de la ciencia y la tecnología, al grupo de Sanidad Animal por prestarme generosa y desinteresadamente equipos y reactivos fundamentales para mis experimentos, y a Aldana López, mi querida primera compañera de laboratorio, cuando éramos tan pocos becarios...

A Fernando Carrari por recibirme en su laboratorio del Instituto de Biotecnología

del INTA Castelar en varias ocasiones durante los primeros años de mi doctorado, cuando todo era nuevo y desordenado para mí, y dedicarme tiempo en su ocupada agenda para hacerme sentir cómodo y asegurarse de que pudiera cumplir en tiempo y forma mis objetivos de cada viaje. A toda la gente del labo de plantas del IB por la buena onda y el acompañamiento en esas semanas tan precipitadas y donde nada parecía encajar: Sebastián Asurmendi, Luisa Bermúdez, Estanislao Burgos, Carla Coluccio, Gabriela Conti, Andy Venturuzzi...

A todo el equipo forestal del Parque Nacional Lanín (Marcelo Peñalba, Liliana Lozano, Marcelo Fernández, Martín Lara y Carlos Clericci) por recibirnos siempre tan bien, por interesarse en nuestros estudios, y por las caminatas y charlas en el hermoso santuario natural que ellos ayudan día a día a conservar.

A Birgit Ziegenhagen, Katrin Heer y Sascha Liepelt de la Universidad de Marburg (Alemania) por hacerme sentir tan bienvenido y cómodo en mi estadía. Conservo los mejores recuerdos científicos y humanos de ese grupo y esa bella ciudad.

A la Plataforma de Genómica y Transcriptómica de Bordeaux (Francia), en especial a Erwan Guichoux, por su profesionalismo, calidez humana y paciencia en lo que fue mi primera estancia científica en el exterior, y algunas de las manipulaciones de laboratorio más sensibles ypreciadas que me han tocado hacer.

A Anselmo Reggiardo y Florencia Bialade de Tecnolab por el asesoramiento y la paciencia durante la puesta a punto de las técnicas de laboratorio.

Al CONICET, a la ANPCyT, al MINCyT (actual Secretaría), al DAAD y a Trees4Future por financiar mi beca, mis estadías en el exterior, y los subsidios con los cuales pude desarrollar mis investigaciones. A la UBA por contarme entre sus alumnos desde el colegio secundario e inculcarme mucho más que conocimiento (y varios dolores de cabeza burocráticos, a decir verdad): el orgullo de ser un hijo de la educación pública.



Un apartado especial merecen los profesores y profesoras que durante años me inspiraron y fomentaron en mí la vocación científica, el espíritu crítico y el amor por el conocimiento en general. El colegio secundario acompaña una maduración personal y apertura de mente en una etapa clave de la vida, y por eso guardo los mejores recuerdos de aquellos años, junto con la admiración y agradecimiento a los siguientes

docentes: Andrea López (*in memoriam*), Diego Supervielle, Adriana Bilgray, Julio Ferreira (*in memoriam*), Carlos González, Ana Maiorana, Silvina Marsimian, Marta Royo, José Luis Bianchi, Silvia Di Segni.

En la facultad se amplía y afina la vocación, y se conoce el mundo académico y de investigación por dentro por primera vez. De aquellos años, agradezco a todos los docentes que me formaron, pero especialmente a los que me inspiran y que recuerdo constantemente con admiración. Gonzalo Comas y Leszek Szybisz (*in memoriam*) cambiaron mi forma de ver la matemática y su rol en el estudio de los sistemas físicos. Gabriela Amodeo y Jorge Muschietti definieron mi vocación por la investigación y la fisiología vegetal. Sonia Rosenfeldt y Pablo Picca me hicieron conocer en profundidad la diversidad del mundo de las plantas y despertaron una pasión dormida en el camino. Sara Maldonado me hizo conocer como nadie los mecanismos moleculares del desarrollo vegetal. Cecilia Carmarán fue más que una docente; casi una madre para los pocos alumnos de Fitopatología, en el que fuera mi último cuatrimestre de facultad, ya con el último aliento pero queriendo quedarme cada noche hasta el final en el laboratorio. Del otro lado del mostrador, a mis queridos alumnos de Botánica, en lo que fue una de las experiencias más hermosas de mi vida: la docencia de la “ciencia amable de las plantas”, al decir de J. J. Valla, a un grupo tan motivado y abierto.

De la Facultad de Filosofía y Letras, a la querida cátedra Castello, donde literalmente mi forma de pensar, analizar y concebir el conocimiento cambió: Ángel Castello, Leandro Pinkler, Marisa Divenosa, Andrés Racket. También a mis profesoras de sánscrito, donde terminé de madurar y dar un giro mental para toda la vida: Rosalía Vofchuk y Gabriela Müller.



Y porque no todo es ciencia y trabajo en estos años, aunque muchas veces lo parece...

A mi familia, por bancarme muchas veces taciturno en mis viajes a Buenos Aires, poco comunicativo y hasta poco presente (en cuerpo o en mente) a veces, pero nunca dejar por eso de expresarme todo su amor y su apoyo en esto de hacer lo que me gusta. También porque obviamente mi fascinación por la ciencia y la investigación tiene sus raíces en la infancia: las revistas National Geographic de mi viejo, los libros de química

orgánica de mi vieja, mi primer libro de Carl Sagan regalado por un cliente de mi viejo a través suyo a mis 11 años, o las caminatas absorto por el vivero de Marcelo y Verónica. Mucho más cercano en el tiempo, a mi hermano por regalarme mi primer libro de S. J. Gould, que junto con Sagan son mis máximos referentes científicos, y por toda una vida para la que no pueden alcanzar las palabras.

A todos mis amigos por escucharme, por compartir sus vidas conmigo y por darme motivos para cada día, cada semana, cada viaje a Buenos Aires, para no perder la cordura ante cada deadline...en una palabra, por mantenerme sobre la tierra y recordarme (porque sí, a veces lo olvido) que la vida es más que el trabajo.

A los pibes por ser los más intensos y porque con ellos cada juntada o salida es una aventura; son las personas con las que más cosas compartí en mi vida. A los muchachos de la división, por crecer conmigo y estar siempre ahí. En especial a Marquito, mi hermano del alma, porque “me conoce como nadie”, parafraseando a uno de nuestros ídolos en común. A los grupetes de la facultad, por compartir esos años fundamentales conmigo y por la alegría de ver cómo cada uno sigue su camino en lo que empezamos juntos. Finalmente (en el tiempo, no en importancia), este último año en Bariloche no habría sido lo mismo sin mis compañeros (y profesor, pero que es como uno más) del Lab Multimedia de la Escuela La Llave, ni sin mis queridos “petecos”: gracias por ayudarme, aunque sea sin saberlo, a soltarme y descontracturarme en estos meses tan demandantes mentalmente.



Finalmente agradezco a esa fuerza o entidad en la que cada uno crea, que escapa al alcance de las palabras. A ella también un pedido, formulado hace miles de años por los sabios hindúes, para el camino que recién empieza:

असतो मा सद्गमय ।
तमसो मा ज्योतिर्गमय ।
मृत्योर्मा अमृतं गमय ।

De lo que no existe a lo que existe haz que vaya

De la oscuridad a la luz haz que vaya

De lo mortal a lo inmortal haz que vaya

Publicaciones

Durante el transcurso de esta Tesis participé como colaborador en la siguiente publicación, no relacionada directamente con la temática de la Tesis:

- Arana, M. V., Tognacca, R. S., **Estravis-Barcala, M.**, Sánchez, R. A. y Botto, J. F. «Physiological and molecular mechanisms underlying the integration of light and temperature cues in *Arabidopsis thaliana* seeds». En: *Plant, Cell & Environment* 40 (2017), pág. 3113.

Además, gran parte del Capítulo 2 forma parte de un artículo enviado:

- **Estravis-Barcala, M.**, Arana, M. V., Marchelli, P., Heer, K., Ziegenhagen, B. y Bellora, N. «RNA-seq of southern beech *Nothofagus pumilio* in contrasting circadian and temperature conditions» Enviado. 2018.

Convenciones utilizadas en esta Tesis

El estilo bibliográfico de citas para esta Tesis es una modificación del estilo Chicago, por orden de aparición en el texto (numérico).

Todos los sitios web referenciados (tanto en el texto principal como en la Bibliografía) fueron accedidos exitosamente en el mes de Diciembre de 2018.



Las palabras y siglas resaltadas en **VERSALITA** se encuentran definidas en el Glosario, ubicado al final de la Tesis, entre la Bibliografía y los Apéndices. Sólo se resalta la primera aparición de cada palabra en la Tesis.



En el Apéndice D se detallan los descargos legales relacionados a la reimpresión de Figuras provenientes de publicaciones científicas.



Esta Tesis fue escrita en $\text{\LaTeX}$ ($\text{\TeX}$ live versión 2018). Comandos y formato a disposición (maxiestravis@gmail.com)

Capítulo 1

Introducción general

Lo desconocido es una abstracción; lo conocido, un desierto; pero lo conocido a medias, lo vislumbrado, es el lugar perfecto para hacer ondular deseo y alucinación.

Juan José Saer, *El Entenado*

1.1. Los bosques y el Cambio Climático

A nivel global, los bosques ocupan alrededor de un tercio de las superficies emergidas (Figura 1.1), y aportan la mitad de la productividad primaria neta [1]. Para la humanidad, los bosques son proveedores de servicios ecosistémicos (como la protección de suelos y cuencas hídricas), además de importantes económicamente, tanto por la producción y utilización de madera como por los productos forestales no maderables, de gran importancia en bosques naturales (semillas, fibras, exudados, hongos, etc.) [2]. De manera no menos importante, los bosques satisfacen necesidades estéticas y espirituales muchas veces de igual o mayor relevancia que las económicas [3]. Por último, los bosques influyen en el clima por su rol primordial en el intercambio de energía y agua con la atmósfera, y son importantes hospedadores de biodiversidad [4].

Los ambientes boscosos del mundo se encuentran amenazados por el Cambio Climático Global (CCG). A lo largo del siglo XX, se produjo un calentamiento global de alrededor de 1°C [6], y las predicciones del CCG indican un aumento de temperatura y cambios en los regímenes de precipitación en todo el globo (Figura 1.2). Los impactos del CCG en los bosques serán agudos, dando lugar a cambios en el uso de la tierra y

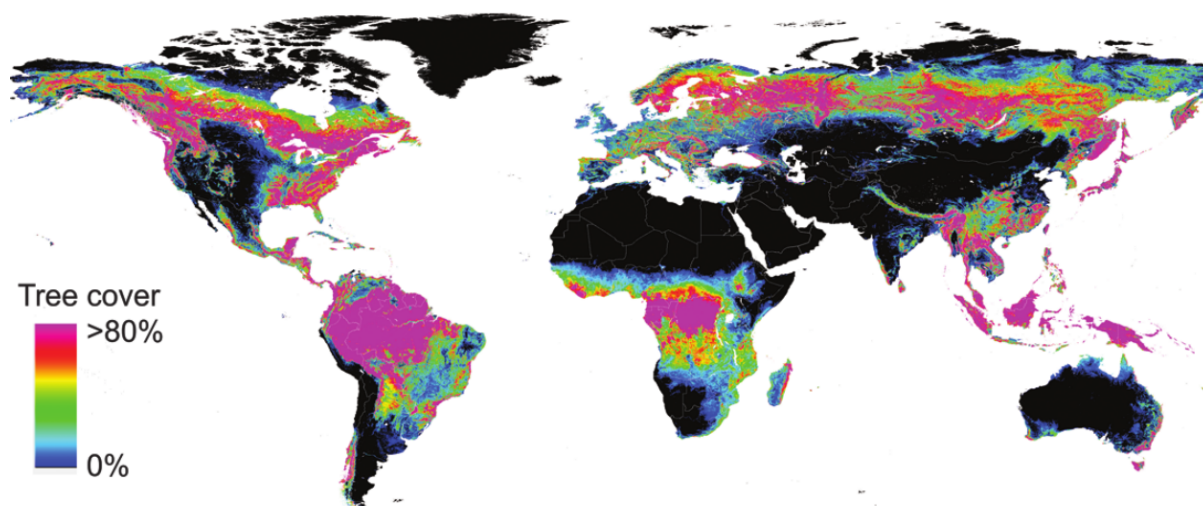
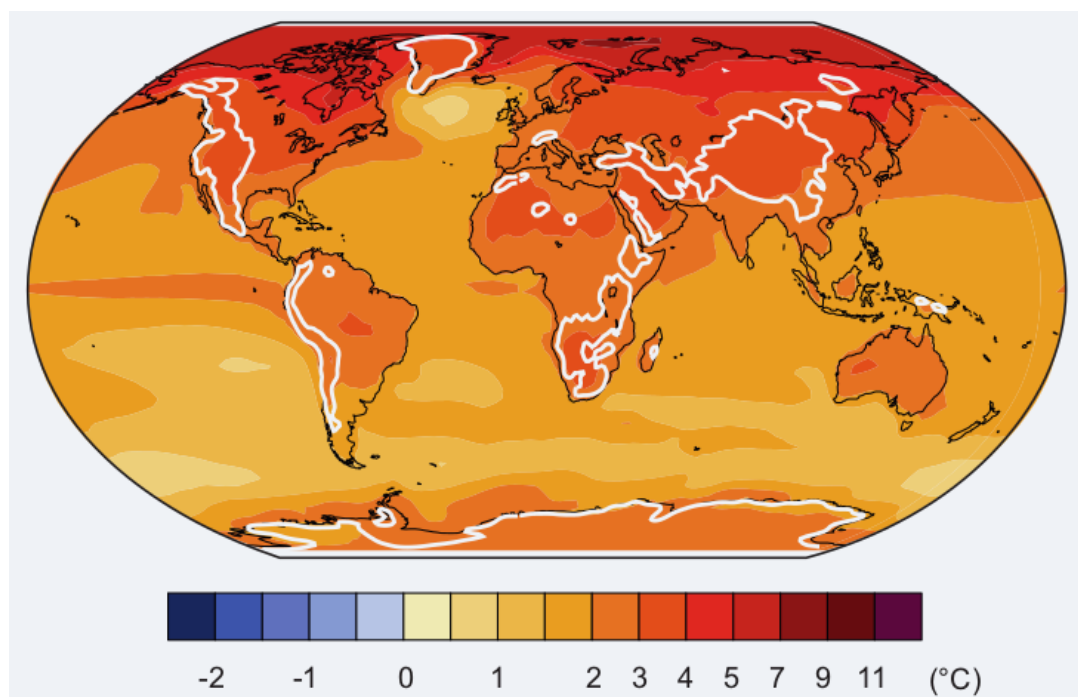


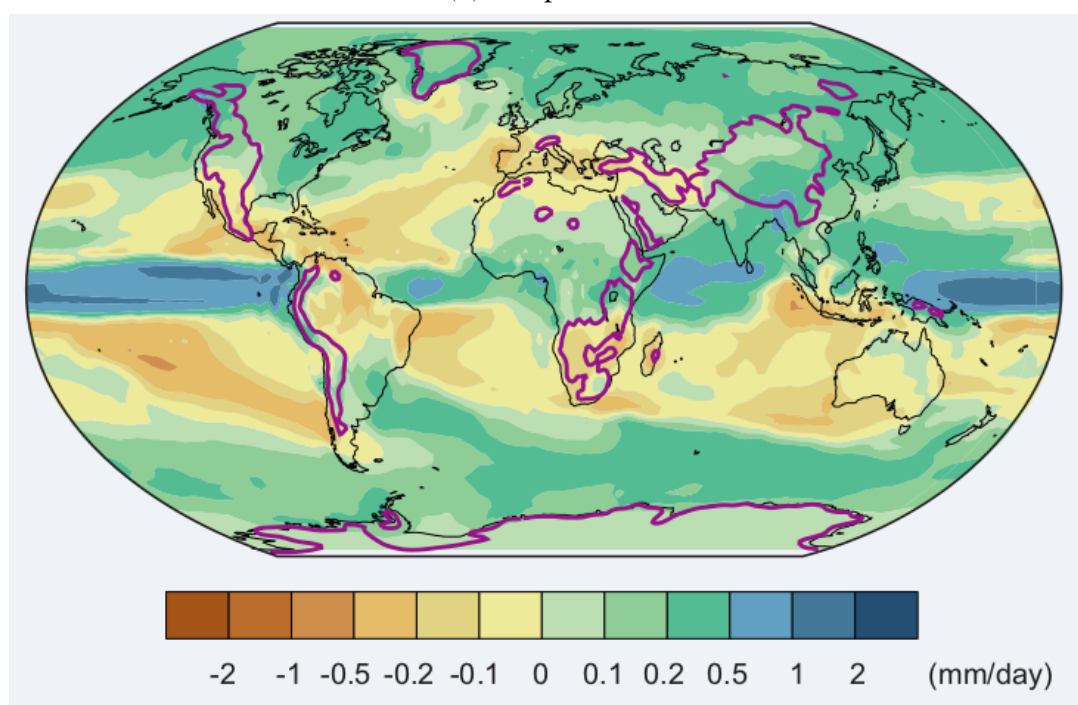
Figura 1.1: Distribución mundial de los bosques. Escala: cobertura de árboles. Fuente: [5]

la distribución de especies entre otros [7]. Las temperaturas altas y las precipitaciones alteradas, junto con la mayor frecuencia de eventos extremos, son amenazas globales a los bosques [8]. Las especies arbóreas, debido a sus largos tiempos generacionales y relativa poca capacidad de dispersión, ven dificultada su migración ante condiciones adversas, debiendo por lo tanto adaptarse localmente para no extinguirse [9, 10]. Las poblaciones pueden no ser capaces de migrar lo suficientemente rápido, resultando en distribuciones más acotadas [11]. En ese mismo sentido, los bosques pueden sufrir mortalidad masiva ante eventos extremos [12], como se ha reportado incluso para Patagonia norte [13]. Otros efectos del CCG sobre los bosques se relacionan con la mayor susceptibilidad de las especies al ataque de insectos y patógenos [14], o con la desventaja competitiva [15].

El cambio climático está actualmente, y seguirá en el futuro, modificando las distribuciones geográficas de las especies vegetales, lo cual no es sorprendente dada la fuerte influencia del clima en la determinación del hábitat de las plantas [18]. Los cambios hacia elevaciones más altas en laderas montañosas son los más frecuentemente reportados [19, 20] (cabe destacar el estudio de la Ref. [21], realizado en Nueva Zelanda en especies de *Nothofagus*, como las de interés para esta Tesis, ver próxima Sección). Las especies vegetales en regiones montañosas suelen estar distribuidas en franjas altitudinales bien delimitadas, comparadas con los rangos más bien amplios



(a) Temperatura



(b) Precipitación

Figura 1.2: Tendencias globales en temperatura (a) y precipitaciones (b) para el siglo XXI según la proyección RCP6.0 (escenario de emisiones moderadas a altas). Las líneas de contorno blancas y violetas encierran alturas mayores a 1000 metros sobre el nivel del mar. Fuente: [16], adaptado de [17]

y poco definidos de distribución latitudinal en las zonas bajas. Esta “compresión” altitudinal de la vegetación de montaña se debe al contraste entre el descenso de la temperatura con la altura respecto de con la latitud, a decir: aproximadamente 1°C cada 160 metros de altura, o cada 145 kilómetros de latitud (a 45°S) [22].

Debido a la presencia de distintos nichos térmicos en distancias relativamente cortas a lo largo de una ladera, los ambientes de montaña podrían funcionar como refugios de biodiversidad ante el CCG [23]. Esta misma característica hace que las montañas se comporten como “laboratorios naturales” [19, 24] para estudiar respuestas y adaptación local a distintos escenarios térmicos en condiciones naturales.

El CCG es, junto con el crecimiento poblacional y económico, una de las presiones más fuertes sobre los ambientes montañosos, que ocupan alrededor del 25% de las superficies continentales, y cuentan con el 40% de la población mundial viviendo en cuencas fluviales originadas en ellos [25]. Dicha presión es debida al impacto del CCG en la distribución de especies, retroceso de glaciares y alteración de ciclos hidrológicos [26]. Como puede verse en la Figura 1.2, las montañas no escapan a las tendencias de aumento de temperatura y cambios en regímenes de precipitación que afectarán al planeta en el siglo XXI. Todos los ambientes de montaña del globo se calentarán en el presente siglo, y en mayor medida que lo registrado para el siglo XX [26].

En este contexto, la Iniciativa para la Investigación en Montañas (<http://www.mountainresearchinitiative.org>), una organización de Universidad de Bern (Suiza) financiada por la FAO y la Academia Suiza de Ciencias, enfatiza la necesidad de abordar cuatro áreas principales de investigación en regiones montañosas [25]:

- Monitoreo de largo plazo de indicadores de cambio ambiental
- Integración de estudios ecológicos, hidrológicos, de uso de la tierra, y económicos para toma de decisiones políticas
- Manejo sustentable de recursos, priorizando cambios en recursos forestales, seguridad alimentaria, y modificaciones de recursos hídricos
- Estudios a lo largo de gradientes altitudinales, que permiten evaluar cambios ambientales en fronteras ecológicas y climatológicas, y así la sensibilidad de la vegetación (y otros recursos naturales) a distintos forzantes

Estas líneas de investigación hacen énfasis y apuntan hacia la integración de varios campos del conocimiento a la hora de abordar un problema tan complejo y multifacto-

rial como el efecto del CCG en especies nativas. La biología molecular es una disciplina promisorio para afrontar este desafío, ya que permite estudiar de manera integral la influencia y el aporte del genotipo a un fenotipo, en un ambiente dado. Sin embargo, los estudios moleculares son de por sí poco comunes en especies no modelo, y lo son menos aún en condiciones naturales, debido a la gran cantidad de variables fuera del control del investigador (ambientales, de competencia con otras especies, etc.). De esta manera, el conocimiento de la biología molecular es enorme, pero prácticamente restringido a condiciones controladas de laboratorio.

El ya mencionado comportamiento de las montañas como “laboratorios naturales” las convierte en un escenario ideal para complementar los tradicionales estudios ecológicos con otros fisiológicos y moleculares en condiciones naturales, aportando a extender el conocimiento desde las condiciones controladas del laboratorio hacia ambientes y preguntas ecológicamente relevantes.

1.2. El bosque andino-patagónico y el Cambio Climático

La República Argentina se extiende ampliamente en sentido norte-sur, desde el Trópico de Capricornio hasta latitudes subantárticas. Los océanos Pacífico y Atlántico, y la Cordillera de los Andes ejercen las mayores influencias físicas sobre el clima argentino. En particular, en la región patagónica (de 40°S hacia el sur), los Andes son progresivamente más bajos, permitiendo que parte de la abundante precipitación que cae en Chile se produzca también al este de la cordillera, en una angosta franja (Figura 1.3), mientras que el resto de la Patagonia argentina es de clima semiárido y árido [27]. En dicha franja de régimen húmedo y subhúmedo se desarrolla la ecorregión del **bosque andino-patagónico**.

Durante la segunda mitad del siglo XX, la región patagónica fue la única del país que vio aumentada su temperatura máxima media estival [29], a la vez que experimentó un calentamiento general [30] y una disminución de la intensidad de los vientos del oeste para el norte de la región [31]. Además, sólo en la región andina de 30°S a 42°S (incluyendo el noroeste de la Patagonia) se produjo una disminución de la precipitación a lo largo del siglo XX [32].

Las proyecciones del CCG para el siglo XXI indican un calentamiento general de

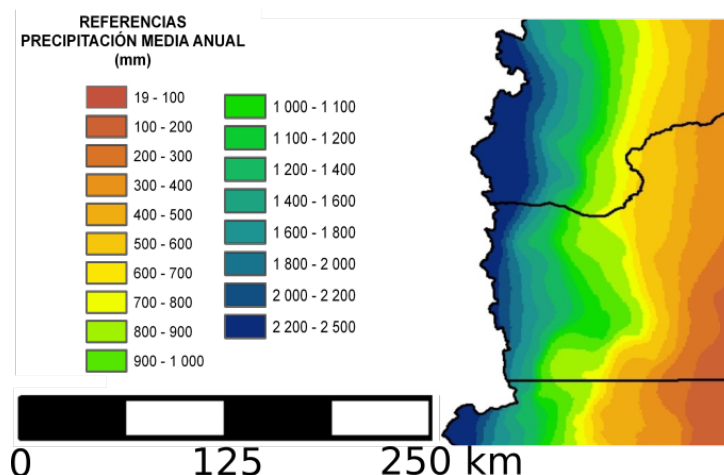
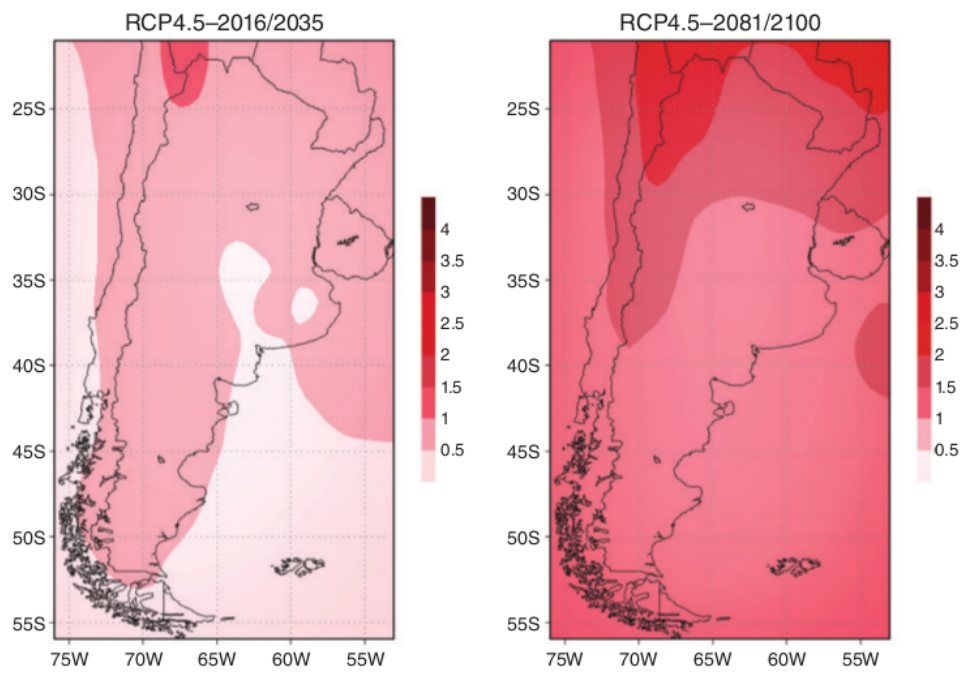


Figura 1.3: Mapa de isohietas de la Patagonia noroccidental. Se muestra el centro-sur de la provincia de Neuquén, el oeste de la provincia de Río Negro, y el extremo noroeste de la provincia de Chubut. Fuente: [28]

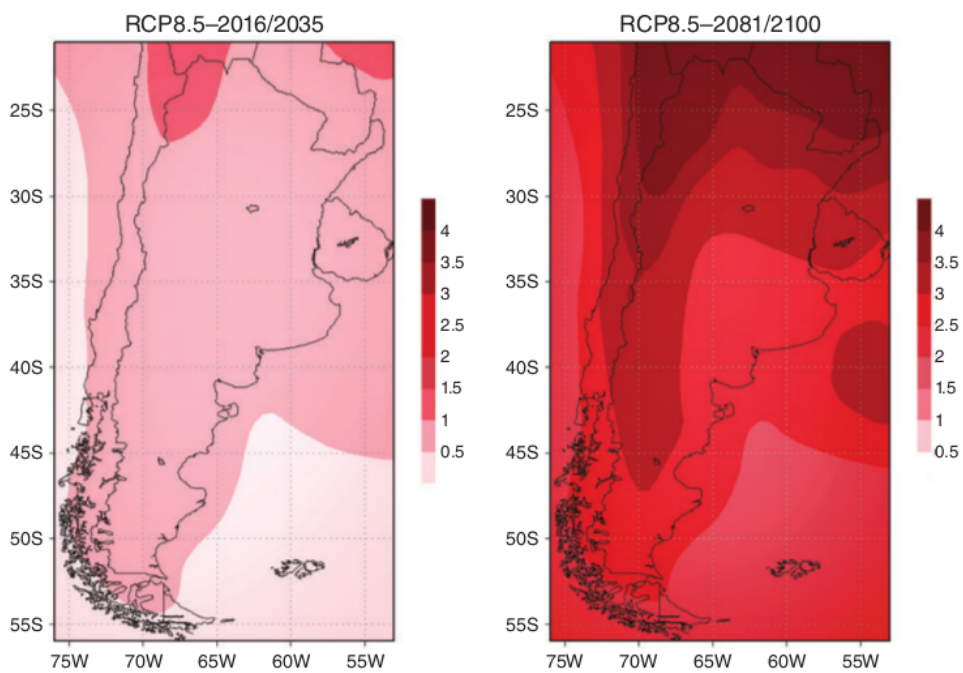
todo el territorio argentino, incluyendo la Patagonia, en el rango de 0.5-1°C (Figura 1.4) [27]. Para las regiones central y norte del país, los modelos indican un leve aumento de las precipitaciones, pero para la región andina central y, más severamente, para toda la Patagonia, la proyección es contraria (disminución de las precipitaciones en un 10 % aproximadamente) [27]. Como puede verse en la Figura 1.5, a lo largo del siglo XXI el mayor impacto de esta disminución será en la Patagonia norte (40°S-45°S). También para la región patagónica se proyectan importantes aumentos en extremos de temperatura: menor cantidad de días y noches frías, y mayor cantidad de días y noches cálidas a lo largo de todo el año [29].

En este contexto, es de fundamental importancia estudiar las respuestas a las temperaturas altas de las especies nativas, en vistas de su conservación y uso sustentable futuros. Dentro de las especies del bosque andino-patagónico, los árboles del género *Nothofagus* son los más característicos y dominan gran diversidad de ambientes (ver más abajo) [33].

Nothofagus es el único género de la familia Nothofagaceae y está compuesto por 35 especies de distribución típicamente gondwánica (ver Capítulo 3, Figura 3.1, panel 3.1a en la página 76, en el contexto de las relaciones entre familias del orden Fagales). La actual distribución disjunta de las especies (en Oceanía y el sur de Sudamérica) y la gran presencia de fósiles (principalmente hojas y polen) en la llamada provincia

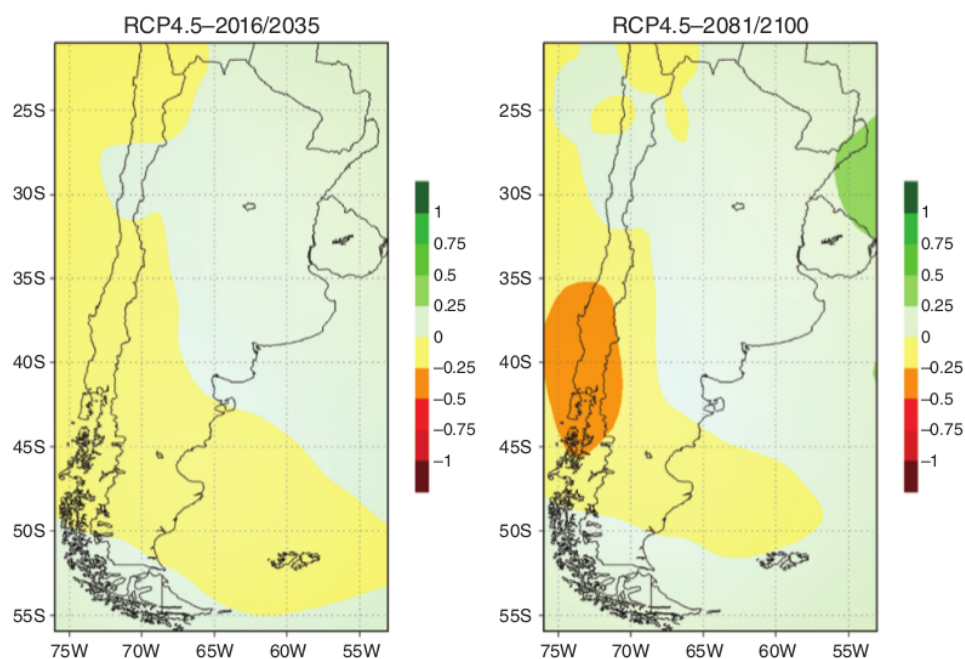


(a) Escenario moderado

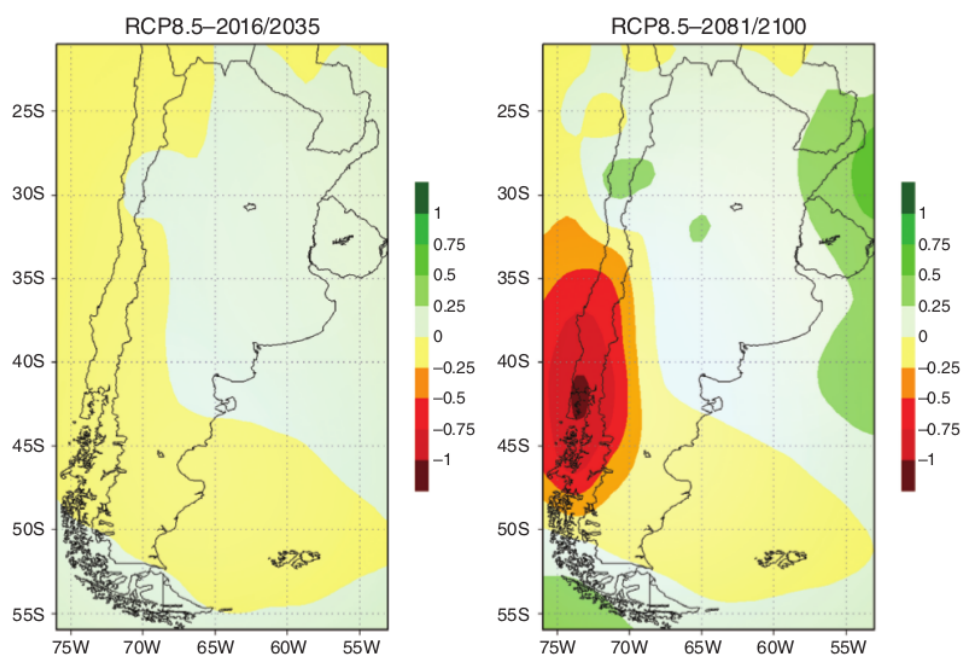


(b) Escenario severo

Figura 1.4: Tendencias para Argentina en temperatura para corto (izquierda) y largo plazo (derecha) del siglo XXI. Escala: grados centígrados en el período señalado. (a): escenario de emisiones moderado (RCP4.5). (b): escenario de emisiones severo (RCP8.5). Fuente: [27]



(a) Escenario moderado



(b) Escenario severo

Figura 1.5: Tendencias para Argentina en precipitaciones para corto (izquierda) y largo plazo (derecha) del siglo XXI. Escala: milímetros por día. (a): escenario de emisiones moderado (RCP4.5). (b): escenario de emisiones severo (RCP8.5). Fuente:

[27]

de Weddell del antiguo continente de Gondwana hizo que históricamente *Nothofagus* fuera un género intensamente estudiado por biogeógrafos, sistemáticos y ecólogos [34, 35]. Este gran interés no fue complementado tradicionalmente por estudios fisiológicos y moleculares, aspectos en los cuales estas importantes especies siguen siendo en gran medida desconocidas. En particular, la información genómica es reducida para el género [36-38], y uno de los objetivos de esta Tesis (ver más adelante) es generar los recursos genómicos más completos para el género hasta el momento, de manera de contar con información confiable que ayude a abordar una gran diversidad de problemas biológicos de forma interdisciplinaria.

En Chile y Argentina habitan 9 especies de *Nothofagus*, de las cuales 3 son siempreverdes (*N. dombeyi*, *N. betuloides* y *N. nitida*), y 6 son caducifolias (*N. pumilio*, *N. antarctica*, *N. obliqua*, *N. alpina*, *N. alessandri* y *N. glauca*) [39]. En Chile, algunas de estas especies se distribuyen desde el centro del país, a 33°S, pero en Argentina el límite septentrional de *Nothofagus* son las Lagunas de Epulauquen, en el norte neuquino, a 36°40'S. Hacia el sur de la Patagonia, *N. pumilio*, *N. antarctica* y *N. betuloides* crecen hasta Tierra del Fuego e islas circundantes, a 55°S.

En Argentina, los bosques dominados por especies de *Nothofagus* cubren gran parte del gradiente pluviométrico y todo el gradiente altitudinal de la Patagonia norte andina. En particular, hacia los 40°S, en las cuencas lacustres del centro del Parque Nacional Lanín, pueden encontrarse zonas de convivencia de *N. pumilio* (lenga), *N. obliqua* (roble pellín) y *N. alpina* (sin. *N. nervosa*, raulí). Si bien existen zonas de simpatria, el roble pellín domina los pisos altitudinales entre 600 y 800 metros sobre el nivel del mar (msnm), el raulí desde los 800 hasta los 1000 msnm, y la lenga desde los 1000 msnm y hasta el ecotono con el ecosistema alpino de alta montaña, conformando a esta latitud el límite altitudinal del bosque a aproximadamente 1700 msnm (acompañada a veces por el ñire, *N. antarctica*) [40, 41]. Esta marcada distribución en estratos altitudinales sugiere una preferencia diferencial de nichos térmicos por parte de las especies, lo cual las convierte en potencialmente sensibles en su adaptabilidad y distribución futura en el contexto del CCG, como se explicó en la Sección anterior.

En esta Tesis nos concentraremos en las especies extremas del gradiente altitudinal, es decir el roble pellín y la lenga. Las Figuras 1.6, 1.7 y 1.8 muestran las distribuciones respectivas de estas especies en Argentina. Dada la gran extensión latitudinal de la

lenga, se presenta por claridad su distribución en dos Figuras distintas.

Además de la importancia de estos árboles como sostenedores del ecosistema boscoso más austral del mundo, ambas especies tienen importancia económica regional y nacional. La madera del roble pellín, al igual que la del raulí, es de muy buena calidad y apreciada tanto con fines estructurales, como para mueblería, aberturas y artesanías [44-47]. Por otra parte, los bosques de lenga ocupan más de un millón y medio de hectáreas en Argentina, a lo largo de más de 2000 kilómetros de extensión latitudinal [48, 49] y su madera es utilizada para revestimientos, aberturas, mueblería, pisos, etc. [45, 46, 50]. En nuestro país, es explotada por su alta calidad desde el siglo pasado [51, 52].

1.3. El reloj circadiano y los procesos biológicos

1.3.1. Abordaje propuesto

En la Sección 1.1 en la página 15 se enfatizó la necesidad de abordar problemas biológicos complejos de forma interdisciplinar, de manera de entender, en última instancia, cómo una regulación molecular contribuye a la fisiología y ventaja adaptativa de un organismo en un ambiente relevante ecológicamente. La Sección anterior presentó nuestro sistema físico y biológico de estudio, el cual constituye un modelo ideal para estudiar la respuesta de las especies nativas al CCG, en particular a un aumento de temperatura. Estas conexiones complejas entre disciplinas requieren de ejemplos manejables [53], y en parte de esta Tesis nos enfocaremos en la influencia de la temperatura sobre el crecimiento de nuestras especies de interés estudiando su efecto sobre el funcionamiento del reloj circadiano. El reloj circadiano es un regulador maestro del crecimiento y desarrollo vegetal que por un lado está bien caracterizado molecularmente en especies modelo, y por otro cuenta con creciente evidencia de estar involucrado en procesos de relevancia adaptativa y de escala ecológica (ver más abajo y las Ref. [53-55]).

En el caso de las especies de *Nothofagus* de interés para esta Tesis, fue necesario desarrollar el conocimiento de todas las disciplinas involucradas, desde la medición de caracteres regulados por el reloj hasta la puesta a punto de técnicas de biología molecular, pasando por la generación y el análisis de los primeros conjuntos de datos



Figura 1.6: Distribución natural de *Nothofagus obliqua* (roble pellín) en Argentina.

Fuente: [42]

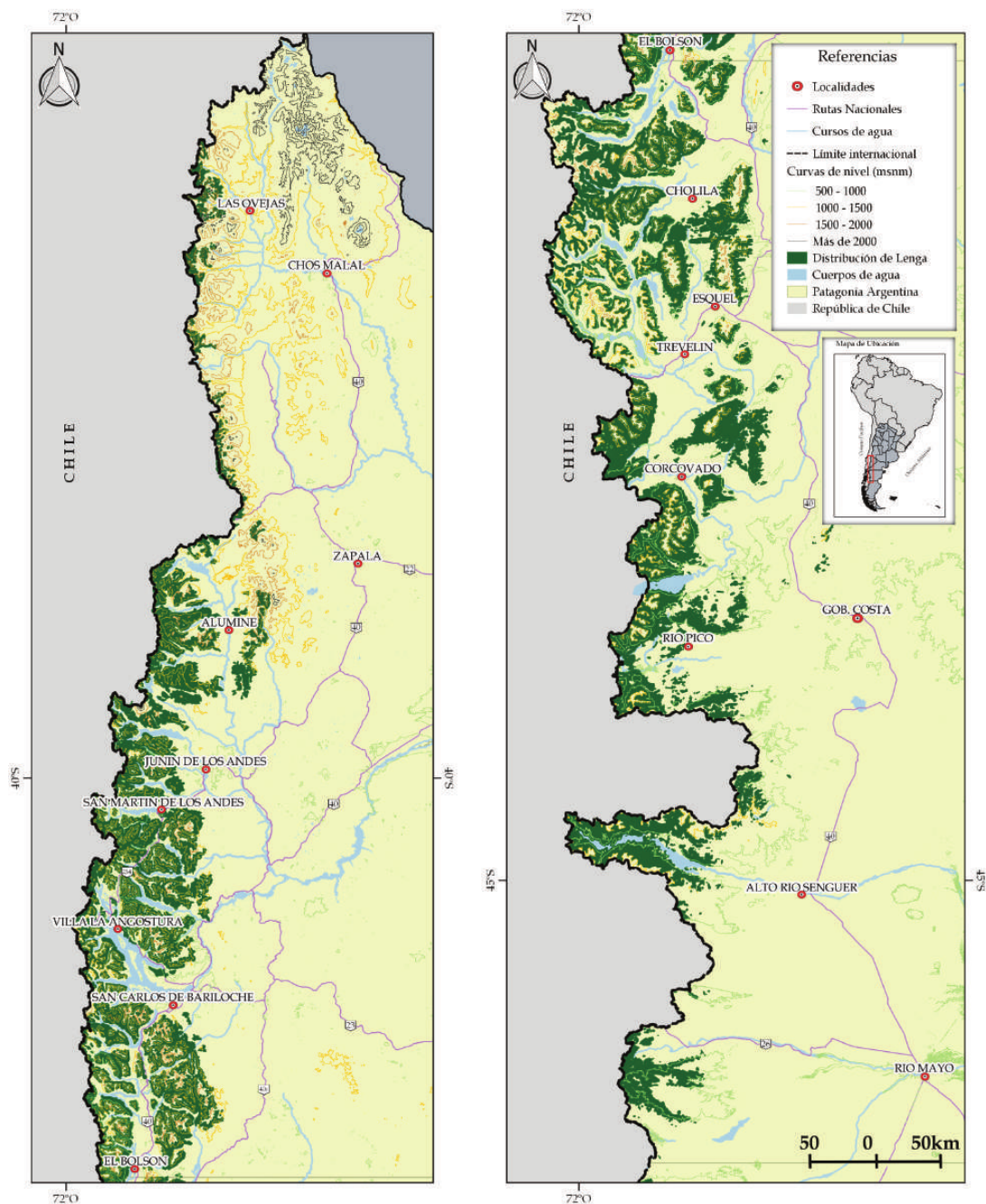


Figura 1.7: Distribución natural de *Nothofagus pumilio* (lenga) en Patagonia norte argentina. Fuente: [43]

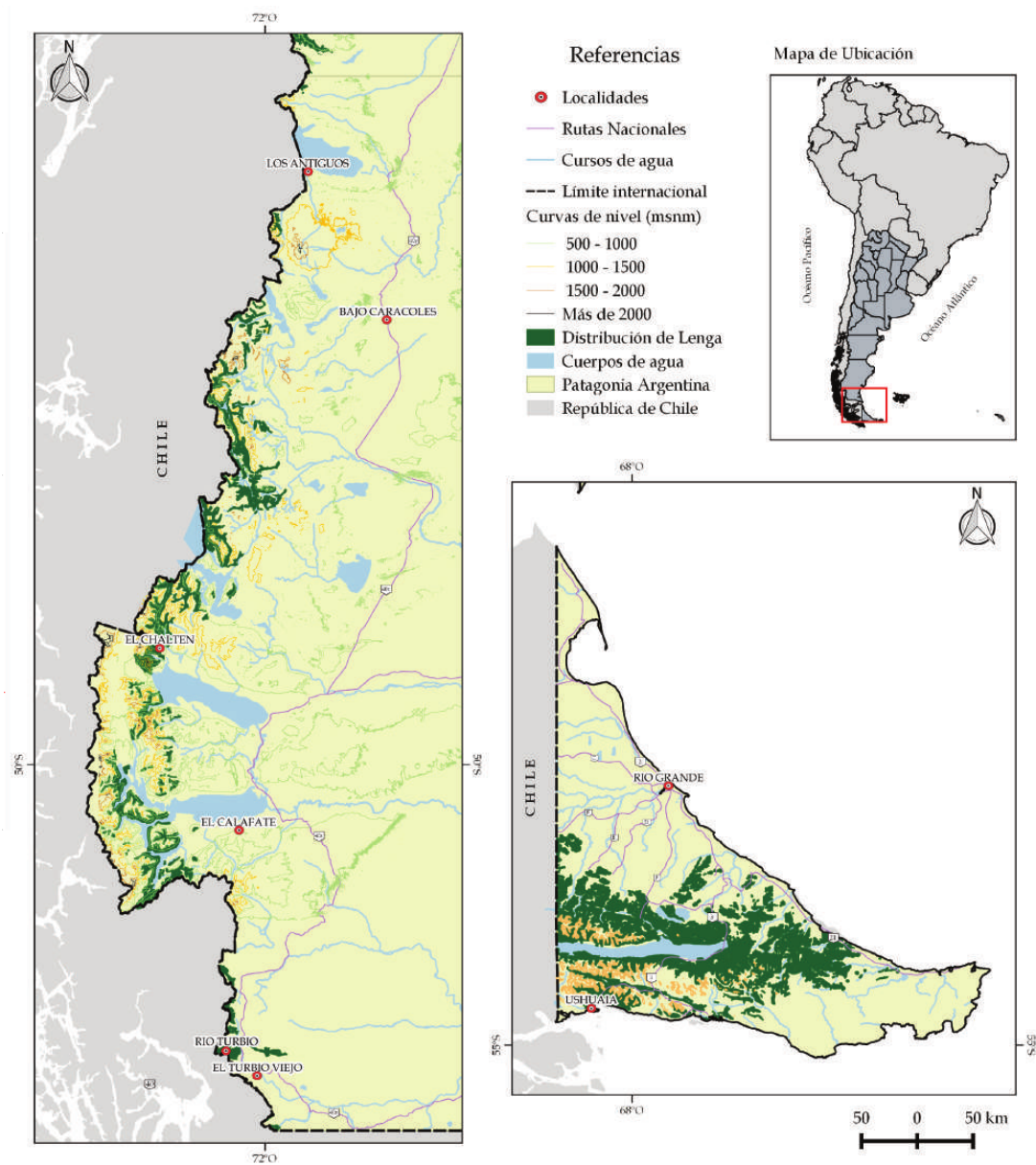


Figura 1.8: Distribución natural de *Nothofagus pumilio* (lenga) en Patagonia austral argentina. Fuente: [43]

genómicos para una de las especies (ver Objetivos a continuación, Secciones 1.4 y 1.5).

1.3.2. Ritmos circadianos

La rotación de la Tierra alrededor de su propio eje impone sobre la superficie del planeta ciclos predecibles de luz y oscuridad. Esta periodicidad es fácilmente observable también en los seres vivos, a través de procesos como el ciclo sueño/vigilia de los animales o cambios en la posición de órganos en las plantas. Muchos de estos ritmos, en prácticamente todos los organismos estudiados, están controlados por un “marcapasos” interno, el **reloj circadiano**.

El reloj circadiano es un oscilador molecular que mantiene ritmos de aproximadamente 24 horas de período, y permite a los organismos anticipar cambios cíclicos del ambiente [56]. En el caso de los organismos fotosintéticos, adaptar la fisiología al momento del día es muy importante ya que la luz solar (fuente de energía para sus vías metabólicas) está predeciblemente ausente durante parte de cada ciclo, y además de manera variable a lo largo del año [53]. De manera menos intuitiva, el reloj también controla procesos estacionales y anuales, como la inducción de la floración en plantas o la hibernación en mamíferos.

Los **ritmos circadianos** son ritmos moleculares, celulares, fisiológicos, metabólicos o de todo el organismo que son producidos en última instancia por un reloj circadiano. Para distinguir este tipo de ritmos de otros que sean sólo respuestas a señales ambientales, se define como ritmo circadiano a aquel que [56]:

1. Persiste con un período cercano a 24 horas luego de que el organismo fue transferido a un ambiente sin variaciones ambientales (condiciones de libre curso o *free-running*),
2. Puede resincronizar su período y su fase ante una nueva señal ambiental cíclica de luz o temperatura, y
3. Presenta **compensación por temperatura**, es decir, que mantiene su período en un rango térmico amplio.

El reloj circadiano se encuentra caracterizado molecularmente en gran diversidad de organismos, desde bacterias hasta árboles, pasando por hongos, insectos y mamíferos [57]. La Figura 1.9 muestra esquemáticamente los componentes mejor estudiados

del reloj circadiano vegetal. Casi todas las descripciones de elementos génicos del reloj y sus redes de regulación fueron estudiadas originalmente en *Arabidopsis thaliana*, la planta modelo de biología molecular [58].

Básicamente, el reloj circadiano de las plantas está compuesto por bucles transcripcionales que conforman un oscilador central, con genes como *TIMING OF CAB EXPRESSION 1* (*TOC1*), *LATE ELONGATED HYPOCOTYL* (*LHY*), *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*), *GIGANTEA* (*GI*), y miembros de las familias *REVEILLE* (*RVE*), *PSEUDO-RESPONSE REGULATOR* (*PRR*; entre ellos *TOC1*, también llamado *PRR1* para explicitar su pertenencia a esta familia génica) y *EARLY FLOWERING* (*ELF*). Este oscilador genera a su vez ritmicidad en la expresión de otros genes que eventualmente regulan procesos celulares y fisiológicos de manera cíclica a lo largo del día.

El primer oscilador transcripcional descrito molecularmente para plantas fue publicado en el año 2001 y estaba compuesto por *TOC1* y *LHY/CCA1* [59]. Desde ese momento, no sólo se caracterizaron nuevos bucles y elementos génicos del oscilador central y osciladores satélites [60-62], sino que se ahondó en los mecanismos de regulación postranscripcional y postraducciona del reloj circadiano, como la estabilización de proteínas o el *SPLICING* alternativo [63, 64].

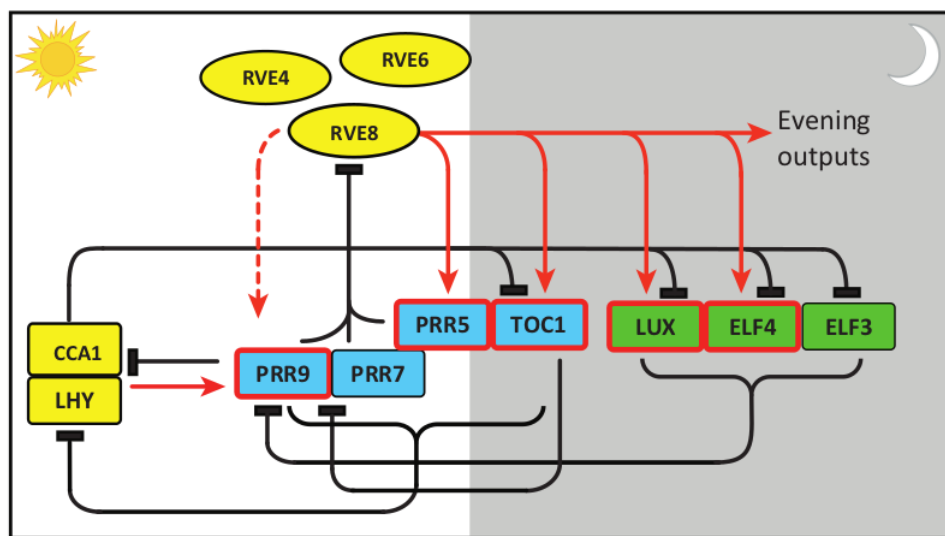


Figura 1.9: Esquema molecular del reloj circadiano en plantas. Obsérvese la regulación recíproca de *TOC1* y *LHY* (a través de los *PRR*s). Fuente: [65]

En plantas, el reloj regula procesos en todos los órganos y a lo largo de todo el ciclo de vida, como la germinación, el crecimiento vegetativo, el intercambio gaseoso, y la

floración [53-55, 66, 67]. En el Capítulo 4, Sección 4.1 en la página 96 se profundiza en los procesos fisiológicos regulados por el reloj, llamados **salidas rítmicas** o **outputs**. Se ha demostrado en plántulas de *A. thaliana* que poseer un reloj no sincronizado con los ciclos ambientales reduce el contenido de clorofila, la asimilación de carbono, el crecimiento y la supervivencia [54]. Esto es indicativo de que el reloj tiene un rol importante en el **FITNESS** y competitividad de las plantas.

La gran diversidad de ritmos en las plantas sugiere que el reloj debe regular muchos procesos bioquímicos y moleculares distintos [53], y de hecho estudios transcriptómicos han encontrado que, en *A. thaliana*, alrededor de un tercio del genoma está bajo control transcripcional del reloj circadiano [68]. Cabe destacar que el primer estudio de escala transcriptómica en organismos eucariotas para caracterizar la regulación génica por parte del reloj fue realizado en plantas, en el año 2000 por Stacey Harmer y colegas, en el grupo de Steve Kay [69]. Desde ese momento, se han realizado estudios transcriptómicos de variación diurna y circadiana en especies no modelo, como lechuga [70] y pino Oregon [71] (por citar ejemplos recientes en especies vegetales).

1.3.3. Compensación por temperatura

Como se mencionó en la Sección anterior, una de las características de los ritmos circadianos es que presentan compensación por temperatura, es decir que mantienen su oscilación y su período robustamente en un rango amplio de temperaturas. Bioquímicamente, esto resulta llamativo, ya que en última instancia los ritmos circadianos se sostienen por reacciones enzimáticas e interacciones entre proteínas, y los procesos bioquímicos aumentan en general su velocidad con la temperatura, del orden de dos o tres veces con un incremento de 10°C [72]. Además, la compensación por temperatura se da no sólo en organismos capaces de regular su temperatura interna (animales homeotermos), sino también en animales poiquilotermos, plantas, hongos, etc. Una explicación teórica para la compensación por temperatura en osciladores biológicos [73] propone que existe un balance entre reacciones que acortan y reacciones que alargan su período con la temperatura. Otra explicación, que cuenta con evidencia empírica en cianobacterias [74] y mamíferos [75], es que una reacción fundamental para mantener el período del reloj es insensible a la temperatura. Final-

mente, otros estudios en hongos [76] y mamíferos [77] proponen que la amplitud de los ritmos se acopla con las variaciones en temperatura de forma de redundar en un período estable.

Más allá del mecanismo físico-químico que permita la compensación por temperatura del reloj circadiano, en *Arabidopsis thaliana* se conocen las bases moleculares de este fenómeno. Diversos artículos estudian los roles de componentes del reloj como *CCA1*, *LHY*, *PRR7* y *9*, *GI*, *ZTL* y *FLC* (*FLOWERING LOCUS C*, un inhibidor de la vía de la floración) en la compensación por temperatura [78-84]. Cabe destacar que la proteína quinasa *CK2*, uno de los pocos componentes moleculares conservados del reloj circadiano entre los reinos de la vida, juega un rol importante en la compensación por temperatura en *A. thaliana* como antagonista de *CCA1* [82].

Son pocos los estudios sobre funcionamiento y desempeño del reloj en distintos escenarios térmicos en especies fuera de *A. thaliana*. Para el castaño común (*Castanea sativa*, familia Fagaceae), un importante árbol de los bosques templados europeos, estudios pioneros del grupo de Isabel Allona de la Universidad Politécnica de Madrid encontraron que las temperaturas bajas afectan la oscilación de los componentes del oscilador central, *TOC1* y *LHY* [85], y también de los otros miembros de la familia de los *PRR* (*PRR5*, *7*, y *9*) [86]. Un patrón idéntico fue observado en *Populus tremula* x *P. tremuloides*, un híbrido de álamo productor de fibra de papel, y se vio que este mecanismo de disrupción del reloj regula la entrada y salida de la dormición invernal, un importante rasgo adaptativo de los árboles de latitudes medias y altas [87]. Estos estudios dan cuenta de la potencial importancia del funcionamiento del reloj circadiano en la fisiología y el *fitness* de las especies arbóreas.

El rango térmico en el cual los ritmos circadianos pueden mantener su período es característico de cada especie [55]. Dado que, como se explicó anteriormente, un mal funcionamiento del reloj impacta negativamente en la fisiología y la competitividad de los individuos [54], se propone que los rangos térmicos que permiten oscilaciones robustas del reloj circadiano podrían influir en la determinación de nichos ecológicos de especies o ecotipos [55]. Estudiar las respuestas a la temperatura de especies nativas mediante el análisis de los rangos térmicos de funcionamiento de sus relojes biológicos constituye un enfoque novedoso, y toma gran importancia en ambientes con transiciones ambientales marcadas como las montañas, y en un contexto de CCG

(ver Sección 1.1 en la página 15).

1.4. Objetivo general

Recapitulando lo expuesto en las distintas Secciones de esta Introducción, podemos decir que:

- El bosque andino-patagónico es un ecosistema que se encuentra afectado por el Cambio Climático Global, y seguirá sufriendo un aumento de temperatura a lo largo del presente siglo.
- En dicho ambiente, especies arbóreas del género *Nothofagus* se distribuyen en nichos térmicos diferenciados a lo largo de un gradiente altitudinal.
- Las especies de *Nothofagus* están poco estudiadas molecularmente, lo cual dificulta el abordaje multidisciplinario de sus respuestas complejas al ambiente.
- El reloj circadiano es un regulador maestro del crecimiento y desarrollo vegetal, y fuera de un rango térmico determinado no es capaz de acoplar correctamente el funcionamiento del individuo con los ciclos ambientales, lo cual redundará en una desventaja competitiva.
- El rango de compensación por temperatura del reloj circadiano es propio de cada especie, y se propone que podría contribuir a la determinación de nichos ecológicos, así como a la capacidad de respuesta de las especies al Cambio Climático Global.

Entonces, el **Objetivo general** de esta Tesis es generar y analizar los primeros recursos genómicos para *Nothofagus pumilio*, y estudiar la respuesta de *N. obliqua* y *N. pumilio* a la temperatura a través de la exploración de su influencia sobre el funcionamiento del reloj circadiano, y su posible injerencia en la determinación de nichos ecológicos en condiciones naturales.

1.5. Objetivos específicos

Los Objetivos específicos de esta Tesis son:

1. Generar y caracterizar los primeros recursos genómicos para *N. pumilio* (lenga).
Este objetivo comprende la secuenciación, ensamblado y anotación del trans-

criptoma de hojas provenientes de ensayos circadianos, muestreados a distintos momentos del día y distintas temperaturas.

2. Construir el primer árbol filogenómico para especies del orden Fagales, como aplicación de herramientas bioinformáticas y recursos genómicos de especies no modelo.
3. Caracterizar el funcionamiento de los relojes biológicos circadianos de *N. obliqua* (roble pellín) y *N. pumilio* (lenga) a través del aislamiento de algunos genes que forman parte del oscilador central, y el estudio de su expresión en condiciones control (20°C) y en temperaturas cálidas (31-34°C). Explorar posibles *outputs* del reloj circadiano en estas especies, y estudiar la influencia de la temperatura sobre estos *outputs*.
4. Explorar la influencia del reloj circadiano sobre la expresión global de genes de lenga, y el efecto de la temperatura sobre esta regulación, a través del análisis bioinformático de expresión diferencial de genes.
5. Explorar en qué medida el comportamiento del reloj circadiano está vinculado con el establecimiento de nichos ecológicos de las especies comparando la expresión de genes del oscilador central y la acumulación de biomasa, clorofila y mortalidad de plántulas de roble y lenga en trasplantes recíprocos a lo largo de un gradiente altitudinal, donde un factor que varía es la temperatura.

Capítulo 2

Secuenciación, ensamblado y anotación de transcriptomas de *Nothofagus pumilio*

Dexter Jettser: *Those analysis droids only focus on symbols. I should think that you Jedi would have more respect for the difference between knowledge and...wisdom.*

Obi-Wan Kenobi: *Well, if droids could **think**, there'd be none of us here, would there?*

Star Wars, Episode II: Attack of the Clones

2.1. Introducción

A mediados de la década de 1970 se desarrolló un método de secuenciación del ADN [88] que fue, con diversas mejoras técnicas y de rendimiento, la fuente de secuencias nuevas de ADN durante 30 años (incluyendo los primeros genomas eucariotas secuenciados, entre ellos el de *Drosophila melanogaster* [89], el de *Arabidopsis thaliana* [90], y el genoma humano [91, 92]). La llamada “secuenciación Sanger” (por el primer autor de los papers originales que describen la técnica), a finales de la década de 1990 y principios del 2000, era realizada por electroforesis capilar en lugar de geles verticales de acrilamida, con marcadores fluorescentes en lugar de radiactivos, y con la lectura de secuencias automatizada en vez de manual. Aún con todas estas mejoras acumuladas a lo largo de décadas, el rendimiento (en número de bases de ADN secuenciadas) es varios órdenes de magnitud menor que para las técnicas desarrolladas

en los últimos diez años.

Las técnicas de secuenciación de ADN conocidas colectivamente como NGS (por sus siglas en inglés, *Next Generation Sequencing*, “secuenciación de próxima generación”) revolucionaron a mediados de la década del 2000 el campo de la genómica y permitieron estudios en profundidad de especies no modelo [93]. A diferencia de la separación electroforética de cadenas individuales de ADN (secuenciación Sanger), la secuenciación NGS determina la secuencia paso a paso a través de la adición alterna e iterativa de nucleótidos, y de manera paralelamente masiva [94]. La desventaja de esta gran cantidad de secuencias es su menor longitud y su mayor tasa de error respecto de la secuenciación Sanger. De esta manera, cabe destacar que el método de Sanger se sigue utilizando, principalmente por su gran fidelidad (baja tasa de error) y secuencias relativamente largas (entre 500 y 1000 bases) para confirmar resultados de NGS o en proyectos de pequeña escala (estudios de uno o unos pocos loci).

Son varios los tipos de secuenciación masiva (Figura 2.1), entre los cuales podemos citar (con unos pocos ejemplos en organismos vegetales y de los últimos años para cada técnica) la secuenciación de genomas completos (DNA-seq o WGS: *Whole Genome Sequencing*) [95-97], sitios sensibles a DNasa (DNase-seq) [98, 99], interacciones de proteínas con ADN (Chip-seq: *Chromatin Immunoprecipitation and Sequencing*) [100, 101], o sitios metilados (*Bisulfite sequencing*) [102, 103].

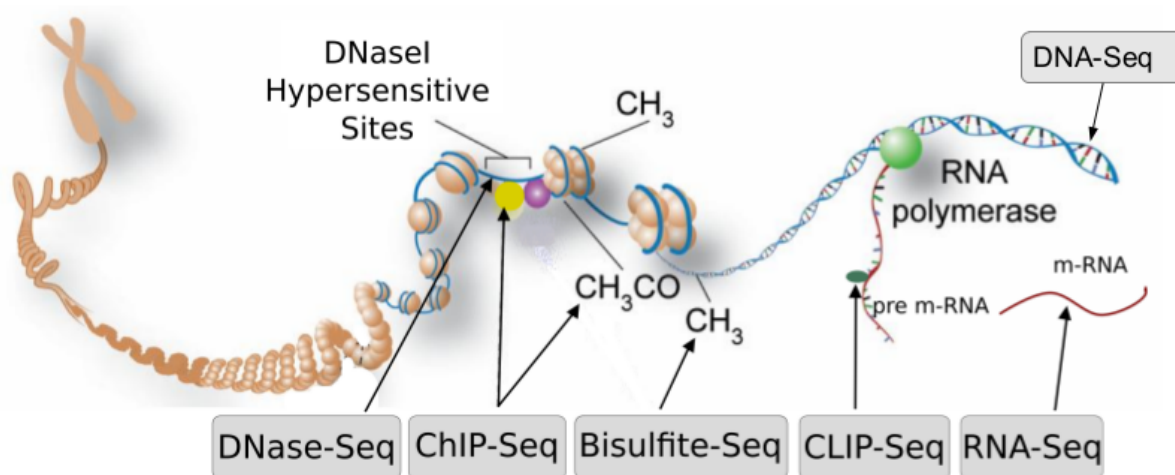


Figura 2.1: Esquema de distintas técnicas de NGS. Fuente: adaptado de [104]

En particular, la secuenciación del ARN mensajero maduro (*RNA-seq*) da como re-

sultado conjuntos de regiones transcriptas y su cantidad (expresión relativa) en una dada condición ontológica o fisiológica [105]. Un estudio como este abre las puertas a análisis de cambios globales en la expresión y su regulación (por ejemplo bajo diversas condiciones de estrés [106, 107], tejidos [108-110] o estadios fenológicos [71]), o de evaluación de polimorfismo en poblaciones naturales [36, 111, 112], y facilita en general el trabajo molecular con la especie secuenciada (diseño específico de *primers* para PCR, caracterización de familias génicas [108, 113], etc.).

Los experimentos de RNA-seq han prácticamente reemplazado al análisis de microarreglos (*microarrays*), que protagonizaron los últimos años del siglo XX y los primeros del XXI [114]. Los microarreglos son básicamente matrices de oligonucleótidos provenientes de miles de genes distintos de la especie en estudio, contra los cuales hibridan las muestras de interés, produciendo distintas intensidades de fluorescencia que son captadas por una cámara. De esta manera pueden identificarse los genes regulados en cada condición biológica cuyos ADNc fueron “presentados” a la matriz de oligonucleótidos. Si bien comparado con el RNA-seq los microarreglos son relativamente poco costosos, las desventajas son muchas [105]. En primer lugar, la fabricación de la matriz requiere conocer secuencias genómicas previamente, con el fin de sintetizar los oligonucleótidos. Esto limitó el estudio a especies con genomas secuenciados, como *Arabidopsis thaliana* [90, 115] o arroz (*Oryza sativa*, [116, 117]), o bien a otras especies tras una cuidadosa selección y curación de secuencias expresadas (ESTs), y el diseño personalizado de una matriz de oligonucleótidos, como fue hecho por investigadores de nuestro país para el girasol [118]. Por otro lado, el nivel de ruido debido a hibridación cruzada es alto, y el rango de detección es más bajo que en el RNA-seq.

El análisis computacional de un experimento típico de RNA-seq para una especie sin genoma de referencia comienza con la limpieza de las secuencias crudas (LECTURAS o *reads*), luego el **ensamblado** DE NOVO de estas en secuencias más largas (CONTIGS), y finalmente la **anotación** de dichos contigs contra una base de datos de secuencias de identidad y función conocidas. De esta manera se llega a un inventario de genes expresados en distintas condiciones experimentales, y una medida de similitud de algunas de ellas con respecto a una o varias especies de referencia, mejor estudiadas genéticamente. Los contigs anotados pueden ser utilizados para describir vías metabólicas, realizar estudios de variación genética, y todos los usos citados anteriormente en es-

ta Introducción. En caso de querer realizar un estudio de **expresión diferencial**, cada contig debe ser cuantificado en base a la cantidad de lecturas con las que fue ensamblado, y sometido a un test estadístico para detectar diferencias entre condiciones (para más detalle, ver Capítulo 4, Sección 4.2.6 en la página 104).

Las estrategias de secuenciación más comunes (como las ofrecidas por los diversos equipos de Illumina® o IonTorrent®) dan como resultado una enorme cantidad de lecturas cortas, que pueden ser decenas de millones por cada condición para un experimento típico de RNA-seq. El ensamblado de lecturas cortas para obtener contigs es una tarea muy demandante computacionalmente, y más aún en especies sin un genoma de referencia (ensamblado *de novo*).

La estrategia seguida por los ensambladores (programas específicos para ensamblar lecturas cortas) *de novo* consiste en romper las lecturas en fragmentos más pequeños de longitud k , llamados **k-MEROS**, que se solapan parcialmente a lo largo de cada lectura. La ventaja de trabajar con k-meros en vez de con las lecturas enteras radica en que el conjunto de k-meros tiene mayor representatividad a escala genómica que el conjunto de lecturas enteras, y por lo tanto es mejor para reconstruir el genoma (o transcriptoma) a ensamblar (por los supuestos teóricos de la construcción de los grafos de de Bruijn, ver más abajo y Ref. [119]). Los primeros proyectos genómicos, antes del advenimiento de las tecnologías NGS, utilizaron ensambladores que recorrían de forma hamiltoniana (visitando cada nodo una vez) un grafo cuyos nodos son los k-meros y cuyos ejes representan el solapamiento entre los nodos que conectan [120]. De esta manera, el recorrido del grafo resulta en un contig. Este método no pudo ser utilizado para los nuevos proyectos de NGS, debido a que recorrer un grafo de forma hamiltoniana es un proceso de muy difícil y lenta solución [121].

En el año 2001 se presentó una nueva forma de pensar los algoritmos de ensamblado genómico, que consiste en no ubicar a los k-meros como los nodos del grafo a recorrer, sino como los ejes del mismo. Esto permite armar contigs recorriendo una vez cada eje, lo cual es llamado un camino euleriano [122]. Recorrer un grafo de manera euleriana es mucho más fácil computacionalmente que recorrerlo de forma hamiltoniana (una vez cada nodo), y hace uso de un resultado teórico obtenido a mediados del siglo XX por el matemático holandés Nicolaas de Bruijn [123], sin relación alguna en su concepción con la genética ni la biología en general. Desde la adopción masiva de

las técnicas NGS, la gran mayoría de los ensambladores (ver citas más adelante) usa el enfoque de resolver eulerianamente un grafo de de Bruijn, como fue propuesto por Pevzner y colaboradores [122].

La Figura 2.2 muestra un esquema de un grafo de de Bruijn, en el marco de un ensamblado transcriptómico. El ensamblador comienza a recorrer el grafo, y en caso de encontrar una ramificación (diferencias en los k-meros que van formando el futuro transcripto) produce dos transcriptos separados. Dichas ramificaciones pueden tener causas variadas, como ilustra la Figura 2.2. Las llamadas *burbujas* (número “2” en el esquema) son de 1 o unos pocos nucleótidos de largo, y pueden ser debidas a un SNP o un INDEL. También puede haber sitios alternativos de comienzo de la transcripción (números “1” y “3”) o utilización alternativa de exones (números “4” y “5”). Todos estos fenómenos tienen también su posible explicación en errores técnicos, ya sea de secuenciación, contaminación con ADN de otra fuente, o unión quimérica de transcriptos de secuencia parecida pero biológicamente independientes. En todos los casos, el problema se agrava para el ensamblado *de novo*, al no contar con secuencias de referencia que puedan considerarse “correctas” técnicamente tras un proceso de validación y curación.

Son muchos los ensambladores que se pueden utilizar para RNA-seq [125-130], pero la falta de secuencias de referencia (esto es, recursos genómicos disponibles para la especie en estudio o especies cercanas) y la alta variabilidad de la expresión génica dentro y entre las muestras conlleva varios inconvenientes, incluyendo el mal ensamblado de contigs (tanto contigs quiméricos conteniendo más de

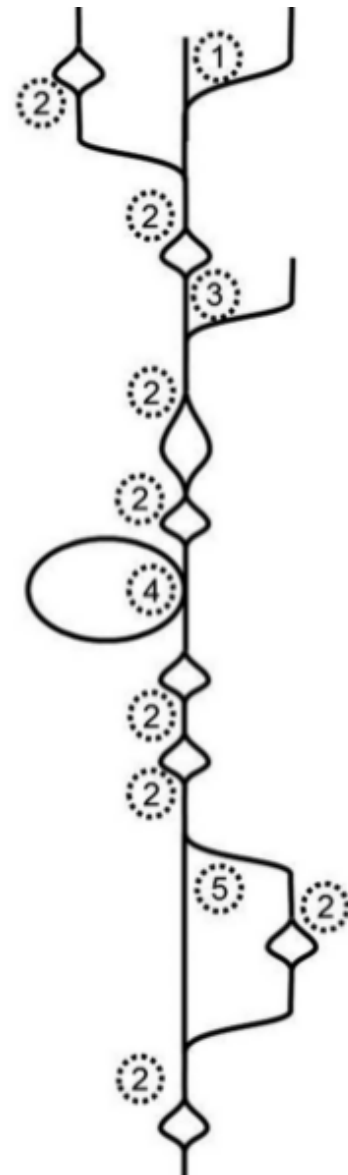


Figura 2.2: Esquema de un grafo de de Bruijn. Ver explicación en el texto.

Fuente: [124]

un transcripto, como un transcripto dividido en más de un contig), la redundancia de contigs y la especificidad de cadena [131, 132]. En consecuencia, es muy importante mejorar la calidad del ensamblado *de novo* para obtener la más confiable y mayor información posible del estudio transcriptómico de especies no modelo.

En este Capítulo se presenta la secuenciación, ensamblado y anotación del transcriptoma de *Nothofagus pumilio* en diversas condiciones fisiológicas (ver más abajo, Sección 2.2). Entre las especies del género que estudiamos (ver Capítulo 1, Sección 1.2 en la página 19) se eligió a *N. pumilio* (lenga) por ser la de mayor abundancia y distribución latitudinal, lo cual la hace estar más expuesta a condiciones climáticas contrastantes, como temperatura y fotoperíodo.

A pesar de su importancia ecológica y económica, los estudios fisiológicos y moleculares en *N. pumilio* son escasos, y los recursos genómicos para todo el género son limitados (ver Sección 1.2 en la página 19). En este Capítulo se presenta la base de datos genómica más completa para la especie hasta el momento: cuatro transcriptomas asociados a condiciones específicas de temperatura y momento del día en condiciones circadianas (ver a continuación, Sección 2.2.1).

Para el análisis y anotación de los transcriptomas se desarrolló una PIPELINE que combina programas publicados y desarrollos propios con el fin de mejorar la calidad de los ensamblados transcriptómicos *de novo*. Usando esta *pipeline*, Lengua [133], es posible mejorar el ensamblado inicial (ver Resultados, Sección 2.3.3 en la página 49) y anotarlo contra cualquier conjunto de secuencias peptídicas de las bases de datos de UniProt [134] (ver Resultados, Sección 2.3.4 en la página 54).

2.2. Materiales y Métodos

2.2.1. Material vegetal y diseño experimental

Plántulas de *N. pumilio* fueron crecidas en vivero durante 2 años en bandejas con 40 tubetes de 90 cm³ cada una, con una mezcla de partes iguales de tierra negra, arena volcánica y turba. Los experimentos fueron realizados en cámaras de crecimiento (SCE BD/600, San Carlos de Bariloche, Argentina) según el siguiente protocolo (Figura 2.3): las plantas fueron aclimatadas durante 7-10 días a 20°C con 12 horas de luz (200 $\mu\text{mol m}^{-2} \text{s}^{-1}$, Osram DULUX® L 36W) y 12 horas de oscuridad, antes de ser so-

medidas a luz constante ($100 \mu\text{mol m}^{-2} \text{s}^{-1}$, condiciones de libre curso o *free-running*) y la correspondiente temperatura del tratamiento: 20°C o 34°C . El comienzo de las condiciones de libre curso es denominado ZT0.

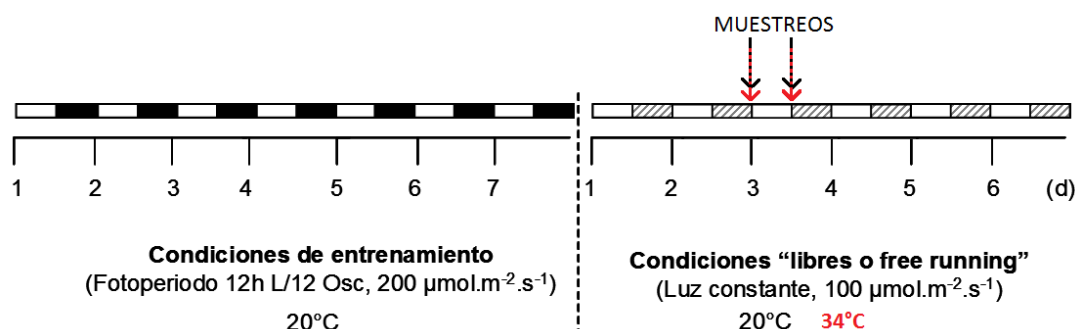


Figura 2.3: Diseño experimental y condiciones de muestreo para experimentos de RNA-seq. Rectángulos blancos: luz. Rectángulos negros: oscuridad. Rectángulos grises rayados: luz (noche subjetiva en condiciones de libre curso)

En ZT48 y ZT60 (amanecer y atardecer subjetivo respectivamente del tercer día de condiciones libres), una hoja entera de 10 individuos distintos fue cortada. Las hojas fueron inmediatamente puestas en un tubo de laboratorio (un tubo por muestreo, conteniendo las 10 hojas) y congeladas en nitrógeno líquido para evitar la degradación del ARN. A continuación los tubos fueron conservados en -80°C hasta el momento de la extracción de ARN. Ambos tratamientos de temperatura fueron realizados por duplicado en las mismas cámaras con distintas plántulas, constituyendo réplicas biológicas independientes.

2.2.2. Extracción de ARN, construcción de *libraries* y secuenciación

Las muestras fueron molidas manualmente con mortero bajo nitrógeno líquido con el objetivo de evitar el descongelamiento en todo momento. Tanto los morteros como las espátulas, mesadas, gradillas y pipetas fueron tratados con una solución de SDS 0.1 % y NaOH 0.1 M para inactivar RNasas. Para la puesta a punto de la extracción y purificación de ARN, véase el Capítulo 4, Sección 4.2.4 en la página 101. El ARN total fue extraído siguiendo el método de Chang [135], tratado con la enzima DNase I Amplification Grade de Sigma-Aldrich (catálogo AMPD1), y purificado con el kit comercial RNeasy® Plant Mini Kit (Qiagen, catálogo 74904). La cantidad y calidad del

ARN fueron comprobadas en geles de agarosa 0.8% p/V, con un espectrofotómetro NanoDrop™ 2000 (ThermoFisher Scientific, catálogo ND-2000), y con un chip Plant RNA Pico en un equipo BioAnalyzer 2100 (Agilent Technologies, catálogo G2939BA), antes de proceder a la preparación de las LIBRARIES.

El ARN mensajero maduro fue seleccionado con el kit comercial Dynabeads® mRNA DIRECT™ Micro Purification Kit (ThermoFisher Scientific, catálogo 61021), añadiendo ERCC RNA Spike-In Mix del mismo fabricante (catálogo 4456740). Las *libraries* transcriptómicas fueron construidas con el kit Ion Total RNA-seq Kit v2 (ThermoFisher Scientific, catálogo 4475936), seguido por PCR en emulsión en un Ion OneTouch™ 2 System (ThermoFisher Scientific, catálogo 4474779), usando el kit Ion PI™ Hi-Q™ OT2 200 (ThermoFisher Scientific, catálogo A26434). En total se construyeron 8 *libraries* (2 temperaturas, 2 momentos del día, 2 réplicas biológicas).

La secuenciación se llevó a cabo en el equipo ionTorrent Proton™ System (ThermoFisher Scientific, catálogo 4476610) de la Plataforma de Genómica y Transcriptómica de Bordeaux, Francia (INRA – Universidad de Bordeaux), en un total de 3 corridas (3 *libraries*, 3 *libraries*, y 2 *libraries* respectivamente) con el fin de asegurar COBERTURA y número de lecturas suficientes para el ensamblado y anotación de cada condición experimental [136-138]. Este sistema de secuenciación por semiconducción [139] produce de 60 a 80 millones de lecturas SINGLE-END por corrida, de hasta 200 pares de bases.

2.2.3. Procesamiento de datos y ensamblado *de novo*

El flujo de trabajo una vez obtenidas las lecturas crudas del secuenciador se resume en la Figura 2.4. Las lecturas de las réplicas biológicas fueron combinadas, dando lugar a 4 conjuntos de datos: ZT48_20C, ZT60_20C, ZT48_34C, y ZT60_34C (la “C” simboliza “grados Celsius”). Al tratar a cada condición biológica como un conjunto de datos por separado (con su propio ensamblado y anotaciones), genes únicos de esa condición o isoformas que intervengan en la regulación génica pueden ser estudiados, o incluso descubiertos. La combinación de todas las condiciones en un único ensamblado es útil para estudios de expresión diferencial (ver Capítulo 4, Sección 4.2.6 en la página 104), pero conlleva el riesgo de fusionar o no distinguir adecuadamente entre isoformas.

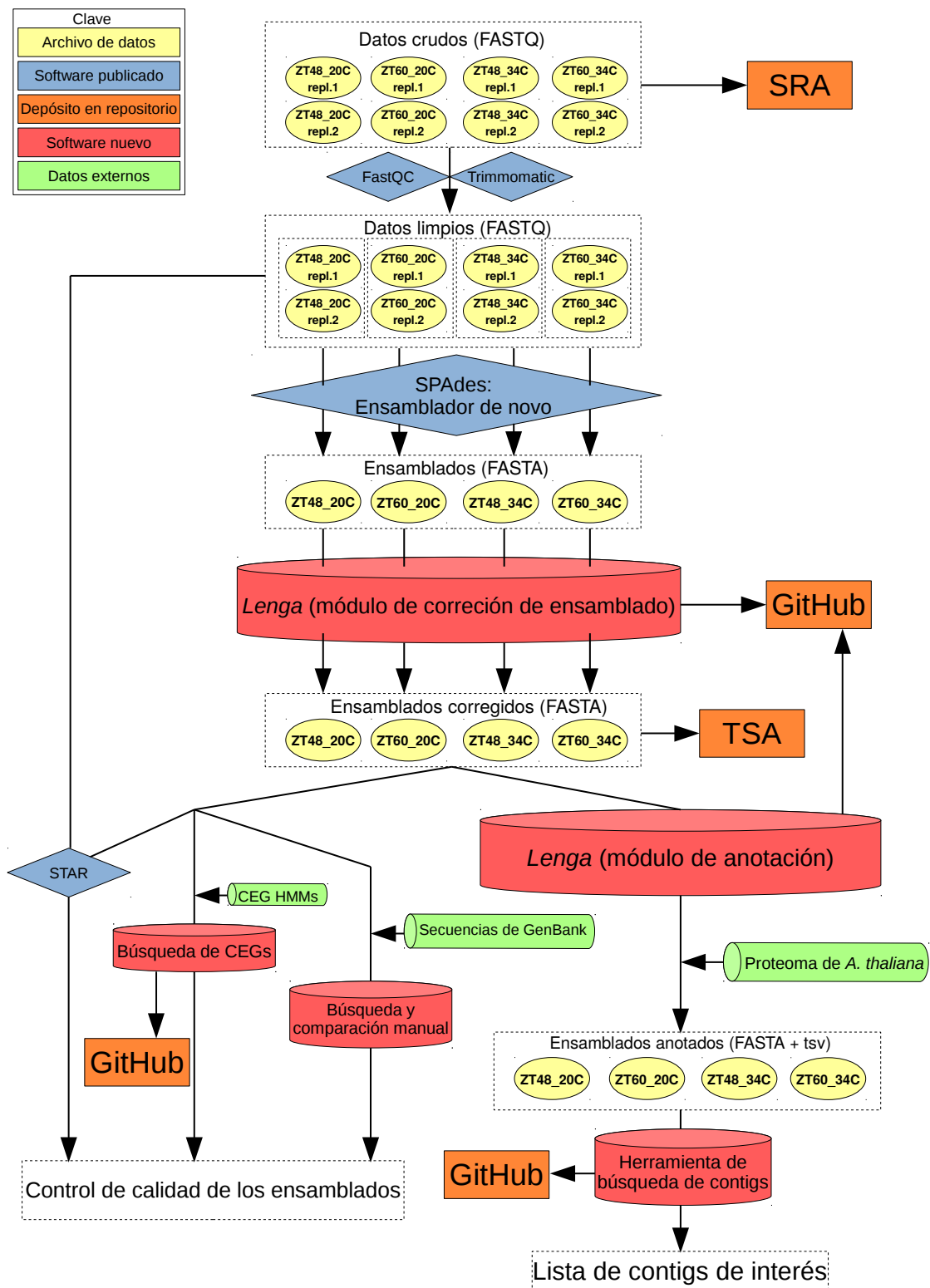


Figura 2.4: Resumen esquemático del Capítulo

Sobre cada conjunto de datos se efectuó un control de calidad con FastQC [140] y un corte (*trimming*) con Trimmomatic [141] (versión 0.33; parámetros: LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:36) y el programa trimmer de la suite Fastx [142] (versión 0.0.13; parámetros: -Q33 -l 250).

Los 4 conjuntos de datos limpios fueron ensamblados por separado con los programas Trinity [127, 143] (versión 2.6.6; parámetros: --SS_lib_type F --KMER.SIZE 32) y SPAdes [125] (versión 3.11.0; parámetros: --rna -iontorrent -k67 --ss-fr). El valor de *k* para Trinity fue elegido por ser el más alto posible en dicho software, y para SPAdes fue elegido luego de varias pruebas con diferentes longitudes de k-mero (cinco en total, de 21 a 77). Los contigs ensamblados por SPAdes (ver Resultados, Sección 2.3.3 en la página 49 y Discusión, Sección 2.4.1 en la página 60, para los motivos de elección de este ensamblador frente a Trinity) fueron analizados con *Lenga*, con el fin de resolver algunos problemas típicos de los ensamblados transcriptómicos *de novo* (ver Introducción y Ref. [131] y [132]). El módulo de corrección de ensamblado de *Lenga* alarga contigs que se superponen considerablemente, elimina contigs altamente redundantes, y orienta a todas las secuencias en sentido *forward*.

2.2.4. Control de calidad de los ensamblados

Se usaron estadísticas básicas tales como número y largo promedio de contigs para comparar las salidas de Trinity con las de SPAdes, y luego de la elección de este último frente a Trinity, las salidas de SPAdes con las de *Lenga* para cada conjunto de datos. A continuación, la calidad de los ensamblados fue comprobada por cuatro métodos diferentes.

En primer lugar se cuantificaron, para cada ensamblado, las lecturas de las dos *libraries* utilizadas, con el programa Salmon [144] (versión 0.8.1; parámetros --seqBias -l U). Así, cada contig tiene, para cada una de las dos *libraries* a partir de las cuales fue ensamblado, una medida de cuantificación, llamada TPM (por sus siglas en inglés, *Transcripts Per Million*, transcritos por millón). El número de TPM de un contig para un juego de lecturas dado (en este caso, provenientes de una *library*) indica cuán abundante sería el transcripto representado por dicho contig en el conjunto de todos los transcriptos de la condición experimental cuantificada. Por esto, se espera que los TPM sean similares para los mismos contigs en réplicas biológicas. De esta manera, se

evaluó la correlación de las réplicas biológicas con las que fue hecho cada ensamblado, mediante la prueba de correlación de Pearson en el software R [145].

A continuación, se realizó un MAPEO de las lecturas limpias contra los ensamblados usando el programa STAR [146] (versión 2.4.2a; parámetros para indexado: `--runMode genomeGenerate --genomeSAindexNbases 11`; parámetros para mapeo: `--alignIntronMax 21 --outFilterScoreMinOverLread 0.4 --outFilterMatchNminOverLread 0.4`) como una medida del porcentaje de lecturas usadas para los ensamblados.

Para evaluar la completitud funcional de los ensamblados, estos fueron comparados contra 3 bases de datos de genes conservados. Por un lado, un conjunto de 248 genes conservados eucariotas (CEGs, por sus siglas en inglés *Core Eukaryotic Genes*, tomados de la *pipeline* CEGMA [147]) fue buscado en los ensamblados con un SCRIPT de Python que corre los 248 modelos génicos (ver más abajo) en las secuencias de entrada y comprueba cuántos (y cuáles) contigs superan el umbral de similitud para cada CEG [148]. Los modelos génicos y los umbrales se obtuvieron de http://korflab.ucdavis.edu/datasets/genome_completeness/. Por otro lado, se utilizó el programa BUSCO [149] (por sus siglas en inglés *Benchmarking Universal Single-Copy Orthologs*) para buscar dos bases de datos (una de 1440 genes conservados de plantas, y otra de 303 genes conservados eucariotas) de modelos de genes ORTÓLOGOS de copia única en los ensamblados. Las bases de datos BUSCO fueron descargadas de <https://busco.ezlab.org/>.

Las bases de datos de genes conservados mencionadas en el párrafo anterior están compuestas de modelos génicos llamados HMMs (por sus siglas en inglés, *Hidden Markov Models*, modelos ocultos de Markov). En pocas palabras, un modelo oculto de Markov contiene el “perfil” de una molécula (en nuestro caso, proteínas), realizado a partir de un ALINEAMIENTO múltiple de especies donde dicha proteína está descrita [150, 151]. Los HMM constituyen una alternativa a los métodos de búsqueda de HOMOLOGÍA por alineamiento pareado (*pairwise*), como BLAST [152]. La diferencia radica en que mientras una herramienta como BLAST asigna un puntaje y una similitud entre la secuencia de entrada y una (o cada una de muchas) secuencia de la base de datos, un método con HMM computa la similitud con un perfil molecular generado a partir de secuencias homólogas en varias especies. Además, un HMM modela la dependencia de las posiciones (aminoácidos en nuestro caso) entre sí, computando la probabilidad

de encontrar un cierto aminoácido dados los anteriores en la secuencia.

Finalmente, cada ensamblado fue usado como referencia para una búsqueda de 5 secuencias de *N. pumilio* disponibles en la base de datos GenBank, para comprobar completitud e identidad de secuencias (Tabla 2.1).

ID	Nombre de la secuencia	Lar- go	Ref.
MF446382	<i>Nothofagus pumilio</i> 4-coumarate-CoA ligase activity (Cum) gene, partial cds	481	[153]
MF446370	<i>Nothofagus pumilio</i> chloroplastic drought-induced stress protein of 32kD (cdsp32) gene, partial cds	234	[153]
MF446388	<i>Nothofagus pumilio</i> protein kinase (SOS2) gene, partial cds	451	[153]
MF446379	<i>Nothofagus pumilio</i> late embryogenesis abundant protein (Lea2) gene, partial cds	717	[153]
NFGCPRBCLS	<i>Nothofagus pumilio</i> ribulose 1,5-bisphosphate carboxylase large subunit (rbcL) gene, partial cds	1345	[154]

Tabla 2.1: Secuencias de GenBank utilizadas para analizar la completitud e identidad de secuencias ensambladas de *N. pumilio*

2.2.5. Anotación

Una vez que los cuatro ensamblados fueron corregidos por el módulo de corrección de ensamblados de *Lenga*, fueron anotados contra el proteoma de *Arabidopsis thaliana* (<http://www.uniprot.org/proteomes/UP000006548>), utilizando como entrada el marco de lectura abierto (ORF, por sus siglas en inglés *Open Reading Frame*) más largo por cada uno de los 6 posibles marcos por cada contig. Se utilizaron secuencias aminoacídicas (potenciales proteínas) por tratarse de regiones codificantes y por anotarse contra una especie relativamente lejana. Se eligió *A. thaliana* por su fuerte tradición como especie de planta modelo [58] y la gran cantidad de recursos disponibles en la red, como atlas de expresión bajo diversas condiciones de estrés [155] o curvas de expresión temporal en diversas condiciones lumínicas y circadianas [156]. Ver Capítulo 4, Secciones 4.2.5 en la página 103 y 4.2.6 en la página 104 para algunas aplicaciones posibles gracias a la elección de este modelo.

El módulo de anotación de *Lenga* busca en la base de datos de referencia (en nuestro caso, el proteoma de *A. thaliana*) utilizando BLAT [157] y genera un archivo de anotaciones con el mejor resultado para cada contig exitosamente anotado contra la referencia. El archivo incluye el nombre del contig, su cobertura, marco de lectura y longitud, y el nombre y longitud del gen correspondiente de la referencia. Además, cada resultado recibe un puntaje (*score*) de anotación que relativiza la similitud entre el contig de entrada y la referencia respecto de la longitud de la secuencia de entrada.

Finalmente, *Lenga* agrega a cada contig anotado los términos de Gene Ontology [158] correspondientes al gen de *A. thaliana* contra el cual anotó. Los términos de GO fueron obtenidos de la Gene Ontology Annotation Database del Laboratorio Europeo de Bioinformática (EBI; <https://www.ebi.ac.uk/GOA>).

2.2.6. Predicción de microsatélites

Para cada ensamblado se buscaron potenciales MICROSATÉLITES (SSRs) utilizando el programa MISA (MICroSATellite identificación tool) [159] con los siguientes criterios: mínimo de 6 unidades repetidas para dinucleótidos, mínimo de 5 unidades repetidas para tri-, tetra-, penta- y hexanucleótidos, y distancia máxima de 10 bases entre dos SSRs distintos para ser considerados un SSR compuesto.

2.3. Resultados

2.3.1. Depósito público de datos

El proyecto (englobando toda la iniciativa de secuenciación y análisis) fue depositado en BioProject, la base de datos de proyectos de secuenciación del NCBI (National Center for Biotechnology Information, de la Biblioteca Nacional de Medicina de los Institutos Nacionales de Salud de los EE.UU.) con el identificador único PRJ-NA414196.

Cada muestra (*library* secuenciada) constituye una BioSample de dicho proyecto, y las secuencias crudas fueron enviadas al repositorio SRA (Sequence Read Archive), también del NCBI (Tabla 2.2).

Finalmente, cada ensamblado definitivo (ver más abajo, Sección 2.3.3) fue deposi-

BioProject	Estudio	BioSample	Corrida de SRA	Nombre
PRJNA414196	SRP133997	SAMN07782031	SRR6808716	ZT48_20C_1
		SAMN07782032	SRR6808715	ZT48_20C_2
		SAMN07782033	SRR6808714	ZT60_20C_1
		SAMN07782034	SRR6808717	ZT60_20C_2
		SAMN07782035	SRR6808721	ZT48_34C_1
		SAMN07782036	SRR6808720	ZT48_34C_2
		SAMN07782037	SRR6808719	ZT60_34C_1
		SAMN07782038	SRR6808718	ZT60_34C_2

Tabla 2.2: Depósito público de los datos crudos

tado en el repositorio TSA (Transcriptome Shotgun Assembly) del NCBI (Tabla 2.3).

BioProject	BioSamples	Corridas de SRA	Accesión de TSA	Nombre
PRJNA414196	SAMN07782031	SRR6808716	GGJD01000000	ZT48_20C
	SAMN07782032	SRR6808715		
	SAMN07782033	SRR6808714	GGJE01000000	ZT60_20C
	SAMN07782034	SRR6808717		
	SAMN07782035	SRR6808721	GGJV01000000	ZT48_34C
	SAMN07782036	SRR6808720		
	SAMN07782037	SRR6808719	GGJW01000000	ZT60_34C
	SAMN07782038	SRR6808718		

Tabla 2.3: Depósito público de los ensamblados

2.3.2. Control de calidad de la secuenciación

Las lecturas crudas fueron analizadas con FastQC [140] para determinar su contenido de GC y calidad ($>Q20$), y con Python para determinar su número y longitud promedio (ver Tabla 2.4).

Luego del filtrado (ver Sección 2.2.3 en la página 42 y Tabla 2.5) se computaron los mismos parámetros. Las lecturas recortadas tienen idéntico contenido de GC y mejor calidad ($Q20$), a expensas de lecturas más cortas en promedio (Tabla 2.4). Estas

Muestra	Número de lecturas (millones)	Largo promedio de lecturas	Q20(%)	Contenido de GC (%)
ZT48_20C	56.51	131.99	91.46	45
ZT48_20C_recortadas	45.30	109.99	99.68	45
ZT60_20C	44.73	136.59	92.04	45
ZT60_20C_recortadas	31.28	113.24	99.77	45
ZT48_34C	55.13	133.93	93.16	45
ZT48_34C_recortadas	36.99	114.40	99.83	45
ZT60_34C	66.45	130.21	89.17	45
ZT60_34C_recortadas	51.13	104.78	99.40	45

Tabla 2.4: Métricas de calidad de las lecturas antes y después del filtrado y recorte para cada conjunto de datos

lecturas fueron las utilizadas para el ensamblado *de novo*.

Muestra	Número de lecturas crudas (millones)	Número de lecturas limpias (millones)	Lecturas limpias (%)
ZT48_20C	56.51	45.30	80.16
ZT60_20C	44.73	31.28	69.93
ZT48_34C	55.13	36.99	67.09
ZT60_34C	66.45	51.13	76.95

Tabla 2.5: Filtrado de lecturas

2.3.3. Control de calidad del ensamblado

La Tabla 2.6 muestra, para cada ensamblado, el número de contigs y su longitud promedio utilizando Trinity o SPAdes como ensamblador *de novo*. Los ensamblados de Trinity presentan mucha mayor cantidad de contigs (alrededor de 5 veces más), y estos son en promedio más cortos que los ensamblados por SPAdes. Ver Discusión, Sección 2.4 en la página 60 para una posible causa de este fenómeno. A partir de este punto se prosiguió para todos los análisis de este Capítulo con los ensamblados de SPAdes.

Ensamblado	Software	Número de contigs	Longitud promedio
ZT48_20C	Trinity	123526	616.08
	SPAdes	25740	903.83
ZT60_20C	Trinity	169040	468.19
	SPAdes	22382	897.33
ZT48_34C	Trinity	131851	618.56
	SPAdes	25410	887.71
ZT60_34C	Trinity	105786	510.75
	SPAdes	21527	748.11

Tabla 2.6: Comparación entre los ensambladores Trinity y SPAdes

La correlación entre las *libraries* de una misma condición biológica utilizadas para cada ensamblado fue muy buena, entre 70 y 85 % (Figura 2.5). Se observa correlación más fuerte en los rangos medios y altos de expresión, lo cual es consistente con estudios con datos simulados y empíricos respecto de la replicabilidad técnica y la cuantificación de lecturas sobre ensamblados transcriptómicos [137, 138].

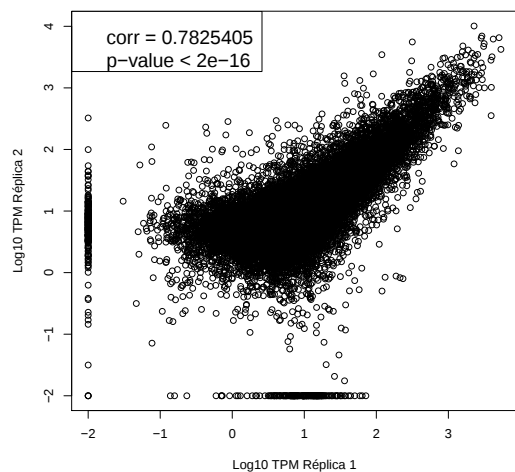
Comparadas con las salidas sin modificar de SPAdes, las salidas de *Lenga* tuvieron menos contigs y más largos, con una mayor longitud promedio, longitud mediana, y un mayor contig más largo (Tabla 2.7). La distribución de longitud de los contigs totales de cada ensamblado se muestra en la Figura 2.6 (líneas negras sólidas). Las salidas de *Lenga* fueron usadas como los ensamblados definitivos para el resto de los análisis.

Se pudieron mapear entre un 85 y un 90 % de las lecturas recortadas a los ensamblados definitivos usando STAR [146] (Tabla 2.8), lo cual constituye un alto porcentaje de utilización.

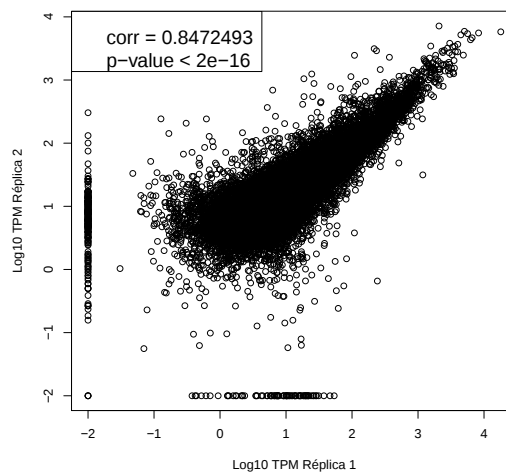
Por otro lado, todos los ensamblados presentaron el 95 % o más de los 248 CEGs (Tabla 2.9). Combinados, los cuatro ensamblados resultaron en un 99 % de los CEGs (246 de 248).

Respecto de los BUSCOs, los porcentajes fueron más bajos (Tabla 2.9; ver Discusión, Sección 2.4.1 en la página 60).

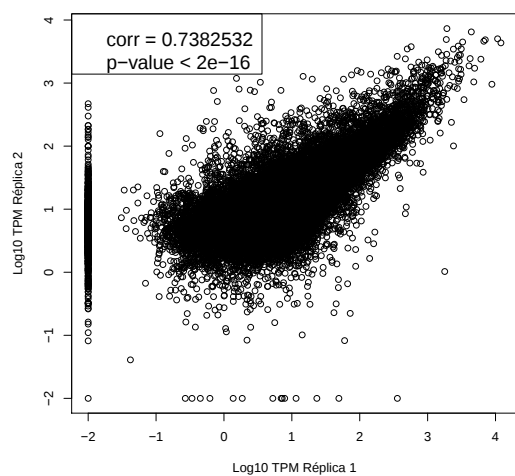
Finalmente, en las cinco comparaciones manuales entre secuencias de GenBank de *N. pumilio* y nuestros cuatro ensamblados, se encontró un contig con alta identidad, y



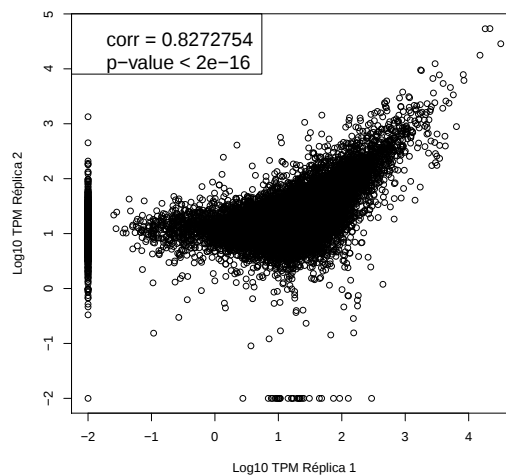
(a) ZT48_20C



(b) ZT60_20C



(c) ZT48_34C



(d) ZT60_340C

Figura 2.5: Correlación entre réplicas biológicas para cada ensamblado. Cada punto representa un contig. Se indica el coeficiente de correlación, y el p-valor (H_0 : no hay correlación entre las variables)

Ensamblado		Número de contigs	Número de bases (millones)	Longitud promedio	Longitud mediana	Contig más corto	Contig más largo
ZT48_20C	S	25740	23.26	903.83	721	333	5453
	L	20765	24.66	1187.37	926	333	7780
ZT60_20C	S	22382	20.08	897.33	712	332	6455
	L	18123	21.22	1170.88	890	332	9220
ZT48_34C	S	25410	22.56	887.71	698	337	5277
	L	20348	24.29	1193.53	920	337	8176
ZT60_34C	S	21527	16.10	748.11	580	337	5413
	L	19140	16.15	843.53	646	337	6766

Tabla 2.7: Métricas de los ensamblados *de novo*. **S** indica el ensamblado de SPAdes sin modificar, y **L** tras el procesamiento con *Lenga*

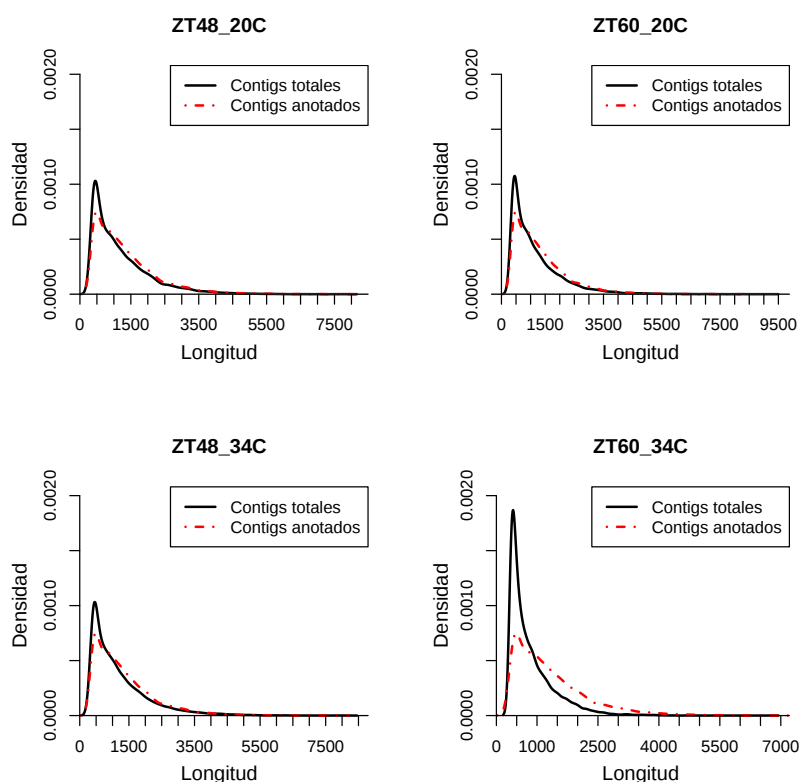


Figura 2.6: Distribución de largo de contigs para cada ensamblado

Ensamblado	Mapeadas único	Mapeadas múltiples	Mapeadas demasiados	Mapeadas total
ZT48_20C	58.59	28.19	3.78	90.56
ZT60_20C	58.06	28.81	2.87	89.74
ZT48_34C	56.24	27.44	5.20	88.88
ZT60_34C	63.23	20.67	2.62	86.52

Tabla 2.8: Métricas de mapeo. **Mapeadas único:** Porcentaje de lecturas mapeadas a un único locus. **Mapeadas múltiples:** Porcentaje de lecturas mapeadas a múltiples loci. **Mapeadas demasiados:** Porcentaje de lecturas mapeadas a demasiados loci.

Mapeadas total: Porcentaje total de lecturas mapeadas

Ensamblado	CEGs (%)	BUSCOs de plantas (%)	BUSCOs de eucariotas (%)
ZT48_20C	97.58	39.10	60.70
ZT60_20C	96.77	34.80	57.40
ZT48_34C	97.98	33.30	60.80
ZT60_34C	97.17	23.30	47.90

Tabla 2.9: Número y porcentaje de genes conservados

más del 98.5% de la secuencia coincidió exactamente (Tabla 2.10).

Ensamblado	<i>Cum</i>	<i>CDSP32</i>	<i>SOS2</i>	<i>Lea2</i>	<i>rbcL</i>
ZT48_20C	99.73	99.57	99.11	98.77	98.72
ZT60_20C	99.76	100	98.89	99.17	98.81
ZT48_34C	99.28	100	98.89	99.16	98.81
ZT60_34C	99.01	100	98.89	99.19	98.81

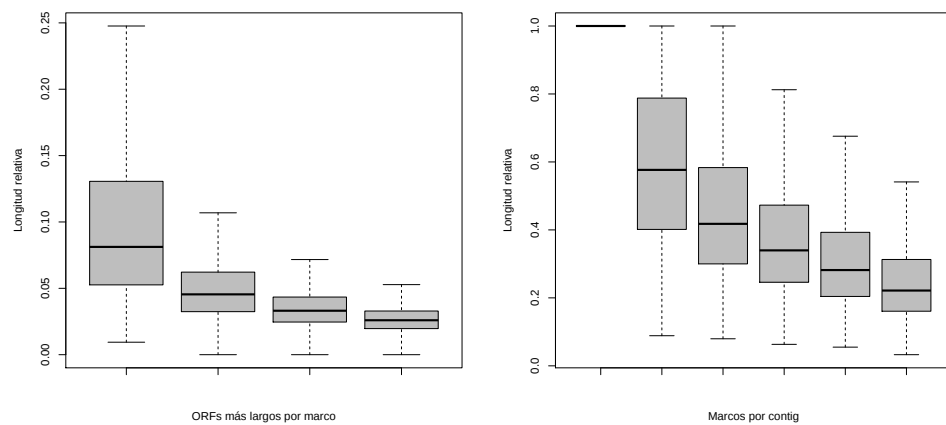
Tabla 2.10: Comparación de secuencias ensambladas y obtenidas de GenBank. Se indica el porcentaje de identidad entre la secuencia de referencia y la región alineada del contig encontrado en cada ensamblado

2.3.4. Control de calidad de la anotación

Para anotar los contigs de lenga contra *Arabidopsis thaliana* se utilizaron como secuencias de búsqueda, para cada contig, el ORF más largo encontrado en cada posible marco de lectura; es decir, un total de 6 secuencias por cada contig (3 marcos de lectura en sentido *forward* y 3 en sentido *reverse*). La Figura 2.7, panel 2.7a muestra el largo relativo al largo total del contig de los 4 ORFs más largos de cada marco para cada contig. Como se ve, en promedio existe un ORF que es más largo que los demás, y esta diferencia es significativa (test de Wilcoxon: p-valor $< 2.2 \times 10^{-16}$ entre la media más alta y la siguiente, calculado con el software R [145]). La Figura 2.7 muestra los resultados obtenidos para el ensamblado ZT48_20C, pero los otros ensamblados muestran exactamente el mismo patrón, como se puede observar en el Apéndice B, Figuras B.1 a B.3 en las páginas 197-198 (para los ensamblados ZT60_20C, ZT48_34C y ZT60_34C respectivamente).

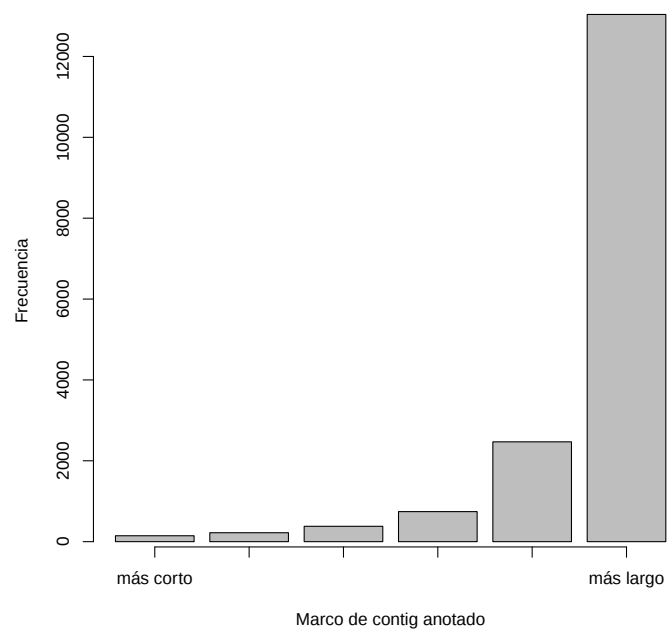
Todos los ensamblados mostraron un alto porcentaje de contigs anotados (mayor al 80%, Tabla 2.11). El puntaje de anotación propuesto (ver Sección 2.2.5 en la página 46) fue en general muy alto (Tabla 2.11 y Figura 2.8), y la distribución de longitud de los contigs anotados no muestra sesgos respecto de la de los contigs totales (Figura 2.6, líneas rojas punteadas).

Si bien para nuestros intereses nos pareció adecuado anotar nuestros contigs contra el proteoma de *A. thaliana* (como fue explicado en la Sección 2.2.5 en la pági-



(a) ORFs dentro de cada marco

(b) Marcos dentro de cada contig



(c) ORFs más largos de cada frame en la anotación

Figura 2.7: Justificación del uso del ORF más largo por marco y por contig para la anotación (ensamblado ZT48_20C). Ver explicación en el texto

Ensam- blado	Con- tigs	Contigs anotados	Contigs anotados (%)	Proteínas únicas de <i>A. thaliana</i>	1° cuartil PA	Media- na PA	3° cuartil PA
ZT48_20C	20765	16965	81.70	10714	0.65	0.85	0.95
ZT60_20C	18123	15388	84.91	9878	0.66	0.85	0.95
ZT48_34C	20348	16566	81.43	10307	0.64	0.84	0.94
ZT60_34C	19140	15650	81.77	10218	0.66	0.86	0.96

Tabla 2.11: Métricas de la anotación de los ensamblados contra *Arabidopsis thaliana*.

PA: Puntaje de Anotación

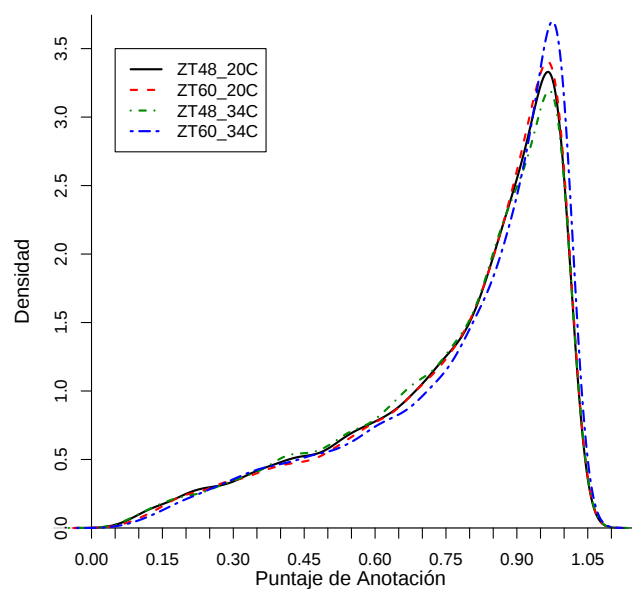


Figura 2.8: Distribución del puntaje de anotación para cada ensamblado

na 46), realizamos también una anotación para cada ensamblado contra la base de datos SwissProt, parte de UniProtKB [134]. Esta base de datos cuenta con más de medio millón de secuencias proteicas, provenientes de todo el árbol de la vida, curadas y anotadas manualmente. Como se ve en la Tabla 2.12, en promedio sólo un 8 % de los contigs no anotados contra *A. thaliana* lo hicieron contra alguna secuencia de SwissProt, añadiendo tan sólo alrededor de 1.5 % a la cuenta de contigs anotados contra *A. thaliana*.

Por último, los mismos contigs que no anotaron contra *A. thaliana* fueron anotados contra toda la base de datos de UniProtKB, compuesta por SwissProt y TrEMBL. Esta última consiste en más de 115 millones de secuencias proteicas anotadas automáticamente, de todos los reinos de la vida. Si bien más de la mitad de los contigs anotaron contra una secuencia de UniProtKB, en el 60 % de los casos se trató de una entrada anotada como **uncharacterized protein** (proteína no caracterizada), lo cual no agrega información biológica valiosa.

Ensamblado	Contigs	No <i>A. thaliana</i> (#)	No <i>A. thaliana</i> (%)	No <i>A. thaliana</i> , sí SP (#)	No <i>A. thaliana</i> , sí SP (%)	Suma a la cuenta de anotados (%)
ZT48_20C	20765	3818	18.39	219	5.74	1.05
ZT60_20C	18123	2753	15.19	155	5.63	0.86
ZT48_34C	20348	3794	18.65	490	12.92	2.41
ZT60_34C	19140	3507	18.32	348	9.92	1.82

Tabla 2.12: Métricas de la anotación de los ensamblados contra SwissProt. **No *A. thaliana* (#)**: Número de contigs no anotados contra *A. thaliana*. **No *A. thaliana* (%)**: Porcentaje de contigs no anotados contra *A. thaliana*. **No *A. thaliana*, sí SP (#)**: Número de contigs no anotados contra *A. thaliana* pero sí contra SwissProt. **No *A. thaliana*, sí SP (%)**: Porcentaje de contigs no anotados contra *A. thaliana* pero sí contra SwissProt

La Figura 2.7, panel 2.7b (y en el Apéndice B, Figuras B.1 a B.3 en las páginas 197-198 para los otros ensamblados) muestra que, de los 6 ORFs usados como secuencias de entrada para cada contig (el más largo de cada uno de los 6 marcos), hay uno que es significativamente más largo (test de Wilcoxon: p-valor $< 2.2 \times 10^{-16}$ entre la media más alta y la siguiente, calculado con el software R [145]). Además, como muestra la Figura 2.7c, es ese ORF más largo entre los 6 el que anota contra *Arabidopsis thaliana* la gran mayoría de las veces (entre 75 y 80 % de las anotaciones para los 4 ensamblados).

2.3.5. Predicciones de microsatélites

La Tabla 2.13 muestra un resumen de los microsatélites (SSRs) predichos *IN SILICO* para los 4 ensamblados. En la Referencia [160] se encuentran los archivos con la información necesaria para recuperar las regiones microsatélites de cada uno de los cuatro ensamblados presentados en este Capítulo. Cada archivo es una tabla delimitada por tabuladores con el identificador del contig, y el número, tipo, tamaño, inicio y final del SSR predicho por MISA [159].

Ensamblado	SSRs	Contigs con SSRs	Contigs con más de 1 SSR	SSRs comunes	SSRs dinucleótidos	SSRs trinucleótidos	SSRs tetranucleótidos	SSRs pentanucleótidos	SSRs hexanucleótidos
1	2839	2410 11.61	359 14.90	88 3.10	1193 42.02	1538 54.17	43 1.51	7 0.25	58 2.04
2	2466	2132 11.71	296 13.94	54 2.19	1070 43.39	1293 52.43	45 1.82	2 0.08	56 2.27
3	2963	2444 12.01	448 18.33	63 2.13	1146 38.68	1710 57.71	52 1.75	11 0.37	44 1.48
4	1934	1704 8.90	205 12.03	36 1.86	867 44.83	993 51.34	34 1.76	5 0.26	35 1.81

Tabla 2.13: Resumen de SSRs predichos en ensamblados de *N. pumilio*. 1: ZT48_20C. 2: ZT60_20C. 3: ZT48_34C. 4: ZT60_34C. Las filas superiores indican número total y las inferiores porcentaje

2.3.6. Código desarrollado

La *pipeline* generada para mejora de ensamblados *de novo* y anotación, *Lenga*, consiste de un WRAPPER escrito en Python que procesa un archivo de inicialización y realiza una serie de llamadas a programas, tanto previamente publicados como desarrollados en el marco de esta tesis. Los nuevos desarrollos fueron escritos en los lenguajes de programación Python [161] y PHP [162]. El archivo de inicialización contiene las direcciones URL para descargar los programas publicados que se utilizan en la *pipeline* (ver Métodos, Secciones 2.2.3 en la página 42 y 2.2.5 en la página 46), y tanto el *wrapper* como los nuevos programas están profusamente comentados para ayudar a la

interpretación de los resultados y el uso en otros conjuntos de datos. *Lenga* está disponible públicamente en la plataforma de proyectos de software GitHub [133] y cuenta con una licencia MIT de software abierto.

El *script* de Python para búsqueda de CEGs compara la base de datos de 248 modelos ocultos de Markov con un ensamblado traducido a aminoácidos (o cualquier conjunto de secuencias peptídicas) y escribe en el disco un archivo con los nombres de todas las secuencias (por ejemplo contigs de un ensamblado) que superan el umbral de similaridad para algún CEG. Este programa está disponible igualmente en GitHub [148] y cuenta con una licencia MIT de software abierto.

Finalmente, como producto terminado para la reutilización pública de los datos y el conocimiento generados por este estudio, se desarrolló un sencillo *script* de Python que realiza una búsqueda de contigs en cualquiera de los 4 ensamblados presentados en este Capítulo, a partir de una lista de interés, provista por el usuario, de cualquiera de los siguientes:

- Identidad de contigs (por ejemplo **NODE_2202**)
- Identidad de *loci* de *Arabidopsis thaliana* (por ejemplo **At5g61380**)
- Identidad de proteínas de *A. thaliana* (por ejemplo **Q9LKL2**)
- Nombres de genes (por ejemplo **Transcription factor PIF3**)
- Términos de Gene Ontology numéricos (por ejemplo **GO:0016021**)
- Términos de Gene Ontology textuales (por ejemplo **integral component of membrane**)

Como salida el programa genera un archivo amigable con el usuario que contiene todos los contigs resultado de la búsqueda, incluyendo de cada uno su nombre, longitud, cobertura, blanco de anotación (nombre, locus y proteína de *A. thaliana*), puntaje de anotación, términos de Gene Ontology (todos, en números y en texto), secuencia nucleotídica y secuencia peptídica (del ORF más largo del marco anotado). En la Figura 2.9 se muestra un ejemplo de visualización de salida. Al igual que los otros programas, este *script* se encuentra disponible en GitHub [163] y cuenta con una licencia MIT de software abierto.

```

### New item: Q9ZWC6 ###

CONTIG:
  NODE_22842
FRAME:
  1
LONG:
  392
COVERAGE:
  5.492308
SCORE:
  0.80198019802
UNIPROT:
  Q9ZWC6
TAIR:
  AT1G55590
NAME:
  F-box protein At-B 05=Arabidopsis
GO:
  cytoplasm      GO:0005737      C
SEQ:
  TTTTAAAGAAAAGTATGTATTATAATTAAATAAGAGCATTTAATGCTTCTTAATGTGTAAGAGTAGCTG
  CGATTAGAGGATTTCTTTGAGTTATCACCGAGATCGTTATGGCAGGATTGTTGATTTAGGTTGTAGCAA
  CCCAGAGTTTCATGTTGTTTGTGCAGAACCTGCTACGTTGGTGACCTTCACCCCAATGTGTGGAAGTAT
  CTGTATGAGAGTTGATTTCTCCATGATTCATCTGCTGATTTGAATTTGTATCACATATATCTTACTAC
  TATTTGGCCATTTGGCTATATGCCTATGTGTACTACAGTGACCAGTATGCCATGAATGAGCACTTTG
  ATTTAGAGAGTATGAGTTAGGCCTTTAAGGGTAAATAT
ORF:
  NVTDYSVEALAMKRSFQDGSKLLRMLDLFNCIGLSADLLRSLKRPSFPLHLWLIGIGHTRLSSKRNVVSTEICKERPWLTLCFDGCCEMGCHDGWQYHRPEYH
#

```

Figura 2.9: Ejemplo de salida del programa de búsqueda de contigs

2.4. Discusión

En este Capítulo se presentan los primeros datos -ómicos para *Nothofagus pumilio*, y los más completos para el género: 4 transcriptomas en condiciones contrastantes de temperatura y momento del día, en condiciones circadianas. Mediante una combinación de programas publicados y desarrollos propios, fue posible llegar a transcriptomas confiables en cuanto a su ensamblado y posterior anotación.

En promedio, cada condición biológica arrojó un ensamblado de unos 20 mil contigs, con gran completitud funcional (más de 95% de genes conservados CEG) y altos porcentajes de anotación contra la especie modelo vegetal más utilizada, *Arabidopsis thaliana* (más de 80% de contigs anotados) en todos los casos. Cada ensamblado contiene alrededor de 10 mil proteínas únicas de *A. thaliana*, y unos 2500 marcadores microsatélites predichos *in silico*.

2.4.1. Análisis bioinformático

El análisis de un experimento de RNA-seq consiste de varios pasos (ver por ejemplo los implementados en este Capítulo, Figura 2.4 en la página 43) dependientes entre sí,

en el sentido de que, por ejemplo, una mala secuenciación puede llevar a un mal ensamblado, un mal ensamblado puede llevar a malas anotaciones, y malas anotaciones pueden llevar a que sea incorrecta la interpretación biológica o fallida la comprobación *in vivo* de las hipótesis de interés. De esta manera, es sumamente importante realizar cada paso cautelosamente y comprobando la calidad de las secuencias en todo momento (lecturas crudas, ensamblados, anotaciones, etc.). En esta Sección se profundiza en algunos desafíos encontrados a lo largo del flujo de trabajo bioinformático.

Para empezar, el hecho de haber realizado la secuenciación en un equipo de Ion Torrent [139], cuya química de secuenciación es distinta a la de los diversos equipos de Illumina, que son más frecuentemente utilizados por la comunidad, trajo aparejado el desafío computacional de que los programas de uso más extendido (correctores de lecturas, ensambladores, etc.) no están puestos a punto para los errores y sesgos de los equipos de Ion Torrent, que difieren de los de Illumina. Básicamente, todas las químicas de secuenciación tienen dificultades en tener buena cobertura en regiones con muy alto o muy bajo contenido de GC [164, 165], pero la química de Ion Torrent, al detectar cambios de pH en el medio en lugar de luminiscencia por la hidrólisis producida al agregar un nucleótido marcado, tiene un sesgo particularmente fuerte hacia secuenciar incorrectamente regiones de homopolímeros (por ejemplo AAAAAA o GGGGGG), dando lugar a errores de tipo *indel* [166]. Por esto es importante que los programas utilizados tengan en cuenta este tipo de errores, particularmente fuertes en una química de secuenciación y no en otras.

Un paso clave para nuestro flujo de trabajo bioinformático fue el ensamblado *de novo*. El ensamblador más utilizado en la comunidad de RNA-seq de especies no modelo, Trinity [127, 143], posee básicamente 3 desventajas que hicieron que no lo utilizáramos para nuestras secuencias: no posee una opción explícita para aclarar la química de secuenciación, no posee un módulo de corrección de lecturas desacoplado del ensamblador propiamente dicho, y tiene un límite muy bajo de largo de k-mero (32). Por el contrario, SPAdes [125] permite explicitar que las secuencias provienen de un equipo Ion Torrent (mediante la opción `-iontorrent`), permite desacoplar la corrección de las lecturas del ensamblado propiamente dicho (por ejemplo para comparar sesgos y estadísticas antes y después de la corrección, o para utilizar las lecturas corregidas en otro ensamblador, o en combinación con otras condiciones biológicas para alivianar la

carga computacional del paso del ensamblado propiamente dicho), y tiene un límite para el largo del k-mero de 128.

La corrección y ensamblado de las lecturas teniendo en cuenta la química con la que fueron secuenciadas son fundamentales, ya que el paso previo de selección y limpieza de lecturas (por ejemplo con FastQC [140] y Trimmomatic [141] como se hizo en este Capítulo) sólo tiene en cuenta factores como el largo de la secuencia, la presencia de oligonucleótidos sobrerrepresentados (por ejemplo adaptadores para secuenciación) o la calidad de la asignación de bases (*base-calling*), y no factores específicos de la química de secuenciación.

De esta manera, al ensamblar con Trinity es mucho más probable que los contigs tengan errores “arrastrados” de las lecturas con errores *indel* de Ion Torrent, mientras que SPAdes contempla la química de secuenciación en la corrección y ensamblado de las lecturas. Esto puede producir directamente corrimientos en los marcos de lectura (*frame shifts*) de los contigs y así dar lugar a anotaciones defectuosas.

Por otro lado, al ser las secuencias de Ion Torrent más largas que las de la mayoría de los equipos de Illumina disponibles al momento de la secuenciación (y del desarrollo de los softwares en cuestión), es recomendable utilizar un k-mero a su vez más largo. El tamaño del k-mero es el único parámetro de un grafo de de Bruijn, y su elección representa un compromiso [167]: un valor de k bajo puede hacer colapsar los grafos (y así romper contigs) si existen repeticiones más largas que k nucleótidos, pero un valor muy alto puede contener en su interior errores, y así no representar correctamente la cantidad de k-meros distintos presentes en el transcriptoma. En todo caso, como regla general se suele probar con varios valores, y el valor final casi siempre se encuentra entre $\frac{1}{2}$ y $\frac{2}{3}$ del largo promedio de las lecturas (como fue en nuestro caso, ver Tabla 2.4 en la página 49). De esta manera, el límite de Trinity para el valor de k (32) es muy bajo para nuestras lecturas.

Otro paso importante en el análisis bioinformático es la comprobación de la completitud funcional de los ensamblados. Un conjunto de contigs puede tener buenas métricas como número de contigs, largo promedio, etc., pero no ser representativo de la diversidad de transcritos de la condición experimental en cuestión. Por eso existen herramientas informáticas que permiten evaluar la completitud génica de ensamblados transcriptómicos o genómicos. La aproximación más común para lograr esto es

“buscar” genes conservados en los ensamblados en estudio (mediante similitud entre las secuencias de interés y modelos ocultos de Markov, ver Sección 2.2.4 en la página 44). Como se puede ver en la Tabla 2.9 en la página 53, para todos nuestros ensamblados el porcentaje de genes BUSCO encontrados fue considerablemente menor que el porcentaje de genes CEG. Esto mismo sucedió con ensamblados transcriptómicos de otras especies arbóreas en el estudio filogenómico tratado en el Capítulo 3. Las razones por las cuales los genes CEG son detectados en mucha mayor medida seguramente tenga que ver con la especificidad de los genes elegidos para cada base de datos.

En el caso de los genes BUSCO, en un trabajo reciente [168] los autores de estas bases de datos advierten que existe un sesgo filogenético en la elección de las especies con las que se construyen los modelos ocultos de Markov, debido al fuerte desbalance en disponibilidad de datos genómicos en favor de especies modelo. Para el caso de la base de datos específica de plantas, los modelos están contruidos a partir de 30 especies vegetales, de las cuales más de la mitad (17) son pastos (Poaceae), y sólo 3 pertenecen al enorme clado *Fabids* (*sensu* APG IV [169]), al cual pertenece el orden Fagales (que incluye a *Nothofagus*). Este sesgo podría producir un sobreajuste (*overfitting*) en la búsqueda de genes conservados en favor de grupos de organismos bien representados. En ese mismo trabajo, los autores de BUSCO notan que dicha herramienta es indicada para aplicar en transcriptomas obtenidos a partir de muestras combinadas de varios tejidos y etapas ontológicas. En nuestro caso, las *libraries* fueron contruidas a partir de un único tejido y en una misma etapa ontológica (hojas expandidas de plántulas). En resumen, no es esperable que en transcriptomas provenientes de especies de grupos poco estudiados, y de un único tejido y estadio de vida, se detecte un porcentaje alto de BUSCOs.

Sin embargo, en el caso de *Nothofagus pumilio*, por contar con anotaciones contra *Arabidopsis thaliana*, pudimos comprobar que existen 410 genes BUSCO no encontrados por el programa mediante la búsqueda con modelos de Markov, pero anotados por nuestra *pipeline* como el correspondiente *locus* de *A. thaliana* con el cual fue armado el modelo proteico. Esto casi duplica el porcentaje de genes encontrados por BUSCO. Estos números en particular están tomados del ensamblado total (combinando las 8 *libraries*) que se utiliza en los Capítulos 3 y 4, pero los patrones con los ensamblados por separado de este Capítulo son idénticos. Este resultado podría indicar que el problema

de BUSCO no sería solamente la distribución desigual (filogenética e histológica; ver también más abajo) de sus modelos constituyentes, sino un sobreajuste que le hace no encontrar genes en ensamblados que los incluyen.

Por otra parte, también por contar con anotaciones contra una especie modelo, pudimos explorar uno de los sesgos presentados por los autores de BUSCO mencionados anteriormente, a decir, que para transcriptomas se espera mayor completitud en muestras obtenidas de diversos tejidos (y etapas ontológicas, de ser posible). Utilizando nuestras anotaciones (igual que en el párrafo anterior, del ensamblado total, pero con un patrón idéntico para los ensamblados por separado) y el atlas de expresión del Instituto Europeo de Bioinformática (EBI-GXA, experimento E-GEOD-38612) pudimos analizar los perfiles de expresión en hojas y raíces de *Arabidopsis thaliana* correspondientes a los genes anotados en *N. pumilio* no encontrados por BUSCO respecto de todos los genes de la base de datos BUSCO de plantas (Figura 2.10). Los genes no encontrados por BUSCO en *N. pumilio* están relativamente más expresados en raíz y menos en hoja (tejido con el cual trabajamos), lo cual confirma para nuestro caso el sesgo planteado por los autores de BUSCO [168].

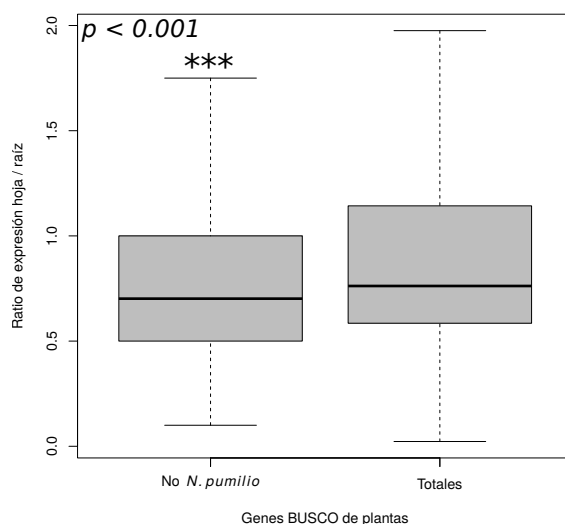


Figura 2.10: Relación entre expresión en hojas y raíces de genes BUSCO de plantas.

No *N. pumilio*: genes no encontrados por el software BUSCO en *Nothofagus pumilio*.

Test estadístico de *t* de Welch, dos colas

Por último, la anotación de los contigs ensamblados abre las puertas a estudios

de genómica funcional, caracterización de familias génicas y vías metabólicas, etc. Como se mencionó en la Sección 2.2.5 en la página 46, para esta Tesis se realizaron anotaciones contra *Arabidopsis thaliana*. Pudimos comprobar que se pierde muy poca información biológica al anotar contra bases de datos abarcativas, como SwissProt o TrEMBL (ver Sección 2.3.4, Tablas 2.11 en la página 56 y 2.12 en la página 57). En el caso de las anotaciones contra TrEMBL, se trata en la mayoría de los casos de proteínas no caracterizadas, o bien con un nombre asignado pero predichas computacionalmente sin curación manual. Las entradas de la base de datos SwissProt, por el contrario, están curadas manualmente y en la mayoría de los casos tienen evidencia a nivel de transcripto o de proteína. La poca ganancia biológica de contigs que no pudieron ser anotados contra *A. thaliana* pero sí contra SwissProt (o contra las entradas de TrEMBL caracterizadas biológicamente) podría explicarse, en el mejor de los casos, como proteínas no presentes o no descriptas en *A. thaliana*, o bien por cuestiones algorítmicas (por ejemplo, estar en el límite de lo que BLAT considera o no un resultado cuando realiza la búsqueda que desencadena el algoritmo de anotación). La ganancia computacional, por otra parte, es considerable, ya que la base de datos del proteoma de *A. thaliana* tiene casi 3000 veces menos secuencias que UniProtKB (SwissProt y TrEMBL combinadas), haciendo mucho más rápida la búsqueda con BLAT (ya que cada secuencia de entrada es comparada con toda la base de datos de referencia). Por otro lado, el tener todas las secuencias anotadas contra la misma especie facilita el trabajo posterior al poder cruzar fácilmente bases de datos, anotaciones, etc. Ver el Capítulo 4, Secciones 4.2.5 en la página 103 y 4.2.6 en la página 104 para algunas aplicaciones posibles.

En resumen, fue posible identificar y resolver varios problemas típicos del análisis bioinformático (calidad de las secuencias, ensamblado *de novo*, especificidad de las anotaciones) y llegar a ensamblados de calidad con anotaciones confiables para todo tipo de análisis posteriores.

2.4.2. Aportes a la comunidad bioinformática

Varias de las dificultades planteadas en la Sección anterior fueron sorteadas utilizando una combinación de programas preexistentes y desarrollos propios. La *pipeline* desarrollada en este Capítulo, *Lenga*, pudo mejorar los ensamblados de SPAdes, alar-

gando contigs con solapamiento parcial considerable y descartando contigs altamente redundantes. Este último punto puede ser más o menos deseable, ya que contigs con un alto porcentaje de similitud podrían en realidad ser isoformas de un mismo gen (por ejemplo por *splicing* alternativo), y no necesariamente errores en el ensamblado, y puede ser de interés para el usuario conservarlas. Es por esto que el código de *Lenga* [133] permite modificar el porcentaje de similitud entre contigs a partir del cual el programa los toma como redundantes.

Cabe destacar que tanto el módulo de corrección de ensamblados como el módulo de anotación de *Lenga* (ver Figura 2.4 en la página 43) tienen un gran potencial de reutilización y pueden ser adaptados a virtualmente cualquier proyecto de transcriptómica *de novo*. Además del porcentaje de redundancia ya mencionado en el párrafo anterior, el módulo de corrección de ensamblados de *Lenga* permite modificar los parámetros de los programas Trimmomatic y fastx (para limpieza y filtrado de lecturas crudas), del ensamblador SPAdes, del alineador BLAT y del programa de mapeo STAR, y cambiar la longitud mínima de contigs a conservar luego del ensamblado.

Respecto del módulo de anotación de *Lenga*, este cuenta por un lado con un puntaje de anotación personalizable (según el usuario quiera priorizar el largo de la región alineada, el largo de la secuencia buscada o el largo de la secuencia de referencia) y por otro con la posibilidad de anotar el ensamblado de entrada contra cualquier conjunto de secuencias de UniProt [134] (<http://www.uniprot.org>). Entre las posibles referencias podemos citar proteomas individuales o en conjunto, clusters de proteínas, o las bases de datos de todo el árbol de la vida SwissProt (anotadas manualmente y/o con evidencia experimental) y TrEMBL (de anotación automática) [170].

Otro desarrollo de este Capítulo [148] es el *script* de Python para encontrar genes conservados eucariotas (CEGs [147]) en ensamblados transcriptómicos. Este sencillo programa produce, para cada ensamblado de entrada, una lista con los contigs que representan los CEGs de la base de datos de referencia (estrictamente, que superan el umbral de similitud establecido por los autores de la base de datos para cada modelo oculto de Markov). Este archivo de salida puede ser usado directamente como entrada del programa *fastafetch* (parte de la suite Exonerate, <http://www.ebi.ac.uk/~guy/exonerate> [171]). De esta manera se puede disponer de las secuencias de los genes conservados para cada especie de interés, y así realizar alineamientos, caracterización

de genes o dominios proteicos, o estudios filogenéticos. El Capítulo 3 hace extenso uso de esta última aplicación.

Por último, el *script* de Python para búsqueda de contigs representa un producto de fácil utilización que resume los recursos generados en este Capítulo para genes de interés del usuario. La información contenida en este archivo puede ser utilizada fácilmente para diseño de cebadores de PCR (*primers*), comparaciones interespecíficas de secuencias, estudios de enriquecimiento de términos de Gene Ontology, etc.

En este Capítulo también se presentan las primeras predicciones de microsatélites (SSRs) a nivel transcriptómico para *Nothofagus pumilio*. Estos SSRs predichos pueden ser probados *in vivo* para detectar si presentan polimorfismo, aumentando los recursos de marcadores genéticos para la especie [37, 172-174] y siendo de gran ayuda para estudios de genética de poblaciones en una especie con una distribución geográfica muy heterogénea y una historia evolutiva compleja [175].

Capítulo 3

Análisis filogenómico del orden Fagales: aplicación de recursos y herramientas bioinformáticas

En sus remotas páginas está escrito que los animales se dividen en (a) pertenecientes al Emperador, (b) embalsamados, (c) amaestrados, (d) lechones, (e) sirenas, (f) fabulosos, (g) perros sueltos, (h) incluidos en esta clasificación, (i) que se agitan como locos, (j) innumerables, (k) dibujados con un pincel finísimo de pelo de camello, (l) etcétera, (m) que acaban de romper el jarrón, (n) que de lejos parecen moscas.

Jorge Luis Borges, *El idioma analítico de John Wilkins*

3.1. Introducción

3.1.1. La filogenética y la filogenómica

La filogenética se encarga de estudiar y reconstruir las relaciones evolutivas (de parentesco) entre taxones: especies, familias, y demás grupos de organismos del árbol de la vida [176]. Al tratarse la evolución de la acumulación de cambios genéticos en la descendencia, a lo largo del tiempo se produce la separación de linajes, y el criterio filogenético es el de establecer **grupos naturales**, es decir, que la clasificación refleje las relaciones evolutivas entre taxones. Ya en el año 1859, Charles Darwin decía [177]: *Our classifications will come to be, as far as they can be so made, genealogies* (“Nuestras cla-

sificaciones llegarán a ser, en la medida en que puedan ser así hechas, genealogías”). El resultado final de un estudio filogenético es un (o varios) árbol filogenético, que presenta una TOPOLOGÍA. Más importante, un árbol filogenético es una hipótesis que se propone para la historia evolutiva del grupo en cuestión. Históricamente, las reconstrucciones filogenéticas estuvieron únicamente basadas en CARACTERES morfológicos o estructurales. Si bien muchos de dichos caracteres tienen la ventaja de poder observarse y cuantificarse en el registro fósil además de en especies vivientes, poseen varias desventajas en comparación con los datos de secuencias de ADN o proteínas que son cada vez más comunes desde finales de los años '70 del siglo pasado. Entre las ventajas de los datos moleculares respecto de los morfológicos, podemos mencionar (para más detalle ver Ref. [178], Capítulo 5) la asignación más directa de homología, su utilidad para reconstruir relaciones entre organismos muy lejanos mediante la comparación de *loci* altamente conservados, y su heredabilidad estricta, entre otras. En estudios de filogenética molecular, los caracteres pasan a ser cada posición de ADN o proteína (es decir, nucleótidos o aminoácidos), ya que cada una es informativa para la reconstrucción de la filogenia. De esta manera se pasó de contados caracteres morfológicos a cientos o miles de caracteres moleculares (y más recientemente a decenas o cientos de miles, y hasta millones, con la nueva generación de secuenciación; ver más abajo).

Desde el advenimiento de los datos moleculares para análisis filogenéticos, unos pocos *loci* se convirtieron en referencia (por ejemplo, los genes de subunidades ribosomales, el espaciador ITS o el gen de la subunidad grande de la enzima RuBisCO en organismos fotosintéticos). Para las plantas con flor en general (angiospermas), se propusieron filogenias durante la última década del siglo XX utilizando uno, dos, o tres marcadores (*loci*) [179-182]. Un trabajo más reciente [183] aumenta el número de genes a 17.

En la última década, las técnicas de secuenciación de próxima generación (NGS) permitieron aumentar enormemente el número de caracteres moleculares a analizar en un estudio filogenético. Esto dio origen a un nuevo campo de investigación, la **filogenómica**, que se encarga de establecer filogenias a partir de datos *-ómicos* [184, 185]. Mediante este tipo de análisis se han reportado recientemente importantes avances en varios grupos de organismos [186-191].

Independientemente de la cantidad de *loci* analizados y del advenimiento de la

filogenómica, los cuellos de botella del análisis filogenético clásico se mantienen: la selección e identificación de caracteres homólogos, y los métodos de reconstrucción de árboles. Respecto de lo primero, en el análisis filogenético es preciso enfrentar dos tipos distintos de sesgo [185]: el **error de muestreo** o **estocástico** se origina en el ruido al azar producido por trabajar con una pequeña cantidad de sitios o genes, y que influye en la reconstrucción de los nodos del árbol. Este tipo de error disminuye a medida que se incrementa la cantidad de datos. Limitaciones técnicas y monetarias anteriores a la adopción masiva de las tecnologías de NGS confinaron a la filogenia molecular al uso de unos pocos marcadores, como ya se mencionó en esta Introducción. En los albores de la filogenómica había un gran optimismo acerca del “fin de la incongruencia” [192] debido a la utilización de cientos o miles de *loci* en un mismo análisis filogenético. Sin embargo, si bien los estudios filogenómicos alivian considerablemente el sesgo de muestreo al tratar simultáneamente con cientos de *loci* o hasta genomas enteros, el **error sistemático** producido por la interferencia con señales no-filogenéticas en los datos sigue presente y no se soluciona, sino que más bien se agrava, con la adición de caracteres [193].

Básicamente, los sesgos sistemáticos pueden ser originados por una tasa variable de evolución entre linajes, una composición heterogénea de nucleótidos (o aminoácidos si se trabaja con secuencias peptídicas), o cambios en la tasa evolutiva entre posiciones o regiones del alineamiento [185]. En todos los casos, los errores sistemáticos (o **señal no filogenética**) consisten en malas interpretaciones, por parte de los métodos de reconstrucción, de verdaderos cambios producidos en la historia evolutiva estudiada, y que conducen a una topología alternativa.

Como consecuencia del incremento fenomenal de caracteres (en el caso de datos moleculares, nucleótidos o aminoácidos) en la era filogenómica, pero la relativa poca atención prestada a cómo detectar y reparar las señales no filogenéticas, los estudios filogenómicos de la última década se caracterizan por presentar árboles con soportes máximos para todas sus ramas (es decir, con gran consistencia interna), pero con topologías contradictorias entre sí (ver ejemplos en [193-198]). En los últimos años se han propuesto métodos concretos para cuantificar la señal no filogenética y corregir o filtrar los datos de manera de minimizar la información ruidosa o espuria de un alineamiento (ver Sección 3.2.4 en la página 80 y Ref. [195, 196]).

Respecto de los métodos de reconstrucción filogenética, los estudios filogenómicos se sirven de los mismos que han sido históricamente utilizados por la filogenética clásica (para una revisión ver [199], y para un detalle de cada método [200]). Básicamente, los **métodos de distancia** computan distancias entre pares de taxones, y luego encuentran el árbol que mejor explique dicha matriz de distancias, mediante algoritmos como la unión de vecinos (*neighbor-joining*) [201] o de evolución mínima [202]. Los **métodos cladísticos**, en cambio, no trabajan con distancias pareadas sino con los caracteres totales crudos (en el caso de la filogenia molecular, con los alineamientos de ADN o aminoácidos). Dentro de estos, el método de **máxima parsimonia** [203] toma como criterio que el mejor árbol será aquel que contenga la menor cantidad de cambios (para datos moleculares, mutaciones). Por el contrario, los métodos de **máxima verosimilitud** [204] y de **inferencia bayesiana** [205] toman en cuenta explícitamente los procesos de evolución molecular particulares del conjunto de datos de interés (que pueden perfectamente no ser parsimoniosos), mediante la adopción de un **modelo probabilístico de evolución molecular**. Es decir, no realizan una asunción *a priori* respecto de la evolución de las secuencias, pero incorporan gran cantidad de parámetros y se enlentece la búsqueda.

Los modelos de evolución molecular consisten básicamente en tasas de sustitución diferencial entre nucleótidos o aminoácidos. Para estos últimos, el trabajo pionero fue realizado por Margaret Dayhoff [206], quien en el año 1972 modeló el proceso de sustitución con cadenas de Markov independientes sitio-específicas. El llamado modelo **Dayhoff** sigue siendo utilizado hoy en día. Algunas extensiones y modificaciones de dicho modelo son el **VT** (por el inglés, *Variable Time*, “tiempo variable”, ya que está optimizado para comparar taxones con tiempos de divergencia muy dispares) [207] y el **JTT** (por las iniciales de los apellidos de sus autores) [208]. Otro enfoque, utilizado por ejemplo por el popular modelo **Blosum** (abreviatura del inglés, *BLOCKS SUBstitution Matrix*, “matriz de sustitución de bloques”, por el tipo de matriz que utiliza el modelo) [209] y similares, consiste en modelar la variación observada en grandes alineamientos de proteínas homólogas. Otros modelos utilizan muestras de diversas familias proteicas (modelo **WAG** [210] o su modificación, el modelo **LG** [211], ambos por las iniciales de los apellidos de sus autores), o incluso secuencias específicas de genes (genes de retrovirus: modelo **RtREV** [212]; genes mitocondriales: modelo **mtREV** [213]; genes

plastídicos: modelo **CpREV** [214]; todos por el modelo de Markov *RE*Versible que utilizan, y cada uno con el prefijo del grupo de genes que modela) o grupos de estudio (genes del VIH: modelos **HIV** [215]).

Las principales diferencias entre los métodos de máxima verosimilitud y de inferencia bayesiana son de índole teórica, ya que el primero se basa en el cálculo explícito de parámetros en un marco probabilístico clásico (por eso a veces se habla de “método probabilístico” refiriéndose al método de máxima verosimilitud en un análisis filogenético), mientras que el segundo utiliza la estadística bayesiana y no se preocupa por encontrar parámetros globales sino por estimar el soporte de las distintas partes del árbol.

La Tabla 3.1 presenta un resumen de los métodos de reconstrucción filogenética aquí explicados.

Método	Ventajas	Desventajas
Distancias	Rapidez	Pérdida de información al convertir secuencias en distancias
Máxima parsimonia	Rapidez y robustez para ramas cortas	Débil para datos con RAMAS LARGAS O variables (cortas y largas)
Máxima verosimilitud	Completitud y solidez probabilística	Lentitud y demanda computacional
Inferencia bayesiana	Similar a Máxima verosimilitud	Dificultad de determinar si el análisis convergió

Tabla 3.1: Resumen de ventajas y desventajas de los métodos de reconstrucción filogenética más comunes. Adaptado de [199]

3.1.2. Filogenómica en organismos vegetales

Los primeros trabajos de filogenómica en plantas fueron realizados por investigadores del Instituto Sackler de Genómica Comparativa, del Museo de Historia Natural de Nueva York [216-219], y se enfocaron en el origen y diversificación temprana de los grandes grupos de plantas terrestres (helechos, gimnospermas y angiospermas), haciendo uso de bases de datos de ESTs y de los primeros genomas de plantas se-

cuenciados (ver Capítulo 2, Sección 2.1 en la página 35). A lo largo de dichos trabajos pioneros el número de taxones analizados fue en aumento (desde 7 en [216] hasta 101 en [219]), y se echó luz sobre temas controversiales de la filogenia de las plantas, como las relaciones internas de las gimnospermas.

En los últimos años, varios estudios han aumentado considerablemente el número de taxones, abarcando no sólo las plantas con semilla sino también sus parientes las briofitas, e incluso los linajes de algas verdes más cercanos a las plantas [220, 221]. El objetivo de estos trabajos fue responder preguntas de gran arraigo en la filogenética del reino vegetal, como cuál es el grupo de algas verdes que dio origen a las plantas, cuáles son los grupos basales de angiospermas, o cuál es la relación de las monocotiledóneas con las eudicotiledóneas.

Los trabajos filogenómicos en plantas, en resumen, se han enfocado en responder preguntas acerca del origen y las relaciones entre los grandes grupos del reino vegetal, y no tanto en estudiar grupos particulares. Sin embargo, existen limitaciones teóricas [222-224] y prácticas (dificultad para asimilar y encontrar patrones en cantidades enormes y desbalanceadas de datos) de las filogenias de gran escala. Dos trabajos recientes se enfocan en grupos particulares (orden Caryophyllales, que incluye a los cactus, las plantas carnívoras, y cultivos como la quinoa, la remolacha y la acelga [225]; y el grupo Campanulidae, uno de los más importantes de eudicotiledóneas, que incluye órdenes como Apiales o Asterales [226]). Trabajar con un grupo particular de organismos permite mayor profundidad en el estudio de cada taxón, conlleva mayor homogeneidad de procesos evolutivos entre taxones (importante, por ejemplo, a la hora de modelar la evolución molecular de las secuencias), y los análisis son más sencillos de llevar a cabo, permitiendo una más rápida y mejor inspección y corrección de errores y sesgos en el análisis [226].

3.1.3. El orden Fagales

El orden Fagales es uno de los más importantes de angiospermas leñosas. Entre los miembros de este orden se encuentran árboles valiosos tanto por su madera como por sus productos alimenticios, además de poseer gran importancia ecológica como árboles dominantes de bosques perennes y caducifolios de latitudes templadas en ambos hemisferios. Como ejemplos bien conocidos podemos citar a los robles, encinas, cas-

taños y hayas (familia Fagaceae), a los abedules, avellanos y alisos (familia Betulaceae), a los nogales (familia Juglandaceae), y a las casuarinas (familia Casuarinaceae). Según la clasificación del Grupo de Filogenia de las Angiospermas (*Angiosperm Phylogeny Group*, APG) [169], el orden contiene 8 familias. La Figura 3.1 muestra la distribución geográfica actual (y pasada para algunas familias de gran presencia en el registro fósil) de las 8 familias del orden. Como se puede apreciar, sólo Nothofagaceae (panel 3.1a) y Casuarinaceae (panel 3.1h) tienen una típica distribución gondwánica, y de ellas sólo Nothofagaceae se encuentra actualmente en Sudamérica (en particular, en los bosques templados y fríos de la Patagonia). Las dos familias más numerosas del orden, Fagaceae (panel 3.1b) y Betulaceae (panel 3.1g), tienen distribución mayormente eurasiática.

No hay hasta el momento estudios filogenómicos publicados enfocados en el orden Fagales. Para este grupo, tanto los trabajos clásicos [228-230] como los más recientes [231-236] utilizan un número bajo de marcadores (hasta 5), aunque el número de especies muestreadas ha ido aumentando considerablemente.

La filogenia aceptada del orden, según APG, se ilustra en la Figura 3.2 y es presentada como poco conflictiva. En particular, la familia Nothofagaceae es hermana del resto de las familias del orden (es decir, es el clado basal del orden). En el sitio web de la última versión del APG [227] se menciona abundante bibliografía, alguna citada en esta Introducción, que soporta la topología de la Figura 3.2.

El objetivo de este Capítulo es construir el primer árbol filogenómico para especies del orden Fagales, a partir de una combinación de algoritmos generados en esta Tesis para la construcción de la supermatriz de caracteres (de secuencias peptídicas en nuestro caso), y programas bien establecidos para el análisis filogenético propiamente dicho (ver a continuación, Materiales y Métodos, Sección 3.2.3). La búsqueda y alineamiento automáticos de genes aceleran considerablemente y quitan sesgo a la construcción de la matriz de caracteres, y fueron posibles gracias a las competencias bioinformáticas adquiridas en esta Tesis. De esta manera, el presente Capítulo no sólo da un puntapié a la filogenómica de un importante grupo de organismos, sino que también sirve de ejemplo de aplicación de herramientas bioinformáticas y reutilización de datos abiertos disponibles *online*, tema de actualidad en la comunidad científica (ver por ejemplo [237, 238]).

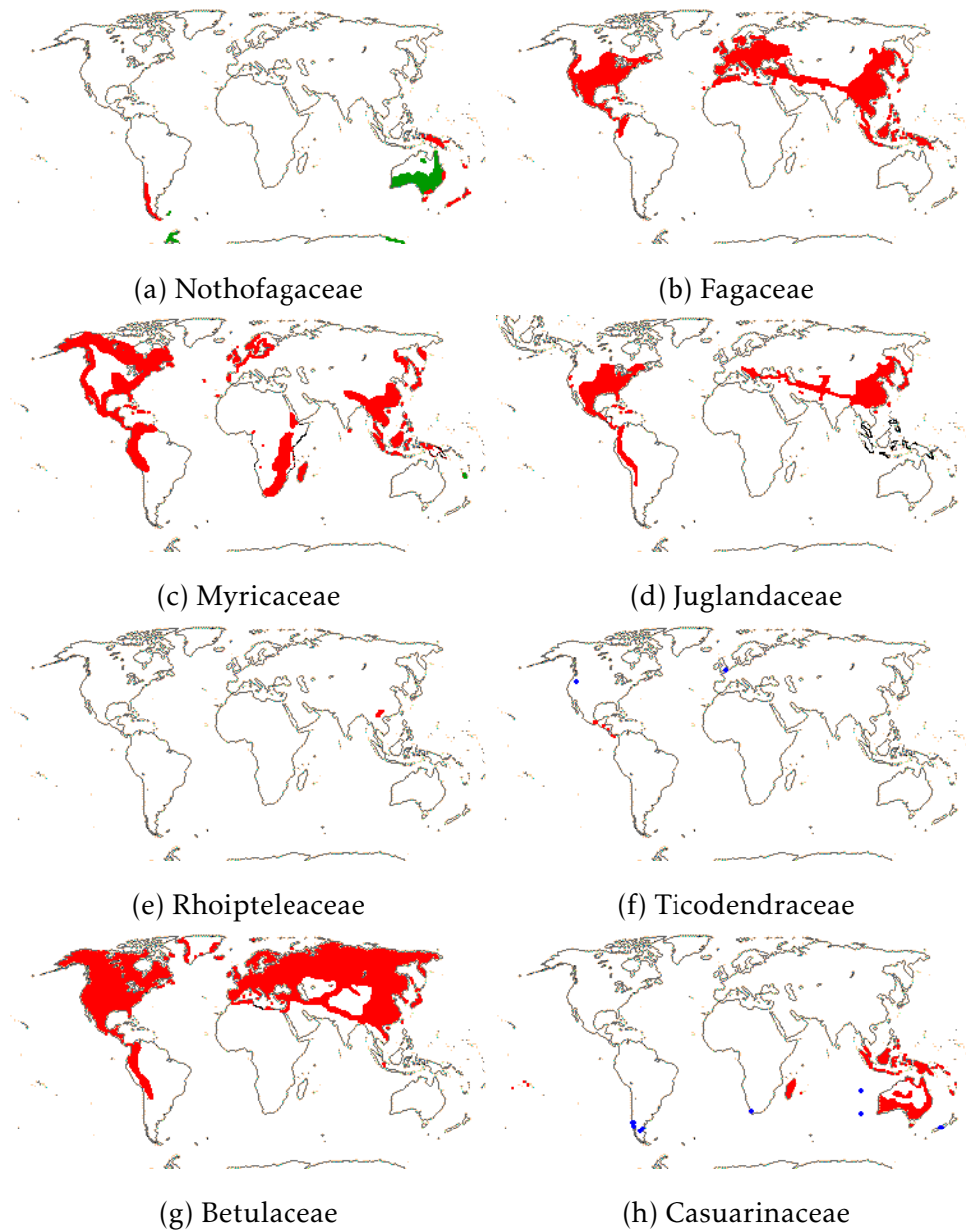


Figura 3.1: Distribución geográfica de las familias del orden Fagales. En color rojo, la distribución presente. En verde (Nothofagaceae) y azul (Ticodendraceae y Casuarinaceae), presencia de fósiles. Fuente: APG IV [227]

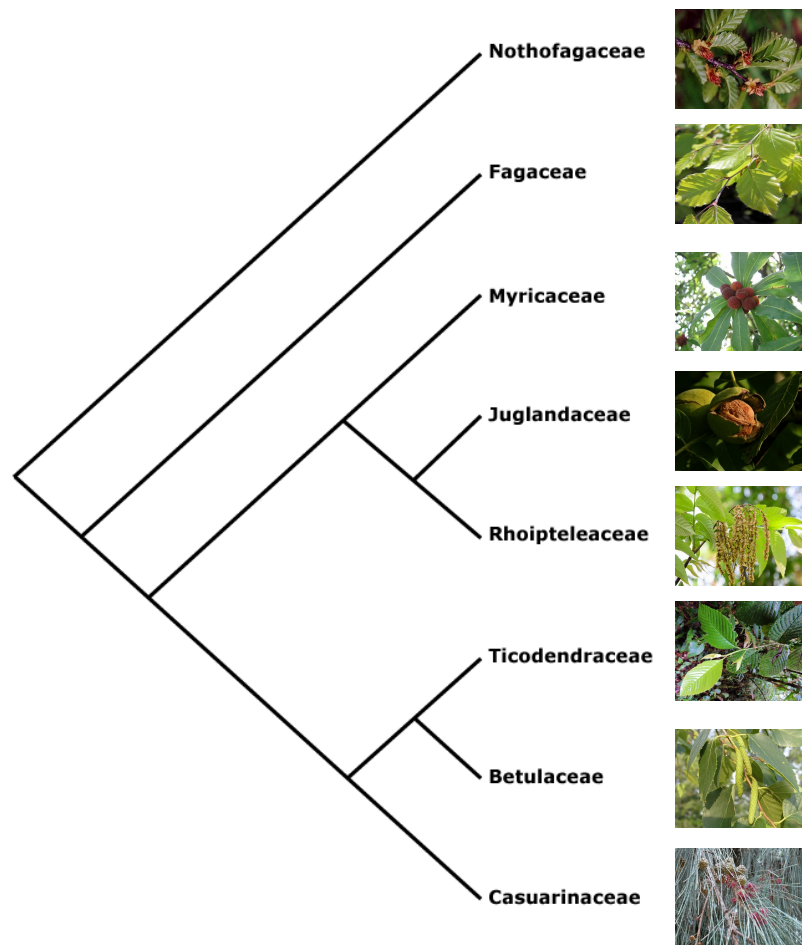


Figura 3.2: Filogenia aceptada para las familias del orden Fagales. Fuente: APG IV

[227]

3.2. Materiales y Métodos

3.2.1. Especies y secuencias utilizadas

Para los análisis filogenómicos se utilizaron ensamblados transcriptómicos, de manera de asegurar que se trata de secuencias transcriptas como las estudiadas en esta tesis para *N. pumilio* (Capítulos 2 y 4), y no por ejemplo predicciones de genes sobre un genoma.

Debido a la cantidad y diversidad de bases de datos, para este Capítulo se intentó maximizar la unificación de la proveniencia de las secuencias. El proyecto Hardwood Genomics (HWG; <https://www.hardwoodgenomics.org>) es una iniciativa de la Universidad de Tennessee y la National Science Foundation de EE.UU. para el estudio genómico de árboles forestales latifoliados, en su mayoría norteamericanos. En su sitio web se encuentran disponibles gran cantidad de datos biológicos (genomas, transcriptomas, regiones microsatélites) y herramientas *online* como BLAST o visualización de expresión génica. En la Tabla 3.2 se muestran los ensamblados transcriptómicos activos (no archivados) del orden Fagales disponibles en HWG. También se muestra a *Gleditsia triacanthos* (orden Fabales, familia Fabaceae, subfamilia Caesalpinioideae), utilizada como grupo externo para los análisis filogenéticos (ver más abajo, Sección 3.2.3).

Además de los ensamblados descritos en la Tabla 3.2, se trabajó con el ensamblado de *Nothofagus pumilio* proveniente de combinar todas las *libraries* en un único ensamblado (ver Capítulo 4, Sección 4.2.6 en la página 104), con el ensamblado transcriptómico de *Nothofagus alpina* (BioProject PRJNA81795), y con un ensamblado transcriptómico de *Betula platyphylla* (familia Betulaceae), un árbol del orden Fagales relativamente bien estudiado (descargado de la base de datos PlantGDB).

3.2.2. Obtención de genes conservados en común

Los caracteres homólogos se obtuvieron mediante la búsqueda, para cada transcriptoma, de genes conservados y de copia única. Para este Capítulo se desarrollaron *scripts* de Python que realizan automáticamente la búsqueda de dichos genes (tanto CEGs como BUSCOs, ver Capítulo 2, Sección 2.2.4 en la página 44) comunes entre las

Especie	Familia	Secuencias	Referencia
<i>Alnus rhombifolia</i>	Betulaceae	26186	BioProject PRJNA320832
<i>Alnus rubra</i>	Betulaceae	29113	BioProject PRJNA320831
<i>Fagus grandifolia</i>	Fagaceae	8319	SRX001797 SRX001798
<i>Gleditsia triacanthos</i>	Fabaceae	56845	BioProject PRJNA273269
<i>Juglans nigra</i>	Juglandaceae	78834	BioProject PRJNA273267
<i>Quercus alba</i>	Fagaceae	22102	SRX001811 SRX001812
<i>Quercus rubra</i>	Fagaceae	52662	BioProject PRJNA273270

Tabla 3.2: Descripción de los ensamblados del orden Fagales del proyecto Hardwood Genomics. Se menciona también a *Gleditsia triacanthos*, del orden Fabales, utilizado como grupo externo

especies de interés, su alineamiento, limpieza, concatenado, y preparación para cada tipo de análisis (ver más abajo, Sección 3.2.3).

En primer lugar, para cada especie se buscaron los modelos ocultos de Markov (ver Capítulo 2, Sección 2.2.4 en la página 44) de genes CEG y BUSCO, y se procedió a elegir las especies definitivas para los estudios en base a la completitud funcional de sus ensamblados (y en consecuencia, a la cantidad de genes comunes al intersectar los encontrados en cada especie por separado; ver Resultados, Sección 3.3.1 en la página 82).

Cabe destacar que estos *scripts* revisten carácter general y pueden ser utilizados para cualquier conjunto de ensamblados transcriptómicos. Esto hace que sean fácilmente utilizables en distintos números y combinaciones de taxones.

A partir de los ensamblados transcriptómicos, los *scripts* llevan a cabo los siguientes pasos:

- Traducción en los 6 marcos de cada contig
- Búsqueda de los modelos ocultos de Markov de genes conservados (CEGs, BUSCOs eucariotas o BUSCOs de plantas) sobre las secuencias aminoacídicas
- Intersección para encontrar genes conservados comunes entre las especies a es-

tudiar

- Aislamiento de los contigs correspondientes a cada gen conservado en cada ensamblado
- Alineamiento de genes conservados entre especies
- Limpieza de las secuencias para conservar sólo las posiciones que tienen información en todos los taxones. Esto permite eliminar regiones mal alineadas (ver Discusión, Sección 3.4.1 en la página 90)
- Concatenado para producir una matriz final a analizar

La matriz con X taxones y n aminoácidos provenientes de Y genes comunes es el producto de entrada para los análisis filogenéticos (ver a continuación).

3.2.3. Análisis filogenéticos

Se utilizaron métodos cladísticos por su mejor aprovechamiento de la información contenida en caracteres moleculares (ver Introducción, Sección 3.1.1 en la página 69). Antes de realizar las reconstrucciones filogenéticas, se utilizó el programa ProtTest [239, 240] (versión 3.4.2) para la evaluación de modelos de evolución molecular para cada matriz. En todos los casos se eligió el modelo que mejor ajustara a los datos.

En particular, la reconstrucción de filogenias se realizó por el método de **máxima verosimilitud**, con el programa RAxML [241] (versión 8.2.12; parámetros -f a -N 10000). El parámetro -f a le indica al programa que realice en una misma corrida un análisis de soporte (*bootstrap*) con N réplicas (en este caso 10000) y luego busque el árbol (o los árboles) con mejor puntaje de verosimilitud.

Los árboles fueron importados, editados y exportados como imágenes vectoriales usando el programa FigTree [242].

Se utilizó a *Gleditsia triacanthos* como grupo externo dado que pertenece a un orden cercano a Fagales (Fabales), y cuenta con un ensamblado transcriptómico en la base de datos HWG (Tabla 3.2).

3.2.4. Análisis de incongruencias

Para encontrar regiones del alineamiento analizado que se desvíen fuertemente de la señal filogenética media, se siguió el método propuesto recientemente por Shen y

colaboradores [196]. Este método se basa en el enfoque probabilístico, según el cual cada sitio (carácter) del alineamiento de entrada tiene una verosimilitud para cada árbol obtenido, es decir una probabilidad de observar dicho carácter dada la topología. Entonces, el método compara dos topologías contrastantes (sean obtenidas por métodos distintos de reconstrucción, o por matrices distintas, o en otros estudios previos, filogenómicos o no) y asigna a cada sitio y cada gen del alineamiento una diferencia en verosimilitud entre las dos hipótesis. Se llama **T1** a la topología obtenida con una matriz y un método de reconstrucción dados, y **T2** a la topología forzada a mantener el nodo o la rama “en disputa”, con la misma matriz y el mismo método de reconstrucción.

Para cada sitio se calcula entonces (con la opción `-f G` de RAxML [241]) la **diferencia por sitio en el logaritmo de la verosimilitud entre hipótesis** (topologías), ΔSLS_i :

$$\Delta SLS_i = \ln L(S_i|T1) - \ln L(S_i|T2) \quad (3.1)$$

donde S_i representa el sitio i -ésimo de la matriz total. Sumando los valores de ΔSLS_i para todos los sitios del gen j -ésimo se obtiene la **diferencia por gen en el logaritmo de la verosimilitud entre hipótesis**, ΔGLS_j :

$$\Delta GLS_j = \ln L(G_j|T1) - \ln L(G_j|T2) \quad (3.2)$$

Una vez obtenida la diferencia de verosimilitud entre hipótesis contrastantes gen a gen, se definen los **genes anómalos (outliers)** como los que tengan un valor absoluto de ΔGLS más grande que el límite superior (*Sup*) o más pequeño que el límite inferior (*Inf*), definidos de la siguiente forma:

$$Sup = \min(\max(x), Q_3 + 1.5 \times IQR) \quad (3.3)$$

$$Inf = \max(\min(x), Q_1 - 1.5 \times IQR) \quad (3.4)$$

donde $\max(x)$ y $\min(x)$ son el valor máximo y mínimo, respectivamente, entre los ΔGLS , Q_1 y Q_3 son el primer cuartil y el tercer cuartil respectivamente de la distribución de valores de ΔGLS , e IQR (rango intercuartiles) es la diferencia entre Q_3 y Q_1 .

Luego de identificar los genes *outliers*, se descartan de la matriz y se vuelven a realizar los análisis filogenéticos, como fue descripto en la Sección anterior.

3.3. Resultados

3.3.1. Especies y genes conservados comunes elegidos

En la Tabla 3.3 se muestran los genes conservados encontrados para cada ensamblado y cada base de datos. Las posibles razones e implicancias de las diferencias entre CEGs y BUSCOs encontrados para un mismo ensamblado se discuten en parte en el Capítulo 2, Sección 2.4.1 en la página 60.

Especie	BUSCOs (Plantas)	BUSCOs (Plantas, %)	BUSCOs (Eucariotas)	BUSCOs (Eucariotas, %)	CEGs (%)	CEGs
<i>Alnus rhombifolia</i>	318	22.1	108	35.6	237	95.6
<i>Alnus rubra</i>	291	20.2	104	34.3	234	94.4
<i>Betula platyphylla</i>	628	43.6	192	63.4	240	96.77
<i>Fagus grandifolia</i>	10	0.7	7	2.3	177	71.4
<i>Gleditsia triacanthos</i>	546	37.9	189	62.4	247	99.6
<i>Juglans nigra</i>	856	59.4	198	65.3	248	100
<i>Nothofagus alpina</i>	62	4.3	36	11.9	215	86.7
<i>Nothofagus pumilio</i>	475	33.0	156	51.5	245	98.8
<i>Quercus alba</i>	60	4.2	53	17.5	236	95.2
<i>Quercus rubra</i>	875	60.8	224	73.9	248	100

Tabla 3.3: Número y porcentaje de genes conservados de cada ensamblado. En el caso de los BUSCO se indican los genes completos y de copia única

Teniendo en cuenta los resultados mostrados en la Tabla 3.3, se decidió dejar afuera del análisis a *Fagus grandifolia*, *Nothofagus alpina*, y *Quercus alba*, debido a la poca cantidad de genes conservados presentes en relación a las otras especies (más notorio

en las bases de datos BUSCO).

Para las 7 especies restantes, se procedió a encontrar los genes conservados comunes. La Tabla 3.4 muestra la información final de las matrices de caracteres para cada base de datos utilizada. Como es de esperar en base a lo expuesto anteriormente, fueron encontrados muchos más genes CEG comunes que BUSCOs, y dentro de estos mayor porcentaje de genes eucariotas que exclusivos de plantas. Al ser esta última base de datos mucho más grande que la primera (1440 genes contra 303), la cantidad absoluta de genes comunes termina siendo mayor.

Base de datos	Información	% de genes de la base de datos
CEGs	224 genes	90.32
BUSCOs Plantas	34 genes	2.36
BUSCOs Eucariotas	14 genes	4.62

Tabla 3.4: Genes conservados comunes entre ensamblados utilizados para el análisis filogenético

En vistas de estos resultados, para los análisis filogenéticos se decidió trabajar con la matriz de genes CEG comunes, debido al bajo número de genes y representatividad de los genes BUSCO comunes, tanto específicos de plantas como eucariotas.

De los 224 genes CEG comunes entre las 7 especies analizadas, 199 forman parte del alineamiento final, y se detallan (con los identificadores de proteína y locus correspondientes a *Arabidopsis thaliana*, una de las especies con las que fueron contruidos los modelos génicos CEG) en la Tabla A.1 en la página 191 del Apéndice A. Los 25 genes comunes que no se encuentran en la matriz final no presentan ninguna posición con información en todas las especies, y por lo tanto no son incorporados al alineamiento concatenado por el *script* desarrollado en este Capítulo (ver Materiales y Métodos, Sección 3.2.2 en la página 78).

3.3.2. Análisis filogenéticos

Análisis con datos crudos

La matriz total de 199 genes CEG comunes se ajustó a un modelo de evolución VT. Para evaluar la posibilidad de HETEROTAQUIA se sometió al alineamiento de cada gen

por separado al mismo ajuste de modelos del software ProtTest que al alineamiento total. La Figura 3.3 muestra que el modelo prevalente entre los genes individuales fue el VT, y que solamente 4 modelos (VT, JTT, LG, y Blosum62) explican la evolución molecular de más del 80% de los genes. El modelo VT se encontró entre los primeros 3 mejor rankeados por ProtTest en un 63% de los genes (126 de 199). Todos los modelos prevalentes son de carácter general (no específicos de un grupo de genes u organismos; ver Introducción, Sección 3.1.1 en la página 69). Estos resultados indican la baja probabilidad de heterotaquia en nuestro conjunto de datos.

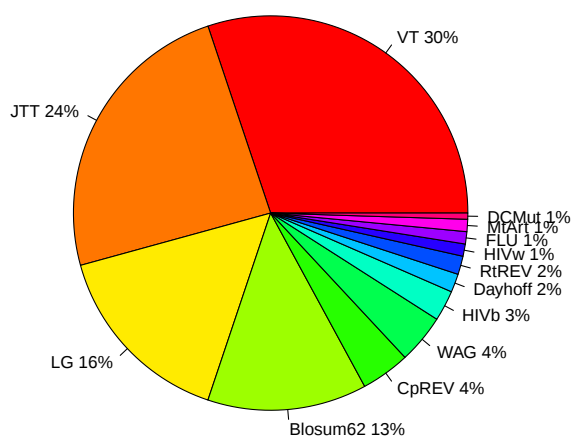


Figura 3.3: Frecuencia de modelos de evolución molecular que mejor ajustan a cada uno de los 199 genes CEG comunes

El árbol obtenido por el método de máxima verosimilitud (Figura 3.4) muestra a Nothofagaceae como hermana a Betulaceae (*Betula platyphylla* y *Alnus* spp.), una topología distinta a la aceptada para el orden Fagales, con Nothofagaceae como clado basal (hermana al resto de las familias del orden; ver Figura 3.2 en la página 77).

Debido al resultado conflictivo con el aceptado para el orden, decidimos realizar un análisis de incongruencias para detectar regiones con señal filogenética anormal, y poder excluirlas del análisis.

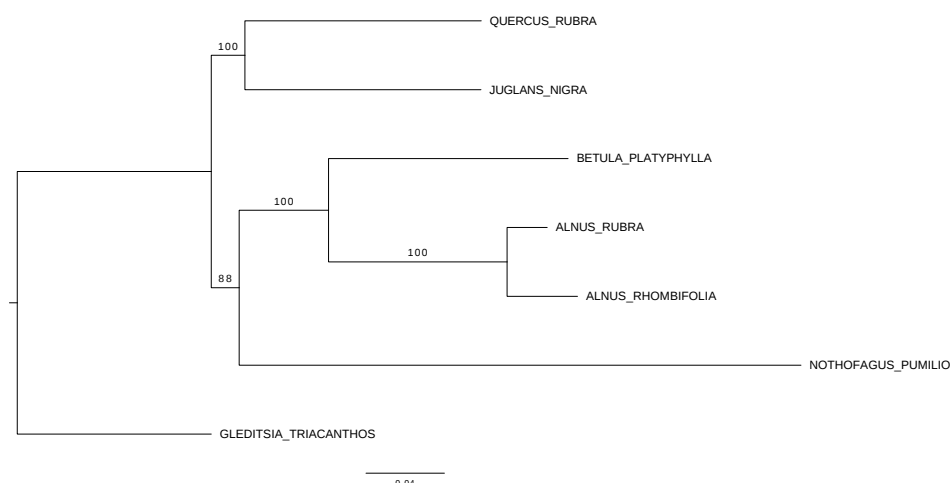


Figura 3.4: Árbol filogenético de máxima verosimilitud obtenido a partir de genes CEGs. Matriz de 7 taxones y 199 genes (28499 aminoácidos). Se indica el soporte de *bootstrap* cada nodo. El largo de las ramas es una estimación de las sustituciones esperadas por sitio

Análisis de incongruencias

Siguiendo el método probabilístico de Shen y colaboradores (ver Materiales y Métodos, Sección 3.2.4 en la página 80 y Ref. [196]), definimos **T1** como la topología obtenida con nuestra matriz de datos mediante el método de máxima verosimilitud (Figura 3.4), y **T2** como la aceptada para el orden según APG. La **BIPARTICIÓN** en disputa, en nuestro caso, es si Nothofagaceae es hermana de Betulaceae o hermana del resto del orden.

Una vez obtenido el árbol de máxima verosimilitud forzado a recuperar a Nothofagaceae como clado basal del orden (T2), se calcularon las diferencias entre hipótesis en verosimilitudes sitio a sitio y gen a gen (Ecuaciones 3.1 a 3.2 en la página 81). La Tabla 3.5 muestra, para los 199 genes CEGs del alineamiento final, el valor de ΔGLS . Como se deduce de la Ecuación 3.2, valores positivos de ΔGLS indican apoyo a T1, y valores negativos apoyo a T2.

Del total de 199 genes, 128 (un 64%) apoyan a T2, es decir, a Nothofagaceae como clado basal del orden Fagales. Los dos genes con valor absoluto de ΔGLS más alto (KOG2948 y KOG1466), sin embargo, apoyan a T1.

La Figura 3.5 muestra la distribución de ΔGLS en función del largo de los 199

CEG	ΔGLS	CEG	ΔGLS	CEG	ΔGLS	CEG	ΔGLS
KOG2948	132.44	KOG0524	0.85	KOG0181	-0.48	KOG1556	-2.80
KOG1466	130.41	KOG2906	0.73	KOG2726	-0.49	KOG3174	-2.91
KOG1390	21.70	KOG3459	0.67	KOG0559	-0.51	KOG1185	-2.92
KOG1688	11.54	KOG0180	0.66	KOG0400	-0.51	KOG2967	-2.96
KOG1394	10.61	KOG3404	0.63	KOG0077	-0.52	KOG0359	-3.02
KOG1211	9.12	KOG2613	0.63	KOG1816	-0.62	KOG0177	-3.07
KOG1353	7.21	KOG3405	0.62	KOG0174	-0.64	KOG1349	-3.09
KOG1643	6.97	KOG1272	0.45	KOG0567	-0.67	KOG2909	-3.17
KOG1723	6.05	KOG3400	0.39	KOG2728	-0.68	KOG2653	-3.17
KOG3180	5.86	KOG0025	0.36	KOG3052	-0.70	KOG1335	-3.22
KOG1112	5.37	KOG0276	0.23	KOG2825	-0.71	KOG0741	-3.30
KOG2916	5.19	KOG2481	0.23	KOG0225	-0.72	KOG2472	-3.49
KOG4655	4.39	KOG2303	0.22	KOG3295	-0.78	KOG0862	-3.60
KOG0211	4.31	KOG3349	0.22	KOG1533	-0.84	KOG0462	-3.66
KOG3157	4.08	KOG0780	0.18	KOG0477	-0.86	KOG2775	-3.70
KOG3189	3.88	KOG2104	0.14	KOG0188	-0.86	KOG3185	-3.71
KOG0142	3.67	KOG3343	0.10	KOG2531	-0.92	KOG2519	-3.72
KOG1637	3.47	KOG0271	0.09	KOG1662	-0.938	KOG3049	-3.83
KOG1322	3.33	KOG3318	0.06	KOG2833	-0.96	KOG2792	-3.88
KOG1235	3.25	KOG3387	0.01	KOG0361	-1.01	KOG3482	-3.91
KOG3013	3.11	KOG1299	0.01	KOG0434	-1.12	KOG0376	-4.13
KOG3479	3.00	KOG0419	-0.001	KOG1058	-1.21	KOG0894	-4.26
KOG0969	2.99	KOG0002	-0.002	KOG0960	-1.32	KOG1541	-4.32
KOG0366	2.71	KOG2311	-0.03	KOG0018	-1.39	KOG1636	-4.35
KOG0122	2.55	KOG0329	-0.03	KOG2638	-1.41	KOG1775	-4.53
KOG2529	2.47	KOG3497	-0.04	KOG2851	-1.43	KOG2035	-4.68
KOG1733	2.47	KOG3297	-0.06	KOG0285	-1.44	KOG0179	-4.79
KOG0292	2.37	KOG1760	-0.07	KOG4392	-1.49	KOG2785	-4.96
KOG3147	2.26	KOG2446	-0.07	KOG2036	-1.51	KOG1664	-5.11
KOG1758	2.26	KOG3432	-0.07	KOG0556	-1.66	KOG3205	-5.13
KOG1795	2.16	KOG1498	-0.08	KOG3330	-1.68	KOG1123	-5.17
KOG0650	2.15	KOG3448	-0.08	KOG1597	-1.76	KOG3954	-5.29
KOG3237	2.03	KOG2908	-0.12	KOG1540	-1.77	KOG3239	-5.38
KOG1393	1.99	KOG2784	-0.13	KOG1936	-1.85	KOG1727	-5.46
KOG2535	1.93	KOG3499	-0.13	KOG1774	-2.022	KOG0424	-5.46
KOG1980	1.91	KOG1872	-0.16	KOG0820	-2.03	KOG0182	-5.60
KOG0233	1.86	KOG1532	-0.16	KOG0964	-2.13	KOG1468	-5.75
KOG1800	1.84	KOG3222	-0.18	KOG1523	-2.19	KOG0302	-5.83
KOG2719	1.82	KOG0815	-0.21	KOG2415	-2.23	KOG0346	-6.31
KOG0985	1.72	KOG2623	-0.22	KOG1068	-2.32	KOG1367	-6.73
KOG1555	1.59	KOG0563	-0.22	KOG0688	-2.36	KOG1549	-7.20
KOG2874	1.48	KOG1535	-0.25	KOG0363	-2.38	KOG1746	-7.38
KOG2017	1.40	KOG0184	-0.25	KOG1770	-2.44	KOG0279	-8.39
KOG3275	1.39	KOG1145	-0.26	KOG1769	-2.44	KOG2707	-8.68
KOG3285	1.30	KOG2537	-0.28	KOG1647	-2.62	KOG2807	-8.71
KOG1358	1.26	KOG1885	-0.33	KOG1712	-2.68	KOG3167	-9.12
KOG1355	1.18	KOG0888	-0.37	KOG0062	-2.69	KOG2004	-9.95
KOG3232	1.07	KOG3361	-0.38	KOG0861	-2.73	KOG1241	-11.25
KOG1463	1.01	KOG3048	-0.43	KOG0687	-2.77	KOG2703	-11.51
KOG3503	0.87	KOG3163	-0.45	KOG3218	-2.78		

Tabla 3.5: Valor de ΔGLS para cada gen CEG conservado común. Valores positivos indican apoyo a T1, y valores negativos apoyo a T2

fragmentos génicos presentes en la matriz analizada. Se señalan con sus respectivos nombres los 15 *outliers* encontrados por tener sus valores de ΔGLS fuera de los límites superior e inferior establecidos por las Ecuaciones 3.3 a 3.4 en la página 81 (Sección 3.2.4). Las descripciones y puntajes de dichos genes *outliers* se detallan en la Tabla 3.6. Interesantemente, en la Figura 3.5 se puede apreciar que de los 15 genes más largos, 12 (un 80%) apoyan a T2.

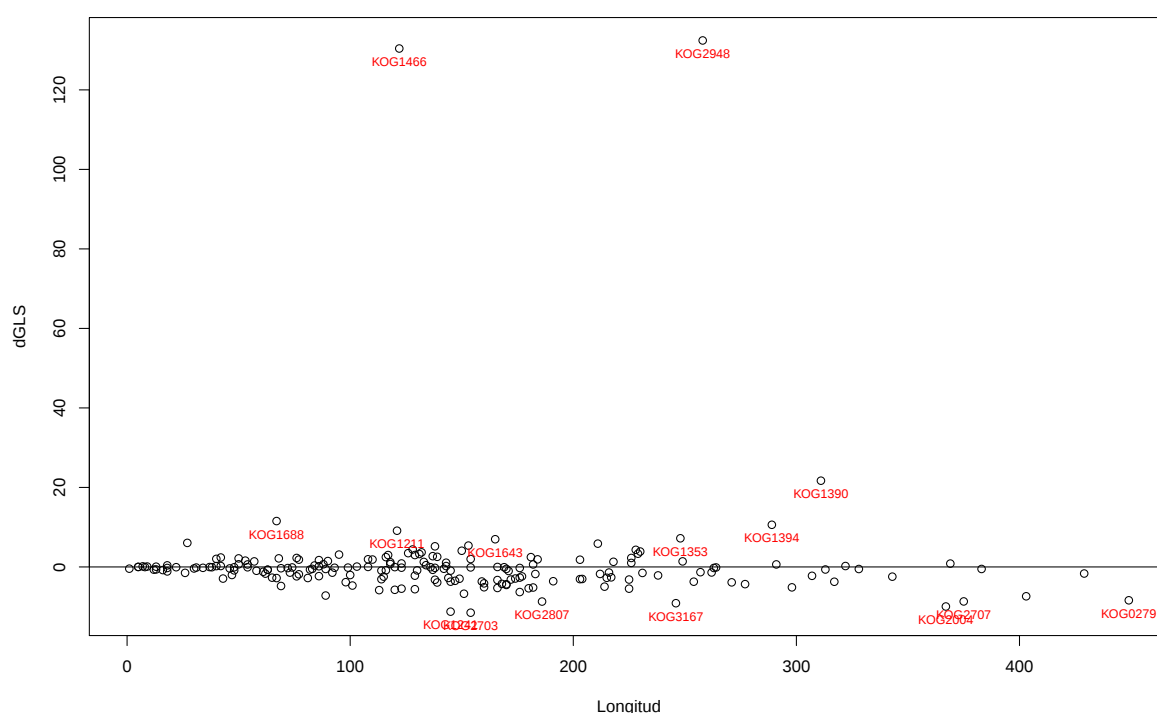


Figura 3.5: Distribución por largo y ΔGLS de los 199 CEGs conservados entre especies

Para cada gen *outlier* se ajustaron modelos de evolución molecular y se realizó un análisis de máxima verosimilitud iguales a lo realizado para la matriz total. En 14 de los 15 casos, el modelo que mejor ajustó fue VT o JTT (los dos modelos preferidos por los genes individuales del análisis en general; ver Figura 3.3), lo cual es indicativo de que las anomalías en señal filogenética de estos genes no estarían originadas por heterotaquia.

Por otro lado, dentro de los 15 *outliers*, 7 de los 8 árboles individuales de genes que apoyan a T1 en el análisis global (valores de ΔGLS positivos; ver Tabla 3.6) presentaron

Proteína	Locus	Descripción	CEG	ΔGLS
Q9SG17	AT3G49320	Metal-dependent protein hydrolase	KOG2948	132.44
Q8W4G0	AT1G72340	Similarity:Belongs to the eIF-2B alpha/beta/delta subunits family.	KOG1466	130.41
Q3E8F0	AT5G47720	Active Site:Acyl-thioester intermediate	KOG1390	21.70
Q9ZWI7	AT2G23310	Protein RER1C	KOG1688	11.54
Q8L3X9	AT2G04540	Beta-ketoacyl-ACP synthase	KOG1394	10.61
Q9LI77	AT3G25660	Glutamyl-tRNA(Gln) amidotransferase subunit A, chloroplastic/mitochondrial	KOG1211	9.12
F4IMB5	AT2G07698	ATP synthase subunit alpha	KOG1353	7.21
P48491	AT3G55440	Triosephosphate isomerase, cytosolic	KOG1643	6.97
O24456	AT1G18080	Guanine nucleotide-binding protein subunit beta-like protein A	KOG0279	-8.39
O22145	AT2G45270	Glycoprotease 1	KOG2707	-8.68
Q9ZVN9	AT1G05055	general transcription factor II H2	KOG2807	-8.71
Q9LEY9	AT5G08180	Nhp2-like protein	KOG3167	-9.12
P93655	AT5G26860	Lon protease homolog 1, mitochondrial	KOG2004	-9.95
Q9FJD4	AT5G53480	Karyopherin subunit beta-1	KOG1241	-11.25
F4K9W6	AT5G22480	ZPR1 zinc-finger domain protein	KOG2703	-11.51

Tabla 3.6: Genes CEG *outliers*. Se indica el producto proteico y el locus de cada uno en *Arabidopsis thaliana*

a Nothofagaceae como hermana de Betulaceae o de alguno de sus géneros presentes en nuestro estudio (*Betula* o *Alnus*; coherente con la hipótesis global T1), y 6 de los 7 árboles individuales de genes que apoyan a T2 en el análisis global (valores de ΔGLS negativos) presentan a Nothofagaceae como clado basal del orden (coherente con la hipótesis global T2). Para los árboles individuales, ver Apéndice B, Figuras B.4 a B.5 en las páginas 199-200. Estos resultados son indicativos del peso que los genes *outliers* tienen sobre el alineamiento global (ver Discusión, Sección 3.4.2 en la página 92).

Análisis con datos corregidos

Una vez filtrados los genes *outliers* de la matriz de caracteres, se analizó la nueva matriz (de 184 genes y 24996 aminoácidos) por máxima verosimilitud. El método para detección y corrección de incongruencias utilizado en este Capítulo se basa en variables y métodos probabilísticos, pero existen desarrollos recientes orientados a cuantificar y disectar la señal filogenética en datos filogenómicos mediante métodos bayesianos [195].

El árbol filogenético obtenido para la nueva matriz sin *outliers* se muestra en la Figura 3.6. Nothofagaceae no se presenta ya como hermana de Betulaceae, sino for-

mando junto a Juglandaceae un clado, aunque poco soportado. Cabe destacar que si bien Nothofagaceae deja su posición de hermana de Betulaceae, no se recupera luego del análisis de incongruencias la topología T2, sino que Fagaceae aparece como clado basal del orden, bien soportado. Ver Discusión, Sección 3.4.2 en la página 92 para posibles interpretaciones.

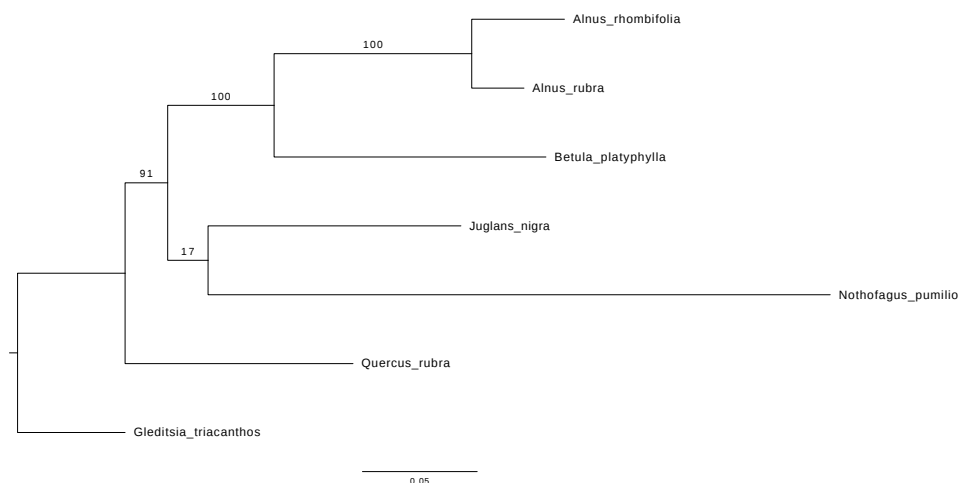


Figura 3.6: Árbol filogenético de máxima verosimilitud obtenido a partir de genes CEGs. Matriz de 7 taxones y 184 genes (24996 aminoácidos). Se indica el soporte de *bootstrap* cada nodo. El largo de las ramas es una estimación de las sustituciones esperadas por sitio

3.4. Discusión

En este Capítulo se presentaron resultados preliminares que conforman el primer estudio filogenómico del orden Fagales. Además de las implicancias evolutivas y para la sistemática de este grupo, discutidas más abajo (Sección 3.4.2 en la página 92), cabe destacar el desarrollo de una herramienta de generación de matrices de genes conservados, listas para el análisis filogenético en los formatos más utilizados. A nuestro entender, este desarrollo novedoso ayuda a cumplir los lineamientos planteados recientemente para construir ordenada y coherentemente un árbol de la vida de las plantas (ver a continuación).

3.4.1. Genes conservados y construcción de la matriz de caracteres

Al trabajar con ensamblados transcriptómicos *de novo* (no habiendo especies modelo de biología molecular en los Fagales), los contigs identificados como genes conservados pueden perfectamente no representar dichos genes completos ni de manera exclusiva (por ejemplo, puede haber partes de más de un gen en un mismo contig quimérico). Por otro lado, aunque los contigs estuvieran perfectamente ensamblados y correspondieran a un gen completo cada uno, las regiones que flanquean a la zona codificante no están conservadas entre especies, y podrían agregar ruido a la señal filogenética.

Ante esto, la herramienta desarrollada en este Capítulo conserva únicamente, para cada gen común, las posiciones del alineamiento con información para todas las especies, de manera de minimizar la presencia de regiones “conflictivas”, muy frecuentes en los extremos de los contigs. La Figura 3.7 ilustra estos problemas con un caso típico, tomado al azar entre los genes CEGs conservados del análisis presentado en este Capítulo. La región que se muestra en la imagen corresponde al final del alineamiento. Hacia la izquierda puede verse una zona altamente conservada, pero sin información para una especie (la tercera de arriba hacia abajo). Esto probablemente se deba a que para esa especie el contig detectado como correspondiente al CEG en cuestión es más corto, ya sea por un error del ensamblado o porque realmente la proteína en esa especie es más corta. Si bien esa región podría agregar información valiosa a la filogenia (computándose como “datos faltantes” las posiciones de la tercera especie), se decidió no tenerla en cuenta, con el fin de evitar la incorporación de regiones como la que se observa a la derecha del alineamiento, que muy probablemente se deba a un mal ensamblado o a la presencia de regiones no codificantes (3'UTR en este caso).

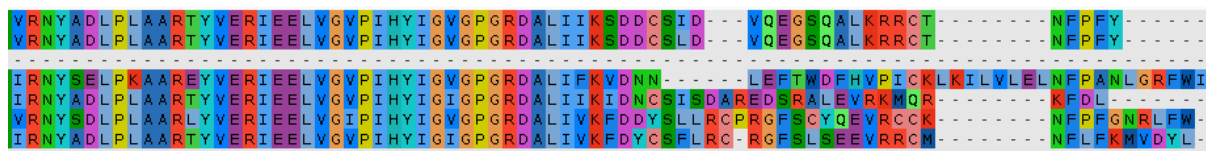


Figura 3.7: Ejemplo de posible conflicto en alineamientos de genes provenientes de ensamblados *de novo*. Ver explicación en el texto

El inconveniente de sacrificar información valiosa en pos de no incorporar ruido

puede ser solucionado en parte mejorando el algoritmo. El código se podría modificar para conservar las regiones muy bien alineadas pero sin información en alguna especie, y descartar únicamente las regiones mal alineadas. Sin embargo, una solución de este tipo requiere muchas pruebas y una cuidadosa puesta a punto ya que la línea divisoria entre lo que es información y lo que es ruido siempre es delgada.

Por otro lado, dicha preocupación puede ni siquiera ser necesaria, ya que el cuello de botella en la longitud de la matriz de caracteres final no es la cantidad de posiciones que tienen información para todas las especies en cada gen, sino el número de genes conservados entre las especies a comparar. Como se ve en las Tablas 3.3 a 3.4 en las páginas 82-83, las bases de datos de BUSCO arrojan menos genes comunes entre especies que la base de datos CEG: en todas nuestras especies en estudio se detectan más de 90 % de los genes CEG, mientras que como máximo se encuentra un 60 % para BUSCOs de plantas y 75 % para BUSCOs eucariotas. Ver Capítulo 2, Sección 2.4.1 en la página 60 para posibles explicaciones al respecto.

Por último, la dicotomía entre si es mejor incluir más *loci* o más taxones, clásica del análisis filogenético, toma gran importancia en la era filogenómica, en la cual hay un enorme incremento en la cantidad de datos disponibles, pero un relativo retraso en el número de especies estudiadas. Por ejemplo, un estudio que incluya transcriptomas o genomas completos de los 37 taxones incluidos en el análisis de Larson-Johnson [236] (o más aún para todas las angiospermas, como los 640 taxones de Soltis *et al.* [183]) es difícil de llevar a cabo fuera de grandes consorcios internacionales, y por el momento es sólo realizable para grupos de gran interés, como insectos o aves (estudios citados en la Introducción). Este tópico ha sido recurrentemente discutido en años recientes [196, 243, 244], y, en resumen, la conclusión es que no es sólo importante la cantidad de genes, sino la señal o ruido que cada uno aporte a la supermatriz.

Respecto de esto último, para el caso de las plantas, como fue presentado en la Introducción, Sección 3.1.2 en la página 73, los casos en los que se trabaja con muchos taxones tratan sobre relaciones profundas y no sobre órdenes o familias particulares. En el afán de aumentar el número de taxones a analizar y la profundidad (en el sentido de ancestralidad) de las preguntas evolutivas estudiadas, al igual que en otros grupos de organismos se descuidaron aspectos relacionados al sesgo de muestreo, sesgo sistemático, y asimetría de información entre grupos (ver Sección 3.1.1 en la página 69).

Algunos trabajos llaman la atención sobre las limitaciones y fuerte desbalance de las secuencias disponibles [245, 246]. Por ejemplo, Hinchliff y Smith [245] utilizan secuencias de GenBank (no sólo -ómicas) de más de 10.000 taxones de plantas e infieren una filogenia de las plantas terrestres. Sin embargo, llaman la atención sobre el gran problema de que más del 99% de las relaciones filogenéticas inferidas descansa sobre menos de 15 *loci*, y alrededor del 50% sobre menos de 4. Los mismos autores, junto con otros colegas, en un trabajo reciente [246] proponen un método para distinguir mal soporte por causa de falta de información, de mal soporte por causa de conflictos entre los datos. Otro trabajo reciente, de Eisenhardt y colaboradores [247], propone un consenso de acciones y pasos a seguir para armar prolijamente el árbol de la vida de las plantas, incluyendo un control de calidad estricto en las secuencias, disponibilidad de métodos, datos crudos y metadatos en repositorios abiertos, y la implementación de *pipelines* automatizadas para servir en diversos grupos de investigación y facilitar la coordinación dentro de la comunidad científica.

El estudio realizado para esta Tesis representa a nuestro entender una prueba de concepto en la dirección de estos lineamientos, ya que presenta información balanceada entre taxones (transcriptomas completos para todas las especies) tomada de repositorios públicos, un control de calidad conservador (búsqueda de genes ortólogos de copia única y filtrado de regiones potencialmente mal alineadas), y la utilización de una sencilla *pipeline* automatizada en conjunto con los programas más utilizados por la comunidad científica para los análisis filogenéticos.

3.4.2. Aportes a la filogenia del orden Fagales

Como se mencionó en la Introducción, para las plantas son pocos los estudios filogenómicos enfocados en grupos particulares. En el marco de esta Tesis (ver Capítulo 1, Secciones 1.4 en la página 32 y 1.5) nos interesa generar y analizar recursos genómicos para *Nothofagus* y profundizar en la aplicación interdisciplinar de herramientas y recursos bioinformáticos para estas importantes especies, poco estudiadas en estos aspectos. Este Capítulo presenta el primer aporte de la filogenómica al estudio de las relaciones filogenéticas entre *Nothofagus* y sus taxones cercanos (orden Fagales), lo cual abre las puertas a investigaciones sobre procesos evolutivos y adaptaciones particulares de este grupo, muy importante dentro de los bosques templados del mundo.

El estudio filogenómico presentado en este Capítulo representa un incremento en el número de genes y caracteres (aminoácidos en nuestro caso) analizados respecto de reportes anteriores para el orden Fagales (citados en la Introducción, Sección 3.1.3 en la página 74). El estudio más actual enfocado en dicho orden fue publicado en el año 2016 por Kathryn Larson-Johnson [236] y utiliza una matriz de 7428 nucleótidos correspondientes a 5 genes (más 89 caracteres morfológicos). Dicha cantidad de nucleótidos equivale a 2476 aminoácidos (con la salvedad de que el análisis de Larson-Johnson utiliza la región ITS, que es transcripta pero no traducida; en todo caso, se podría decir que las matrices de n aminoácidos presentadas en este Capítulo provienen de $n \times 3$ nucleótidos). El alineamiento utilizado para los análisis filogenéticos de este Capítulo consiste de 199 genes (28499 aminoácidos) antes del filtrado de *outliers*, y 184 genes (24996 aminoácidos) después, superando ampliamente la cantidad de caracteres moleculares analizados respecto del estudio de Larson-Johnson y los anteriores enfocados en el orden Fagales. Como fue resaltado en la Introducción (Sección 3.1.1 en la página 69), este incremento en el número de caracteres disminuye fuertemente el sesgo de muestreo o estocástico, pero no el sesgo sistemático (ver más abajo). Por otra parte, cabe destacar que los caracteres morfológicos, estructurales y del registro fósil son extremadamente útiles para la calibración y la interpretación ecológica y evolutiva de los análisis filogenéticos, y en este Capítulo nos limitamos a los aportes desde la filogenómica.

Los resultados presentados en este Capítulo mostraron en un principio a Nothofagaceae como familia hermana de Betulaceae (Figura 3.4 en la página 85). Esta topología no es mencionada en la extensa revisión del sitio web de APG IV [227], pero una búsqueda bibliográfica exhaustiva arrojó que fue sugerida por Nixon en el año 1989, en un volumen colectivo sobre las relaciones del obsoleto grupo polifilético Hamamelidae [248]. Considerando estos resultados conflictivos, se decidió llevar a cabo un análisis de incongruencias, siguiendo el método propuesto por Shen y colaboradores recientemente [196], en línea con las advertencias cada vez más visibilizadas por la comunidad científica respecto del peligro de los sesgos sistemáticos en análisis filogenómicos (ver Introducción, Sección 3.1.1 en la página 69).

Para nuestra matriz de datos, pudimos comprobar que unos pocos genes presentan un gran peso en la construcción de la topología final (ver Sección 3.3.2 en la página 83).

El efecto de remover dichos genes *outliers* modifica la reconstrucción (Figura 3.6 en la página 89) pero no de manera contundente: si bien Nothofagaceae deja de aparecer como hermana a Betulaceae, se presenta como formando un clado junto con Juglandaceae en lugar de como grupo basal (topología T2).

Los autores del método utilizado sugieren que si se encuentran pocos sitios o genes que influyen fuertemente en la topología, las biparticiones en disputa (en nuestro caso, la posición de Nothofagaceae dentro del orden Fagales) deberían pensarse como no resueltas momentáneamente, hasta que haya mayor disponibilidad de datos y/o métodos nuevos para analizarlos. Sin embargo, nuestros análisis (tanto antes como después del filtrado de genes anómalos) coinciden en que Nothofagaceae no ocuparía la posición basal del orden Fagales, como se acepta hoy en día (Figura 3.2 en la página 77), lo cual puede tener importantes implicancias en estudios evolutivos de este diverso grupo de árboles.

En conclusión, los resultados de este Capítulo sugieren que un análisis filogenómico podría cambiar las relaciones aceptadas para este importante grupo de organismos, como lo ha hecho para otros grupos en el árbol de la vida. Si bien hay una falta importante de datos *-ómicos* para muchas especies (y familias enteras) del orden Fagales, la disponibilidad de ensamblados actual y el desarrollo en esta Tesis de un nuevo transcriptoma para la familia Nothofagaceae proveyeron una oportunidad para examinar la calidad de dichos datos y realizar una primera aproximación a la genómica comparativa y filogenómica de este grupo de árboles.

Capítulo 4

Caracterización y respuesta a la temperatura del reloj circadiano de *Nothofagus* spp.

*Cuando se abre en la mañana,
roja como sangre está;
el rocío no la toca
porque se teme quemar.*

*Abierta en el mediodía
es dura como el coral;
el sol se asoma a los vidrios
para verla relumbrar.*

*Cuando en las ramas empiezan
los pájaros a cantar,
y se desmaya la tarde
en las violetas del mar,
se pone blanca, con blanco
de una mejilla de sal.*

*Y cuando toca la noche
blanco cuerno de metal,
y las estrellas avanzan
mientras los aires se van,
en la raya de lo oscuro,
se comienza a deshojar.*

Federico García Lorca, *Doña Rosita la soltera*, Acto I

4.1. Introducción

Como se mencionó en la Introducción General (Capítulo 1, Sección 1.3 en la página 24), el **reloj circadiano** es un oscilador molecular que permite a los organismos anticipar cambios cíclicos en el ambiente, ajustando el funcionamiento de los individuos a dichos cambios. Además de la oscilación de genes del reloj, y de genes involucrados en vías reguladas circadianamente (de hormonas, floración, etc.), existe un gran número de respuestas reguladas por el reloj (llamadas *outputs* o salidas rítmicas del reloj), y que han sido estudiadas en plantas. La caracterización y estudio de estas salidas rítmicas permiten, por un lado, integrar el nivel génico con el nivel de organización fisiológico o de todo el organismo y por lo tanto entender mejor el rol del reloj circadiano en el funcionamiento del individuo como un todo, y por el otro, gracias a su relativo bajo costo económico, escalar el estudio de fenómenos circadianos para incorporar modelados matemáticos, estudios fenómicos, mediciones *in vivo* a campo, etc. En su rol como regulador maestro del crecimiento y desarrollo vegetal, se han descrito muchas vías fisiológicas y a nivel de organismo reguladas por el reloj, entre ellas: crecimiento de hojas [249], movimiento de cotiledones [250, 251], conductancia estomática [54], concentración de clorofilas *a* y *b* [252], fluorescencia de clorofila normal [253, 254] y retrasada [255], concentración de calcio citoplasmático [256], cambio de forma de frutos [257], acumulación de jasmonatos y ácido salicílico en respuesta a herbivoría (también rítmica) [258], o emisión de compuestos volátiles por flores [259, 260].

Como se puede ver, el reloj circadiano regula procesos muy variados, entre ellos varios de importancia agronómica. Además, como se trató también en la Introducción General, los ritmos circadianos son sincronizados por señales ambientales, de manera que son sensibles a las condiciones climáticas presentes y futuras, a través por ejemplo del rango de compensación por temperatura del reloj, propio de cada especie. Así, las vías reguladas por el reloj podrían ser modificadas por cambios en, entre otras variables, temperatura, y verse afectado todo el organismo. El reloj circadiano es, entonces, un sistema interesante para estudiar la relación entre la regulación molecular de procesos celulares y el desenvolvimiento de un individuo o especie fisiológicamente en un ambiente natural. En los últimos años, cada vez más, se está avanzando en la in-

tegración de estos niveles de estudio, y las revisiones al respecto son abundantes [53, 67, 261-266].

En el caso de nuestras especies de interés, *Nothofagus obliqua* y *N. pumilio*, no se cuenta con ninguna caracterización molecular del reloj circadiano ni de las vías reguladas por el mismo. Por presentar estas especies nichos térmicos diferenciados, y ante el presente (y futuro) cambio en las condiciones de temperatura, nos parece interesante estudiar la influencia de la temperatura sobre el funcionamiento del reloj circadiano y su relación con el establecimiento de nichos ecológicos en condiciones naturales (ver Introducción General, Secciones 1.2 en la página 19 y 1.4 en la página 32).

Hasta el momento de esta Tesis no había información sobre estos temas, ni molecular ni ecofisiológica. Así, para cumplir el objetivo general de este Capítulo nos planteamos 3 objetivos específicos (ver Introducción, Sección 1.5 en la página 32): (1) caracterizar el funcionamiento de los relojes biológicos circadianos de *N. obliqua* y *N. pumilio* en condiciones control (20°C) y en temperaturas cálidas (31-34°C), (2) explorar la influencia del reloj circadiano sobre la expresión global de genes *N. pumilio*, y el efecto de la temperatura sobre esta regulación, y (3) explorar en qué medida el comportamiento del reloj circadiano está vinculado con el establecimiento de nichos ecológicos a lo largo de un gradiente altitudinal.

Los recursos -ómicos fueron generados para *N. pumilio* (Capítulo 2), con lo cual el análisis de expresión global de genes se restringe a esa especie. Sin embargo, *N. obliqua* se utilizó como especie “complementaria” para los estudios de laboratorio y de campo por habitar en un nicho térmico contrastante y no solapado con el de la lenga (ver Introducción, Sección 1.2 en la página 19).

4.2. Materiales y Métodos

4.2.1. Material vegetal y diseño experimental en condiciones controladas

El material vegetal y las condiciones de crecimiento y muestreo fueron idénticos a lo expuesto en el Capítulo 2, Sección 2.2.1 en la página 40.

Además de muestreos en condiciones circadianas (ver Figura 4.1, panel A y Figu-

ra 2.3 en la página 41) se realizaron muestreos en condiciones **diurnas** (Figura 4.1, panel B), es decir manteniendo el fotoperíodo luz/oscuridad de la fase de entrenamiento durante el muestreo. Por otro lado, los muestreos para este Capítulo no se realizaron solamente a 20°C y 34°C, sino además a 31°C para condiciones diurnas (y a 28°C para *N. pumilio*, Figura 4.1, panel B).

El muestreo se realizó cada 3 horas luego de un mínimo de 7 días de entrenamiento (más 24 hs. en condiciones de libre curso para los muestreos circadianos) y durante 2 a 4 días. Las muestras fueron tomadas y congeladas de inmediato como se describe en el Capítulo 2, Sección 2.2.1 en la página 40. Todos los muestreos se realizaron por duplicado con plántulas distintas, constituyendo réplicas biológicas independientes. Para *N. pumilio* en condiciones diurnas y circadianas, y para *N. obliqua* en condiciones circadianas, las muestras de ambas réplicas fueron utilizadas para la cuantificación de la expresión génica por qPCR (ver más abajo, Sección 4.2.4 en la página 101), y en las Figuras de este Capítulo se muestra una de ellas. Sin embargo, para el caso de *N. obliqua* en condiciones diurnas, se cuantificaron las muestras provenientes de una única réplica debido a un retraso en fondos asignados para estas líneas de investigación. Dado que los muestreos están hechos y conservados en el freezer, a lo largo de los primeros meses de 2019 se cuantificarán las segundas réplicas de dichos muestreos para la versión definitiva de esta Tesis.

4.2.2. Medición de *outputs* fisiológicos

Las plantas utilizadas para medición de *outputs* fisiológicos siguieron el protocolo expuesto en la Sección anterior y el Capítulo 2, Sección 2.2.1 en la página 40 para los muestreos en condiciones circadianas a 20°C, es decir: entrenamiento de 7 a 10 días con fotoperíodo 12:12 a 20°C, transición de 1 día en luz constante, y muestreo a partir del ZT24.

La conductancia estomática se midió en 4-8 hojas de individuos distintos cada 3 horas durante 2 días utilizando un porómetro Decagon SC-1 (actual Meter Environment SC-1).

La concentración de clorofilas se midió en 3 hojas de individuos distintos cada 3 horas durante 2 días siguiendo el método de Gould [267]: se tomó de cada individuo un disco foliar de 1 centímetro de diámetro y se dejó en oscuridad y 4°C durante 24

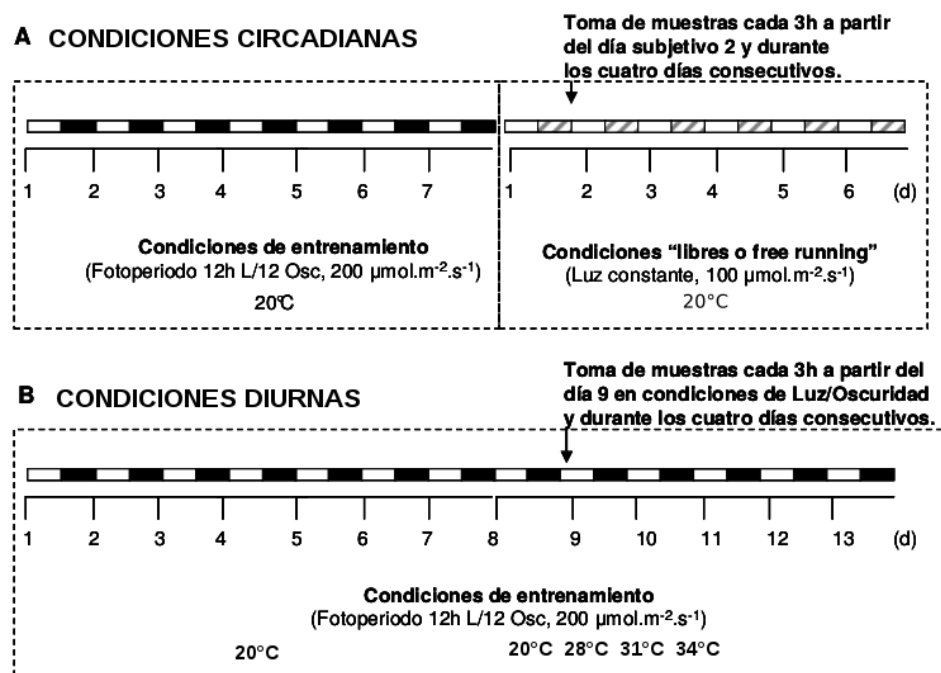


Figura 4.1: Diseño experimental y condiciones de muestreo para experimentos de reloj circadiano en condiciones controladas. Rectángulos blancos: luz. Rectángulos negros: oscuridad. Rectángulos grises rayados: luz (noche subjetiva en condiciones de *free-running*)

horas en 1 mililitro de solución acetona:agua (4:1 V/V). Al término de la extracción, se quitaron los discos foliares de los tubos y se midió la absorbancia a 470, 647 y 663 nm en un espectrofotómetro Ultrospec® 1000 (Amersham Pharmacia Biotech). Las concentraciones de clorofilas se estimaron con las ecuaciones planteadas por Lichtenthaler [268]:

$$[Clorofila_a] = 12.25 \times A_{663} - 2.79 \times A_{647}$$

$$[Clorofila_b] = 21.50 \times A_{647} - 5.10 \times A_{663}$$

El crecimiento de hojas verdaderas se midió a partir de fotos a intervalos regulares (cada una hora, durante 4 días) de una cámara Panasonic DMC-TS25 fijada en la parte superior de la cámara de crecimiento. Con las fotos de 16 megapíxeles de definición se procedió a medir el área de 4 hojas de manera manual con el programa ImageJ [269], y computar el crecimiento como tasa horaria de diferencia de área.

En todos los casos el análisis de los datos y los gráficos se realizaron mediante *scripts* propios en lenguaje R, versión 3.2.3 [145].

4.2.3. Caracterización *in silico* del reloj circadiano

Para la búsqueda de genes del reloj circadiano en *N. pumilio* se acudió a una revisión publicada recientemente en *Trends in Plant Science* [65] que describe los componentes del reloj en *Arabidopsis thaliana*. A partir de la Tabla 1 de dicho artículo, se procedió a la búsqueda de los identificadores de *A. thaliana* con un programa idéntico al descrito en el Capítulo 2, Sección 2.3.6 en la página 58, pero que no busca en los 4 ensamblados separados estudiados en dicho Capítulo, sino en el ensamblado total utilizado en este (ver más adelante, Sección 4.2.6 en la página 104).

Para los contigs de lenga que mejor puntaje de anotación presentaron contra los genes de reloj de *A. thaliana* se realizó una búsqueda de dominios proteicos *in silico* a partir de modelos ocultos de Markov (ver Capítulo 2, Sección 2.2.4 en la página 44) de la base de datos Pfam [270], utilizando el programa HMMER [271] (versión 3.1b1, opción `--nobias`).

4.2.4. Medición de la expresión génica

No habiendo antecedentes respecto de la extracción de ARN y medición de la expresión génica en especies de *Nothofagus*, se procedió a realizar una puesta a punto de la extracción, purificación y retrotranscripción del ARN de hojas de *Nothofagus*, y la medición de la expresión génica por PCR cuantitativa (qPCR).

Las muestras fueron molidas manualmente con mortero bajo nitrógeno líquido con el objetivo de evitar el descongelamiento en todo momento. Tanto los morteros como las espátulas, mesadas, gradillas y pipetas fueron tratados con una solución de SDS 0.1 % y NaOH 0.1 M para inactivar RNasas. Para la extracción de ARN se probaron los siguientes métodos:

- Kit comercial de Sigma-Aldrich (Spectrum™ Plant Total RNA Kit, catálogo STRN50).
- Reactivo TRIzol™ (ThermoFisher Scientific, catálogo 15596026) en conjunto con cloroformo, isopropanol y etanol de nuestro laboratorio.
- Método de Louime [272], puesto a punto originalmente en hojas de vid.
- Método de Chang [135], puesto a punto originalmente en plántulas de pino.

Para la eliminación del ADN genómico residual se utilizó la enzima DNase I Amplification Grade de Sigma-Aldrich (catálogo AMPD1).

Para la purificación del ARN extraído se probaron los siguientes métodos:

- Filtrado del ARN a través de columnas de PVPP (poli-vinil-poli-pirrolidona) al 10 %.
- Reprecipitación con acetato de amonio [273].
- Reprecipitación con cloruro de litio [273].
- Kit comercial de Qiagen (RNeasy® Plant Mini Kit, catálogo 74904).

Para la retrotranscripción se probaron los siguientes métodos:

- Kit comercial de ThermoFisher Scientific (SuperScript® III First-Strand Synthesis System for RT-PCR, catálogo 18080-051).
- Enzimas comerciales de Promega: M-MLV Reverse Transcriptase (catálogo M1701) y Recombinant RNasin® Ribonuclease Inhibitor (catálogo N2511).

Para la preparación de la PCR cuantitativa se utilizaron las *mix* comerciales FastStart Universal SYBR Green Master (Roche, catálogo 4913850001) y SsoAdvanced™

Universal SYBR® Green Supermix (Bio-Rad, catálogo 1725271). Las reacciones se llevaron a cabo en los equipos StepOnePlus™ (Applied Biosystems™, catálogo 4376600) y CFX96 Touch™ (Bio-Rad, catálogo 1855195). En todos los casos, el protocolo de ciclado fue: 3 minutos a 95°C, y 40 ciclos de 10 segundos a 95°C y 30 segundos a 60°C. Al final de cada uno de los ciclos el equipo tomó la fotografía para medir fluorescencia. A este protocolo de ciclado y amplificación se agregó una etapa de análisis de curva de disociación (*melting curve analysis*) para evaluar la especificidad de la amplificación y la presencia de contaminaciones (por ejemplo dímeros de *primers*): gradiente de 65°C a 95°C, cada 0.5°C, durante 5 segundos seguido de la fotografía para detección de fluorescencia.

Al finalizar la puesta a punto (ver Resultados, Sección 4.3.3 en la página 112) se llegó al siguiente protocolo para la medición sistemática de la expresión génica:

PROTOCOLO DEFINITIVO PARA MEDICIÓN DE EXPRESIÓN GÉNICA EN
HOJAS DE *NOTHOFAGUS*

1. Extracción de ARN según Chang [135].
2. Eliminación de ADN genómico con DNase I de Sigma-Aldrich.
3. Purificación del ARN con kit comercial RNeasy® Plant Mini Kit de Qiagen.
4. Retrotranscripción con M-MLV Reverse Transcriptase y RNasin® Ribonuclease Inhibitor de Promega.
5. PCR cuantitativa con SsoAdvanced™ Universal SYBR® Green Supermix y en un equipo CFX96 Touch™ de Bio-Rad.

Para la relativización de la expresión de los genes de interés en la PCR cuantitativa se utilizaron, salvo indicación contraria, dos genes de referencia (ver a continuación, Sección 4.2.5, para la búsqueda y diseño de *primers*). Cada reacción de qPCR se realizó por triplicado técnico. El análisis de expresión relativa de los genes de interés se realizó en planillas de cálculo siguiendo el método $\Delta\Delta C_t$ [274]. Los gráficos se realizaron en el programa Prism, versión 5.03 (GraphPad Software) y se exportaron como imágenes vectoriales.

Para la preparación de la PCR de punto final se utilizó la enzima comercial (y su correspondiente *buffer*) GoTaq® DNA polymerase (Bio-Rad, catálogo M3005). El protocolo de ciclado fue: 3 minutos a 94°C, 35 ciclos de 40 segundos a 94°C, 40 segundos

a 60°C y 40 segundos a 72°, y 7 minutos a 72°C. Las reacciones se llevaron a cabo en un equipo MyCycler™ (Bio-Rad, catálogo 170-9703, discontinuado). Los geles de agarosa fueron teñidos con SYBR™ Safe DNA Gel Stain (ThermoFisher Scientific, catálogo S33102) y observados con un equipo Safe Imager™ Blue-Light Transilluminator (ThermoFisher Scientific, discontinuado, modelo actual catálogo G6600). Como marcador de peso molecular se utilizó el estándar 1Kb Plus DNA Ladder (ThermoFisher Scientific, catálogo 10787018).

4.2.5. Diseño de primers

Para el diseño de los *primers* de PCR (tanto de punto final como cuantitativa) se utilizaron secuencias del transcriptoma de *N. pumilio* (ver más adelante, Sección 4.2.6 en la página siguiente).

Para los genes del oscilador central *TOC1* y *LHY* se buscaron las correspondientes secuencias publicadas de un árbol cercano (el castaño europeo *Castanea sativa*, en el cual se realizaron estudios de reloj circadiano [85, 86]) en el transcriptoma ensamblado de la lenga, utilizando el programa BLAT [157]. Los *primers* se diseñaron sobre las regiones más conservadas entre especies utilizando el programa online Primer3 [275]. Para otro gen del reloj, *GIGANTEA (GI)*, se procedió de igual manera (alineamiento con BLAST-BLAT y restricción a zonas conservadas entre especies para el diseño de primers), pero sin la ayuda de una especie tan cercana como el castaño.

Para el gen *PP2A*, utilizado como referencia de qPCR al principio de esta tesis (ver más abajo para la mejora aplicada a partir del análisis de RNA-seq), se diseñaron *primers* a partir del transcriptoma base del raulí [36] para secuenciar por electroforesis capilar (método de Sanger) fragmentos en las dos especies de *Nothofagus* de interés (roble pellín y lenga), y a partir de un consenso entre ellas diseñar primers para qPCR.

Para los genes de referencia de qPCR se partió del análisis de expresión diferencial del transcriptoma de la lenga (ver a continuación, Sección 4.2.6 en la página siguiente). Allí se buscaron contigs que cumplieran con todas las siguientes condiciones:

- No cambien su expresión en ninguna de las 4 comparaciones realizadas (entre momentos del día y entre temperaturas; ver siguiente Sección)
- Tengan una expresión de media a alta en las 4 condiciones estudiadas
- Estuvieran anotados contra *Arabidopsis thaliana*

- Los genes correspondientes de *A. thaliana* tengan una expresión de media a alta
- Los genes correspondientes de *A. thaliana* no oscilen en condiciones circadianas y entrenamiento con fotoperíodo 12:12 [156]
- Los genes correspondientes de *A. thaliana* tengan expresión estable ante estrés térmico (3 horas a 38°C) [155]

Se encontraron 5 contigs de lenga correspondientes a 3 genes de *A. thaliana* (Tabla 4.1). Se diseñaron *primers* con el programa online Primer3 [275] sobre las regiones más conservadas entre lenga, *A. thaliana* y otras especies. Luego de probados los *primers* mediante PCR de punto final, se decidió usar dos genes como referencia: *P2C22* y *DER2.2*.

Nombre de la proteína	Nombre abreviado	Identificador del locus	Identificador de la proteína
Probable lysine-specific demethylase JM14	JM14	AT4G20400	Q8GUI6
Probable protein phosphatase 2C 22	P2C22	AT2G25620	Q9SLA1
Derlin-2.2	DER2.2	AT4G04860	Q9ZS88

Tabla 4.1: Genes de *Arabidopsis thaliana* candidatos para referencia de qPCR

Las secuencias de los *primers* utilizados se detallan en la Tabla 4.2.

4.2.6. Análisis de expresión diferencial

Para realizar un análisis de expresión diferencial a partir de un experimento de transcriptómica, es necesario partir de lecturas limpias y un ensamblado robusto. El proceso para llegar a dicho punto fue descrito en detalle en el Capítulo 2. A diferencia de dicho Capítulo, en el cual se describieron 4 ensamblados separados para cada condición biológica (y se expusieron las razones por las cuales este enfoque es bueno para el estudio detallado de genes y vías metabólicas), para realizar un análisis de expresión diferencial es más conveniente tener un único ensamblado sobre el cual se cuantificarán todas las lecturas provenientes de las distintas condiciones (ver más abajo).

Gen	Sentido	Secuencia
<i>TOC1</i>	F	5' AATGCTTGGCCATTATACGG 3'
	R	5' GCCTTCTTTCAGCAAGTCGT 3'
<i>LHY</i>	F	5' TGGCCTTTCAAGCACTCTTT 3'
	R	5' TGCCTCCCAAATTTGTGTTT 3'
<i>GI</i>	F	5' GCCTACCAGCTACAATCCGA 3'
	R	5' TTTCAATGTCGGCTTGCCAA 3'
<i>PP2A</i>	F	5' TGGGAATGGCACCAGTGTTA 3'
	R	5' TTGCTGGCAAGAGAGACTGA 3'
<i>DER2.2</i>	F	5' TGCTTGGGTGGATCTACTGG 3'
	R	5' AATCTCACATTTGCAGGCCG 3'
<i>P2C22</i>	F	5' GAGCCAAAATGCTGTGGATT 3'
	R	5' AGCTTTGCAACTCCCTCAAA 3'

Tabla 4.2: Secuencias de *primers* de PCR de punto final y cuantitativa utilizados en este Capítulo

Las mismas lecturas cuyo tratamiento se trata en la Sección 2.2.3 en la página 42 fueron combinadas y ensambladas con SPAdes [125] (versión 3.10.0; parámetros: --rna -iontorrent -k67). A continuación, el ensamblado fue analizado con *Lenga* de igual manera a lo descrito en las Secciones 2.2.3 a 2.2.5 en las páginas 42-46.

Volviendo al análisis de expresión diferencial, el primer paso una vez obtenido un buen ensamblado y lecturas limpias consiste en cuantificar cada juego de lecturas (provenientes de una única *library*) sobre el ensamblado. Las cuantificaciones de nuestras 8 *libraries* (ver Capítulo 2 respecto de las condiciones biológicas estudiadas) fueron realizadas con el programa Salmon [144] (versión 0.8.1; parámetros --seqBias -1 U).

A partir de las cuantificaciones de cada condición biológica, se armó una tabla que condensara la cuantificación de cada contig en cada condición. Esta tabla es la entrada para el análisis de expresión diferencial, para el cual se utilizó el paquete DESeq [276], parte del proyecto Bioconductor, dentro del entorno R [145]. Este paquete, en pocas palabras, compara la variabilidad entre réplicas biológicas de una condición respecto de la variabilidad entre condiciones biológicas. Se realizaron las siguientes compara-

ciones pareadas, de interés biológico para nuestros estudios:

- 20°C ZT48 vs. 20°C ZT60 (*entre momentos del día, a 20°C*)
- 20°C ZT48 vs. 34°C ZT48 (*entre temperaturas, en el amanecer subjetivo*)
- 34°C ZT48 vs. 34°C ZT60 (*entre momentos del día, a 34°C*)
- 20°C ZT60 vs. 34°C ZT60 (*entre temperaturas, en el atardecer subjetivo*)

Cada una de las condiciones enumeradas cuenta con dos *libraries* cuantificadas, provenientes de réplicas biológicas. Para evaluar el efecto de la temperatura sobre la expresión global de genes, independientemente del momento del día, se realizó un quinto análisis, tomando las 4 muestras de 20°C (2 del amanecer subjetivo, y 2 del atardecer subjetivo) como réplicas, y las 4 muestras de 34°C de igual manera.

La salida del análisis de expresión diferencial muestra, para cada contig, su expresión media en cada condición comparada, el *fold change* (cuántas veces más o menos se expresa en una condición o en la otra) tanto en escala lineal como logarítmica, y un valor de probabilidad (*p-value* o p-valor) de que dicho cambio sea significativo estadísticamente. Los contigs cuyo p-valor es menor que 0.05 son llamados **regulados** o **diferencialmente expresados** entre las condiciones comparadas.

A partir de los contigs diferencialmente expresados en una o en alguna de todas las condiciones, se pueden realizar agrupamientos (*clustering*) para encontrar patrones de expresión. Los agrupamientos (sin supervisión, es decir sin proveerle al algoritmo ninguna información acerca de los contigs más que su expresión) y gráficos fueron realizados en los lenguajes R (versión 3.2.3) [145] y Python (versión 2.7.14) [161], combinando paquetes y funciones existentes con desarrollos propios.

Para encontrar patrones de expresión se sometió a los genes de *Arabidopsis thaliana* correspondientes a los contigs anotados para cada comparación a un análisis de enriquecimiento de términos de Gene Ontology [158] utilizando el programa PANTHER [277].

Además del criterio de p-valor de DESeq menor a 0.05, adoptamos un criterio *ad hoc* debido a la mucha menor intensidad de la regulación génica diurna en relación al efecto de la temperatura, y a la exigencia estadística producto de contar con sólo dos réplicas biológicas (ver Secciones 4.3.5 en la página 116 y 4.4.2 en la página 139). Este criterio consistió en considerar como “regulados” aquellos contigs cuyo *fold chan-*

ge entre condiciones fuera mayor a 2, y que tuvieran expresión intermedia (entre la mediana y el tercer cuartil de la distribución de la expresión).

4.2.7. Muestreos en condiciones naturales

Sitios de estudio

Los experimentos de campo se llevaron a cabo en parcelas permanentes que posee el grupo de Genética Ecológica y Mejoramiento Forestal del INTA Bariloche, mediante un convenio con la Administración de Parques Nacionales. Dichas parcelas se encuentran ubicadas en la cuenca Lácar del Parque Nacional Lanín (provincia de Neuquén), en las cercanías de San Martín de los Andes (Figura 4.2).



Figura 4.2: Cuenca del Lago Lácar. Se destacan con un punto rojo el cerro Quilanlahue, en la cabecera este del lago la localidad de San Martín de los Andes, y el cordón Chapelco al sudeste

El sistema de parcelas permanentes consiste en un gradiente altitudinal de 3 sitios de 2 ha. cada uno, ubicados a 680, 930, y 1100 metros sobre el nivel del mar (msnm) en la ladera sur del cerro Quilanlahue. Estas parcelas se encuentran inventariadas por el equipo de Técnicos Forestales del PN Lanín. Dicho trabajo muestra la dominancia de *Nothofagus obliqua* en el estrato inferior, *N. alpina* en el estrato medio, y *N. pumilio* en el estrato superior. Resultados previos del grupo [278] muestran que la temperatura es la principal variable que cambia con la altura en este sistema, sobre todo en la estación de crecimiento (Ref. [278], Figura S1).

Expresión génica

Las plantas utilizadas para medir la expresión de genes de reloj circadiano en condiciones naturales fueron crecidas en vivero durante 2 meses en bandejas con 40 tubetes de 90 cm³ cada una, con una mezcla de partes iguales de tierra negra, arena volcánica y turba, y fueron llevadas al gradiente altitudinal. Luego de dos meses de entrenamiento lumínico y de temperatura en el campo, en enero de 2013 (antes del inicio de esta Tesis) se muestrearon entre 5 y 10 hojas de individuos distintos por especie por estrato cada 3 horas durante 2 días. Las hojas fueron congeladas inmediatamente en nitrógeno líquido y, una vez en el laboratorio, conservadas a -80°C.

Para la medición de la expresión génica se procedió según el protocolo puesto a punto en esta tesis (ver Sección 4.2.4 en la página 101 y Resultados a continuación).

Emergencia y supervivencia de plántulas

Para las mediciones de emergencia y supervivencia se llevaron a los sitios del gradiente altitudinal semillas de *N. pumilio* y *N. obliqua* en junio de 2013 (antes del inicio de esta Tesis). Cada especie en cada estrato altitudinal contaba con entre 8 y 14 líneas de 40 o 25 (para lenga y roble respectivamente) semillas por línea.

La emergencia se monitoreó en octubre y noviembre de 2013, y la supervivencia se contó en febrero de 2014 sobre la base de las plántulas emergidas totales (de 3 a 10 por línea, es decir de 24 a 140 plántulas por especie y por estrato).

Área foliar y clorofilas

Las plantas utilizadas para medición de clorofilas y área foliar fueron criadas en invernadero durante una temporada en bandejas con 28 tubetes de 265 cm³ cada una. El sustrato fue inerte, de partes iguales de arena volcánica y turba, y se siguió un protocolo de crecimiento y riego con nutrientes (fertirrigación) puesto a punto especialmente para *Nothofagus* en el Campo Forestal Trevelin del INTA – EEA Esquel (Chubut) [279]. Al final de la temporada de crecimiento, 1200 plantas (400 por especie, de *N. pumilio*, *N. obliqua* y *N. alpina*) fueron llevadas al gradiente altitudinal en junio de 2017 y plantadas, dando un total de 133 plantas por especie y por estrato.

Al final de la siguiente temporada de crecimiento (habiendo pasado un invierno y una temporada en el campo) se muestrearon 30 hojas por especie y por estrato, que

fueron congeladas inmediatamente en nitrógeno líquido para evitar la degradación de las clorofilas. Una vez en el laboratorio las hojas fueron conservadas en -80°C hasta la medición de área foliar y clorofilas.

Para la medición de clorofilas se procedió idénticamente a la Sección 4.2.2 en la página 98. Para área foliar, se escanearon las hojas enteras recién descongeladas con un escáner de escritorio y se realizó una medición de área por contraste utilizando el programa ImageJ [269].

4.3. Resultados

4.3.1. *Outputs* fisiológicos del reloj circadiano

En la Figura 4.3 se muestran las curvas temporales para los 3 *outputs* medidos en condiciones circadianas en *N. pumilio* (y *N. obliqua* para el caso de la conductancia estomática). Ninguna de las variables medidas arrojó un patrón de oscilación diario, a diferencia de lo reportado en la literatura para otras especies (ver Introducción, Sección 4.1 en la página 96, y Discusión, Sección 4.4.2 en la página 139).

4.3.2. Caracterización *in silico* de genes del reloj

En la Tabla 4.3 se presentan los homólogos de genes del reloj circadiano de *Arabidopsis thaliana* encontrados en el ensamblado transcriptómico de *N. pumilio*. Estos genes representan más de la mitad de los listados por Hsu y Harmer [65], lo cual es satisfactorio considerando la redundancia de miembros de familias multigénicas y la baja expresión de los genes centrales del reloj (ver más adelante, Sección 4.3.5 en la página 116 y Discusión, Sección 4.4.2 en la página 139).

Los genes homólogos encontrados en lenga están distribuidos entre genes con funciones centrales en el reloj (como *TOC1* y los otros *PRRs*, *LHY*, *ZTL*, o *CHE*), genes con otras funciones además de la regulación circadiana (como *GI*, involucrado en floración, o *SKIP*, relacionado con respuestas a estrés abiótico), y genes con roles hasta el momento desconocidos en el reloj (como *LWD1*, *LIP1* o *TIC*).

Para cada gen en la Tabla 4.3 se indica el puntaje de anotación obtenido para el contig (de lenga) y el blanco (de *A. thaliana*) indicados. Como se indicó en el Capítulo

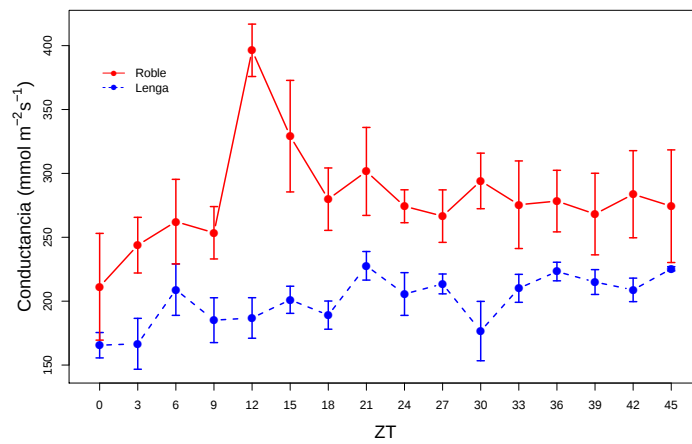
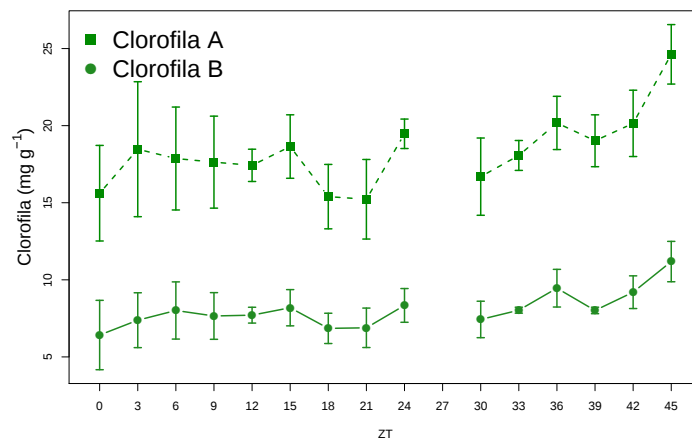
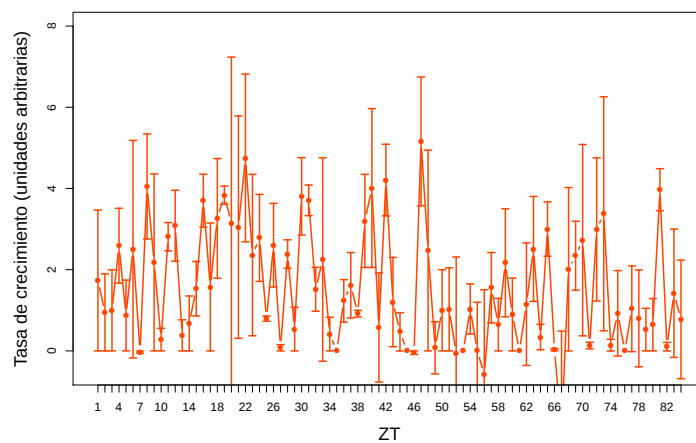
(a) Conductancia estomática ($n = 4-8$)(b) Concentración de clorofilas ($n = 3$)(c) Crecimiento de hojas verdaderas ($n = 4$)

Figura 4.3: Algunos *outputs* fisiológicos del reloj circadiano medidos en *N. pumilio* (y *N. obliqua* para conductancia estomática). Se grafica la media y la desviación estándar para cada punto

2, Sección 2.2.5 en la página 46, dicho puntaje toma en cuenta la similitud entre las secuencias de entrada y de referencia, relativizada a la longitud de la secuencia de entrada. De esta manera, un puntaje alto no indica necesariamente una buena cobertura del gen de *A. thaliana*, sino una buena similitud a lo largo del contig de lenga.

Gen	Locus	Producto proteico	Nombre	Contig	Puntaje de anotación
<i>PRMT5</i>	AT4G31120	Q8GWT4	Protein arginine N-methyltransferase 1.5	NO-DE.12505	1.0
<i>PRR7</i>	AT5G02810	Q93WK5	Two-component response regulator-like APRR7	NO-DE.70999	0.96
<i>PRR3</i>	AT5G60100	Q9LVG4	Two-component response regulator-like APRR3	NO-DE.55466	0.92
<i>GI</i>	AT1G22770	Q9SQI2	Protein GIGANTEA	chain.5147	0.92
<i>ZTL</i> (<i>ADO1</i>)	AT5G57360	Q94BT6	ZEITLUPE (Adagio protein 1)	chain.7038	0.91
<i>TEJ</i>	AT2G31870	Q9SKB3	Poly(ADP-ribose) glycohydrolase 1	chain.14749	0.88
<i>FIO1</i>	AT2G21070	Q5XEU1	U6 small nuclear RNA (adenine-(43)-N(6))-methyltransferase	NO-DE.2604	0.86
<i>SKIP</i>	AT1G77180	O80653	SNW/SKI-interacting protein	NO-DE.1168	0.86
<i>LWD1</i>	AT1G12910	Q9LPV9	WD repeat-containing protein LWD1	NO-DE.1119	0.84
<i>LIP1</i>	AT5G64813	Q9C5J9	Small GTPase LIP1	chain.12026	0.83
<i>ELF3</i>	AT2G25930	O82804	Protein EARLY FLOWERING 3	chain.49983	0.79
<i>TCP21</i> (<i>CHE</i>)	AT5G08330	Q9FTA2	Transcription factor TCP21 (CCA1 Hiking Expedition)	NO-DE.19543	0.71
<i>ELF4</i>	AT2G40080	O04211	Protein EARLY FLOWERING 4	chain.70789	0.66
<i>RVE8</i>	AT3G09600	Q8RWU3	Protein REVEILLE 8	chain.1813	0.65
<i>TIC</i>	AT3G22380	Q94KE2	Protein TIME FOR COFFEE	NO-DE.24579	0.60
<i>CKB4</i>	AT2G44680	O80507	Putative casein kinase II subunit beta-4	chain.16922	0.57
<i>TOC1</i> (<i>PRR1</i>)	AT5G61380	Q9LKL2	Timing of CAB Expression 1 (Two-component response regulator-like APRR1)	chain.71062	0.55
<i>LHY</i>	AT1G01060	Q6R0H1	Protein LHY	chain.6998	0.36
<i>PRR5</i>	AT5G24470	Q6LA42	Two-component response regulator-like APRR5	NO-DE.1012	0.25

Tabla 4.3: Genes del reloj circadiano de *Arabidopsis thaliana* encontrados *in silico* en *N. pumilio*

La Tabla 4.4 muestra los dominios proteicos de la base de datos Pfam encontrados en los contigs correspondientes a cada gen de la Tabla 4.3. Los dominios indicados como “Ausentes” no se encuentran en el contig de *N. pumilio* pero sí en la referencia

de *A. thaliana*. En total, de los 47 dominios presentes en los 19 genes de referencia, los contigs mejor anotados de *N. pumilio* tienen 39, es decir un 83% de presencia y conservación de dominios entre especies. Las proteínas GI, ELF3 y TIC no tienen dominios proteicos descritos, y coherentemente tampoco los contigs anotados contra dichos blancos poseen dominios identificables.

4.3.3. Puesta a punto de protocolos de laboratorio

El kit comercial de Sigma-Aldrich tuvo un rendimiento irregular, dando ARN en algunas muestras pero no en otras (datos no presentados). De los otros métodos de extracción probados, sólo los de Chang y Louime dieron ARN consistentemente en las 3 especies de interés (Figura 4.4). Entre estos se eligió para todos los experimentos posteriores el método de Chang por haber rendido extracciones de ARN con perfiles electroforéticos cuyos ARN ribosomales eran bien visibles y sin ADN genómico en la parte superior del gel. Además, el método de Chang ha sido utilizado en *Nothofagus alpina* [36] y *Quercus petraea* [280] entre otras especies cercanas.

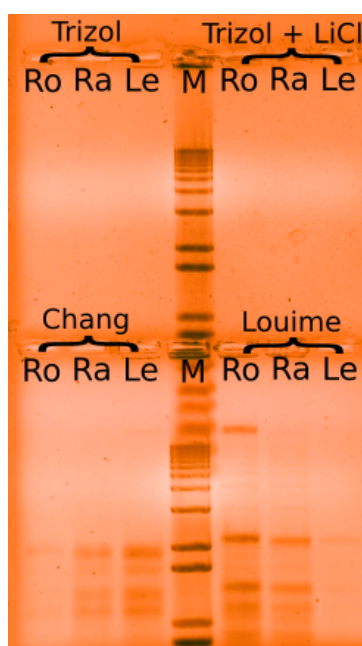


Figura 4.4: Puesta a punto de métodos de extracción de ARN. Gel de agarosa 0.8% p/V. ; M: marcador de peso molecular. Ro: roble; Ra: raulí; Le: lenga

A pesar de contar con ARN extraído en buena cantidad, la purificación fue de gran importancia, como se puede ver en la Figura 4.5. Las muestras no purificadas no

Gen	Contig	Presentes			Ausentes
<i>PRMT5</i>	NODE_12505	PRMT5			-
<i>PRR7</i>	NODE_70999	Response_reg			CCT
<i>PRR3</i>	NODE_55466	Response_reg			CCT DNA_pol_phi
<i>GI</i>	chain_5147	-			-
<i>ZTL</i>	chain_7038	F-box	F-box-like	Kelch_1	PAS_8
		Kelch_2	Kelch_3	Kelch_4	
		Kelch_5	Kelch_6	PAS	
		PAS_3	PAS_9		
<i>TEJ</i>	chain_14749	PARG_cat			-
<i>FIO1</i>	NODE_2604	Methyltransf_10			-
<i>SKIP</i>	NODE_1168	SKIP_SNW			-
<i>LWD1</i>	NODE_1119	WD40			-
<i>LIP1</i>	chain_12026	AAA_16	AAA_24	Arf	Sigma54_activat
		DUF258	GTP_EFTU	Gtr1_RagA	
		Miro	MMR_HSR1	Ras	
			SRPRB		
<i>ELF3</i>	chain_49983	-			-
<i>CHE</i>	NODE_19543	TCP			-
<i>ELF4</i>	chain_70789	DUF1313			SSXT
<i>RVE8</i>	chain_1813	Myb_DNA-binding			-
		Myb_DNA-bind_6			
<i>TIC</i>	NODE_24579	-			-
<i>CKB4</i>	chain_16922	CK_II_beta			-
<i>TOC1</i>	chain_71062	Response_reg			-
		CCT			
<i>LHY</i>	chain_6998	Myb_DNA-binding			-
		Myb_DNA-bind_6			
<i>PRR5</i>	NODE_1012	Response_reg			DUF305
		CCT			HCV_NS5a_C

Tabla 4.4: Dominios proteicos de genes del reloj circadiano de *A. thaliana* encontrados *in silico* en *N. pumilio*

fueron capaces de amplificar fragmentos de 3 genes distintos por PCR, mientras que las muestras purificadas con las columnas comerciales de Qiagen sí. Los otros métodos de purificación probados no fueron capaces de recuperar ARN (datos no presentados).

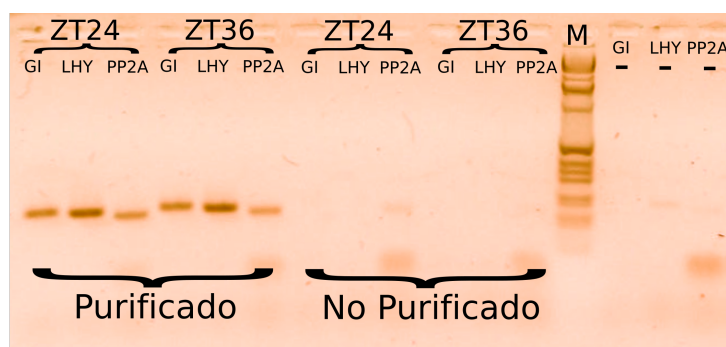


Figura 4.5: Puesta a punto de la purificación del ARN. Gel de agarosa 2% p/V. **M**: marcador de peso molecular. ADNc de lenga, condiciones circadianas, 34°C

Una vez extraído y purificado el ARN, se procedió a probar métodos de retrotranscripción y cuantificación por qPCR. Como se puede ver en la Figura 4.6, la mix de qPCR de Roche no funcionó en nuestras muestras, y no se observó una diferencia entre las muestras retrotranscriptas con M-MLV o con SS-III.

En vistas de esto, se procedió a probar la PCR cuantitativa (usando la mix comercial de BioRad) con las muestras extraídas según Chang, purificadas en las columnas comerciales, y retrotranscriptas con M-MLV.

En la Figura 4.7 se muestra un ejemplo de los resultados de puesta a punto de la qPCR, para el gen *GI* con ADNc de lenga. Si bien no es frecuente en las publicaciones, este tipo de análisis debe ser hecho para cada par de primers previamente a sacar conclusiones respecto de expresión génica, debido a la posibilidad de amplificación inespecífica, formación de dímeros de primers, o no-linealidad de la reacción en el rango de concentración de ADNc estudiado.

En el panel 4.7a se observa la amplificación en escala lineal de 3 réplicas técnicas de cada una de 4 diluciones seriadas. El espaciamiento regular y la similaridad entre réplicas técnicas se aprecia mejor en el panel 4.7b, donde se grafica el ciclo de amplificación de cada muestra contra su concentración inicial relativa. A partir de la pendiente de este gráfico se calcula la eficiencia, que debe ser entre 90 y 110% (en el caso de este par de primers fue de 99.9%). Por último, en el panel 4.7c se observa la

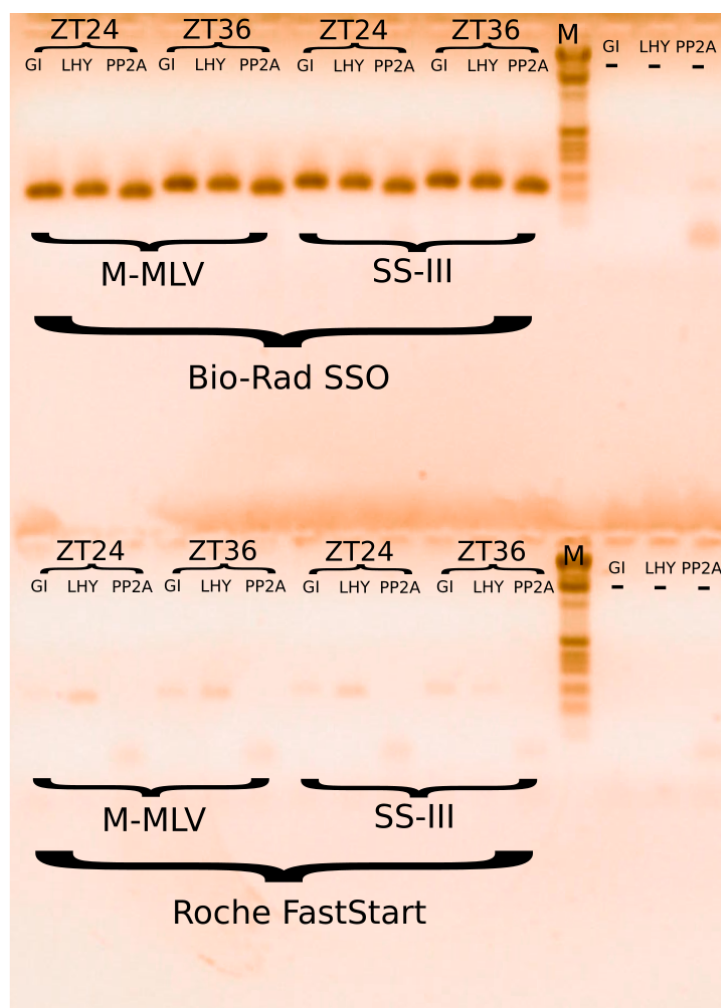


Figura 4.6: Puesta a punto de la retrotranscripción y la PCR cuantitativa. Gel de agarosa 2% p/V. **M**: marcador de peso molecular. ADNc de lenga, condiciones diurnas, 20°C

curva de disociación de los primers. Un único pico para todas las muestras indica la presencia de un único amplicón y la ausencia de dímeros de primers.

4.3.4. Caracterización de la expresión de genes del oscilador central de *Nothofagus* spp. en condiciones controladas

Las Figuras 4.8 y 4.9 muestran la expresión del gen homólogo a *TOC1* a lo largo de 1 día en condiciones diurnas a 20°C (temperatura control). Se puede observar, para ambas especies, una amplia oscilación con una fase coherente con lo reportado en la literatura para este gen (hacia el final del período de luz; ver por ejemplo la Ref. [65]). La Figura 4.8 muestra también la expresión de *NpTOC1* a 28°C, donde se observa igualmente una fuerte oscilación en la misma fase que a 20°C.

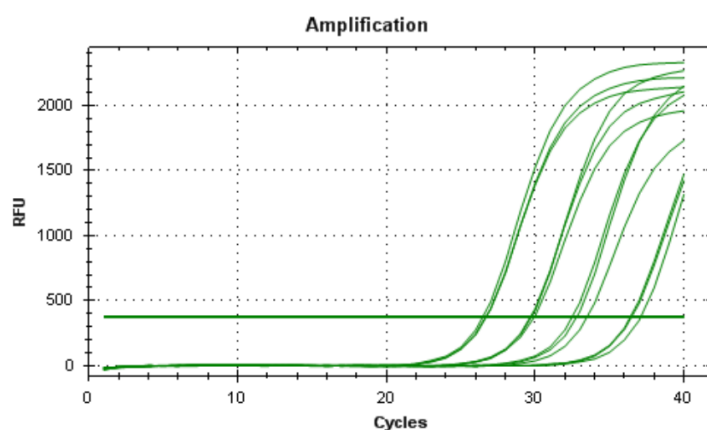
De manera similar, en la Figuras 4.10 y 4.11 se observa el mismo comportamiento para *TOC1*, y un patrón complementario en la expresión de *LHY*, para condiciones circadianas a 20°C en ambas especies.

4.3.5. Expresión diferencial

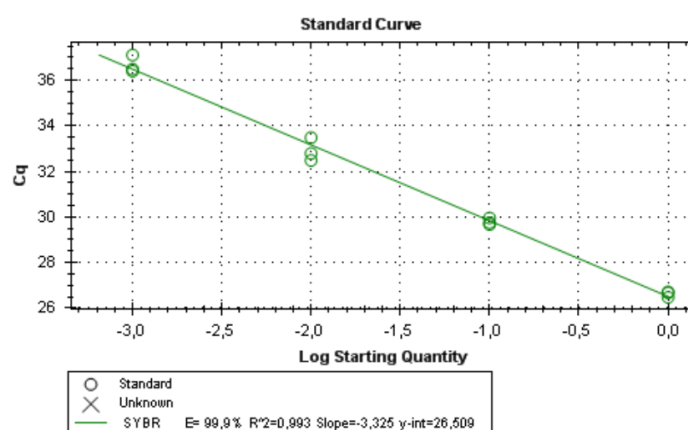
Efecto de la temperatura en la regulación circadiana de la expresión génica global

Con el fin de estudiar los efectos de las altas temperaturas en la regulación de la expresión génica en condiciones circadianas, se analizó la expresión diferencial de genes en el transcriptoma de *N. pumilio* secuenciado a partir de muestreos en dos momentos del día (ZT48 y ZT60) y a dos temperaturas (20°C y 34°C; ver Capítulo 2, Sección 2.2.1 en la página 40). Mientras que a 20°C se observó oscilación de *NpTOC1* y *NpLHY* en fases complementarias (Figura 4.10), a 34°C se pierde dicha coordinación para ambos genes (Figura 4.12), indicando que la temperatura utilizada como tratamiento en nuestro experimento de RNA-seq es una temperatura fuera del rango de compensación del reloj de *N. pumilio*.

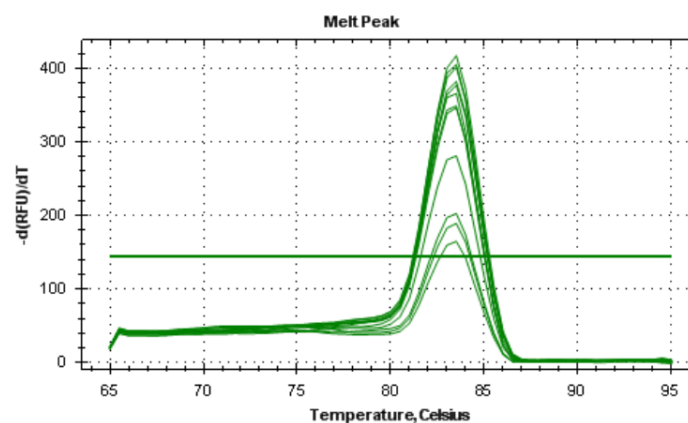
El análisis de expresión diferencial realizado con el paquete DESeq arrojó un total de 2142 contigs regulados en nuestras comparaciones de interés (entre momentos del día a 20°C, entre momentos del día a 34°C, entre temperaturas al amanecer subjetivo, y entre temperaturas al atardecer subjetivo; p-valor < 0.05 para al menos una comparación; ver Sección 4.2.6 en la página 104), de los cuales 535 se presentan en el



(a) Amplificación



(b) Curva de estándares de concentración



(c) Curva de disociación

Figura 4.7: Puesta a punto de la PCR cuantitativa. ADNc de lenga, condiciones diurnas, 20°C, 12 horas, 4 diluciones 10X sucesivas a partir del producto de la retrotranscripción de ARN purificado en columnas de Qiagen y retrotranscripto con enzima M-MLV de Promega. Triplicado técnico. Primer utilizado: GI. Ver texto para mayor explicación

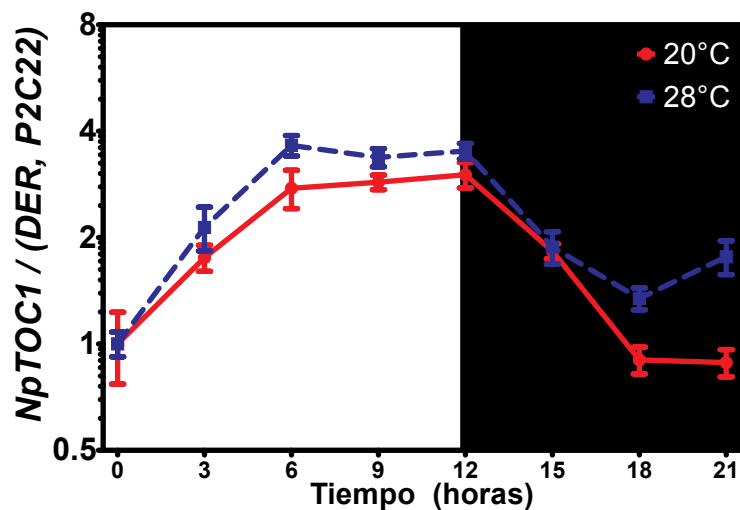


Figura 4.8: Expresión de *NpTOC1* en condiciones diurnas para temperaturas control (20 y 28°C). Cada punto representa la media y la desviación estándar entre 3 réplicas técnicas relativizadas con 2 genes de referencia

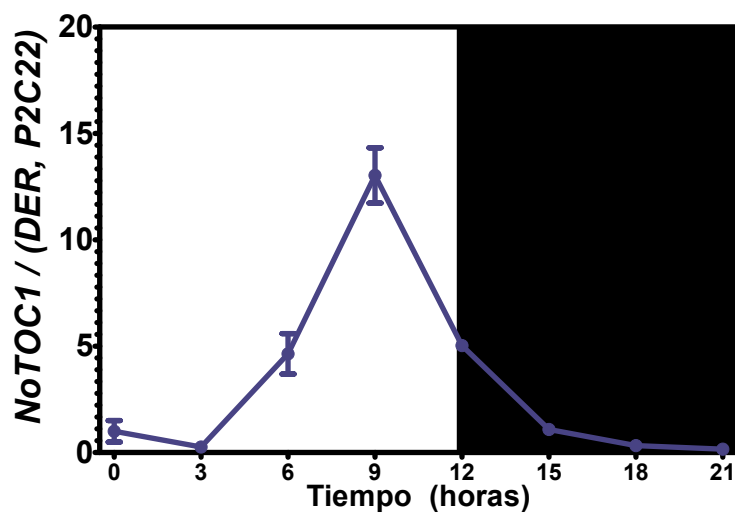


Figura 4.9: Expresión de *NoTOC1* en condiciones diurnas para temperatura control (20°C). Cada punto representa la media y la desviación estándar entre 3 réplicas técnicas relativizadas con 2 genes de referencia

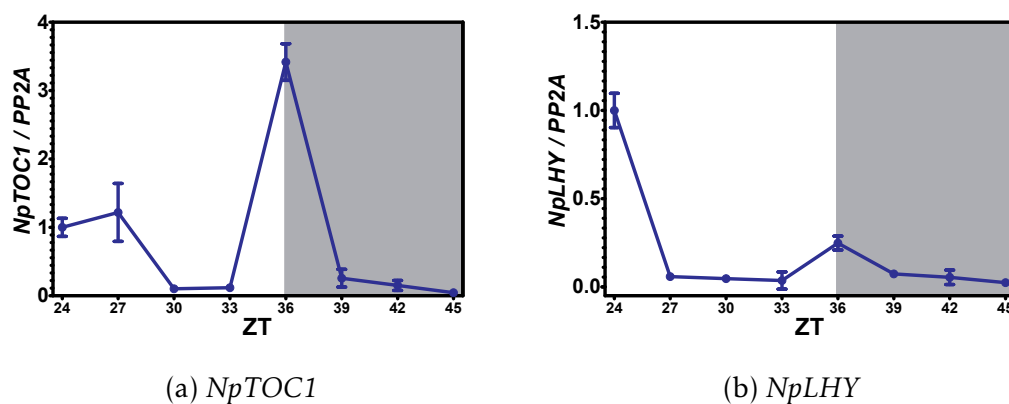


Figura 4.10: Expresión de *NpTOC1* y *NpLHY* en condiciones circadianas para temperatura control (20°C). Cada punto representa la media y la desviación estándar entre 3 réplicas técnicas relativizadas con un gen de referencia

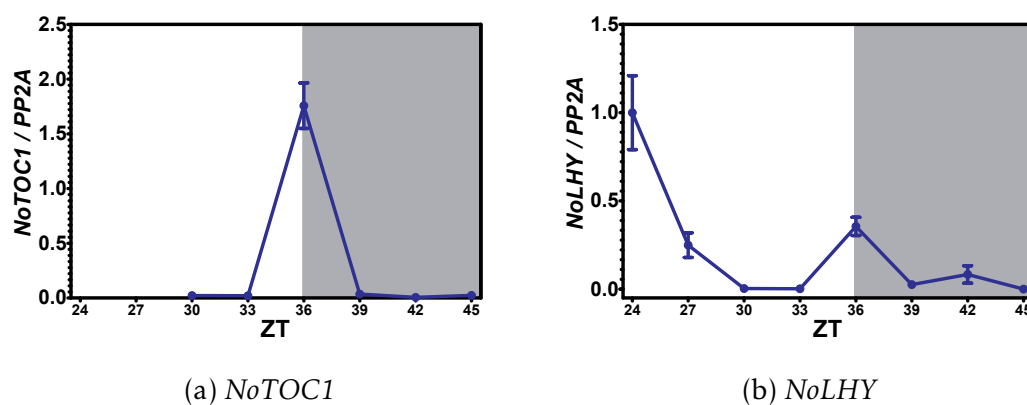


Figura 4.11: Expresión de *NoTOC1* y *NoLHY* en condiciones circadianas para temperatura control (20°C). Cada punto representa la media y la desviación estándar entre 3 réplicas técnicas relativizadas con un gen de referencia

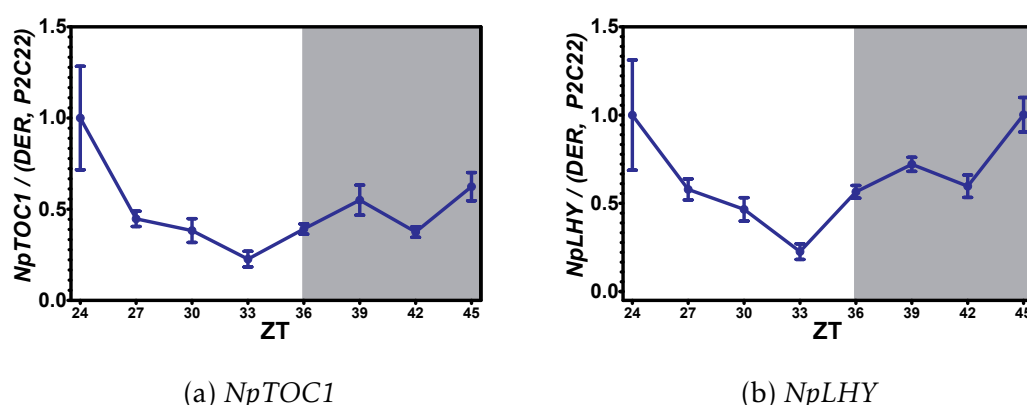


Figura 4.12: Expresión de *NpTOC1* y *NpLHY* en condiciones circadianas a 34°C. Cada punto representa la media y la desviación estándar entre 3 réplicas técnicas relativizadas con dos genes de referencia

heat-map de la Figura 4.13. Esta restricción se debe a que para este Capítulo decidimos concentrarnos en grupos de genes con expresión y cambio entre condiciones moderados, como una primera aproximación a la regulación de la expresión génica global de *N. pumilio*, nunca antes estudiada. Los 535 genes de la Figura 4.13 presentan, entonces, una expresión entre la mediana y el tercer cuartil de la distribución de expresión de los genes regulados (“expresión intermedia”).

El algoritmo de agrupamiento separó dos grandes grupos bien diferenciados, visibles en la Figura 4.13 como productos de la primera bifurcación del árbol de agrupamiento a la izquierda. Un grupo (en la parte superior del gráfico) está compuesto por genes más expresados a 20°C que a 34°C (243 genes), y el otro (en la parte inferior del gráfico) por genes más expresados a 34°C que a 20°C (292 genes). De dichos genes, un 80% presentó anotaciones contra *Arabidopsis thaliana* (192 y 231 respectivamente), y sobre sus identificadores se realizó un análisis de sobrerepresentación de términos de Gene Ontology.

Las Tablas 4.5 y 4.6 muestran algunos términos sobrerepresentados en los genes de los grandes grupos de la Figura 4.13. Dichos términos indican que a 34°C se produce por un lado la interrupción de la fotosíntesis y la producción de energía, y por otro el desencadenamiento del estrés por calor y oxidativo, y la síntesis de metabolitos secundarios. En el Apéndice C, Tablas C.1 a C.2 en las páginas 203-205 se muestran todos los términos de Gene Ontology sobrerepresentados para cada grupo, y sus respectivos

p-valores.

Los genes del reloj circadiano reportados en la Tabla 4.3 no presentaron expresión diferencial significativa en ninguna comparación (con la excepción de dos genes en la comparación entre temperaturas en el atardecer subjetivo: *PRR5* y *LHY*). En la Discusión, Sección 4.4.2 en la página 139 se presentan posibles motivos y explicaciones al respecto. Sin embargo, como se ve en la Figura 4.14, panel 4.14a, los genes diferencialmente expresados entre momentos del día a 20°C no son los mismos que a 34°C, indicando un posible rol del reloj en la regulación de la expresión génica que se ve modificado por las temperaturas altas. Los genes que cambian su expresión significativamente entre 20°C y 34°C (Figura 4.14, panel 4.14b), por un lado, son muchos más que entre momentos del día (puntos coloreados totales), lo cual es indicativo del fuerte reacomodamiento de la expresión global debido al estrés térmico (ver siguiente apartado), y por otro en muchos casos cambian su expresión entre temperaturas independientemente del momento del día (puntos azules). Estos resultados permiten hipotetizar que existen al menos dos vías por las cuales la temperatura influye en la expresión génica global de *N. pumilio*: dependiente e independiente del reloj circadiano (ver más abajo y Discusión, Sección 4.4.4 en la página 144)

Aplicando el criterio de restricción para análisis de expresión diferencial de DESeq (p-valor < 0.05), se observa que no hay grupos importantes de genes que presenten regulación a lo largo del día a 20°C y cambien su expresión a 34°C (Figura 4.13), que serían genes de sumo interés para los objetivos de esta Tesis (ver Capítulo 1, Sección 1.5 en la página 32, y más abajo, Discusión, Sección 4.4.2 en la página 139 para posibles motivos). Por el contrario, el análisis de agrupamiento realizado sobre los genes seleccionados por el criterio *ad hoc* de significancia (cambio entre condiciones mayor a 2, y expresión intermedia; ver Materiales y Métodos, Sección 4.2.6 en la página 104) arroja gran cantidad de grupos, como puede verse en la Figura 4.15.

En total, este criterio arroja 4119 genes regulados, de los cuales nos enfocaremos en 4 grupos (destacados en la Figura 4.15). Los grupos **A** y **B** consisten de genes que presentan regulación a lo largo del día a 20°C, y que ven reprimida su expresión a 34°C; en particular, a 20°C, los genes del grupo **A** se expresan más en el amanecer subjetivo, y los genes del grupo **B** se expresan más en el atardecer subjetivo. Por el contrario, los grupos **C** y **D** consisten de genes que no cambian su expresión durante

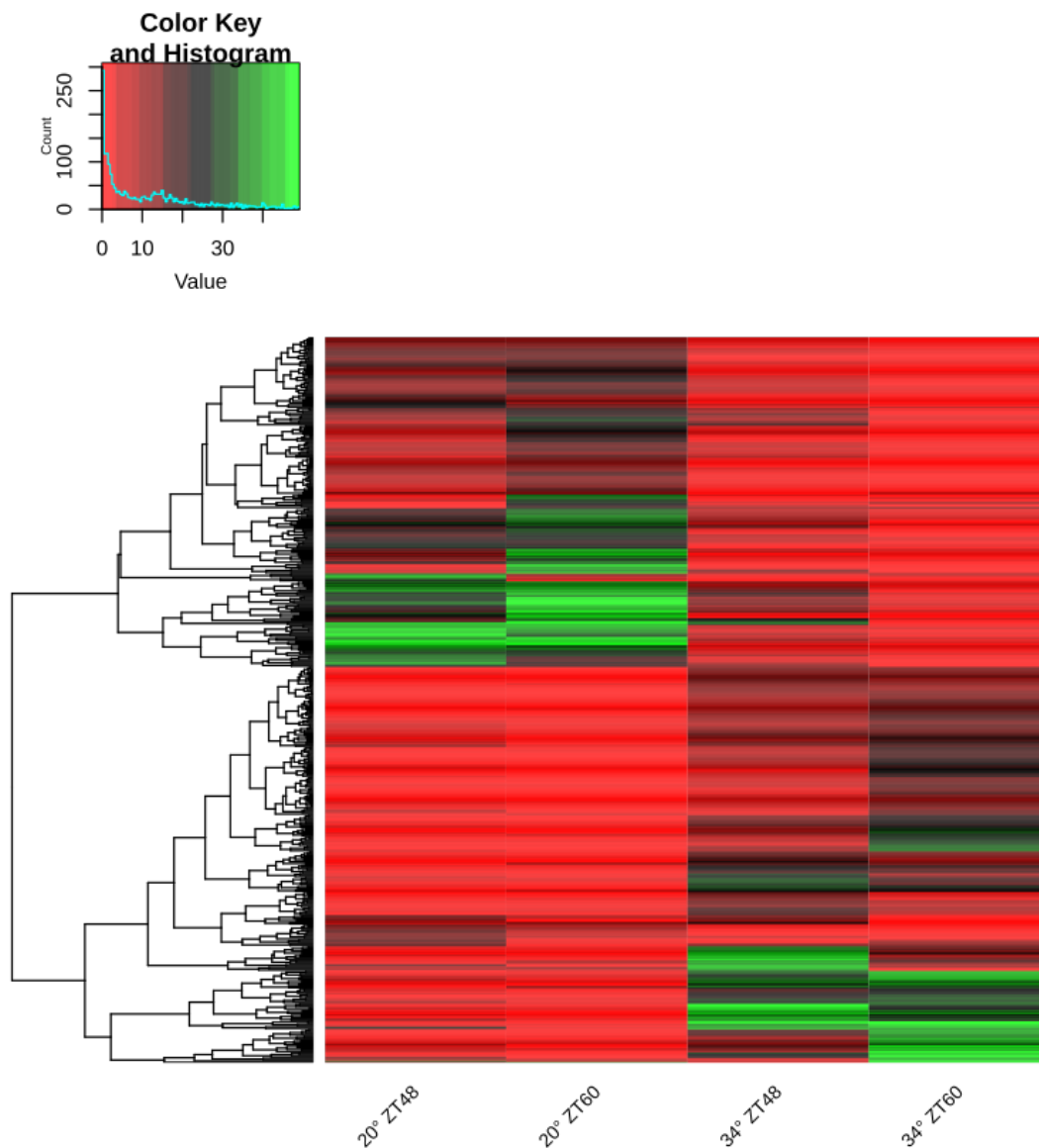


Figura 4.13: *Heat-map* para genes diferencialmente expresados según DESeq en condiciones contrastantes de temperatura y momento del día. Cada columna representa una condición, cada fila un contig, y el color la expresión (clave arriba a la izquierda). Se representan los contigs con un p-valor < 0.05 según DESeq y expresión intermedia (ver texto)

Nombre	Referencia	p-valor
Proceso biológico		
photosynthesis	GO:0015979	2.47×10^{-9}
generation of precursor metabolites and energy	GO:0006091	1.18×10^{-5}
carbon fixation	GO:0015977	2.93×10^{-3}
photosynthesis, light reaction	GO:0019684	1.48×10^{-2}
ATP generation from ADP	GO:0006757	3.72×10^{-2}
Componente celular		
chloroplast envelope	GO:0009941	4.14×10^{-20}
thylakoid	GO:0009579	6.83×10^{-19}
chloroplast stroma	GO:0009570	7.23×10^{-19}
chloroplast	GO:0009507	1.47×10^{-15}
photosynthetic membrane	GO:0034357	3.84×10^{-14}
photosystem	GO:0009521	7.46×10^{-5}
photosystem I	GO:0009522	4.07×10^{-4}

Tabla 4.5: Algunos términos de Gene Ontology sobrerrepresentados en genes más expresados a 20°C que a 34°C (parte superior de la Figura 4.13)

Nombre	Referencia	p-valor
Proceso biológico		
response to stress	GO:0006950	6.30×10^{-6}
response to heat	GO:0009408	3.98×10^{-4}
response to reactive oxygen species	GO:0000302	1.84×10^{-3}
response to oxidative stress	GO:0006979	5.60×10^{-3}
secondary metabolite biosynthetic process	GO:0044550	3.68×10^{-2}
Componente celular		
intracellular organelle part	GO:0044446	9.41×10^{-4}
cytosolic small ribosomal subunit	GO:0022627	1.87×10^{-2}

Tabla 4.6: Algunos términos de Gene Ontology sobrerrepresentados en genes más expresados a 34°C que a 20°C (parte superior de la Figura 4.13)

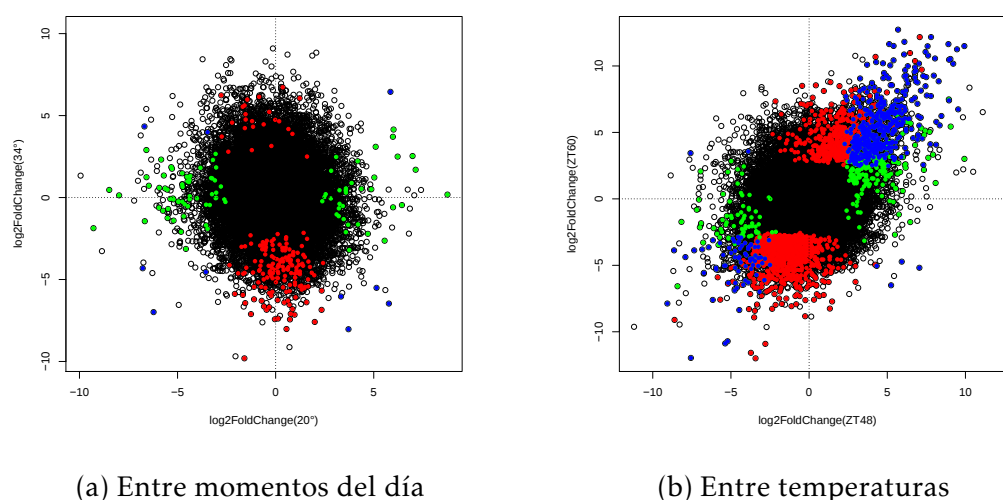


Figura 4.14: Genes regulados ordenados por cambio en expresión. Izquierda: cambio entre momentos del día a 20°C (eje x) vs. cambio entre momentos del día a 34°C (eje y). Derecha: cambio entre temperaturas al amanecer subjetivo (eje x) vs. cambio entre temperaturas al atardecer subjetivo (eje y). Puntos verdes: genes regulados en la comparación del eje x . Puntos rojos: genes regulados en la comparación del eje y .

Puntos azules: genes regulados en ambas comparaciones

el día a 20°C, pero sí entre temperaturas: el grupo **C** contiene genes que se expresan más a 34°C que a 20°C, y el grupo **D** contiene genes que se expresan más a 20°C que a 34°C, en ambos casos independientemente del momento del día. Los grupos **A**, **B**, **C** y **D** tienen 279 (99 anotados), 816 (367 anotados), 327 (161 anotados), y 352 (188 anotados) genes respectivamente. La Figura 4.16 muestra algunos términos de Gene Ontology sobrerrepresentados en dichos grupos de interés. En el Apéndice C, Tablas C.3 a C.6 en las páginas 206-211 se muestran todos los términos sobrerrepresentados para cada grupo y sus respectivos p-valores.

Este análisis evidencia la fuerte regulación circadiana de la fotosíntesis y el metabolismo primario (grupo **B**: término **photosynthesis**, término **chloroplast thylakoid membrane**, término **primary metabolic process**). La regulación de la fotosíntesis también se da independientemente del momento del día, pero en mucha menor medida (grupo **D**). Por el contrario, la respuesta a estrés y un reordenamiento de la expresión génica y el estado iónico de las células se ven disparados a 34°C independientemente del momento del día (grupo **C**: término **response to jasmonic acid**, término **cytosolic**

large ribosomal subunit, término **vacuole**).

En la Figura 4.16 se destacan visualmente los procesos anteriores, y otros. La represión de la fotosíntesis y el metabolismo a 34°C es fuerte independientemente del momento del día (grupo **D**), pero la comparación de los grupos **A** y **B** evidencia que se trata de un proceso regulado circadianamente a 20°C, con un fuerte incremento de procesos metabólicos y fotosintéticos en el atardecer subjetivo. Algunos de los genes del grupo **B** codifican para varias de las importantes enzimas del ciclo de Calvin-Benson: RuBisCO, RuBisCO activasa, sedoheptulosa-1,7-bisfosfatasa, fosforibulokinasa, y fructosa-bisfosfato aldolasa. En el amanecer subjetivo, por el contrario, se expresan genes relacionados con el crecimiento, como el cofactor de plegamiento de tubulina *TFCB*, la expansina *EXPA16*, o los genes de respuesta a auxinas *SAUR36* y *AXR1*. Este balance es indicativo de la regulación circadiana del crecimiento y desarrollo presente a 20°C pero reprimida a 34°C.

Por último, cabe destacar que no se observan en la Figura 4.15 grupos de genes que tengan regulación circadiana a 20°C y que disparen su expresión a 34°C (de la misma manera que hay contados puntos azules en la Figura 4.14 en la página opuesta, panel 4.14a; ver Discusión, Sección 4.4.4 en la página 144). Esto apoya la coherencia entre el criterio *ad hoc* y el criterio estadístico de DESeq, junto con la presencia de los mismos grupos de términos de Gene Ontology regulados (fotosíntesis, estrés).

Efecto del estrés térmico en la expresión génica global

En el apartado anterior se destacó que el análisis de expresión diferencial no distinguió grandes grupos de genes de acuerdo a su variación en expresión a lo largo del día, sino que el efecto de la temperatura fue mucho más fuerte, siendo el factor dominante en la separación en grupos del algoritmo de agrupamiento. Por eso se llevó a cabo el análisis *ad hoc*, que desentrañó el posible rol del reloj circadiano en la expresión génica en condiciones control y su disrupción a temperaturas altas (ver Discusión, Sección 4.4.4 en la página 144).

Dada la evidente preponderancia de la temperatura como factor de reacomodamiento de la expresión en nuestras condiciones experimentales, se decidió llevar a cabo un análisis de expresión diferencial que se independizara del momento del día al cual fueron tomadas las muestras, y considerara únicamente la temperatura a la

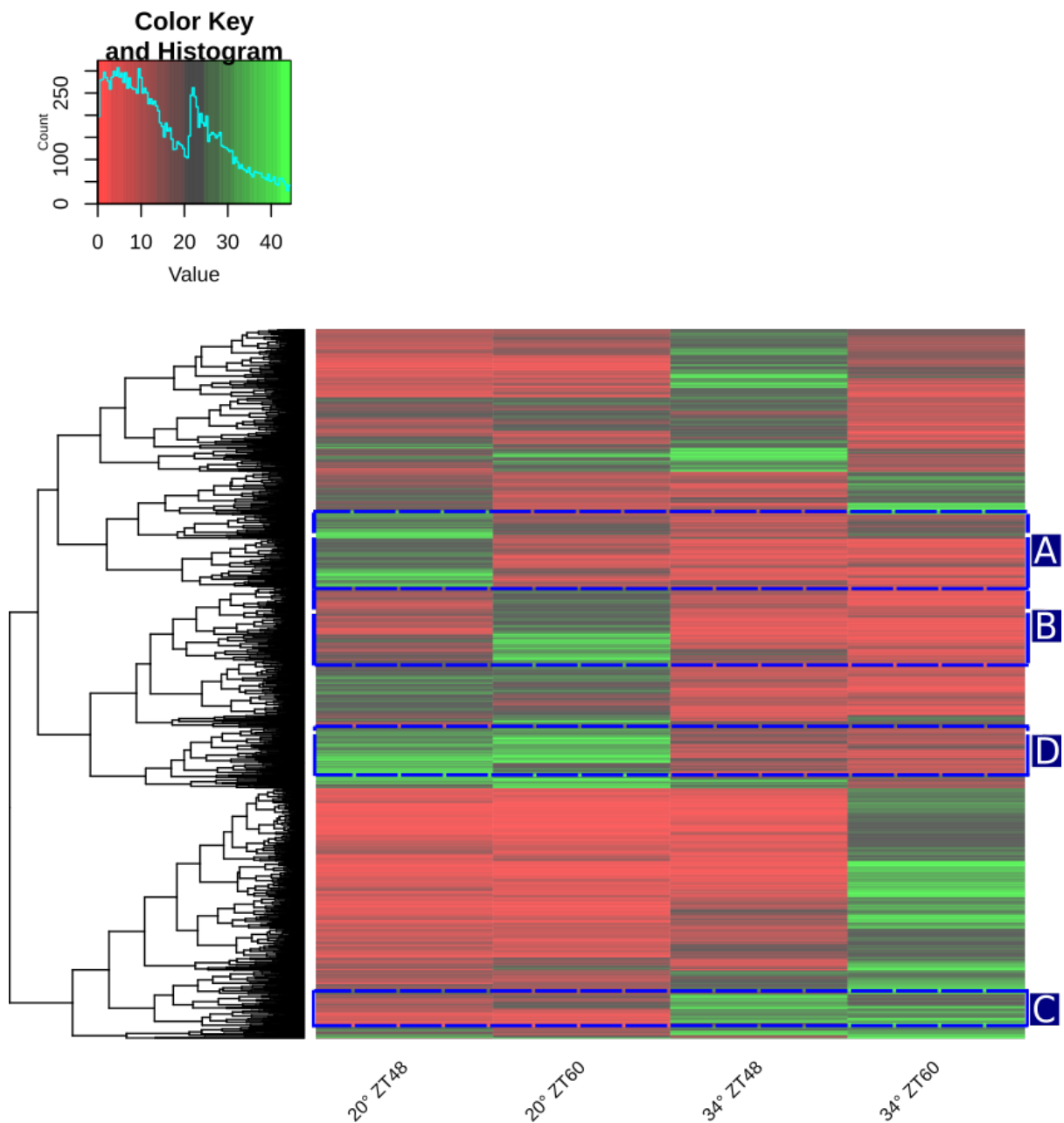


Figura 4.15: *Heat-map* para genes diferencialmente expresados según criterio *ad hoc* de significancia en condiciones contrastantes de temperatura y momento del día. Cada columna representa una condición, cada fila un contig, y el color la expresión (clave arriba a la izquierda). Se representan los contigs con expresión intermedia y que superan el criterio de significancia explicado en el texto

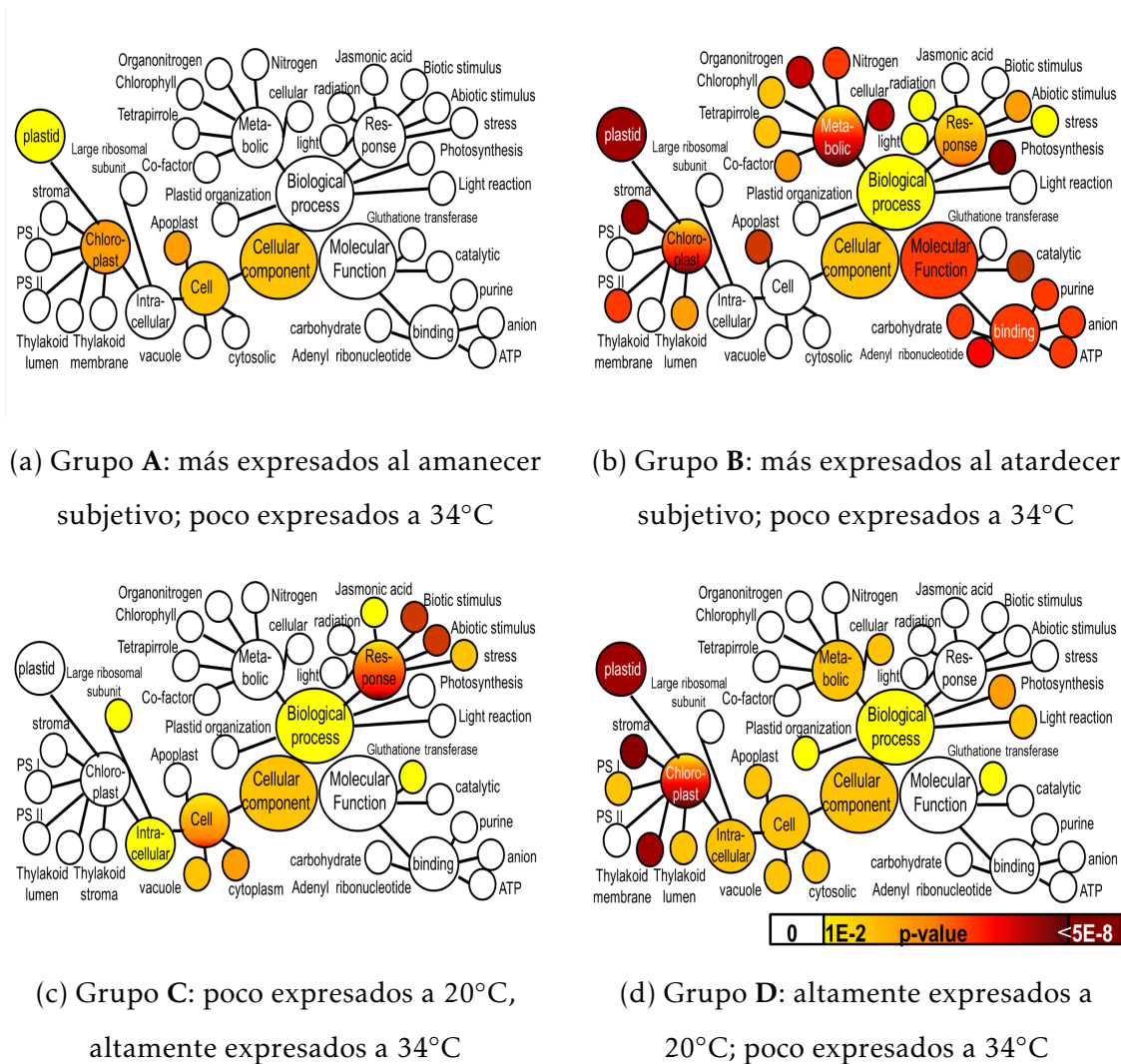


Figura 4.16: Algunos términos de Gene Ontology sobrerrepresentados en genes de los grupos destacados según el criterio *ad hoc* de significancia

cual se realizó el experimento. Entonces, como se explicó en Materiales y Métodos, Sección 4.2.6 en la página 104, se tomaron las 4 *libraries* de 20°C como réplicas entre sí, y las 4 *libraries* de 34°C de igual manera.

Este análisis de expresión diferencial por temperatura con DESeq arrojó 6536 genes regulados, un 8% de los contigs totales ensamblados. De esos, 1634 presentaron expresión intermedia (entre la mediana y el tercer cuartil de la distribución de la expresión) y se presentan en un *heat-map* en el Apéndice B, Figura B.6 en la página 201. El algoritmo de agrupamiento separa, como en el caso de la Figura 4.13, un gran grupo con genes más expresados a 34°C que a 20°C (parte superior del gráfico), y otro gran grupo, menos numeroso, de genes más expresados a 20°C que a 34°C (parte inferior del gráfico). Dentro de estos grupos, en la Figura B.6 se indican los grupos A y B, en los cuales las diferencias de expresión entre temperaturas están exacerbadas. También en el Apéndice C, Tablas C.7 a C.8 en las páginas 213-214 se presentan todos los términos de Gene Ontology, con sus respectivos p-valores, sobrerrepresentados en los genes anotados de dichos grupos.

Al igual que en análisis del apartado anterior, se manifiesta la situación de estrés abiótico y reacomodamiento del metabolismo y el estado iónico de las células a 34°C, y la regulación de la fotosíntesis y el metabolismo primario en general a 20°C.

A diferencia de un *heat-map*, que permite comparar visualmente varias condiciones biológicas a la vez, y presenta a los contigs agrupados por patrones similares de expresión, un gráfico del tipo *MA-plot* como el de la Figura 4.17 compara únicamente dos condiciones, y por eso se presenta en este contexto (20°C contra 34°C). Cada punto del gráfico representa un contig. El eje horizontal indica la media de expresión entre las dos condiciones (es decir, genes más expresados a la derecha y genes menos expresados a la izquierda), y el eje vertical indica el cambio (*fold change*) en expresión entre las condiciones: hacia arriba del 0 genes más expresados a 34°C que a 20°C, y hacia abajo del 0 genes más expresados a 20°C que a 34°C. En la Figura 4.17 se presentan destacados en cuadrados rojos contigs anotados como proteínas de shock térmico (*Heat Shock Proteins*; HSPs) en *A. thaliana* y expresados diferencialmente entre temperaturas (p-valor < 0.05), la gran mayoría más a 34°C que a 20°C (33 de 35, un 94%). Esto confirma con un grupo particular de genes lo observado mediante el enriquecimiento funcional de grupos de contigs expuesto en párrafos anteriores.

Por otro lado, en triángulos amarillos vemos 451 transcritos regulados entre temperaturas, que no presentaron similaridad contra ninguna secuencia de UniProtKB (una base de datos de secuencias proteicas con más de 115 millones de entradas, ver Capítulo 2, Sección 2.3.4 en la página 54), y que no poseen ningún dominio proteico conocido (buscados con la base de datos Pfam [270]). Ver Discusión, Sección 4.4.3 en la página 142, para la potencial importancia de este conjunto de contigs.

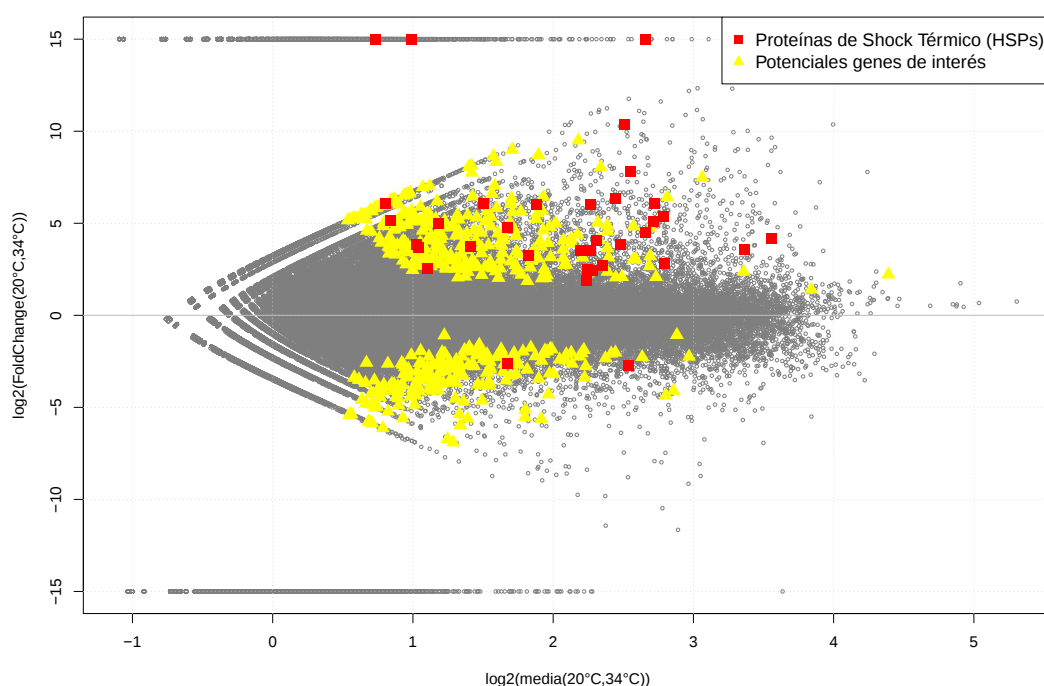


Figura 4.17: MA-plot entre 20°C y 34°C. Los *fold change* muy altos se representan juntos en los extremos del eje vertical para mayor claridad. Se destacan algunos contigs importantes (ver explicación en el texto)

4.3.6. Efecto de la temperatura en el funcionamiento del reloj y su relación con la distribución natural de *Nothofagus* spp.

Expresión de *NpTOC1* y *NoTOC1* en condiciones diurnas a temperaturas altas

En la Figura 4.18 se muestra la expresión de *NpTOC1* y *NoTOC1* a lo largo de 1 día en condiciones diurnas para distintas temperaturas. Debido a que *NpTOC1* presentó oscilación a 28°C (Figura 4.8 en la página 118), no se midió la expresión de *NoTOC1*

a esa temperatura, sino directamente en una temperatura en la que *NpTOC1* ya no oscilara. Para *N. pumilio*, se observa que el patrón de oscilación en condiciones control se pierde a 31°C y a 34°C, mientras que para *N. obliqua* el patrón de expresión de 20°C se mantiene a 31°C, al igual que ocurre para *NpTOC1* entre 20°C y 28°C, perdiéndose la oscilación a 34°C.

Este resultado sugiere, en línea con nuestras hipótesis de trabajo, una preferencia térmica diferencial entre las especies en estudio, en particular debida al rango de funcionamiento de sus relojes circadianos. Por esto a continuación estudiamos caracteres génicos y cuantitativos de *N. pumilio* y *N. obliqua* en condiciones naturales, a lo largo del gradiente altitudinal en estudio (ver Materiales y Métodos, Sección 4.2.7 en la página 107).

Diferencias térmicas en el gradiente altitudinal

Cada parcela de nuestro sistema altitudinal en el cerro Quilanlahue cuenta con datos, tomados en el lugar, de temperatura y humedad del aire y del suelo cada hora. Con algunas interrupciones, dichos datos abarcan desde el año 2010 hasta el presente. Nuestros dos estratos altitudinales de interés son los de 680 y 1100 msnm. Como se mencionó en la Introducción General (Sección 1.2 en la página 19), en el estrato más bajo predomina naturalmente *N. obliqua*, y en el más alto *N. pumilio*, sin zonas de simpatría. En la Figura 4.19 se muestra la temperatura media y el índice de sequía diarios a lo largo de 3 años a 680 y 1100 msnm. El índice de sequía se calcula como $((3 \times T_{max}) + T_{min}) / (1 + HR_m)$, donde T_{max} es la temperatura máxima diaria, T_{min} la temperatura mínima diaria, y HR_m la humedad relativa media diaria. Mientras que no se observa un sesgo hacia un estrato con mayor índice de sequía, en el 97% de los días de dicho período se registró una temperatura media más alta en el estrato de 680 msnm que en el de 1100 msnm, indicando una importante diferencia térmica entre estratos.

Estos Resultados ponen de manifiesto la existencia de nichos térmicos diferentes entre los estratos de 680 y 1100 msnm, dando pie a los estudios de expresión génica y caracteres cuantitativos (a continuación) usando a la montaña como un “laboratorio natural” (ver Sección 1.1 en la página 15).

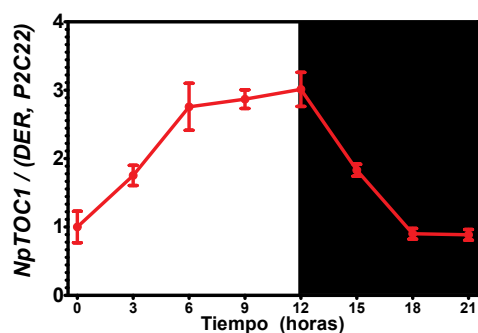
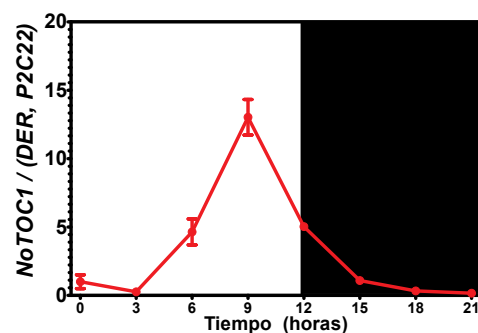
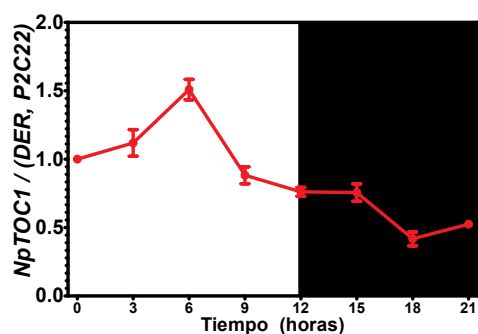
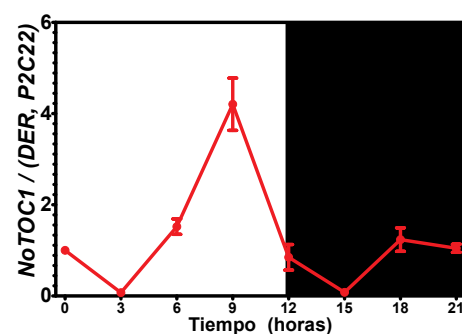
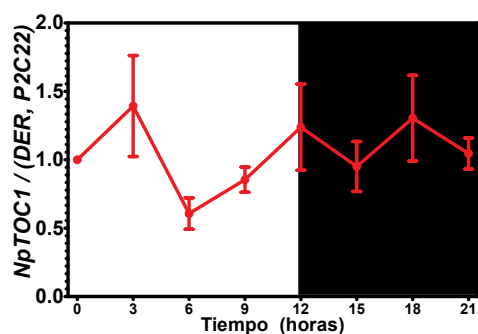
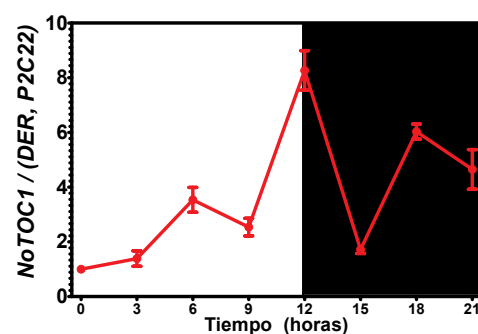
(a) *N. pumilio*, 20°C(b) *N. obliqua*, 20°C(c) *N. pumilio*, 31°C(d) *N. obliqua*, 31°C(e) *N. pumilio*, 34°C(f) *N. obliqua*, 34°C

Figura 4.18: Expresión de *NpTOC1* y *NoTOC1* en condiciones diurnas para temperaturas control y tratamiento. Cada punto representa la media y la desviación estándar entre 3 réplicas técnicas relativizadas con 2 genes de referencia. Cada serie de tiempo está a su vez relativizada a su primer punto

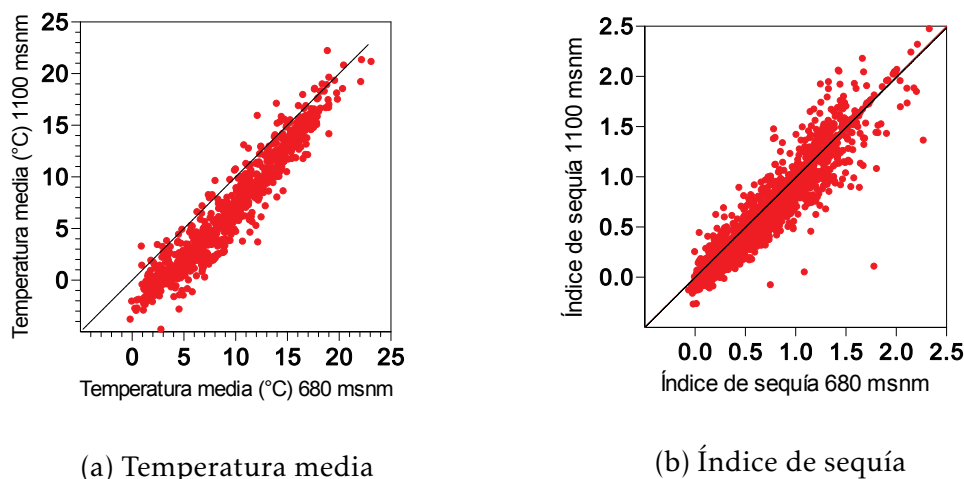


Figura 4.19: Comparación de la temperatura media y el índice de sequía a lo largo de 3 años a 680 y 1100 msnm. Cada punto representa un día de los años 2015 a 2017. Las líneas rectas diagonales están formadas por puntos de igual temperatura o índice de sequía entre estratos (pendiente igual a 1)

Expresión de *NpTOC1* y *NoTOC1* dentro y fuera del área de distribución natural de las especies

La Figura 4.20 muestra las temperaturas máximas diarias a lo largo de un año y medio en ambos estratos altitudinales. En línea con lo mencionado en el apartado anterior, se observa en la Figura que en el estrato de 680 msnm apenas 1 día en 2 inviernos tuvo una temperatura máxima menor a 0°C, y que en el estrato de 1100 msnm son menos los días de verano con temperatura máxima mayor a 30°C. Ambos paneles están centrados en enero de 2013, cuando se realizó el muestreo cada 3 horas durante 2 días para medir expresión génica (ver Materiales y Métodos, Sección 4.2.7 en la página 107). A simple vista se observa que dicho verano fue más caluroso en el estrato inferior (680 msnm) que en el superior (1100 msnm). Este resultado pone en contexto, para las mediciones de expresión génica reportadas a continuación, las diferencias térmicas entre nuestros estratos de interés.

En las Figuras 4.21 y 4.22 se muestra la expresión de *NpTOC1* y *NoTOC1* a lo largo de 2 días de enero de 2013 en condiciones naturales. Para cada gen se presentan curvas de expresión en el estrato altitudinal donde naturalmente crece la lenga (1100 msnm), y un estrato contrastante (680 msnm, donde naturalmente predomina el roble). Se

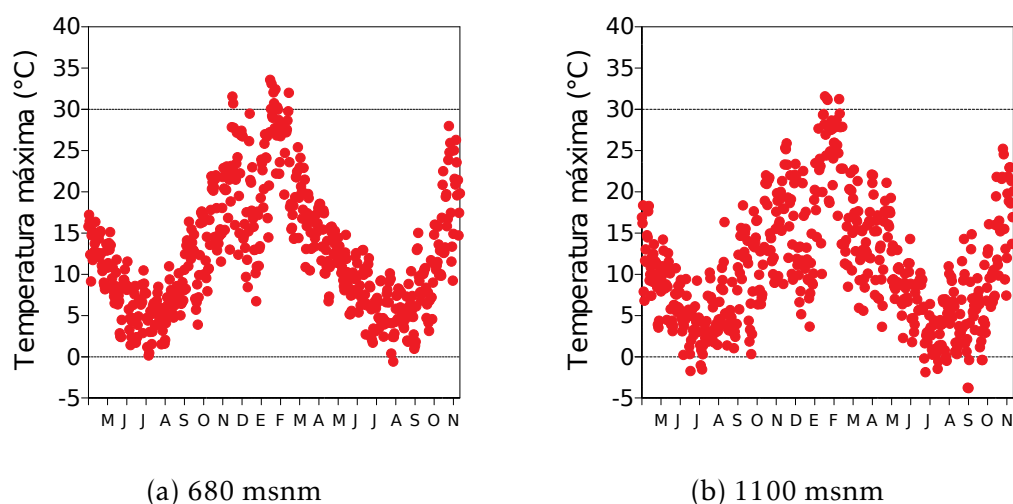


Figura 4.20: Temperatura máxima diaria para 680 y 1100 msnm de mayo de 2012 a noviembre de 2013. Cada punto representa un día

observa para ambas especies una fuerte oscilación y, en el caso de la lenga, un atraso en el pico máximo en el estrato de 680 msnm respecto del de 1100 msnm. Para el roble, el desfase entre estratos no es igual de evidente (ver Discusión, Sección 4.4.4 en la página 144).

Evaluación de caracteres cuantitativos a lo largo del gradiente altitudinal

Estudios en diversas especies indican que el reloj circadiano controla gran cantidad de caracteres cuantitativos, como la fotosíntesis o el crecimiento (ver estudios citados en la Introducción, Sección 4.1 en la página 96). Con el fin de evaluar la asociación entre los cambios en el funcionamiento del reloj entre estratos reportados en el apartado anterior, y el comportamiento de las plantas en su ambiente natural, se evaluaron caracteres fenotípicos reportados como controlados por el reloj circadiano para ambas especies en ambos estratos.

La Figura 4.23 muestra la mortalidad de plántulas de roble y lenga en condiciones naturales, dentro y fuera de sus respectivas áreas de distribución. La lenga presenta significativamente mayor supervivencia dentro de su estrato altitudinal natural, mientras que para el roble la diferencia de mortalidad entre estratos no es significativa. El mismo patrón se observa para la concentración de clorofilas, como se ve en la Figura 4.24: la lenga presenta mayor concentración de dichos pigmentos en su área natural de distribución, y el roble no presenta diferencias significativas. También el peso se-

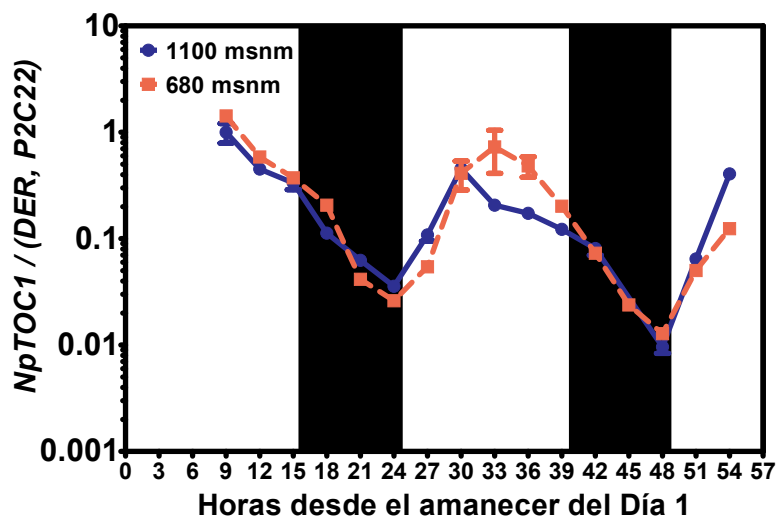


Figura 4.21: Expresión de *NpTOC1* en condiciones naturales. Cada punto representa la media y la desviación estándar entre 3 réplicas técnicas relativizadas con dos genes de referencia

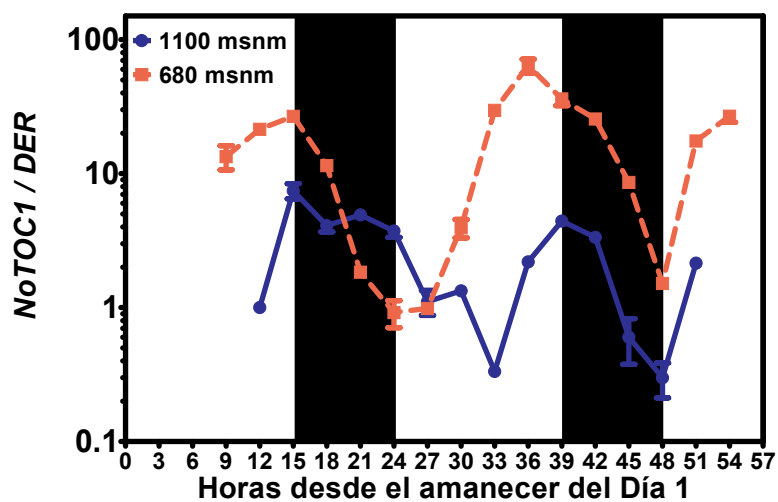


Figura 4.22: Expresión de *NoTOC1* en condiciones naturales. Cada punto representa la media y la desviación estándar entre 3 réplicas técnicas relativizadas con un gen de referencia

co (acumulación de biomasa) es mayor para la lenga dentro de su estrato altitudinal natural (Figura 4.25). Para este caracter no contamos con datos de roble, ya que no sobrevivió un número adecuado de plantas en cada estrato. Para el área foliar (Figura 4.26) el patrón es distinto: el roble presenta significativamente mayor área foliar fuera de su área de distribución natural, mientras que para la lenga la diferencia entre estratos no es significativa (ver Discusión, Sección 4.4.4 en la página 144).

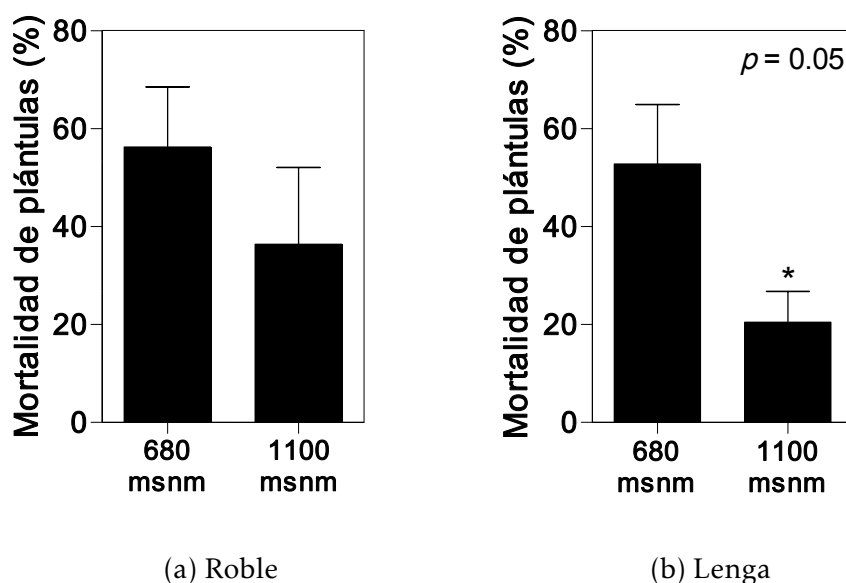


Figura 4.23: Mortalidad de plántulas de roble y lenga en condiciones naturales. n : 24 a 140 por especie y por estrato. Se grafica la media y el error estándar. Test estadístico de t de Student, dos colas

4.4. Discusión

En este Capítulo se describe por primera vez el aislamiento de genes homólogos del reloj circadiano en *Nothofagus*, su expresión en condiciones diurnas y circadianas a distintas temperaturas, y la asociación entre dicha expresión génica y el comportamiento de las plantas en su ambiente natural. Además, se presentan los primeros resultados de expresión diferencial a nivel transcriptómico para especies de este importante género. Estos estudios son los primeros en especies arbóreas que abordan el efecto de las temperaturas altas en el reloj circadiano. Los resultados de este Capítulo proveen las primeras evidencias a favor de que diferencias entre especies respecto

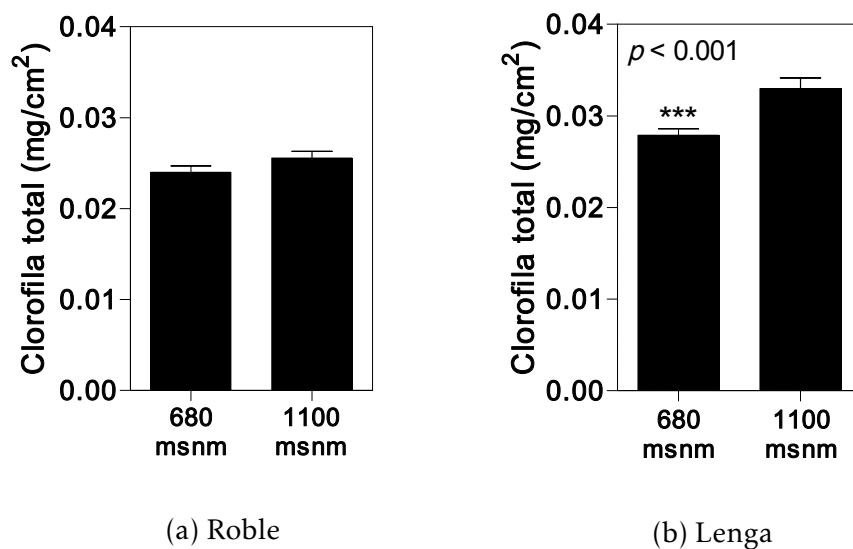


Figura 4.24: Concentración de clorofila de plántulas de roble y lenga en condiciones naturales. n : 30 plántulas por especie por estrato. Se grafica la media y el error estándar. Test estadístico de t de Student, dos colas

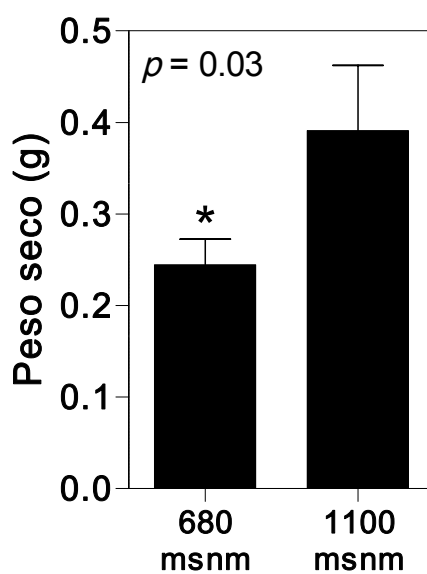


Figura 4.25: Peso seco de plántulas de lenga en condiciones naturales. n : 7-14 plántulas por estrato. Se grafica la media y el error estándar. Test estadístico de t de Student, dos colas

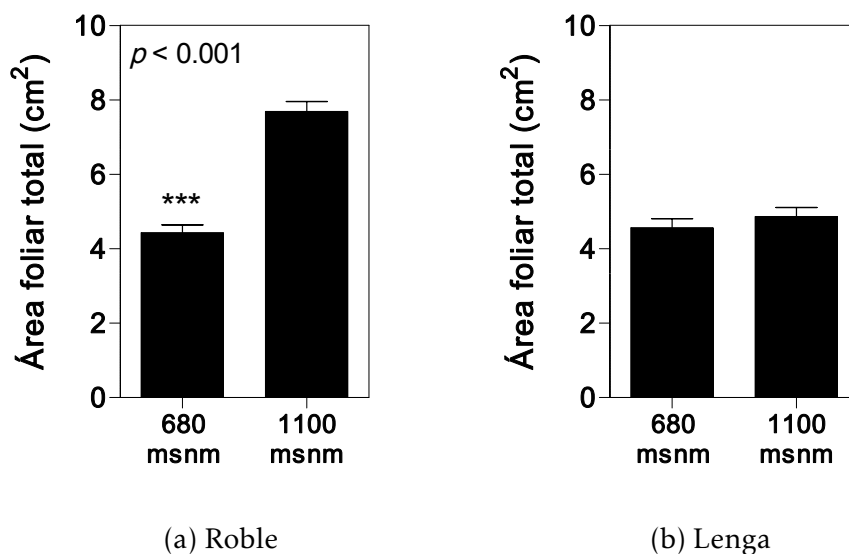


Figura 4.26: Área foliar de plántulas de roble y lenga en condiciones naturales. n : 30 plántulas por especie por estrato. Se grafica la media y el error estándar. Test estadístico de t de Student, dos colas

de la influencia de la temperatura sobre el funcionamiento del reloj circadiano contribuirían a la adaptación fisiológica a ambientes térmicos locales, y por lo tanto a diferencias en rangos de distribución altitudinal.

4.4.1. Puesta a punto de la medición de expresión génica en *Nothofagus* spp.

En tejidos vegetales, la extracción de ácidos nucleicos es dificultosa debido principalmente a la presencia de carbohidratos y de compuestos fenólicos y polifenólicos. Los carbohidratos suelen co-precipitar con los ácidos nucleicos, interfiriendo por ejemplo con su cuantificación espectrofotométrica (la sacarosa, junto con la glucosa y otros monosacáridos, presenta un pico de absorción en UV alrededor de 270 nm, en la misma zona que los ácidos nucleicos), o formando agregados que dificultan las corridas electroforéticas o reacciones enzimáticas posteriores a la extracción [281]. Los compuestos fenólicos, por otra parte, son fácilmente oxidados y forman quinonas que se unen fuertemente a los ácidos nucleicos [282]. Si bien el uso de polímeros vinílicos y la especificidad en la elección de alcoholes, sales, y temperaturas para precipitar ayudan a deshacerse de impurezas y compuestos no deseados, el metabolismo secun-

dario vegetal es muy complejo y cada especie puede tener decenas o cientos de compuestos no descritos o de comportamiento desconocido. No habiendo antecedentes publicados respecto de perfiles generales de metabolitos secundarios para *Nothofagus*, se debió realizar una puesta a punto para la adecuada extracción de ARN y posterior medición de la expresión génica a partir de hojas frescas. No sólo resulta desconocida en la gran mayoría de las especies la composición química de los tejidos a extraer, sino que los productos comerciales están puestos a punto para un número reducido de organismos modelo, y en muchos casos no se dan a conocer sus componentes exactos por razones legales y de patentes, como en el caso de las columnas de purificación utilizadas en esta Tesis, o los dos productos ensayados para PCR cuantitativa (FastStart de Roche y SSO de BioRad), con lo cual es imposible conocer las potenciales interacciones con los tejidos de interés.

Para nuestras especies en estudio, pudimos comprobar (Sección 4.3.3 en la página 112) que no sólo algunos métodos de extracción son incapaces de recuperar ARN (Figura 4.4), sino que la purificación es fundamental (Figura 4.5) y que aún con ARN de buena cantidad y calidad, la PCR cuantitativa puede no ser exitosa, dependiendo de la enzima y *buffers* utilizados (Figura 4.6).

Contar con un experimento de RNA-seq en *Nothofagus* para condiciones de temperatura y momentos del día contrastantes, además de las completas anotaciones génicas y recursos online para *Arabidopsis thaliana* [155, 156], nos permitió encontrar genes adecuados para usar como referencia en la PCR cuantitativa, es decir genes que no varíen su expresión en las condiciones experimentales de interés. Se han reportado casos de genes de referencia altamente utilizados, como la enzima gliceraldehído-3-fosfato deshidrogenasa (GAPDH) o la actina en animales, que no tienen expresión constante y que están involucrados en procesos celulares altamente regulados, con lo que no son adecuados para referencia de PCR cuantitativa [283, 284]. La literatura acerca de normalización y utilización de genes de referencia en qPCR es amplia [285-287], y, además de la preocupación por encontrar genes de referencia específicos de las condiciones de interés, existe consenso dentro de la comunidad respecto de las mejoras obtenidas al utilizar más de un gen de referencia. En nuestro caso, para algunas muestras de roble (las provenientes del gradiente altitudinal, Figura 4.22) no funcionó uno de los genes de referencia (*P2C22*), lo cual podría indicar especificidad

de secuencia por haber sido diseñado a partir del transcriptoma de la lenga, sumado a particularidades químicas de las muestras. Sin embargo, siempre que fue posible se utilizaron dos genes de referencia.

En conclusión, los métodos desarrollados en esta Tesis permitieron establecer un protocolo de extracción de ARN y medición de la expresión génica para hojas de *Nothofagus* (ver caja destacada en la Sección 4.2.4 en la página 101), y en particular un conjunto de dos genes de referencia (*DER2.2* y *P2C22*, ver Tablas 4.1 y 4.2) para estudios de respuesta a temperaturas altas y/o en distintos momentos del día.

4.4.2. Caracterización del funcionamiento del reloj circadiano en *Nothofagus* spp.

Contar con un ensamblado transcriptómico para *N. pumilio* nos permitió un primer abordaje *in silico*, a través del aislamiento y caracterización de genes homólogos del reloj circadiano. En particular, en nuestras anotaciones contra *Arabidopsis thaliana* fueron encontrados 19 genes del reloj circadiano (Tabla 4.3 en la página 111). Dentro de estos están incluidos homólogos de los importantes genes (ver Figura 1.9 en la página 29 de la Introducción General) *TOC1*, *LHY*, *GI*, *ZTL*, un miembro de la familia *RVE*, *ELF3*, *ELF4*, y varios miembros de la familia *PRR*, todos ellos con roles centrales en la generación de ritmicidad por parte del reloj circadiano. En varios de los casos hubo más de un contig cuyo blanco de anotación era el mismo gen de *A. thaliana*, debido a la redundancia propia de un ensamblado *de novo* y de realizar anotaciones contra una especie lejana (ver por ejemplo la Tabla 2.11 en la página 56 como medida de la posible redundancia de los ensamblados estudiados en el Capítulo 2). Sin embargo, siempre un contig fue el que presentó un puntaje de anotación más alto, y en todos los casos este contig presentaba todos o algunos de los dominios proteicos conservados identificados en la referencia de *A. thaliana* (Tabla 4.4 en la página 113). Estos resultados indican la buena representatividad de genes del reloj en nuestro ensamblado, y su completitud funcional mediante la conservación de dominios proteicos, lo cual abre las puertas a futuros estudios de evolución o variabilidad inter- e intraespecífica de genes del reloj circadiano.

Respecto al estudio de salidas rítmicas del reloj en *Nothofagus*, se observó ausencia de ritmicidad en caracteres reportados como *outputs* del reloj en otras especies (Sec-

ción 4.3.1 en la página 109 y estudios citados en la Introducción). En cambio, la oscilación de la expresión del homólogo del gen del oscilador central *TOC1* fue robusta a 20°C para nuestras dos especies, y tanto en condiciones controladas (diurnas y circadianas) como en condiciones de campo (dos estratos altitudinales). De esta manera, la medición de la expresión de dicho gen constituye un indicador robusto para estudiar el funcionamiento del reloj circadiano en *Nothofagus*. Como se puede observar en las Figuras de la Sección 4.3.4 en la página 116, en condiciones diurnas la oscilación de *NpTOC1* y *NoTOC1* es más robusta y de mayor amplitud que en condiciones circadianas.

En varias especies está reportado el debilitamiento de los ritmos circadianos en condiciones constantes (ver Figura 4.27). En especies de *Populus* (álamos), utilizadas como modelo de biología molecular de árboles, no se encontró ritmicidad en varios caracteres utilizados rutinariamente como *outputs* del reloj en *A. thaliana* (M. Eriksson, com. pers.). La literatura acerca de dicho debilitamiento o interrupción de los ritmos no es amplia, pero se hipotetiza que la causa podría ser la ligera y paulatina desincronización entre osciladores celulares independientes ante la ausencia de un ciclo ambiental [288], o la distinta estructura molecular del reloj entre especies [289]. En todo caso, la definición clásica de ritmos circadianos como persistentes en condiciones constantes (por ejemplo Ref. [56]) está siendo cuestionada a medida que el número de caracteres (génicos, fisiológicos u orgánicos) y de especies estudiados aumenta.

Respecto de los estudios de expresión diferencial *in silico*, estos no indicaron fuertes cambios de expresión de genes en condiciones circadianas a una misma temperatura, lo cual apoya nuestra observación de la existencia de debilitamiento de los ritmos circadianos en *Nothofagus*.

Los resultados de expresión diferencial presentados en este Capítulo son los primeros para el género *Nothofagus*, permitiendo estudiar la variación global de la expresión génica ante un estrés abiótico (en particular térmico) y su interacción con la regulación circadiana del crecimiento y desarrollo. Como quedó reflejado en el orden de los Capítulos de esta Tesis, la secuenciación de los transcriptomas se realizó al principio del período de trabajo, antes de la caracterización de la expresión génica, los experimentos de temperatura alta y el análisis de los estudios en condiciones naturales. Así, no fue posible para nosotros prever la dificultad de estudiar la oscilación circa-

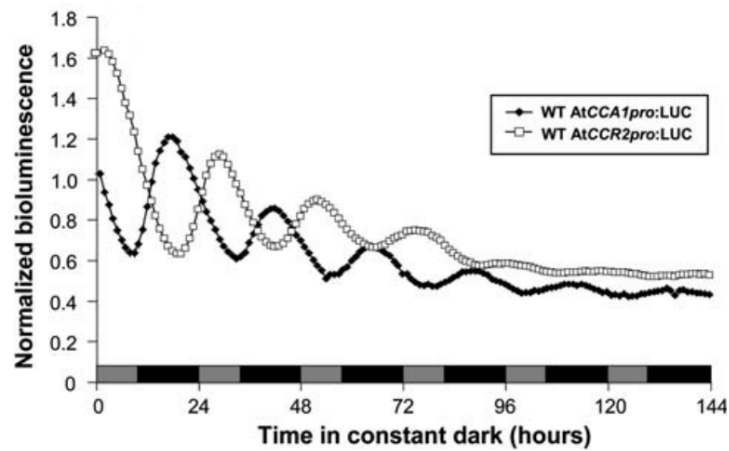
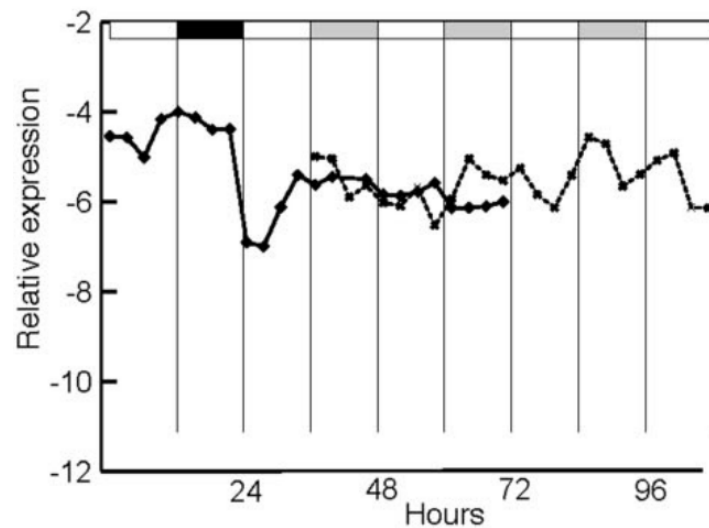
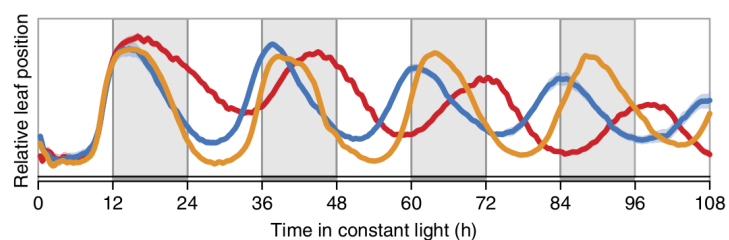
(a) Bioluminiscencia en *Populus tremula* x *P. tremuloides*(b) Expresión génica de *TOC1* en *Picea abies*(c) Posición relativa de hojas en *Solanum lycopersicum*

Figura 4.27: Algunos ejemplos de debilitamiento de ritmos circadianos en condiciones constantes. Fuente: (a) [290]; (b) [289]; (c) [251]

diana *in silico* debido al debilitamiento de la ritmicidad en ZT48 y ZT60 (ver párrafos anteriores). Además, los estudios transcriptómicos que analizan ritmos circadianos generalmente no se basan en la detección de cambios en la expresión de genes por su *fold change* entre máximo y mínimo de expresión, sino por patrones diarios. En el ya citado artículo pionero de Stacey Harmer y colaboradores [69] se desarrolló el método CORRCOS (explicado en detalle en las Referencias [291, 292]), que correlaciona perfiles de expresión a lo largo del tiempo con funciones coseno [293, 294]. Otros métodos no requieren necesariamente el espaciado regular de los datos [295] o usan autorregresión [294, 296], pero en todos los casos se trata de detección de patrones diarios y no análisis directo de cambio de expresión entre dos puntos.

Por otro lado, si bien nuestro diseño experimental contaba con dos réplicas biológicas para cada condición, cuando para este tipo de estudios se suele recomendar un mínimo de tres [137, 138, 297, 298], dado todo lo expuesto anteriormente dudamos de que para nuestros objetivos contar con más réplicas hubiera ayudado.

En resumen, en este Capítulo se dieron los primeros pasos en la caracterización del reloj circadiano (y las limitaciones para su estudio) de especies de *Nothofagus* combinando aproximaciones de campo, *in silico* y de biología molecular. Quedan asimismo asentados los métodos y herramientas para profundizar en aspectos de interés en futuras líneas de investigación.

4.4.3. Efecto del estrés térmico en la expresión génica global de *N. pumilio*

Todos los análisis de expresión diferencial realizados en este Capítulo pusieron de relieve el efecto de la temperatura alta (34°C) en la expresión génica global de *N. pumilio*. En las Figuras 4.13 en la página 122, 4.15 en la página 126 y B.6 en la página 201 se puede ver que la primera división hecha por el algoritmo de agrupamiento (primer quiebre en el árbol a la izquierda del *heat-map*) corresponde a separar los genes con mayor expresión (en promedio) a 20°C que a 34°C, de los genes con mayor expresión a 34°C que a 20°C. Estos fuertes cambios debidos a la temperatura produjeron, junto con las razones expuestas en la Sección anterior, un enmascaramiento de la influencia del reloj circadiano en la expresión génica global. Es por esto que para estudiar la interacción entre la regulación por parte del reloj y por parte de la temperatura fue

necesario establecer un criterio *ad hoc* de expresión diferencial (ver siguiente Sección).

Sin embargo, más allá del efecto del reloj circadiano y su interacción con la temperatura, en este Capítulo se estudió también el efecto propiamente dicho de la temperatura alta en la expresión génica global de *N. pumilio*. Se pudo comprobar, independientemente del momento del día, el desencadenamiento de vías de estrés abiótico (oxidativo, térmico, osmótico) y la represión de la fotosíntesis y el metabolismo primario. En particular, a 34°C se expresan varias proteínas de shock térmico (HSPs) significativamente más que a 20°C (ver Figura 4.17 en la página 129). Estas proteínas, como está bien estudiado, son fundamentales en la respuesta a las temperaturas altas en plantas [113, 299].

En esa misma Figura aparecen también destacados 451 contigs que no presentaron similaridad contra ninguna secuencia de la base de datos UniProtKB, ni dominios proteicos identificados por Pfam. En primer lugar, cabe aclarar que muchos de dichos contigs pueden estar mal ensamblados, ser quiméricos, o representar transcritos no traducidos. Sin embargo, vale la pena profundizar en su estudio debido a que el conocimiento sobre procesos moleculares vinculados a la respuesta a estrés abiótico involucra mayoritariamente información obtenida a partir de especies modelos, y por lo tanto estos transcritos de *N. pumilio* constituyen un interesante recurso para el hallazgo y caracterización de nuevas vías moleculares (ver Discusión General, Capítulo 5). Estos transcritos podrían también codificar proteínas con dominios intrínsecamente desordenados, es decir regiones aminoacídicas no caracterizadas como dominios conservados en bases de datos como Pfam pero con importantes roles en la señalización celular y las redes de regulación tejido-específicas [300]. Cabe destacar que la proteína PER2, un componente central del reloj circadiano en mamíferos, contiene un dominio intrínsecamente desordenado predicho, y que su flexibilidad conformacional tendría un rol protagónico en la compensación por temperatura [301].

En resumen, nuestros resultados indican que *N. pumilio* se encuentra en una situación de estrés térmico a 34°C, reordenando su expresión génica acordeamente y disminuyendo su actividad fotosintética. Los genes (contigs anotados o no) que presentan mayores diferencias entre las condiciones son candidatos ideales para estudiar su regulación y su rol como indicadores de estrés térmico.

4.4.4. Respuestas a la temperatura y establecimiento de nichos ecológicos

El efecto de la temperatura en el reloj circadiano está poco estudiado, y la mayor parte del trabajo fue realizado en especies modelo, como *Arabidopsis thaliana*, y en condiciones de libre curso (ver artículos citados en la Introducción General, Sección 1.3.3 en la página 30). Dentro de este campo, se encuentra mejor caracterizada la respuesta al frío que al calor. En *A. thaliana*, el reloj regula la expresión rítmica de los genes de respuesta al frío *COR* a través de los *PRRs* [302]. A su vez, los factores de transcripción *CBF*, que activan a los genes *COR*, son regulados por el reloj y la luz a través de los fitocromos [303]. Estos estudios, entre otros, evidencian la íntima relación entre la señalización por luz y por temperatura y el reloj circadiano. Sin embargo, el rol del reloj en la respuesta a temperaturas altas no ha sido investigado con la misma profundidad [55]. Sin ir más lejos, en especies arbóreas los trabajos sobre temperatura y reloj se han concentrado en el efecto de las temperaturas frías en rasgos controlados por el reloj y de gran valor adaptativo como la brotación y la dormición invernal [85-87].

Respecto de nuestras especies, una vez caracterizada la oscilación del homólogo de *TOC1* a 20°C, se procedió a estudiar la respuesta a la temperatura de dicho carácter génico, tanto en condiciones controladas (diurnas) como de campo (dos estratos altitudinales). Las condiciones diurnas son más similares a las de los ambientes naturales, por tener períodos alternados de luz y oscuridad, y esto se alinea mejor con nuestras hipótesis y preguntas biológicas, que se interesan en última instancia por la respuesta de las especies en sus ambientes naturales. Los estudios del reloj circadiano en condiciones de libre curso, en cambio, están orientados a desentrañar los mecanismos moleculares endógenos del reloj, para lo cual es importante descartar la influencia de un ambiente cambiante. Además, como se mencionó en la Sección 4.4.2 en la página 139, en condiciones constantes se observa un debilitamiento de la ritmicidad, dificultándose el estudio de la expresión génica circadiana. Por todo esto fue que se decidió realizar los experimentos de expresión a temperaturas altas sólo en condiciones de luz/oscuridad (Sección 4.3.6 en la página 129).

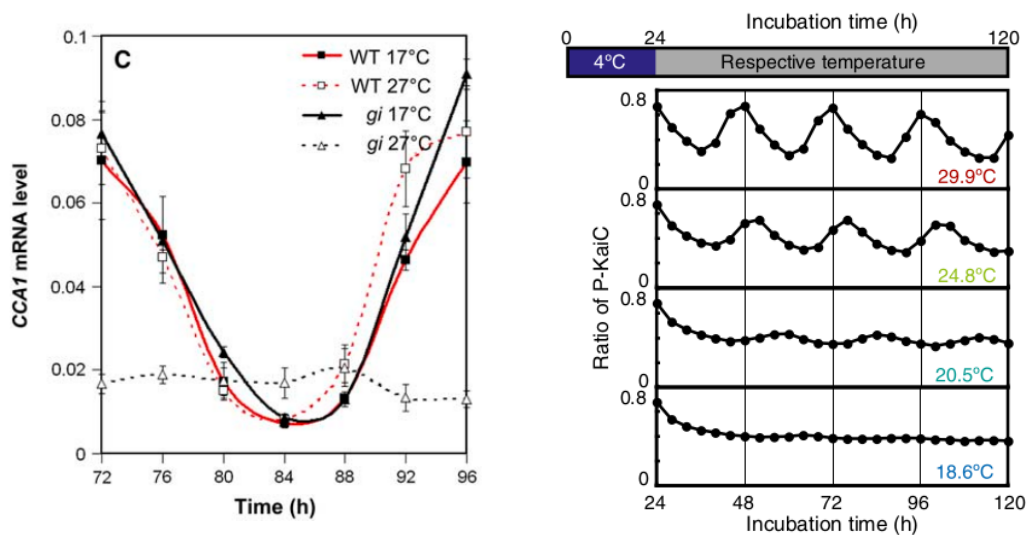
Para *NpTOC1* se observó la pérdida de oscilación a 31°C, y coherentemente también a 34°C (Figura 4.18 en la página 131). En cambio, *NoTOC1* a 31°C presentó un

patrón igual que a 20°C, lo cual nos indica que a esa temperatura no se modifica la oscilación de dicho gen, mientras que esta se interrumpe claramente a 34°C. En estudios de reloj circadiano, la pérdida de oscilación de genes del oscilador central o de salidas rítmicas ante temperaturas altas o bajas se interpreta como incapacidad del reloj de funcionar correctamente (ver ejemplos en la Figura 4.28). De esta manera, se reporta en este Capítulo la disrupción, y consecuente funcionamiento defectuoso, del reloj circadiano por efecto de la temperatura para *N. pumilio* a una temperatura entre 28°C y 31°C, y para *N. obliqua* entre 31°C y 34°C (ver Figura 4.18 en la página 131). Estos resultados constituyen el primer estudio para especies arbóreas del efecto de la temperatura alta (como indicador del Cambio Climático Global) en el funcionamiento del reloj circadiano.

Esta diferencia en el límite superior del rango térmico de funcionamiento del reloj entre especies está asociado con los nichos altitudinales de preferencia natural de las mismas: *N. pumilio* crece a una altitud en la que las temperaturas mayores a 30°C son raras, y donde en general a lo largo del año son menores las temperaturas en comparación con el ambiente preferido por *N. obliqua* (Figuras 4.19 y 4.20).

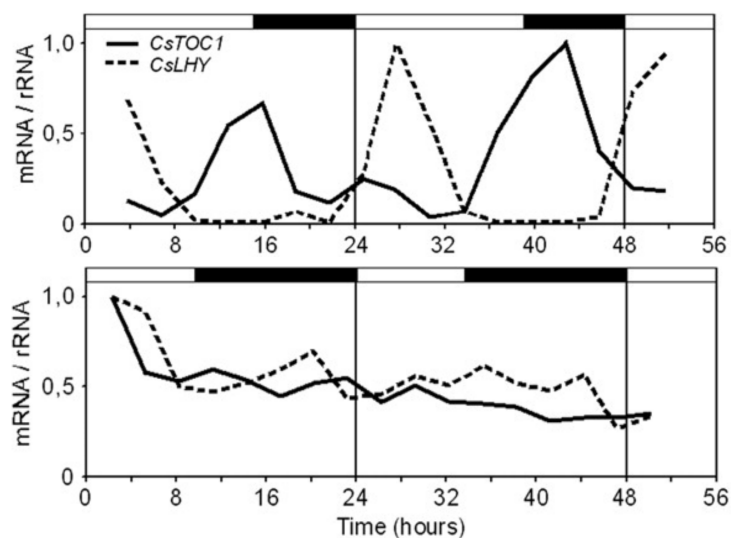
En condiciones naturales (Figuras 4.21 y 4.22 en la página 134) se observó para la lenga un cambio en la fase de oscilación para *NpTOC1* entre estratos altitudinales. En el caso del roble, este desfase no es del todo evidente, ya que en el primer día medido, el pico máximo de expresión coincide entre estratos (a las 15 horas desde el amanecer del día 1), mientras que en el segundo día los picos máximos no coinciden (a las 36 horas desde el amanecer del día 1 para el estrato de 680 msnm, y 3 horas después para el estrato de 1100 msnm; Figura 4.22).

Así como fue distinto el patrón de expresión de *NpTOC1* para la lenga entre estratos, también varios caracteres cuantitativos (descritos como bajo el control del reloj circadiano en *A. thaliana* [54]) presentaron variación según la altitud (Figuras 4.23 a 4.25 en las páginas 135-136). En su área de distribución natural (1100 msnm), la lenga presenta significativamente mayor supervivencia, concentración de clorofilas y peso seco respecto del estrato de 680 msnm. Para el roble, las diferencias entre pisos altitudinales en dichos caracteres no son significativos, lo cual es indicativo de un efecto disruptivo sobre la fisiología de la temperatura alta (para la lenga a 680 msnm) pero no en la misma medida de la temperatura baja (para el roble a 1100 msnm).



(a) Expresión del gen *CCA1* en *Arabidopsis thaliana*

(b) Fosforilación de la proteína KaiC en cianobacterias



(c) Expresión de los genes *TOC1* y *LHY* en *Castanea sativa* en verano (arriba) e invierno (abajo)

Figura 4.28: Algunos ejemplos de disrupción del reloj circadiano por temperatura.

Fuente: (a) [80]; (b) [304]; (c) [85]

Para el área foliar, el comportamiento de las especies entre estratos no fue el esperado (Figura 4.26 en la página 137). Sin embargo, el área foliar es un carácter extremadamente plástico en las plantas [305, 306], con lo cual lo observado para nuestras especies en el gradiente altitudinal puede ser un efecto de micrositio (irradiancia, sombreado, orientación, etc.) y no un indicador del funcionamiento del reloj o el *fitness* de las plantas en respuesta a la temperatura. Para *N. pumilio*, puntualmente, está reportada la variación en área y otros caracteres foliares en respuesta a irradiancia y humedad del suelo [307].

Estos resultados asocian por primera vez cambios en la expresión de genes homólogos del reloj circadiano (lo cual estaría vinculado a su funcionamiento) con el *fitness* de especies vegetales en condiciones naturales. Mediante su rol en la regulación de gran cantidad de procesos fisiológicos y de desarrollo (ver Introducción, Sección 4.1 en la página 96), el reloj podría tener un rol protagónico en explicar las diferencias observadas entre estratos altitudinales. Por ejemplo, la fotosíntesis está regulada por el reloj, y es una de sus salidas rítmicas mejor caracterizadas [262, 263]. Recientemente se ha demostrado en varias especies vegetales que la concentración de clorofilas *a* y *b* constituye un buen indicador de la oscilación diaria de la fotosíntesis [252], lo cual facilita considerablemente el estudio circadiano de este importante proceso. En *A. thaliana* se demostró que los individuos cuyos relojes no estaban sincronizados con el ambiente producían menos clorofila y consecuentemente fijaban menos dióxido de carbono [54]. Así, la diferencia en el comportamiento oscilatorio de *NpTOC1* podría generar una desincronización con el ambiente que lleve a que la lenga tenga menos clorofila en el estrato donde no crece naturalmente (680 msnm; Figura 4.24 en la página 136).

Por otro lado, está bien estudiado en plantas que el crecimiento se encuentra regulado por el reloj, a través de la acumulación y degradación cíclica de almidón a lo largo del día y la noche respectivamente, en última instancia guiados por la fotosíntesis [67]. Un reloj defectuoso, como ya se dijo, genera en plantas menor crecimiento en términos de biomasa [54], lo cual podría explicar la diferencia observada para el peso seco de *N. pumilio* entre estratos (Figura 4.25 en la página 136) en los cuales el patrón de oscilación de *NpTOC1* difiere.

Por último, una integración de todos los procesos y vías metabólicas potencialmente desregulados ante un comportamiento anómalo del reloj podría ser la causa de la

diferencia en supervivencia observada también entre estratos para *N. pumilio* (Figura 4.23 en la página 135).

En el caso de *N. obliqua*, el comportamiento entre estratos altitudinales no presenta las mismas diferencias que para *N. pumilio*. Esto podría indicar que a los fines del *fitness* de los individuos, el efecto de la temperatura alta sobre la lenga es más severo que el efecto de la temperatura baja sobre el roble. En caso de ser cierta esta hipótesis, el hecho de que *N. obliqua* no crezca naturalmente en el estrato donde domina la lenga no sería debido a una pérdida de función de su reloj circadiano. Entre otras causas en las cuales estamos y seguiremos profundizando, se encuentran la posible desventaja competitiva frente a *N. pumilio* en estratos altitudinales altos, o cambios en la fenología, debidos a la temperatura, que repercutan en el establecimiento y desarrollo de las plántulas. En un estudio clásico, para especies arbóreas europeas se encontró que un aumento de 1°C en la primavera temprana adelanta la estación de crecimiento 7 días [308]. Recientemente, para especies de *Nothofagus* se reportó que los patrones temporales de germinación se encuentran fuertemente influidos por la temperatura del suelo, ocurriendo este importante proceso en el momento óptimo del año para cada especie en el estrato altitudinal donde predomina naturalmente [278]. Esto probablemente contribuye a los patrones existentes de distribución de especies a lo largo de un gradiente altitudinal.

Respecto de la expresión génica global en *N. pumilio* (Sección 4.3.5 en la página 116), se puede concluir que a 34°C la lenga se encuentra en una situación de fuerte estrés abiótico (ver Sección anterior), y reacomoda su expresión génica acordeamente. Si bien con el análisis estadístico aplicado no se pudo distinguir un efecto de la temperatura en el funcionamiento del reloj (ver Sección 4.4.2), la Figura 4.14 en la página 124 sugiere que existe una regulación entre momentos del día a 20°C que se pierde a 34°C. Por otro lado, gracias al criterio de significancia *ad hoc* adoptado en esta Tesis, se pudieron distinguir grupos de genes potencialmente relacionados con la regulación de la expresión génica mediada por el reloj circadiano (Figura 4.15 en la página 126, grupos A y B). Este análisis nos permite hipotetizar que la temperatura alta tiene principalmente dos efectos, en principio independientes, sobre la expresión génica en *N. pumilio*: por un lado, reprime la expresión de vías de salida rítmicas relacionadas con crecimiento y metabolismo, y por otro lado promueve la expresión de

genes de estrés y de señalización por jasmonatos de manera independiente del reloj. En particular se comprobó la represión de varios genes codificantes de enzimas del ciclo de Calvin-Benson, y la inducción de la expresión de numerosas proteínas de shock térmico. La fotosíntesis aparece como un proceso reprimido en ambos casos, pero más fuertemente en el grupo regulado circadianamente. Cabe aclarar que si bien la fotosíntesis está reprimida, según nuestros resultados, tanto de manera dependiente como independiente del reloj, esta regulación simultánea se puede deber al involucramiento de genes distintos dentro de este complejo proceso metabólico. Una tercera evidencia a favor de una regulación de la expresión global mediada por el reloj es el hecho de que en la Figura 4.15 no se observan grupos que cambien su expresión entre momentos del día a 20°C y que disparen su expresión a 34°C, apoyando la independencia de las vías de regulación por el reloj y las vías de estrés por calor, además de la disrupción del reloj circadiano a altas temperaturas.

En resumen, se comprobó en condiciones controladas de laboratorio la disrupción del reloj circadiano de *N. obliqua* y *N. pumilio* a temperaturas altas, con una diferencia en el límite superior del rango de compensación entre las especies correlacionado con la temperatura de los ambientes naturales en los que ambas especies viven (el roble en ambientes más cálidos que la lenga). En condiciones de campo, se observó un comportamiento diferencial entre estratos de la oscilación de *NpTOC1* para la lenga, asociada con un mayor *fitness* en su área de distribución natural. Para el roble no se encontraron diferencias significativas de caracteres cuantitativos entre estratos, además de ser menos claro el desfase en la expresión de *NoTOC1*. Además, a nivel de expresión global la lenga se encuentra en un fuerte estrés térmico a 34°C, reprimiéndose la fotosíntesis y vías de salida rítmicas del reloj relacionadas con el crecimiento y el metabolismo. Estos resultados sugieren un rol del reloj circadiano, en tanto regulador maestro del crecimiento y desarrollo vegetal, en el establecimiento de nichos térmicos en condiciones naturales, actuando en particular como limitante la temperatura alta. Considerando que durante el siglo XX la temperatura de la región patagónica aumentó, y la tendencia continuará a lo largo del siglo XXI (ver Introducción General, Sección 1.2 en la página 19), los estudios realizados en este Capítulo alertan sobre el efecto que dicha tendencia podría tener sobre la fisiología y adaptabilidad de las especies arbóreas nativas.

Capítulo 5

Conclusiones generales y perspectivas futuras

*Creció en mi frente un árbol.
Creció hacia dentro.
Sus raíces son venas,
nervios sus ramas,
sus confusos follajes pensamientos.
Tus miradas lo encienden
y sus frutos de sombras
son naranjas de sangre
son granadas de lumbre.

Amanece
en la noche del cuerpo.
Allá adentro, en mi frente,
el árbol habla.

Acércate, ¿lo oyes?*

Octavio Paz, *Árbol adentro*

Esta Tesis presenta los datos genómicos más completos a la fecha para *Nothofagus*, el primer estudio filogenómico del orden Fagales, y los primeros estudios sobre la influencia de la temperatura en el funcionamiento del reloj circadiano para árboles nativos del hemisferio Sur, en condiciones controladas y de campo. Todos estos resultados aportan al abordaje interdisciplinar de las respuestas de las especies nativas al Cambio Climático Global.

Abordar este tipo de estudios en especies no modelo requirió que una parte im-

portante del trabajo fuera la puesta a punto de técnicas y métodos (de laboratorio y computacionales). Por ejemplo, no se cuenta con secuencias genómicas de referencia para *Nothofagus*, con lo cual los ensamblados transcriptómicos tuvieron que ser *de novo* y con estrictos controles de calidad junto con el desarrollo de nuevos métodos bioinformáticos (Capítulo 2), los ensamblados de especies no modelo presentan con mayor frecuencia contigs quiméricos que dificultan los alineamientos entre supuestas proteínas homólogas, debiendo ser necesario desarrollar un cuidadoso método de obtención de matrices filogenómicas (Capítulo 3), y la expresión génica en especies de *Nothofagus* nunca había sido medida, lo cual implicó un laborioso proceso de puesta a punto de laboratorio de biología molecular, desde la extracción de ARN hasta la cuantificación por qPCR (Capítulo 4).

Todas estas técnicas y métodos desarrollados y puestos a punto se encuentran a disposición y permiten gran diversidad de estudios que van mucho más allá de los temas de incumbencia estricta de esta Tesis, por ejemplo: caracterización *in silico* de familias génicas y vías metabólicas, evaluación *in vivo* de polimorfismo de microsatélites predichos computacionalmente, búsqueda y alineamiento de genes conservados eucariotas en ensamblados transcriptómicos para estudios filogenómicos, o cuantificación por qPCR de cualquier gen transcripto en especies de *Nothofagus*. Con respecto a esto último, se presentan también en esta Tesis dos genes de referencia estrictamente optimizados para estudios de variación diurna y/o del efecto de temperaturas altas en especies de *Nothofagus*: *P2C22* y *DER2.2*.

Así, esta Tesis inaugura varias líneas de investigación con promisorio futuro, plausibles de ser profundizadas, relacionadas y ramificadas para continuar con la integración de disciplinas y niveles de organización en vistas de entender, en última instancia, la compleja respuesta de las especies nativas al cambio climático en ambientes naturales.

Un paso importante en pos de dicha integración está siendo llevado a cabo mediante la secuenciación del genoma de la lenga (*Nothofagus pumilio*). Esta secuenciación se inscribe en una reciente colaboración con el grupo del Dr. Rodrigo Gutiérrez, Investigador Principal de la Pontificia Universidad Católica de Chile (PUC) y Director Alternativo del Instituto Milenio de Biología Integrativa de Sistemas y Sintética (MIISB), en Santiago (Chile). El análisis de dicha secuenciación, que se está llevando a cabo a

lo largo del año 2018 en el MIISSB, forma parte del trabajo posdoctoral planeado del Tesista.

La información contenida en el genoma de una especie tiene innumerables aplicaciones. Por un lado, en un aspecto productivo, la secuencia del genoma puede acelerar el mejoramiento y facilitar enormemente el estudio de rasgos cuantitativos complejos [309, 310]. Como ya se dijo en la Introducción General (Sección 1.2 en la página 19), la lenga es una importante especie maderera, y está priorizada por el INTA en su programa de mejoramiento genético y domesticación de árboles forestales nativos (<https://inta.gob.ar/proyectos/PNFOR-1104063>). También es posible, tanto con fines productivos como ecológicos, desarrollar marcadores genómicos para estudios poblacionales o de variedades [310, 311]. Finalmente, la información genómica sirve para la identificación y caracterización de nuevos genes y vías metabólicas, y, en suma, para entender los mecanismos genéticos que regulan las respuestas de los organismos a sus ambientes cambiantes [310, 312]. En particular, el entendimiento del efecto de factores que cambiarán según las predicciones del CCG, como ser la temperatura, sobre el crecimiento y desarrollo requiere de un análisis extensivo de los caracteres fisiológicos relacionados a la respuesta del organismo a dicho factor. Estos datos suelen ser, en la mayoría de los casos, escasos y aislados [313]. Los abordajes genómicos constituyen una alternativa para comenzar a estudiar de forma global los efectos de cambios en el clima sobre los organismos, permitiendo el estudio de procesos múltiples, así como la caracterización de zonas genómicas asociadas a dichas respuestas. Esto es de gran relevancia, por ejemplo, para planes de restauración asistida, ya que este tipo de abordajes en articulación con otras disciplinas (como la ecofisiología o la genética cuantitativa) permitiría la identificación de material de acervo genético adecuado, con buenas chances de establecimiento en condiciones presentes y futuras, como las predichas por el CCG [314-316]. La lenga es una de las especies que forma parte del plan de reforestación ForestAR 2030, del Ministerio de Ambiente y Desarrollo Sustentable (actual Secretaría), que se propone restaurar ecosistemas boscosos degradados.

En combinación con los datos transcriptómicos generados en esta Tesis, se identificarán en el genoma regiones reguladoras de la expresión asociadas a procesos involucrados en la respuesta a estrés térmico con potencial aplicación diagnóstica y productiva en programas de restauración forestal y domesticación. Los contigs que pre-

sentan mayores diferencias entre las condiciones son candidatos ideales para estudiar su regulación y su rol como indicadores de estrés térmico. Entre dichos contigs, los que no presentan anotaciones contra secuencias proteicas conocidas son de gran interés, ya que pueden dar pie al hallazgo y caracterización de nuevas vías moleculares de respuesta a estrés en plantas, no descritas hasta el momento. En última instancia, se propone llegar a un paquete tecnológico que consista de una serie de “genes diagnóstico” para evaluar la susceptibilidad al estrés térmico en rodales naturales o plantaciones de especies forestales.

Otra línea futura que profundiza lo iniciado en esta Tesis consiste en aumentar el número de familias y especies del orden Fagales considerados en el análisis filogenómico. La heterogeneidad en la procedencia de las secuencias (más allá de la base de datos utilizada en el Capítulo 3) requerirá una cuidadosa estandarización de la metodología de ensamblado y construcción de la matriz de caracteres. La secuenciación y ensamblado del genoma de la lenga permitirá incorporar regiones genómicas no transcritas al alineamiento, dado que el orden Fagales cuenta ya con varios genomas secuenciados, algunos muy recientemente, en las familias Juglandaceae [309, 317], Fagaceae [312, 318-320], Myricaceae [321], Betulaceae [311, 322, 323] y Casuarinaceae [323, 324]. Un estudio filogenómico de esta magnitud permitiría comprender más profundamente la evolución de los caracteres distintivos (entre ellos muchos de importancia económica, como la producción de madera y de frutos y semillas comestibles) de este importante grupo de árboles.

Por último, la cuantificación de la expresión de genes, puesta a punto en esta Tesis, no será utilizada solamente como un indicador de estrés térmico (ver párrafos anteriores), sino para profundizar en el estudio del efecto de la temperatura en el reloj circadiano, y en la regulación por parte de este de procesos celulares y fisiológicos, tema al que también esta Tesis dio un puntapié inicial.

En suma, estas perspectivas evidencian la diversidad de temáticas abordadas y en gran parte iniciadas en esta Tesis, y cómo a lo largo de todo el proceso no se dejó de enfatizar en la importancia de integrar conocimiento y marcos conceptuales de distintas disciplinas como acercamiento adecuado al estudio de la adaptación y las respuestas de las especies nativas a ambientes cambiantes y ecológicamente relevantes, en un contexto de Cambio Climático Global.

Bibliografía

- [1] Sabine, C. L., Heimann, M., Artaxo, P., Baker, D. C. E., Chen, C.-T. A. y col. «Current status and past trends of the global carbon cycle». En: *The global carbon cycle: integrating humans, climate and the natural world*. Ed. por Field, C. B. y Raupach, M. R. Washington, D.C.: Island Press, 2004.
- [2] Belcher, B., Ruíz-Pérez, M. y Achdiawan, R. «Global patterns and trends in the use and management of commercial NTFPs: implications for livelihoods and conservation». En: *World development* 33 (2005), pág. 1435.
- [3] Millennium Ecosystem Assessment. *Ecosystems and human well-being: synthesis*. Washington, D.C.: Island Press, 2005.
- [4] Bonan, G. B. «Forests and Climate Change: forcings, feedbacks, and the climate benefits of forests». En: *Science* 320 (2008), pág. 1444.
- [5] Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A. A. y col. «High-resolution global maps of 21st-century forest cover change». En: *Science* 342 (2013), pág. 850.
- [6] Hartmann, D. L., Tank, A. M. K., Rusticucci, M., Alexander, L. V., Brönnimann, S. y col. «Observations: atmosphere and surface». En: *Climate Change 2013: the physical science basis. Working Group I contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, 2013.
- [7] Sitch, S., Huntingford, C., Gedney, N., Levy, P. E., Lomas, M. y col. «Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs)». En: *Global Change Biology* 14 (2008), pág. 2015.

- [8] Dumroese, R. K., Williams, M. I., Stanturf, J. A. y Clair, J. B. S. «Considerations for restoring temperate forests of tomorrow: forest restoration, assisted migration, and bioengineering». En: *New Forests* 46 (2015), pág. 947.
- [9] Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T. y Curtis-McLane, S. «Adaptation, migration or extirpation: climate change outcomes for tree populations». En: *Evolutionary Applications* 1 (2008), pág. 95.
- [10] Oddou-Muratorio, S., Klein, E. K., Vendramin, G. G. y Fady, B. «Spatial vs. temporal effects on demographic and genetic structures: the roles of dispersal, mast seeding and differential mortality on patterns of recruitment in *Fagus sylvatica*». En: *Molecular Ecology* 20 (2011), pág. 1997.
- [11] Zhu, K., Woodall, C. W. y Clark, J. S. «Failure to migrate: lack of tree range expansion in response to climate change». En: *Global Change Biology* 18 (2012), pág. 1042.
- [12] Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N. y col. «A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests». En: *Forest Ecology and Management* 259 (2010), pág. 660.
- [13] Bran, D., Pérez, A., Ghermandi, L. y Barrios-Lamuniere, S. «Evaluación de poblaciones de coihue (*Nothofagus dombeyi*) del Parque Nacional Nahuel Huapi, afectadas por la sequía 98/99, a escala de paisaje». Informe APN. 2001.
- [14] Hogg, E., Brandt, J. P. y Kochtubajda, B. «Growth and dieback of aspen forests in northwestern Alberta, Canada, in relation to climate and insects». En: *Canadian Journal of Forest Research* 32 (2002), pág. 823.
- [15] Peñuelas, J., Ogaya, R., Boada, M. y Jump, A. «Migration, invasion and decline: changes in recruitment and forest structure in a warming-linked shift of european beech forest in Catalonia (NE Spain)». En: *Ecography* 30 (2007), pág. 829.
- [16] Kohler, T. y Universität Bern, eds. *Mountains and Climate Change: a global concern*. Sustainable Mountain Development Series. Bern: Centre for Development and Environment, 2014.

- [17] Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T. y col. «Long-term Climate Change: projections, commitments and irreversibility». En: *Climate Change 2013: the physical science basis. Working Group I contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, 2013.
- [18] Woodward, F. *Climate and plant distribution*. Cambridge: Cambridge University Press, 1987.
- [19] Jump, A. S., Mátyás, C. y Peñuelas, J. «The altitude-for-latitude disparity in the range retractions of woody species». En: *Trends in Ecology & Evolution* 24 (2009), pág. 694.
- [20] Lenoir, J., Gegout, J. C., Marquet, P. A., de Ruffray, P. y Brisse, H. «A significant upward shift in plant species optimum elevation during the 20th century». En: *Science* 320 (2008), pág. 1768.
- [21] Wardle, P. y Coleman, M. «Evidence for rising upper limits of four native New Zealand forest trees». En: *New Zealand Journal of Botany* 30 (1992), pág. 303.
- [22] Körner, C. *Alpine plant life: functional plant ecology of high mountain ecosystems*. Heidelberg: Springer, 2003.
- [23] Loarie, S. R., Duffy, P. B., Hamilton, H., Asner, G. P., Field, C. B. y col. «The velocity of climate change». En: *Nature* 462 (2009), pág. 1052.
- [24] Körner, C. «Why are there global gradients in species richness? Mountains might hold the answer». En: *Trends in Ecology & Evolution* 15 (2000), pág. 513.
- [25] Beniston, M. «Climatic change in mountain regions: a review of possible impacts». En: *Climatic Change* 59 (2003), pág. 5.
- [26] Nogués-Bravo, D., Araújo, M., Errea, M. y Martínez-Rica, J. «Exposure of global mountain systems to climate warming during the 21st century». En: *Global Environmental Change* 17 (2007), pág. 420.
- [27] Barros, V. R., Boninsegna, J. A., Camilloni, I. A., Chidiak, M., Magrín, G. O. y col. «Climate change in Argentina: trends, projections, impacts and adaptation». En: *WIREs Climate Change* 6 (2015), pág. 151.

- [28] INTA. *Atlas Climático Digital de la República Argentina*. 2012. URL: <https://inta.gob.ar/documentos/atlas-climatico-digital-de-la-republica-argentina>.
- [29] Rusticucci, M. y Barrucand, M. «Observed trends and changes in temperature extremes over Argentina». En: *Journal of Climate* 17 (2004), pág. 4099.
- [30] Hoffmann, J. A., Núñez, S. E. y Vargas, W. M. «Temperature, humidity and precipitation variations in Argentina and the adjacent sub-antarctic region during the present century». En: *Meteorologische Zeitschrift* 6 (1997), pág. 3.
- [31] Barros, V., Camilloni, I., Frumento, O. y Rivero, M. «Estudio del potencial eólico en la República Argentina». En: *Programa de Abastecimiento Eléctrico a la Población Rural Dispersa*. Buenos Aires: Subsecretaría de Energía de la Nación, 1997.
- [32] Vera, C. y Camilloni, I. «Vulnerabilidad de la Patagonia y sur de las provincias de Buenos Aires y La Pampa». En: *Enabling Study for the Second National Communication to the UNFCCC*. Buenos Aires: SAyDS - Fundación e Instituto Di Tella, 2006.
- [33] Dimitri, M. J. *La región de los bosques andino-patagónicos: sinopsis general*. Buenos Aires: INTA Ediciones, 1972.
- [34] Hill, R. S. «*Nothofagus*: evolution from a southern perspective». En: *Trends in Ecology & Evolution* 7 (1992), pág. 190.
- [35] Hill, R. S. y Dettmann, M. E. «Origin and diversification of the genus *Nothofagus*». En: *The ecology and biogeography of Nothofagus forests*. Ed. por Veblen, T. T., Hill, R. S. y Read, J. New Haven: Yale University Press, 1996.
- [36] Torales, S., Rivarola, M., Pomponio, M. F., Fernández, P., Acuña, C. V. y col. «Transcriptome survey of patagonian southern beech *Nothofagus nervosa* (= *N. alpina*): assembly, annotation and molecular marker discovery». En: *BMC Genomics* 13 (2012), pág. 291.
- [37] El Mujtar, V. A., Gallo, L. A., Lang, T. y Garnier-Géré, P. «Development of genomic resources for *Nothofagus* species using next-generation sequencing data». En: *Molecular Ecology Resources* 14 (2014), pág. 1281.

- [38] Hasbún, R., González, J., Iturra, C., Fuentes, G., Alarcón, D. y col. «Using genome-wide SNP discovery and genotyping to reveal the main source of population differentiation in *Nothofagus dombeyi* (Mirb.) Oerst. in Chile». En: *International Journal of Genomics* 2016 (2016).
- [39] Donoso, C. «Ecology of *Nothofagus* forests in central Chile». En: *The ecology and biogeography of Nothofagus forests*. Ed. por Veblen, T. T., Hill, R. S. y Read, J. New Haven: Yale University Press, 1996.
- [40] Donoso, C., ed. *Las especies arbóreas de los bosques templados de Chile y Argentina: autoecología*. Valdivia: Marisa Cuneo Ediciones, 2006.
- [41] Veblen, T. T., Donoso, C., Kitzberger, T. y Rebertus, A. J. «Ecology of southern chilean and argentinean *Nothofagus* forests». En: *The ecology and biogeography of Nothofagus forests*. Ed. por Veblen, T. T., Hill, R. S. y Read, J. New Haven: Yale University Press, 1996.
- [42] Azpilicueta, M. M. y Marchelli, P., eds. *Zonas genéticas de raulí y roble pellín en Argentina: herramientas para la conservación y el manejo de la diversidad genética*. San Carlos de Bariloche: Ediciones INTA, 2016.
- [43] Soliani, C. y Marchelli, P., eds. *Zonas genéticas de lenga y ñire en Argentina y su aplicación en la conservación y manejo de los recursos forestales*. San Carlos de Bariloche: Ediciones INTA, 2017.
- [44] Tortorelli, L. *Maderas y bosques argentinos*. Buenos Aires: Editorial ACME, 1956.
- [45] Díaz-Vaz, J., Poblete, H., Juacida, L. y Devlieger, F. *Maderas comerciales de Chile*. 3° ed. Valdivia: Marisa Cuneo Ediciones, 2002.
- [46] Hall, M. y Witte, J. *Maderas del sur de Chile*. 2° ed. Santiago: Editorial Universitaria, 2004.
- [47] Donoso, P., Donoso, C., Gallo, L., Azpilicueta, M. M., Baldini, A. y col. «*Nothofagus obliqua* (Mirb.) Oerst.» En: *Las especies arbóreas de los bosques templados de Chile y Argentina: autoecología*. Ed. por Donoso, C. Valdivia: Marisa Cuneo Ediciones, 2006.

- [48] CIEFAP y Ministerio de Ambiente y Desarrollo Sustentable de la Nación. «Actualización de la clasificación de tipos forestales y cobertura del suelo de la región Bosque Andino-Patagónico. Informe final». 2016.
- [49] Secretaría de Ambiente y Desarrollo Sustentable de la Nación. «Primer inventario nacional de bosques nativos». Proyecto Bosques Nativos y Áreas Protegidas, Préstamo BIRF 4085 – AR. 2005.
- [50] González, M. E., Donoso, C., Ovalle, P. y Martínez-Pastur, G. «*Nothofagus pumilio* (Poep. et Endl) Krasser». En: *Las especies arbóreas de los bosques templados de Chile y Argentina: autoecología*. Ed. por Donoso, C. Valdivia: Marisa Cuneo Ediciones, 2006.
- [51] Bava, J. y Rechene, C. «Dinámica de la regeneración de lenga (*N. pumilio* (Poepp. et Endl. Krasser)) como base para la aplicación de sistemas silvícolas». En: *Ecología y Manejo de los Bosques de Argentina*. Ed. por Goya, J. F., Frangi, J. L. y Arturi, M. F. La Plata: Universidad Nacional de La Plata, 2004.
- [52] Martínez-Pastur, G., Lencinas, M. V., Peri, P. L., Cellini, J. M. y Moretto, A. «Investigación sobre manejo forestal a largo plazo en Patagonia sur - Argentina: lecciones del pasado, desafíos del presente». En: *Revista Chilena de Historia Natural* 83 (2010), pág. 159.
- [53] Millar, A. J. «The intracellular dynamics of circadian clocks reach for the light of ecology and evolution». En: *Annual Review of Plant Biology* 67 (2016), pág. 595.
- [54] Dodd, A. N., Salathia, N., Hall, A., Kévei, E., Tóth, R. y col. «Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage». En: *Science* 309 (2005), pág. 630.
- [55] Resco, V., Hartwell, J. y Hall, A. «Ecological implications of plants' ability to tell the time». En: *Ecology Letters* 12 (2009), pág. 583.
- [56] Harmer, S. L. «The circadian system in higher plants». En: *Annual Review of Plant Biology* 60 (2009), pág. 357.
- [57] Young, M. W. y Kay, S. A. «Time zones: a comparative genetics of circadian clocks». En: *Nature Reviews Genetics* 2 (2001), pág. 702.

- [58] Provart, N. J., Alonso, J., Assmann, S. M., Bergmann, D., Brady, S. M. y col. «50 years of *Arabidopsis* research: highlights and future directions». En: *New Phytologist* 209 (2016), pág. 921.
- [59] Alabadí, D., Oyama, T., Yanovsky, M. J., Harmon, F. G., Más, P. y col. «Reciprocal regulation between TOC1 and LHY/CCA1 within the *Arabidopsis* circadian clock». En: *Science* 293 (2001), pág. 880.
- [60] Pruneda-Paz, J. L., Breton, G., Para, A. y Kay, S. A. «A functional genomics approach reveals CHE as a component of the *Arabidopsis* circadian clock». En: *Science* 323 (2009), pág. 1481.
- [61] Nakamichi, N., Kiba, T., Henriques, R., Mizuno, T., Chua, N.-H. y col. «PSEUDO-RESPONSE REGULATORS 9, 7, and 5 are transcriptional repressors in the *Arabidopsis* circadian clock». En: *The Plant Cell* 22 (2010), pág. 594.
- [62] Helfer, A., Nusinow, D. A., Chow, B. Y., Gehrke, A. R., Bulyk, M. L. y col. «LUX ARRHYTHMO encodes a nighttime repressor of circadian gene expression in the *Arabidopsis* core clock». En: *Current Biology* 21 (2011), pág. 126.
- [63] Más, P., Kim, W.-Y., Somers, D. E. y Kay, S. A. «Targeted degradation of TOC1 by ZTL modulates circadian function in *Arabidopsis thaliana*». En: *Nature* 426 (2003), pág. 567.
- [64] Rugnone, M. L., Faigón Soverna, A., Sánchez, S. E., Schlaen, R. G., Hernando, C. E. y col. «LNK genes integrate light and clock signaling networks at the core of the *Arabidopsis* oscillator». En: *Proceedings of the National Academy of Sciences* 110 (2013), pág. 12120.
- [65] Hsu, P. Y. y Harmer, S. L. «Wheels within wheels: the plant circadian system». En: *Trends in Plant Science* 19 (2014), pág. 240.
- [66] De Montaigu, A., Tóth, R. y Coupland, G. «Plant development goes like clockwork». En: *Trends in Genetics* 26 (2010), pág. 296.
- [67] Mora-García, S., de Leone, M. J. y Yanovsky, M. «Time to grow: circadian regulation of growth and metabolism in photosynthetic organisms». En: *Current Opinion in Plant Biology* 35 (2017), pág. 84.

- [68] Covington, M. F., Maloof, J. N., Straume, M., Kay, S. A. y Harmer, S. L. «Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development». En: *Genome Biology* 9 (2008), R130.
- [69] Harmer, S. L., Hogenesch, J. B., Straume, M., Chang, H.-S., Han, B. y col. «Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock». En: *Science* 290 (2000), pág. 2110.
- [70] Higashi, T., Aoki, K., Nagano, A. J., Honjo, M. N. y Fukuda, H. «Circadian oscillation of the lettuce transcriptome under constant light and light–dark conditions». En: *Frontiers in Plant Science* 7 (2016), pág. 1114.
- [71] Cronn, R., Dolan, P. C., Jogdeo, S., Wegrzyn, J. L., Neale, D. B. y col. «Transcription through the eye of a needle: daily and annual cyclic gene expression variation in Douglas-fir needles». En: *BMC Genomics* 18 (2017), pág. 558.
- [72] Johnson, C., Elliott, J., Foster, R., Honma, K. y Kronauer, R. «Fundamental properties of circadian rhythms». En: *Chronobiology: biological timekeeping*. Ed. por Dunlap, J. C., Loros, J. J. y Decoursey, P. J. Sunderland: Sinauer, 2004.
- [73] Ruoff, P. «Introducing temperature-compensation in any reaction kinetic oscillator model». En: *Journal of Interdisciplinary Cycle Research* 23 (1992), pág. 92.
- [74] Terauchi, K., Kitayama, Y., Nishiwaki, T., Miwa, K., Murayama, Y. y col. «AT-Pase activity of KaiC determines the basic timing for circadian clock of cyanobacteria». En: *Proceedings of the National Academy of Sciences* 104 (2007), pág. 16377.
- [75] Isojima, Y., Nakajima, M., Ukai, H., Fujishima, H., Yamada, R. G. y col. «CKI ϵ / δ -dependent phosphorylation is a temperature-insensitive, period-determining process in the mammalian circadian clock». En: *Proceedings of the National Academy of Sciences* 106 (2009), pág. 15744.
- [76] Lakin-Thomas, P. L., Brody, S. y Côté, G. G. «Amplitude model for the effects of mutations and temperature on period and phase resetting of the *Neurospora* circadian oscillator». En: *Journal of Biological Rhythms* 6 (1991), pág. 281.
- [77] Kurosawa, G., Fujioka, A., Koinuma, S., Mochizuki, A. y Shigeyoshi, Y. «Temperature–amplitude coupling for stable biological rhythms at different temperatures». En: *PLoS Computational Biology* 13 (2017), e1005501.

- [78] Edwards, K. D., Lynn, J. R., Gyula, P., Nagy, F. y Millar, A. J. «Natural allelic variation in the temperature compensation mechanisms of the *Arabidopsis thaliana* circadian clock». En: *Genetics* 170 (2005), pág. 387.
- [79] Edwards, K. D., Anderson, P. E., Hall, A., Salathia, N. S., Locke, J. C. y col. «FLOWERING LOCUS C mediates natural variation in the high-temperature response of the *Arabidopsis* circadian clock». En: *The Plant Cell* 18 (2006), pág. 639.
- [80] Gould, P. D., Locke, J. C. W., Larue, C., Southern, M. M., Davis, S. J. y col. «The molecular basis of temperature compensation in the *Arabidopsis* circadian clock». En: *The Plant Cell* 18 (2006), pág. 1177.
- [81] Salomé, P. A., Weigel, D. y McClung, C. R. «The role of the *Arabidopsis* morning loop components CCA1, LHY, PRR7, and PRR9 in temperature compensation». En: *The Plant Cell* 22 (2010), pág. 3650.
- [82] Portolés, S. y Más, P. «The functional interplay between Protein Kinase CK2 and CCA1 transcriptional activity is essential for clock temperature compensation in *Arabidopsis*». En: *PLoS Genetics* 6 (2010), e1001201.
- [83] Schlaen, R. G., Mancini, E., S´nchez, S. E., Perez-Santángelo, S., Rugnone, M. L. y col. «The spliceosome assembly factor GEMIN2 attenuates the effects of temperature on alternative splicing and circadian rhythms». En: *Proceedings of the National Academy of Sciences* 112 (2015), pág. 9382.
- [84] Shalit-Kaneh, A., Kumimoto, R. W., Filkov, V. y Harmer, S. L. «Multiple feedback loops of the *Arabidopsis* circadian clock provide rhythmic robustness across environmental conditions». En: *Proceedings of the National Academy of Sciences* 115 (2018), pág. 7147.
- [85] Ramos, A., Pérez-Solís, E., Ibáñez, C., Casado, R., Collada, C. y col. «Winter disruption of the circadian clock in chestnut». En: *Proceedings of the National Academy of Sciences* 102 (2005), pág. 7037.
- [86] Ibáñez, C., Ramos, A., Acebo, P., Contreras, A., Casado, R. y col. «Overall alteration of circadian clock gene expression in the chestnut cold response». En: *PLoS One* 3 (2008), e3567.

- [87] Ibáñez, C., Kozarewa, I., Johansson, M., Ogren, E., Rohde, A. y col. «Circadian clock components regulate entry and affect exit of seasonal dormancy as well as winter hardiness in *Populus* trees». En: *Plant Physiology* 153 (2010), pág. 1823.
- [88] Sanger, F., Nicklen, S. y Coulson, A. R. «DNA sequencing with chain-terminating inhibitors». En: *Proceedings of the National Academy of Sciences* 74 (1977), pág. 5463.
- [89] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D. y col. «The genome sequence of *Drosophila melanogaster*». En: *Science* 287 (2000), pág. 2185.
- [90] Arabidopsis Genome Initiative y col. «Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*». En: *Nature* 408 (2000), pág. 796.
- [91] International Human Genome Sequencing Consortium y col. «Initial sequencing and analysis of the human genome». En: *Nature* 409 (2001), pág. 860.
- [92] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J. y col. «The sequence of the human genome». En: *Science* 291 (2001), pág. 1304.
- [93] Metzker, M. L. «Sequencing technologies — the next generation». En: *Nature Reviews Genetics* 11 (2010), pág. 31.
- [94] Holt, R. A. y Jones, S. J. «The new paradigm of flow cell sequencing». En: *Genome Research* 18 (2008), pág. 839.
- [95] Vogel, A., Schwacke, R., Denton, A. K., Usadel, B., Hollmann, J. y col. «Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*». En: *Nature Communications* 9 (2018), pág. 2515.
- [96] Salojärvi, J., Smolander, O.-P., Nieminen, K., Rajaraman, S., Safronov, O. y col. «Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch». En: *Nature Genetics* 49 (2017), pág. 904.
- [97] Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B. y col. «The genome of *Chenopodium quinoa*». En: *Nature* 542 (2017), pág. 307.
- [98] Cumbie, J. S., Filichkin, S. A. y Megraw, M. «Improved DNase-seq protocol facilitates high resolution mapping of DNase I hypersensitive sites in roots in *Arabidopsis thaliana*». En: *Plant Methods* 11 (2015), pág. 42.

- [99] Sullivan, A. M., Bubba, K. L., Sandstrom, R., Stamatoyannopoulos, J. A. y Queitsch, C. «DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants». En: *Current Plant Biology* 3 (2015), pág. 40.
- [100] Sun, T., Zhang, Y., Li, Y., Zhang, Q., Ding, Y. y col. «ChIP-seq reveals broad roles of SARD1 and CBP60g in regulating plant immunity». En: *Nature Communications* 6 (2015), pág. 10159.
- [101] Ricardi, M. M., González, R. M., Zhong, S., Domínguez, P. G., Duffy, T. y col. «Genome-wide data (ChIP-seq) enabled identification of cell wall-related and aquaporin genes as targets of tomato ASR1, a drought stress-responsive transcription factor». En: *BMC Plant Biology* 14 (2014), pág. 29.
- [102] Chen, X., Ge, X., Wang, J., Tan, C., King, G. J. y col. «Genome-wide DNA methylation profiling by modified reduced representation bisulfite sequencing in *Brassica rapa* suggests that epigenetic modifications play a key role in polyploid genome evolution». En: *Frontiers in Plant Science* 6 (2015), pág. 836.
- [103] Susek, K., Braszewska-Zalewska, A., Bewick, A. J., Hasterok, R., Schmitz, R. J. y col. «Epigenomic diversification within the genus *Lupinus*». En: *PLoS One* 12 (2017), e0179821.
- [104] Pennisi, E. «ENCODE Project writes eulogy for junk DNA». En: *Science* 337 (2012), pág. 1159.
- [105] Wang, Z., Gerstein, M. y Snyder, M. «RNA-seq: a revolutionary tool for transcriptomics». En: *Nature Reviews Genetics* 10 (2009), pág. 57.
- [106] Divya Bhanu, B., Ulaganathan, K., Shanker, A. K. y Desai, S. «RNA-seq analysis of irrigated vs. water stressed transcriptomes of *Zea mays* cultivar Z59». En: *Frontiers in Plant Science* 7 (2016), pág. 239.
- [107] Lukoszek, R., Feist, P. e Ignatova, Z. «Insights into the adaptive response of *Arabidopsis thaliana* to prolonged thermal stress by ribosomal profiling and RNA-seq». En: *BMC Plant Biology* 16 (2016), pág. 221.
- [108] Niu, S.-C., Xu, Q., Zhang, G.-Q., Zhang, Y.-Q., Tsai, W.-C. y col. «De novo transcriptome assembly databases for the butterfly orchid *Phalaenopsis equestris*». En: *Scientific Data* 3 (2016), pág. 160083.

- [109] Chakrabarti, M., Dinkins, R. D. y Hunt, A. G. «*De novo* transcriptome assembly and dynamic spatial gene expression analysis in red clover». En: *The Plant Genome* 9 (2016).
- [110] Wang, Y., Li, X., Zhou, W., Li, T. y Tian, C. «*De novo* assembly and transcriptome characterization of spruce dwarf mistletoe *Arceuthobium sichuanense* uncovers gene expression profiling associated with plant development». En: *BMC Genomics* 17 (2016), pág. 771.
- [111] D'Esposito, D., Orrù, L., Dattolo, E., Bernardo, L., Lamontara, A. y col. «Transcriptome characterisation and Simple Sequence Repeat marker discovery in the seagrass *Posidonia oceanica*». En: *Scientific Data* 3 (2016), pág. 160115.
- [112] Mora-Ortiz, M., Swain, M. T., Vickers, M. J., Hegarty, M. J., Kelly, R. y col. «*De novo* transcriptome assembly for gene identification, analysis, annotation, and molecular marker discovery in *Onobrychis viciifolia*». En: *BMC Genomics* 17 (2016), pág. 756.
- [113] Liu, Z.-W., Wu, Z.-J., Li, X.-H., Huang, Y., Li, H. y col. «Identification, classification, and expression profiles of heat shock transcription factors in tea plant (*Camellia sinensis*) under temperature stress». En: *Gene* 576 (2016), pág. 52.
- [114] Schulze, A. y Downward, J. «Navigating gene expression using microarrays—a technology review». En: *Nature Cell Biology* 3 (2001), E190.
- [115] Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P. y col. «Empirical analysis of transcriptional activity in the *Arabidopsis* genome». En: *Science* 302 (2003), pág. 842.
- [116] Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R. y col. «A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)». En: *Science* 296 (2002), pág. 92.
- [117] Lan, L., Chen, W., Lai, Y., Suo, J., Kong, Z. y col. «Monitoring of gene expression profiles and isolation of candidate genes involved in pollination and fertilization in rice (*Oryza sativa* L.) with a 10K cDNA microarray». En: *Plant Molecular Biology* 54 (2004), pág. 471.

- [118] Fernández, P., Soria, M., Blesa, D., DiRienzo, J., Moschen, S. y col. «Development, characterization and experimental validation of a cultivated sunflower (*Helianthus annuus* L.) gene expression oligonucleotide microarray». En: *PLoS One* 7 (2012), e45899.
- [119] Compeau, P. E., Pevzner, P. A. y Tesler, G. «How to apply de Bruijn graphs to genome assembly». En: *Nature Biotechnology* 29 (2011), pág. 987.
- [120] Kececioğlu, J. D. y Myers, E. W. «Combinatorial algorithms for DNA sequence assembly». En: *Algorithmica* 13 (1995), pág. 7.
- [121] Garey, M. R. y Johnson, D. S. *Computers and intractability*. New York: W. H. Freeman, 1979.
- [122] Pevzner, P. A., Tang, H. y Waterman, M. S. «An eulerian path approach to DNA fragment assembly». En: *Proceedings of the National Academy of Sciences* 98 (2001), pág. 9748.
- [123] De Bruijn, N. G. «A combinatorial problem». En: *Koninklijke Nederlandse Akademie van Wetenschappen* 49 (1946), pág. 758.
- [124] Schliesky, S., Gowik, U., Weber, A. P. M. y Bräutigam, A. «RNA-seq assembly – are we there yet?». En: *Frontiers in Plant Science* 3 (2012), pág. 220.
- [125] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M. y col. «SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing». En: *Journal of Computational Biology* 19 (2012), pág. 455.
- [126] Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E. y col. «Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs». En: *Genome Research* 14 (2004), pág. 1147.
- [127] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A. y col. «Full-length transcriptome assembly from RNA-seq data without a reference genome». En: *Nature Biotechnology* 29 (2011), pág. 644.
- [128] Zerbino, D. R. y Birney, E. «Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs». En: *Genome Research* 18 (2008), pág. 821.

- [129] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. y col. «ABYSS: a parallel assembler for short read sequence data». En: *Genome Research* 19 (2009), pág. 1117.
- [130] Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J. y col. «SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-seq reads». En: *Bioinformatics* 30 (2014), pág. 1660.
- [131] Cabau, C., Escudié, F., Djari, A., Guiguen, Y., Bobe, J. y col. «Compacting and correcting Trinity and Oases RNA-seq *de novo* assemblies». En: *PeerJ* 5 (2017), e2988.
- [132] Yang, Y. y Smith, S. A. «Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics». En: *BMC Genomics* 14 (2013), pág. 328.
- [133] Estravis-Barcala, M. y Bellora, N. *Lenga*. 2018. URL: <https://github.com/MaxiEstravis/lenga/blob/0c06aea844e766c0a94dca6e098b6517d1674078/lenga.zip>.
- [134] The UniProt Consortium. «UniProt: the universal protein knowledgebase». En: *Nucleic Acids Research* 45 (2017), pág. D158.
- [135] Chang, S., Puryear, J. y Cairney, J. «A simple and efficient method for isolating RNA from pine trees». En: *Plant Molecular Biology Reporter* 11 (1993), pág. 113.
- [136] Wang, Y., Ghaffari, N., Johnson, C. D., Braga-Neto, U. M., Wang, H. y col. «Evaluation of the coverage and depth of transcriptome by RNA-seq in chickens». En: *BMC Bioinformatics* 12 (2011), S5.
- [137] McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L. y col. «RNA-seq: technical variability and sampling». En: *BMC Genomics* 12 (2011), pág. 293.
- [138] Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J. y col. «Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing». En: *BMC Genomics* 13 (2012), pág. 484.

- [139] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W. y col. «An integrated semiconductor device enabling non-optical genome sequencing». En: *Nature* 475 (2011), pág. 348.
- [140] Andrews, S. *FastQC: a quality control tool for high throughput sequence data*. 2010. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- [141] Bolger, A. M., Lohse, M. y Usadel, B. «Trimmomatic: a flexible trimmer for Illumina sequence data». En: *Bioinformatics* 30 (2014), pág. 2114.
- [142] Gordon, A. y Hannon, G. *FASTQ/A short-reads preprocessing tools*. 2010. URL: http://hannonlab.cshl.edu/fastx_toolkit.
- [143] Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D. y col. «De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis». En: *Nature Protocols* 8 (2013), pág. 1494.
- [144] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. y Kingsford, C. «Salmon provides fast and bias-aware quantification of transcript expression». En: *Nature Methods* 14 (2017), pág. 417.
- [145] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Viena, 2015.
- [146] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C. y col. «STAR: ultrafast universal RNA-seq aligner». En: *Bioinformatics* 29 (2013), pág. 15.
- [147] Parra, G., Bradnam, K. y Korf, I. «CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes». En: *Bioinformatics* 23 (2007), pág. 1061.
- [148] Estravis-Barcala, M. y Bellora, N. *CEG search*. 2018. URL: <https://github.com/MaxiEstravis/lenga/blob/0c06aea844e766c0a94dca6e098b6517d1674078/ceg.zip>.
- [149] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. y Zdobnov, E. M. «BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs». En: *Bioinformatics* 31 (2015), pág. 3210.

- [150] Krogh, A., Brown, M., Mian, I. S., Sjölander, K. y Haussler, D. «Hidden Markov models in computational biology: applications to protein modeling». En: *Journal of Molecular Biology* 235 (1994), pág. 1501.
- [151] Eddy, S. R. «Profile hidden Markov models.» En: *Bioinformatics* 14 (1998), pág. 755.
- [152] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. y Lipman, D. J. «Basic Local Alignment Search Tool». En: *Journal of Molecular Biology* 215 (1990), pág. 403.
- [153] Soliani, C., Azpilicueta, M. M., Arana, M. V. y Marchelli, P. «Signs of selection along precipitation gradients in patagonian temperate forests: the case of three *Nothofagus* spp.» Enviado. 2018.
- [154] Martin, P. y Dowd, J. «Using sequences of *rbcL* to study phylogeny and biogeography of *Nothofagus* species». En: *Australian Systematic Botany* 6 (1993), pág. 441.
- [155] Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S. y col. «The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses». En: *The Plant Journal* 50 (2007), pág. 347.
- [156] Mockler, T., Michael, T., Priest, H., Shen, R., Sullivan, C. y col. «The DIURNAL project: diurnal and circadian expression profiling, model-based pattern matching, and promoter analysis». En: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 72. Cold Spring Harbor Laboratory Press. Nueva York, 2007.
- [157] Kent, W. J. «BLAT—the BLAST-like alignment tool». En: *Genome Research* 12 (2002), pág. 656.
- [158] Botstein, D., Cherry, J. M., Ashburner, M., Ball, C. A., Blake, J. A. y col. «Gene Ontology: tool for the unification of biology». En: *Nature Genetics* 25 (2000), pág. 25.
- [159] Thiel, T. *MISA: microsatellite identification tool*. 2003. URL: <http://pgrc.ipk-gatersleben.de/misa/>.
- [160] Estravis-Barcala, M. *Predicciones de SSRs para Nothofagus pumilio*. 2018. URL: <https://figshare.com/s/86b576e1d6b6fa34d42b>.

- [161] Python Core Team. *Python: a dynamic, open source programming language*. Python Software Foundation. 2017.
- [162] The PHP Documentation Group. *PHP: language reference*. The PHP Group. 2018.
- [163] Estravis-Barcala, M. *Contig retrieval*. 2018. URL: https://github.com/MaxiEstravis/lenga/blob/0c06aea844e766c0a94dca6e098b6517d1674078/contig_retrieval.zip.
- [164] Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J. y col. «Characterizing and measuring bias in sequence data». En: *Genome Biology* 14 (2013), R51.
- [165] Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E. y col. «Performance comparison of benchtop high-throughput sequencing platforms». En: *Nature Biotechnology* 30 (2012), pág. 434.
- [166] Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. y Tyson, G. W. «Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data». En: *PLoS Computational Biology* 9 (2013), e1003031.
- [167] Chikhi, R. y Medvedev, P. «Informed and automated k-mer size selection for genome assembly». En: *Bioinformatics* 30 (2013), pág. 31.
- [168] Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P. y col. «BUSCO applications from quality assessments to gene prediction and phylogenomics». En: *Molecular Biology and Evolution* 35 (2017), pág. 543.
- [169] The Angiosperm Phylogeny Group. «An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV». En: *Botanical Journal of the Linnean Society* 181 (2016), pág. 1.
- [170] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A. y col. «The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003». En: *Nucleic Acids Research* 31 (2003), pág. 365.
- [171] Slater, G. S. C. y Birney, E. «Automated generation of heuristics for biological sequence comparison». En: *BMC Bioinformatics* 6 (2005), pág. 31.

- [172] Azpilicueta, M. M., Caron, H., Bodénès, C. y Gallo, L. A. «SSR markers for analysing south american *Nothofagus* species». En: *Silvae Genetica* 53 (2004), pág. 240.
- [173] Marchelli, P., Caron, H., Azpilicueta, M. M. y Gallo, L. A. «Primer note: a new set of highly polymorphic nuclear microsatellite markers for *Nothofagus nervosa* and related south american species». En: *Silvae Genetica* 57 (2008), pág. 82.
- [174] Soliani, C., Sebastiani, F., Marchelli, P., Gallo, L. A. y Vendramin, G. G. «Development of novel genomic microsatellite markers in the southern beech *Nothofagus pumilio* (Poepp. et Endl.) Krasser». En: *Molecular Ecology Resources* 10 (2010), pág. 404.
- [175] Soliani, C., Tsuda, Y., Bagnoli, F., Gallo, L. A., Vendramin, G. G. y col. «Halfway encounters: meeting points of colonization routes among the southern beeches *Nothofagus pumilio* and *N. antarctica*». En: *Molecular Phylogenetics and Evolution* 85 (2015), pág. 197.
- [176] Futuyma, D. J. *Evolution*. Sunderland: Sinauer, 2005.
- [177] Darwin, C. *On the origin of species*. London: John Murray, 1859.
- [178] Graur, D. y Li, W.-H. *Fundamentals of Molecular Evolution*. 2º ed. Sunderland: Sinauer, 2000.
- [179] Chase, M. W., Soltis, D. E., Olmstead, R. G., Morgan, D., Les, D. H. y col. «Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*». En: *Annals of the Missouri Botanical Garden* 80 (1993), pág. 528.
- [180] Soltis, D. E., Soltis, P. S., Nickrent, D. L., Johnson, L. A., Hahn, W. J. y col. «Angiosperm phylogeny inferred from 18S ribosomal DNA sequences». En: *Annals of the Missouri Botanical Garden* 83 (1997), pág. 1.
- [181] Savolainen, V., Chase, M. W., Hoot, S. B., Morton, C. M., Soltis, D. E. y col. «Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences». En: *Systematic Biology* 49 (2000), pág. 306.
- [182] Soltis, D. E., Soltis, P. S., Chase, M. W., Mort, M. E., Albach, D. C. y col. «Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences». En: *Botanical Journal of the Linnean Society* 133 (2000), pág. 381.

- [183] Soltis, D. E., Smith, S. A., Cellinese, N., Wurdack, K. J., Tank, D. C. y col. «Angiosperm phylogeny: 17 genes, 640 taxa». En: *American Journal of Botany* 98 (2011), pág. 704.
- [184] Delsuc, F., Brinkmann, H. y Philippe, H. «Phylogenomics and the reconstruction of the tree of life». En: *Nature Reviews Genetics* 6 (2005), pág. 361.
- [185] Philippe, H., Delsuc, F., Brinkmann, H. y Lartillot, N. «Phylogenomics». En: *Annual Review of Ecology, Evolution, and Systematics* 36 (2005), pág. 541.
- [186] Fernández, R., Hormiga, G. y Giribet, G. «Phylogenomic analysis of spiders reveals nonmonophyly of orb weavers». En: *Current Biology* 24 (2014), pág. 1772.
- [187] Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A. y col. «Phylogenomics resolves the timing and pattern of insect evolution». En: *Science* 346 (2014), pág. 763.
- [188] Pease, J. B., Haak, D. C., Hahn, M. W. y Moyle, L. C. «Phylogenomics reveals three sources of adaptive variation during a rapid radiation». En: *PLoS Biology* 14 (2016), e1002379.
- [189] Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P. y col. «Whole-genome analyses resolve early branches in the tree of life of modern birds». En: *Science* 346 (2014), pág. 1320.
- [190] Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J.-Y. y col. «Phylotranscriptomic consolidation of the jawed vertebrate timetree». En: *Nature Ecology & Evolution* 1 (2017), pág. 1370.
- [191] Shen, X.-X., Zhou, X., Kominek, J., Kurtzman, C. P., Hittinger, C. T. y col. «Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data». En: *G3: Genes, Genomes, Genetics* 6 (2016), g3.
- [192] Gee, H. «Evolution: ending incongruence». En: *Nature* 425 (2003), pág. 782.
- [193] Jeffroy, O., Brinkmann, H., Delsuc, F. y Philippe, H. «Phylogenomics: the beginning of incongruence?» En: *Trends in Genetics* 22 (2006), pág. 225.
- [194] Phillips, M. J., Delsuc, F. y Penny, D. «Genome-scale phylogeny and the detection of systematic biases». En: *Molecular Biology and Evolution* 21 (2004), pág. 1455.

- [195] Brown, J. M. y Thomson, R. C. «Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses». En: *Systematic Biology* 66 (2016), pág. 517.
- [196] Shen, X.-X., Hittinger, C. T. y Rokas, A. «Contentious relationships in phylogenomic studies can be driven by a handful of genes». En: *Nature Ecology & Evolution* 1 (2017), pág. 0126.
- [197] Betancur-R., R., Naylor, G. J. P. y Orti, G. «Conserved genes, sampling error, and phylogenomic inference». En: *Systematic Biology* 63 (2014), pág. 257.
- [198] Whelan, N. V., Kocot, K. M., Moroz, L. L. y Halanych, K. M. «Error, signal, and the placement of Ctenophora sister to all other animals». En: *Proceedings of the National Academy of Sciences* 112 (2015), pág. 5773.
- [199] Holder, M. y Lewis, P. O. «Phylogeny estimation: traditional and bayesian approaches». En: *Nature Reviews Genetics* 4 (2003), pág. 275.
- [200] Felsenstein, J. *Inferring phylogenies*. Sunderland: Sinauer, 2004.
- [201] Saitou, N. y Nei, M. «The neighbor-joining method: a new method for reconstructing phylogenetic trees.» En: *Molecular Biology and Evolution* 4 (1987), pág. 406.
- [202] Rzhetsky, A. y Nei, M. «Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference». En: *Journal of Molecular Evolution* 35 (1992), pág. 367.
- [203] Edwards, A. W. y Cavalli-Sforza, L. L. «The reconstruction of evolution». En: *Annals of Human Genetics* 27 (1963), pág. 104.
- [204] Felsenstein, J. «Evolutionary trees from DNA sequences: a maximum likelihood approach». En: *Journal of Molecular Evolution* 17 (1981), pág. 368.
- [205] Huelsenbeck, J. P., Ronquist, F., Nielsen, R. y Bollback, J. P. «Bayesian inference of phylogeny and its impact on evolutionary biology». En: *Science* 294 (2001), pág. 2310.
- [206] Dayhoff, M. O. «A model of evolutionary change in proteins». En: *Atlas of Protein Sequence and Structure* 5 (1972), pág. 89.
- [207] Müller, T. y Vingron, M. «Modeling amino acid replacement». En: *Journal of Computational Biology* 7 (2000), pág. 761.

- [208] Jones, D. T., Taylor, W. R. y Thornton, J. M. «The rapid generation of mutation data matrices from protein sequences». En: *Bioinformatics* 8 (1992), pág. 275.
- [209] Henikoff, S. y Henikoff, J. G. «Amino acid substitution matrices from protein blocks». En: *Proceedings of the National Academy of Sciences* 89 (1992), pág. 10915.
- [210] Whelan, S. y Goldman, N. «A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach». En: *Molecular Biology and Evolution* 18 (2001), pág. 691.
- [211] Le, S. Q. y Gascuel, O. «An improved general amino acid replacement matrix». En: *Molecular Biology and Evolution* 25 (2008), pág. 1307.
- [212] Dimmic, M. W., Rest, J. S., Mindell, D. P. y Goldstein, R. A. «rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny». En: *Journal of Molecular Evolution* 55 (2002), pág. 65.
- [213] Adachi, J. y Hasegawa, M. «Model of amino acid substitution in proteins encoded by mitochondrial DNA». En: *Journal of Molecular Evolution* 42 (1996), pág. 459.
- [214] Adachi, J., Waddell, P. J., Martin, W. y Hasegawa, M. «Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA». En: *Journal of Molecular Evolution* 50 (2000), pág. 348.
- [215] Nickle, D. C., Heath, L., Jensen, M. A., Gilbert, P. B., Mullins, J. I. y col. «HIV-specific probabilistic models of protein evolution». En: *PLoS One* 2 (2007), e503.
- [216] De la Torre-Bárcena, J. E., Egan, M. G., Katari, M. S., Brenner, E. D., Stevenson, D. W. y col. «ESTimating plant phylogeny: lessons from partitioning». En: *BMC Evolutionary Biology* 6 (2006), pág. 48.
- [217] De la Torre-Bárcena, J. E., Kolokotronis, S.-O., Lee, E. K., Stevenson, D. W., Brenner, E. D. y col. «The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data». En: *PLoS One* 4 (2009), e5764.

- [218] Cibrián-Jaramillo, A., de la Torre-Bárcena, J. E., Lee, E. K., Katari, M. S., Little, D. P. y col. «Using phylogenomic patterns and gene ontology to identify proteins of importance in plant evolution». En: *Genome Biology and Evolution* 2 (2010), pág. 225.
- [219] Lee, E. K., Cibrian-Jaramillo, A., Kolokotronis, S.-O., Katari, M. S., Stamatakis, A. y col. «A functional phylogenomic view of the seed plants». En: *PLoS Genetics* 7 (2011), e1002411.
- [220] Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E. y col. «Phylo-transcriptomic analysis of the origin and early diversification of land plants». En: *Proceedings of the National Academy of Sciences* 111 (2014), E4859.
- [221] Smith, S. A. y Brown, J. W. «Constructing a broadly inclusive seed plant phylogeny». En: *American Journal of Botany* 105 (2018), pág. 302.
- [222] Sanderson, M. J., McMahon, M. M. y Steel, M. «Phylogenomics with incomplete taxon coverage: the limits to inference». En: *BMC Evolutionary Biology* 10 (2010), pág. 155.
- [223] Sanderson, M. J., McMahon, M. M. y Steel, M. «Terraces in phylogenetic tree space». En: *Science* 333 (2011), pág. 448.
- [224] Sanderson, M. J., McMahon, M. M., Stamatakis, A., Zwickl, D. J. y Steel, M. «Impacts of terraces on phylogenetic inference». En: *Systematic Biology* 64 (2015), pág. 709.
- [225] Walker, J. F., Yang, Y., Feng, T., Timoneda, A., Mikenas, J. y col. «From cacti to carnivores: improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of Caryophyllales». En: *American Journal of Botany* 105 (2018), pág. 446.
- [226] Beaulieu, J. M. y O'Meara, B. C. «Can we build it? Yes we can, but should we use it? Assessing the quality and value of a very large phylogeny of campanulid angiosperms». En: *American Journal of Botany* 105 (2018), pág. 417.
- [227] Stevens, P. *Angiosperm Phylogeny Website*. 2017. URL: <http://www.mobot.org/MOBOT/research/APweb/>.

- [228] Manos, P. S., Nixon, K. C. y Doyle, J. J. «Cladistic analysis of restriction site variation within the chloroplast DNA inverted repeat region of selected Hamamelididae». En: *Systematic Botany* 18 (1993), pág. 551.
- [229] Manos, P. S. y Steele, K. P. «Phylogenetic analyses of “higher” Hamamelididae based on plastid sequence data». En: *American Journal of Botany* 84 (1997), pág. 1407.
- [230] Li, R., Chen, Z., Lu, A., Soltis, D. E., Soltis, P. S. y col. «Phylogenetic relationships in Fagales based on DNA sequences from three genomes». En: *International Journal of Plant Sciences* 165 (2004), pág. 311.
- [231] Xing, Y., Onstein, R. E., Carter, R. J., Stadler, T. y Peter Linder, H. «Fossils and a large molecular phylogeny show that the evolution of species richness, generic diversity, and turnover rates are disconnected». En: *Evolution* 68 (2014), pág. 2821.
- [232] Xiang, X.-G., Wang, W., Li, R.-Q., Lin, L., Liu, Y. y col. «Large-scale phylogenetic analyses reveal fagalean diversification promoted by the interplay of diaspores and environments in the Paleogene». En: *Perspectives in Plant Ecology, Evolution and Systematics* 16 (2014), pág. 101.
- [233] Li, H.-L., Wang, W., Mortimer, P. E., Li, R.-Q., Li, D.-Z. y col. «Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change». En: *Scientific Reports* 5 (2015), pág. 14023.
- [234] Hong-Lei, L., Wei, W., Rui-Qi, L., Jing-Bo, Z., Miao, S. y col. «Global versus chinese perspectives on the phylogeny of the N-fixing clade». En: *Journal of Systematics and Evolution* 54 (2016), pág. 392.
- [235] Zhi-Duan, C., Tuo, Y., Li, L., Li-Min, L., Hong-Lei, L. y col. «Tree of life for the genera of chinese vascular plants». En: *Journal of Systematics and Evolution* 54 (2016), pág. 277.
- [236] Larson-Johnson, K. «Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales». En: *New Phytologist* 209 (2016), pág. 418.

- [237] Culina, A., Crowther, T. W., Ramakers, J. J., Gienapp, P. y Visser, M. E. «How to do meta-analysis of open datasets». En: *Nature Ecology & Evolution* 2 (2018), pág. 1053.
- [238] Culina, A., Baglioni, M., Crowther, T. W., Visser, M. E., Woutersen-Windhouwer, S. y col. «Navigating the unfolding open data landscape in ecology and evolution». En: *Nature Ecology & Evolution* 2 (2018), pág. 420.
- [239] Darriba, D., Taboada, G. L., Doallo, R. y Posada, D. «ProtTest 3: fast selection of best-fit models of protein evolution». En: *Bioinformatics* 27 (2011), pág. 1164.
- [240] Guindon, S. y Gascuel, O. «A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood». En: *Systematic Biology* 52 (2003), pág. 696.
- [241] Stamatakis, A. «RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies». En: *Bioinformatics* 30 (2014), pág. 1312.
- [242] Rambaut, A. *FigTree, a graphical viewer of phylogenetic trees*. 2007. URL: <http://tree.bio.ed.ac.uk/software/figtree>.
- [243] Salichos, L. y Rokas, A. «Inferring ancient divergences requires genes with strong phylogenetic signals». En: *Nature* 497 (2013), pág. 327.
- [244] Salichos, L., Stamatakis, A. y Rokas, A. «Novel information theory-based measures for quantifying incongruence among phylogenetic trees». En: *Molecular Biology and Evolution* 31 (2014), pág. 1261.
- [245] Hinchliff, C. E. y Smith, S. A. «Some limitations of public sequence data for phylogenetic inference (in plants)». En: *PLoS One* 9 (2014), e98986.
- [246] Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E. y Smith, S. A. «Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life». En: *American Journal of Botany* 105 (2018), pág. 385.
- [247] Eiserhardt, W. L., Antonelli, A., Bennett, D. J., Botigué, L. R., Burleigh, J. G. y col. «A roadmap for global synthesis of the plant tree of life». En: *American Journal of Botany* 105 (2018), pág. 614.

- [248] Nixon, K. C. «Origins of Fagaceae». En: *Evolution, Systematics, and Fossil History of the Hamamelidae*, vol. 2. Ed. por Crane, P. R. y Blackmore, S. Oxford: Clarendon Press, 1989.
- [249] Walter, A. y Schurr, U. «Dynamics of leaf and root growth: endogenous control versus environmental impact». En: *Annals of Botany* 95 (2005), pág. 891.
- [250] Salathia, N., Lynn, J. R., Millar, A. J. y King, G. J. «Detection and resolution of genetic loci affecting circadian period in *Brassica oleracea*». En: *Theoretical and Applied Genetics* 114 (2007), pág. 683.
- [251] Müller, N. A., Wijnen, C. L., Srinivasan, A., Ryngajllo, M., Ofner, I. y col. «Domestication selected for deceleration of the circadian clock in cultivated tomato». En: *Nature Genetics* 48 (2015), pág. 89.
- [252] Pan, W.-J., Wang, X., Deng, Y.-R., Li, J.-H., Chen, W. y col. «Nondestructive and intuitive determination of circadian chlorophyll rhythms in soybean leaves using multispectral imaging». En: *Scientific Reports* 5 (2015), pág. 11108.
- [253] Litthauer, S., Battle, M. W., Lawson, T. y Jones, M. A. «Phototropins maintain robust circadian oscillation of PSII operating efficiency under blue light». En: *The Plant Journal* 83 (2015), pág. 1034.
- [254] Dakhiya, Y., Hussien, D., Fridman, E., Kiflawi, M. y Green, R. «Correlations between circadian rhythms and growth in challenging environments». En: *Plant Physiology* 173 (2017), pág. 1724.
- [255] Gould, P. D., Diaz, P., Hogben, C., Kusakina, J., Salem, R. y col. «Delayed fluorescence as a universal tool for the measurement of circadian rhythms in higher plants». En: *The Plant Journal* 58 (2009), pág. 893.
- [256] Dodd, A. N., Love, J. y Webb, A. A. «The plant clock shows its metal: circadian regulation of cytosolic free Ca^{2+} ». En: *Trends in Plant Science* 10 (2005), pág. 15.
- [257] Derbidge, R., Baumgartner, S. y Heusser, P. «Mistletoe berry outline mapping with a path curve function and recording the circadian rhythm of their phenotypic shape change». En: *Frontiers in Plant Science* 7 (2016), pág. 1749.

- [258] Goodspeed, D., Chehab, E. W., Min-Venditti, A., Braam, J. y Covington, M. F. «*Arabidopsis* synchronizes jasmonate-mediated defense with insect circadian behavior». En: *Proceedings of the National Academy of Sciences* 109 (2012), pág. 4674.
- [259] Helsper, J. P., Davies, J. A., Bouwmeester, H. J., Krol, A. F. y van Kampen, M. H. «Circadian rhythmicity in emission of volatile compounds by flowers of *Rosa hybrida* L. cv. Honesty». En: *Planta* 207 (1998), pág. 88.
- [260] Kolosova, N., Gorenstein, N., Kish, C. M. y Dudareva, N. «Regulation of circadian methyl benzoate emission in diurnally and nocturnally emitting plants». En: *The Plant Cell* 13 (2001), pág. 2333.
- [261] Yakir, E., Hilman, D., Harir, Y. y Green, R. M. «Regulation of output from the plant circadian clock». En: *FEBS Journal* 274 (2007), pág. 335.
- [262] Farré, E. M. y Weise, S. E. «The interactions between the circadian clock and primary metabolism». En: *Current Opinion in Plant Biology* 15 (2012), pág. 293.
- [263] Dodd, A. N., Kusakina, J., Hall, A., Gould, P. D. y Hanaoka, M. «The circadian regulation of photosynthesis». En: *Photosynthesis Research* 119 (2014), pág. 181.
- [264] Kinmonth-Schultz, H. A., Golembeski, G. S. e Imaizumi, T. «Circadian clock-regulated physiological outputs: dynamic responses in nature». En: *Seminars in Cell & Developmental Biology* 24 (2013), pág. 407.
- [265] Greenham, K. y McClung, C. R. «Integrating circadian dynamics with physiological processes in plants». En: *Nature Reviews Genetics* 16 (2015), pág. 598.
- [266] Atamian, H. S. y Harmer, S. L. «Circadian regulation of hormone signaling and plant physiology». En: *Plant Molecular Biology* 91 (2016), pág. 691.
- [267] Gould, K. S. «Functional role of anthocyanins in the leaves of *Quintinia serrata* A. Cunn.» En: *Journal of Experimental Botany* 51 (2000), pág. 1107.
- [268] Lichtenthaler, H. K. «Chlorophylls and carotenoids: pigments of photosynthetic biomembranes». En: *Methods in Enzymology*. Ed. por Packer, L. y Douce, R. Cambridge: Academic Press, 1987.
- [269] Schneider, C. A., Rasband, W. S. y Eliceiri, K. W. «NIH Image to ImageJ: 25 years of image analysis». En: *Nature Methods* 9 (2012), pág. 671.

- [270] Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J. y col. «The Pfam protein families database: towards a more sustainable future». En: *Nucleic Acids Research* 44 (2015), pág. D279.
- [271] Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. y Punta, M. «Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions». En: *Nucleic Acids Research* 41 (2013), e121.
- [272] Louime, C., Vasanthaiah, H., Jittayasothorn, Y., Lu, J., Basha, S. y col. «A simple and efficient protocol for high quality RNA extraction and cloning of chalcone synthase partial CDS from muscadine grape cultivars (*Vitis rotundifolia* Michx.)» En: *European Journal of Scientific Research* 22 (2008), pág. 232.
- [273] Walker, S. E. y Lorsch, J. «RNA Purification – Precipitation Methods». En: *Laboratory Methods in Enzymology: RNA*. Ed. por Lorsch, J. Cambridge: Academic Press, 2013.
- [274] Livak, K. J. y Schmittgen, T. D. «Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method». En: *Methods* 25 (2001), pág. 402.
- [275] Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C. y col. «Primer3—new capabilities and interfaces». En: *Nucleic Acids Research* 40 (2012), e115.
- [276] Anders, S. y Huber, W. «Differential expression analysis for sequence count data». En: *Genome Biology* 11 (2010), R106.
- [277] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B. y col. «PANTHER: a library of protein families and subfamilies indexed by function». En: *Genome Research* 13 (2003), pág. 2129.
- [278] Arana, M. V., Gonzalez-Polo, M., Martínez-Meier, A., Gallo, L. A., Benech-Arnold, R. L. y col. «Seed dormancy responses to temperature relate to *Nothofagus* species distribution and determine temporal patterns of germination across altitudes in Patagonia». En: *New Phytologist* 209 (2016), pág. 507.
- [279] Schinelli-Casares, T. *Producción de Nothofagus bajo condiciones controladas*. San Carlos de Bariloche: Ediciones INTA, 2012.

- [280] Derory, J., Léger, P., García, V., Schaeffer, J., Hauser, M.-T. y col. «Transcriptome analysis of bud burst in sessile oak (*Quercus petraea*)». En: *New Phytologist* 170 (2006), pág. 723.
- [281] Tesniere, C. y Vayda, M. E. «Method for the isolation of high-quality RNA from grape berry tissues without contaminating tannins or carbohydrates». En: *Plant Molecular Biology Reporter* 9 (1991), pág. 242.
- [282] Loomis, W. D. «Overcoming problems of phenolics and quinones in the isolation of plant enzymes and organelles». En: *Methods in Enzymology*. Ed. por Fleischer, S. y Packer, L. Cambridge: Academic Press, 1974.
- [283] Zhu, G., Chang, Y., Zuo, J., Dong, X., Zhang, M. y col. «Fudenine, a C-terminal truncated rat homologue of mouse prominin, is blood glucose-regulated and can up-regulate the expression of *GAPDH*». En: *Biochemical and Biophysical Research Communications* 281 (2001), pág. 951.
- [284] Glare, E. M., Divjak, M., Bailey, M. J. y Walters, E. H. « β -Actin and *GAPDH* housekeeping gene expression in asthmatic airways is variable and not suitable for normalising mRNA levels». En: *Thorax* 57 (2002), pág. 765.
- [285] Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N. y col. «Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes». En: *Genome Biology* 3 (2002), research0034.
- [286] Sanders, R., Mason, D. J., Foy, C. A. y Huggett, J. F. «Considerations for accurate gene expression measurement by reverse transcription quantitative PCR when analysing clinical samples». En: *Analytical and Bioanalytical Chemistry* 406 (2014), pág. 6471.
- [287] Huggett, J., Dheda, K., Bustin, S. y Zumla, A. «Real-time RT-PCR normalisation; strategies and considerations». En: *Genes & Immunity* 6 (2005), pág. 279.
- [288] Guerriero, M. L., Pokhilko, A., Fernández, A. P., Halliday, K. J., Millar, A. J. y col. «Stochastic properties of the plant circadian clock». En: *Journal of The Royal Society Interface* 9 (2012), pág. 744.

- [289] Gyllenstrand, N., Karlgren, A., Clapham, D., Holm, K., Hall, A. y col. «No time for spruce: rapid dampening of circadian rhythms in *Picea abies* (L. Karst)». En: *Plant and Cell Physiology* 55 (2014), pág. 535.
- [290] Jizarewa, I., Ibáñez, C., Johansson, M., Ögren, E., Mozley, D. y col. «Alteration of *PHYA* expression change circadian rhythms and timing of bud set in *Populus*». En: *Plant Molecular Biology* 73 (2010), pág. 143.
- [291] Straume, M. «DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning». En: *Methods in Enzymology*. Ed. por Brand, L. y Johnson, M. L. Vol. 383. Cambridge: Academic Press, 2004.
- [292] Yamada, R. y Ueda, H. R. «Microarrays: statistical methods for circadian rhythms». En: *Methods in Molecular Biology*. Ed. por Rosato, E. Totowa: Humana Press, 2007.
- [293] Duffield, G. E., Best, J. D., Meurers, B. H., Bittner, A., Loros, J. J. y col. «Circadian programs of transcriptional activation, signaling, and protein turnover revealed by microarray analysis of mammalian cells». En: *Current Biology* 12 (2002), pág. 551.
- [294] Xu, W., Yang, R., Li, M., Xing, Z., Yang, W. y col. «Transcriptome phase distribution analysis reveals diurnal regulated biological processes and key pathways in rice flag leaves and seedling leaves». En: *PLoS One* 6 (2011), e17613.
- [295] Glynn, E. F., Chen, J. y Mushegian, A. R. «Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms». En: *Bioinformatics* 22 (2006), pág. 310.
- [296] Yang, R. y Su, Z. «Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation». En: *Bioinformatics* 26 (2010), pág. i168.
- [297] Auer, P. L. y Doerge, R. W. «Statistical design and analysis of RNA sequencing data». En: *Genetics* 185 (2010), pág. 405.
- [298] Kvam, V. M., Liu, P. y Si, Y. «A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data». En: *American Journal of Botany* 99 (2012), pág. 248.

- [299] Richter, K., Haslbeck, M. y Buchner, J. «The heat shock response: life on the verge of death». En: *Molecular Cell* 40 (2010), pág. 253.
- [300] Wright, P. E. y Dyson, H. J. «Intrinsically disordered proteins in cellular signalling and regulation». En: *Nature Reviews Molecular Cell Biology* 16 (2015), pág. 18.
- [301] Zhou, M., Kim, J. K., Eng, G. W. L., Forger, D. B. y Virshup, D. M. «A Period2 phosphoswitch regulates and temperature compensates circadian period». En: *Molecular Cell* 60 (2015), pág. 77.
- [302] Chinnusamy, V., Zhu, J. y Zhu, J.-K. «Cold stress regulation of gene expression in plants». En: *Trends in Plant Science* 12 (2007), pág. 444.
- [303] Fowler, S. G., Cook, D. y Thomashow, M. F. «Low temperature induction of *Arabidopsis* CBF1, 2, and 3 is gated by the circadian clock». En: *Plant Physiology* 137 (2005), pág. 961.
- [304] Murayama, Y., Kori, H., Oshima, C., Kondo, T., Iwasaki, H. y col. «Low temperature nullifies the circadian clock in cyanobacteria through Hopf bifurcation». En: *Proceedings of the National Academy of Sciences* 114 (2017), pág. 5641.
- [305] Smith, W. K. y Nobel, P. S. «Influence of irradiation, soil water potential, and leaf temperature on leaf morphology of a desert broadleaf, *Encelia farinosa* Gray (Compositae)». En: *American Journal of Botany* 65 (1978), pág. 429.
- [306] Hanba, Y., Kogami, H. y Terashima, I. «The effect of growth irradiance on leaf anatomy and photosynthesis in *Acer* species differing in light demand». En: *Plant, Cell & Environment* 25 (2002), pág. 1021.
- [307] Ivancich, H. S., Lencinas, M. V., Martínez-Pastur, G. J., Soler-Esteban, R. M., Hernández, L. y col. «Foliar anatomical and morphological variation in *Nothofagus pumilio* seedlings under controlled irradiance and soil moisture levels». En: *Tree Physiology* 32 (2012), pág. 554.
- [308] Chmielewski, F.-M. y Rötzer, T. «Response of tree phenology to climate change across Europe». En: *Agricultural and Forest Meteorology* 108 (2001), pág. 101.

- [309] Martínez-García, P. J., Crepeau, M. W., Puiu, D., Gonzalez-Ibeas, D., Whalen, J. y col. «The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols». En: *The Plant Journal* 87 (2016), pág. 507.
- [310] Yasui, Y., Hirakawa, H., Ueno, M., Matsui, K., Katsube-Tanaka, T. y col. «Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes». En: *DNA Research* 23 (2016), pág. 215.
- [311] Wang, N., Thomson, M., Bodles, W. J. A., Crawford, R. M. M., Hunt, H. V. y col. «Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers». En: *Molecular Ecology* 22 (2013), pág. 3098.
- [312] Plomion, C., Aury, J.-M., Amselem, J., Alaeitabar, T., Barbe, V. y col. «Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies». En: *Molecular Ecology Resources* 16 (2016), pág. 254.
- [313] Bennett, J. M., Calosi, P., Clusella-Trullas, S., Martínez, B., Sunday, J. y col. «GlobTherm, a global database on thermal tolerances for aquatic and terrestrial organisms». En: *Scientific Data* 5 (2018), pág. 180022.
- [314] Food and Agriculture Organization of the United Nations (FAO). *Global guidelines for the restoration of degraded forests and landscapes in drylands. Building resilience and benefiting livelihoods*. Roma: FAO, 2015.
- [315] Wheeler, N. y Sederoff, R. «Role of genomics in the potential restoration of the american chestnut». En: *Tree Genetics & Genomes* 5 (2009), pág. 181.
- [316] Jacobs, D. F., Dalgleish, H. J. y Nelson, C. D. «A conceptual framework for restoration of threatened plants: the effective model of american chestnut (*Castanea dentata*) reintroduction». En: *New Phytologist* 197 (2013), pág. 378.
- [317] Stevens, K. A., Woeste, K., Chakraborty, S., Crepeau, M. W., Leslie, C. A. y col. «Genomic variation among and within six *Juglans* species». En: *G3: Genes, Genomes, Genetics* 8 (2018), pág. 2153.
- [318] Sork, V. L., Fitz-Gibbon, S. T., Puiu, D., Crepeau, M., Gugger, P. F. y col. «First draft assembly and annotation of the genome of a California endemic oak *Quercus lobata* Nee (Fagaceae)». En: *G3: Genes, Genomes, Genetics* 6 (2016), pág. 3485.

- [319] Ramos, A. M., Usié, A., Barbosa, P., Barros, P. M., Capote, T. y col. «The draft genome sequence of cork oak». En: *Scientific Data* 5 (2018), pág. 180069.
- [320] Mishra, B., Gupta, D. K., Pfenninger, M., Hickler, T., Langer, E. y col. «A reference genome of the european beech (*Fagus sylvatica* L.)» En: *GigaScience* 7 (2018), giy063.
- [321] Jia, H.-M., Jia, H.-J., Cai, Q.-L., Wang, Y., Zhao, H.-B. y col. «The red bayberry genome and genetic basis of sex determination». En: *Plant Biotechnology Journal* (2018), Early Online View.
- [322] Salojärvi, J., Smolander, O.-P., Nieminen, K., Rajaraman, S., Safronov, O. y col. «Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch». En: *Nature Genetics* 49 (2017), pág. 904.
- [323] Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G. y col. «Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis». En: *Science* (2018), Early Online View.
- [324] Ye, G., Zhang, H., Chen, B., Nie, S., Liu, H. y col. «De novo genome assembly of the stress tolerant forest species *Casuarina equisetifolia* provides insight into secondary growth». En: *The Plant Journal* (2018), Early Online View.

Glosario

alineamiento Arreglo de secuencias de ADN o proteínas de varios individuos (de la misma o distintas especies) de manera de identificar caracteres homólogos y regiones conservadas. Producido por algoritmos de alineamiento (por ejemplo, Clustal o BLAST). 45

bipartición Producto de cada nodo de un árbol filogenético, que genera una bifurcación. 85

carácter Rasgo utilizado para comparar entre taxones en un análisis filogenético. Ejemplos: longitud de un hueso, presencia o ausencia de estípulas foliares, nucleótido de una secuencia de ADN. 70

cobertura En secuenciación, número total de bases secuenciadas dividido el tamaño del transcriptoma o genoma en estudio. En ensamblado, promedio de lecturas que soportan cada posición nucleotídica a lo largo de un contig. En ambos casos, es una medida del sustento y la representatividad del estudio *-ómico*. Se utiliza frecuentemente el término inglés, *coverage*. 42

contig Abreviatura del inglés *contiguous* (“contiguo”). Secuencia de ADN consensuada a partir de la superposición de lecturas. 37

de novo Del latín , “de nuevo”, “a partir de lo nuevo” (fig. “desde el principio”). Hace referencia a un ensamblado transcriptómico o genómico sin secuencias de referencia. 37

ESTs Del inglés, *Expressed Sequence Tags* (marcador de secuencia expresada). Fragmento de secuencia de ADN transcripta (génica). 37

fitness Del inglés, “aptitud”. Capacidad de un individuo de reproducirse con cierto genotipo. Es proporcional a la cantidad de alelos de dicho individuo en relación

a los alelos totales de la población en la siguiente generación. 30

heterotaquia Derivado del griego, “distinta rapidez”. Señal no filogenética asociada a la distinta tasa de evolución, y por lo tanto distinto modelo evolutivo molecular, entre regiones de un alineamiento. 83

homología Carácter presente en el ancestro común entre los taxones comparados. Se utiliza como indicador confiable de ancestralidad. 45

in silico Simplificación del latín *in silicon*, “en silicio”. Se refiere a un análisis o simulación hecha en una computadora. El nombre alude a los clásicos términos *in vitro* e *in vivo*. 58

indel Abreviatura de “inserción–delección”. Polimorfismo causado por la ausencia de uno o varios nucleótidos en una o varias secuencias (o bien la presencia en las otras). 39

k-mero Fragmento de longitud k de una lectura. El nombre es una generalización de *dímero*, *trímero*, *polímero*, etc. Por ejemplo, se puede hablar de 32-meros o 64-meros. Se utiliza frecuentemente el término inglés, *k-mer*. 38

lectura Fragmento individual de ADN secuenciado. Varía en longitud según el secuenciador utilizado, desde menos de 50 nucleótidos hasta varios miles. Se utiliza frecuentemente la palabra inglesa, *read*. 37

library Del inglés, “biblioteca” (plural *libraries*). Conjunto de secuencias de ADN purificadas y amplificadas para su secuenciación. 42

loci Ver **locus**. 36

locus Del latín, “lugar” (plural *loci*). Posición fija en un cromosoma. Puede contener un gen, un marcador molecular, o cualquier secuencia de interés. 53

mapeo Alineamiento de secuencias cortas (por ejemplo lecturas) contra secuencias de referencia (por ejemplo contigs). 45

microsatélite Secuencia de ADN que contiene repeticiones consecutivas de fragmentos de 2 a 6 nucleótidos. Son utilizados extensamente como marcadores moleculares por ser codominantes y de alta tasa mutacional. También se utiliza la sigla inglesa, SSR *Simple Sequence Repeat*. 47

ortólogos Dos o más genes en especies distintas provenientes de un gen en un ancestro común; es decir, genes homólogos entre especies. 45

pipeline Del inglés, “tubería”. Concatenación en serie de programas de análisis computacional, donde la salida de un paso es la entrada del siguiente. 40

ramas largas Fenómeno producto de comparar en un análisis filogenético taxones muy lejanos. Al haber gran cantidad de cambios entre taxones, las relaciones se visualizan como ramas largas en un árbol filogenético cuyas longitudes de ramas son proporcionales a la cantidad de cambios (filogramas). Lo contrario (*ramas cortas*) sucede cuando se comparan taxones muy cercanos (por ejemplo individuos de una misma población). 73

script Del inglés, “guión”. Porción de código que es interpretado por un programa especial (por ejemplo los intérpretes de Python, PHP o JavaScript) para ser convertido en código fuente, listo para ser ejecutado directamente por una computadora. 45

single-end Del inglés, “único extremo”. Lectura de una única hebra de una molécula de ADN, a partir de uno de sus extremos. Contrariamente, las lecturas *paired-end* (extremos pareados) contienen información de los dos extremos, en hebras complementarias, de una misma molécula de ADN. 42

SNP Del inglés, *Single Nucleotide Polymorphism* (polimorfismo de nucleótido único). Polimorfismo causado por diferencias entre secuencias de un único nucleótido. 39

splicing Del inglés, “empalme”, “juntura”. Proceso post-transcripcional por el cual se eliminan ciertos fragmentos del ARN precursor como parte de la maduración previa a la traducción. 29

SSRs Ver **microsatélite**. 47

topología Representación gráfica de las relaciones filogenéticas entre taxones. 70

wrapper Del inglés, “envoltorio”. En computación, un programa cuya función principal es ejecutar ordenadamente otros programas y controlar sus entradas y salidas. 58

ZT Del alemán, *Zeitgeber* (dador de tiempo, fig. “sincronizador”). Señal ambiental sincronizadora de ritmos biológicos. 41

Apéndice A

Genes conservados para filogenómica

Tabla A.1: Genes CEGs conservados entre especies. Ver explicación en Sección 3.3.1

Proteína	Locus	Descripción	CEG
O22268	AT2G47420	18S nuclear rRNA (adenine(1785)-N(6)/adenine(1786)-N(6))- dimethyltransferase.	KOG0820
Q9LP45	AT1G29150	19S proteasome subunit 9	KOG1463
O23715	AT2G27020	20S proteasome alpha subunit G-1	KOG0184
O81146	AT5G35590	20S proteasome subunit alpha A-1	KOG0182
O81153	AT1G77440	20S proteasome beta subunit C-2	KOG0180
P42742	AT3G60820	20S proteasome beta subunit F-1	KOG0179
Q8L4A7	AT1G79210	20S proteasome alpha subunit B-2	KOG0181
Q94AU2	AT1G11890	25.3 kDa vesicle transport protein	KOG0862
B3H6B0	AT5G09900	26S proteasome non-ATPase regulatory subunit 12 homolog A	KOG1498
O24412	AT5G05780	26S proteasome regulatory subunit RPN8a	KOG1556
Q8RWF0	AT5G45620	26S proteasome regulatory subunit RPN9a	KOG2908
Q93Y35	AT4G24820	26S proteasome regulatory subunit RPN7	KOG0687
Q9LT08	AT5G23540	26S proteasome regulatory subunit RPN11	KOG1555
Q9LF46	AT5G17380	2-hydroxyphytanoyl-CoA lyase	KOG1185
Q8H107	AT4G26910	2-oxoglutarate dehydrogenase complex component E2-2	KOG0559
Q9SP08	AT3G61620	3'-5'-exoribonuclease family protein	KOG1068
P54873	AT4G11820	3-hydroxy-3-methylglutaryl coenzyme A synthase	KOG1393
P59223	AT3G60770	40S ribosomal protein S13-1	KOG0400
O22860	AT2G43460	60S ribosomal protein L38	KOG3499
P57691	AT3G11250	60S acidic ribosomal protein P0-3	KOG0815
Q8L8W6	AT3G02190	60S ribosomal protein L39-2	KOG0002
Q9FF90	AT5G23900	60S ribosomal protein L13-3	KOG3295

continúa en la siguiente página

Genes CEG conservados (continuación)

Proteína	Locus	Descripción	CEG
Q9FWA3	AT3G02360	6-phosphogluconate dehydrogenase, decarboxylating 3	KOG2653
Q9SYA6	AT1G61700	ABC10	KOG3497
Q9M4B5	AT1G08780	ABI3-interacting protein 3	KOG1760
P57750	AT5G18070	Acetylglucosamine phosphomutase	KOG2537
Q9SJW6	AT2G31300	Actin-related protein C1	KOG1523
Q8VXX4	AT5G27740	Activator 1 subunit 3	KOG2035
Q3E8F0	AT5G47720	Active Site:Acyl-thioester intermediate	KOG1390
Q9SUW2	AT4G22570	Adenine phosphoribosyltransferase 3	KOG1712
P36428	AT1G50200	Alanyl-tRNA synthetase	KOG0188
Q9SJT9	AT2G21390	Alpha-coat protein 2	KOG0292
Q683D4	AT5G15750	Alpha-L RNA-binding motif/Ribosomal protein S4 family protein	KOG4655
Q949M9	AT1G01910	Arsenical pump-driving ATPase homolog	KOG2825
Q8H104	AT4G26870	Aspartyl-tRNA synthetase	KOG0556
F4IMB5	AT2G07698	ATP synthase subunit alpha	KOG1353
O04630	AT5G26830	AtSYT1	KOG1637
Q9SV21	AT4G31480	Beta-coat protein 1	KOG1058
Q8L3X9	AT2G04540	Beta-ketoacyl-ACP synthase	KOG1394
Q9M9G7	AT1G71790	CapZ-beta	KOG3174
Q940P8	AT5G20890	CCT-beta	KOG0363
Q9SF16	AT3G11830	CCT-eta	KOG0361
Q9M888	AT3G02530	CCT-zeta 1	KOG0359
Q56YN8	AT2G27170	Chromosome segregation protein SMC-3	KOG0964
Q6Q1P4	AT3G54670	Chromosome segregation protein SMC-1	KOG0018
Q0WLB5	AT3G08530	Clathrin heavy chain 2	KOG0985
Q9LK29	AT3G27240	Complex III subunit 4-1	KOG3052
F4JWF6	AT5G11200	DEAD-box ATP-dependent RNA helicase 56	KOG0329
Q9SW44	AT4G34910	DEAD-box ATP-dependent RNA helicase 16	KOG0346
Q39080	AT1G32210	Defender against cell death 1	KOG1746
B9DGP9	AT3G58180	Deoxyhypusine dioxygenase	KOG0567
Q84WC6	AT3G21280	Deubiquitinating enzyme 7	KOG1872
Q9M5K3	AT1G48030	Dihydrolipoamide dehydrogenase 1	KOG1335
F4JZZ1	AT5G41880	DNA primase	KOG2851
F4KC69	AT5G63960	DNA polymerase	KOG0969
A0A1P8BFF9	AT5G02080	DNA / pantothenate metabolism flavoprotein	KOG2728
F4J5R0	AT3G52090	DNA-directed RNA polymerases II, IV and V subunit 11	KOG4392
O81098	AT3G22320	DNA-directed RNA polymerase II subunit E	KOG3218
Q9SJ96	AT2G04630	DNA-directed RNA polymerases II and V subunit 6B	KOG3405
Q9ZQC0	AT4G07950	DNA-directed RNA polymerase subunit	KOG2906
P42746	AT5G62540	E2 ubiquitin-conjugating enzyme 3	KOG0419
Q9FK29	AT5G50430	E2 ubiquitin-conjugating enzyme 33	KOG0894

continúa en la siguiente página

Genes CEG conservados (*continuación*)

Proteína	Locus	Descripción	CEG
O22854	AT2G43400	Electron transfer flavoprotein-ubiquinone oxidoreductase, mitochondrial	KOG2415
Q9C6I6	AT1G50940	Electron transfer flavoprotein subunit alpha, mitochondrial	KOG3954
Q9LSW8	AT5G43430	Electron transfer flavoprotein subunit beta, mitochondrial	KOG3180
Q9FNM5	AT5G08650	Elongation factor 4 homolog	KOG0462
Q93ZR1	AT5G50320	Elongator component 3	KOG2535
Q9FUG4	AT5G41360	ERCC3 homolog 2	KOG1123
Q9FE78	AT5G05470	eukaryotic translation initiation factor 2 alpha subunit	KOG2916
Q9FI86	AT5G06000	Eukaryotic translation initiation factor 3 RNA-binding subunit	KOG0122
Q9M060	AT3G55620	Eukaryotic translation initiation factor 6-2	KOG3185
Q8RX88	AT4G01320	Farnesylated proteins-converting enzyme 1	KOG2719
Q96252	AT5G47030	F-ATPase delta' subunit	KOG1758
F4K191	AT5G26680	Flap endonuclease 1	KOG2519
O23443	AT4G15940	Fumarylacetoacetate (FAA) hydrolase family	KOG1535
Q9ASU6	AT1G42440	FUNCTIONS IN: molecular_function unknown	KOG1980
Q9FGI9	AT5G47680	FUNCTIONS IN: molecular_function unknown	KOG2967
O49553	AT4G21110	G10 family protein	KOG3404
Q8H0V6	AT1G64550	GCN20-type ATP-binding cassette protein GCN3	KOG0062
O22287	AT2G39770	GDP-mannose pyrophosphorylase 1	KOG1322
Q9ZVN9	AT1G05055	general transcription factor II H2	KOG2807
F4JCU3	AT3G54250	GHMP kinase family protein	KOG2833
Q43727	AT5G35790	Glucose-6-phosphate 1-dehydrogenase 1, chloroplastic	KOG0563
Q9SHS2	AT2G13440	glucose-inhibited division family A protein	KOG2311
Q9LI77	AT3G25660	Glutamyl-tRNA(Gln) amidotransferase subunit A, chloroplastic/mitochondrial	KOG1211
O22145	AT2G45270	Glycoprotease 1	KOG2707
O23514	AT4G16710	glycosyltransferase family protein 28	KOG3349
O04834	AT4G02080	GTP-binding protein SAR1A	KOG0077
O24456	AT1G18080	Guanine nucleotide-binding protein subunit beta-like protein A	KOG0279
C0SVF3	AT3G57150	H/ACA ribonucleoprotein complex subunit 4	KOG2529
Q8GUN2	AT3G56490	HIS triad family protein 3	KOG3275
F4IYF8	AT3G02760	Histidyl-tRNA synthetase	KOG1936
Q8VYP0	AT3G08950	Homolog of the copper chaperone SCO1 member 1	KOG2792
Q96529	AT3G57610	IMP-aspartate ligase	KOG1355
Q8LBZ7	AT3G27380	Iron-sulfur subunit of complex II	KOG3049

continúa en la siguiente página

Genes CEG conservados (continuación)

Proteína	Locus	Descripción	CEG
F4JLM5	AT4G10320	Isoleucyl-tRNA synthetase	KOG0434
Q42553	AT3G02780	Isopentenyl pyrophosphate isomerase II	KOG0142
Q9FJD4	AT5G53480	Karyopherin subunit beta-1	KOG1241
Q9FT93	AT5G08420	KRR-R motif-containing protein 1	KOG2874
P93655	AT5G26860	Lon protease homolog 1, mitochondrial	KOG2004
Q9ZPI1	AT3G11710	Lysyl-tRNA synthetase	KOG1885
Q9SG17	AT3G49320	Metal-dependent protein hydrolase	KOG2948
Q9LPD9	AT1G44900	Minichromosome maintenance protein 2	KOG0477
Q8W3L1	AT4G32360	Mitochondrial ferredoxin reductase	KOG1800
Q9XH48	AT1G61570	Mitochondrial import inner membrane translocase subunit TIM13	KOG1733
Q9LU93	AT3G25980	Mitotic arrest deficient protein 2	KOG3285
Q9ZNW0	AT5G55130	Molybdenum cofactor synthesis protein 3	KOG2017
Q9C723	AT1G55090	NAD(+) synthase	KOG2303
Q9M0Y8	AT4G04910	N-ethylmaleimide-sensitive fusion protein	KOG0741
Q9LEY9	AT5G08180	Nhp2-like protein	KOG3167
O49543	AT5G65720	NIFS homolog 1	KOG1549
O81433	AT4G04080	NifU-like N-terminal domain-containing protein ISU3	KOG3361
Q8L968	AT4G13720	Non-canonical purine NTP pyrophosphatase	KOG3222
Q9SI58	AT2G03820	nonsense-mediated mRNA decay NMD3 family protein	KOG2613
Q9FLX9	AT5G52820	Notchless protein homolog	KOG0271
Q9FZK4	AT1G27310	nuclear transport factor 2A	KOG2104
Q93XX8	AT2G20490	Nucleolar protein 10	KOG3503
P39207	AT4G09320	Nucleoside diphosphate kinase I	KOG0888
Q96251	AT5G13450	Oligomycin sensitivity conferral protein	KOG1662
Q39097	AT5G47880	Omnipotent suppressor protein 1 homolog 1	KOG0688
Q56Y85	AT3G59990	Peptidase M	KOG2775
Q8GYI6	AT1G08750	Peptidase C13 family	KOG1349
A2RVR7	AT2G47020	Peptide chain release factor 1	KOG2726
Q9LYK7	AT5G14520	Pescadillo homolog	KOG2481
Q9SGE9	AT1G72550	Phenylalanyl-tRNA synthetase beta subunit	KOG2472
P34795	AT5G42740	Phosphoglucose isomerase	KOG2446
Q9SAJ4	AT1G79550	Phosphoglycerate kinase	KOG1367
O80840	AT2G45790	Phosphomannomutase	KOG3189
Q56XY2	AT5G22370	P-loop containing nucleoside triphosphate hydrolases superfamily protein	KOG1533
Q8W586	AT4G21800	P-loop containing nucleoside triphosphate hydrolases superfamily protein	KOG1532
F4JUG5	AT4G38780	Pre-mRNA-processing-splicing factor	KOG1795
B9DGX5	AT2G44860	Probable ribosome biogenesis protein RLP24	KOG1723
P57742	AT5G23290	Probable prefoldin subunit 5	KOG3048
Q56Z73	AT3G02090	Probable mitochondrial-processing peptidase subunit beta	KOG0960

continúa en la siguiente página

Genes CEG conservados (continuación)

Proteína	Locus	Descripción	CEG
Q8LCU7	AT3G45770	Probable trans-2-enoyl-CoA reductase, mitochondrial	KOG0025
Q9M223	AT3G60360	Probable U3 small nucleolar RNA-associated protein 11	KOG3237
Q9M7Q7	AT1G07370	Proliferating cellular nuclear antigen 1	KOG1636
F4JIF9	AT4G14800	Proteasome subunit beta type	KOG0177
F4JRY2	AT4G31300	Proteasome subunit beta type	KOG0174
Q08A56	AT4G15900	Protein pleiotropic regulatory locus 1	KOG0285
Q38799	AT5G50850	Protein MACCI-BOU	KOG0524
Q38845	AT1G25490	Protein enhancer of ethylene-response 1	KOG0211
Q39258	AT4G11150	Protein EMBRYO DEFECTIVE 2448	KOG1664
Q42551	AT3G57870	Protein EMBRYO DEFECTIVE 1637	KOG0424
Q84WW2	AT5G24400	Protein EMBRYO DEFECTIVE 2024	KOG3147
Q94IB8	AT4G36480	Protein EMBRYO DEFECTIVE 2779	KOG1358
Q9FKB0	AT5G48870	Protein SUPERSENSITIVE TO ABA AND DROUGHT 1	KOG1775
Q9M876	AT3G02660	Protein EMBRYO DEFECTIVE 2768	KOG2623
Q9SJ20	AT2G21790	Protein DEFECTIVE IN POLLEN DNA DEGRADATION 2	KOG1112
Q9SSM4	AT1G73030	Protein CHARGED MULTIVESICULAR BODY PROTEIN 1B	KOG3232
Q9XGX9	AT3G46560	Protein EMBRYO DEFECTIVE 2474	KOG3479
Q9ZWI7	AT2G23310	Protein RER1C	KOG1688
O04212	AT2G40090	Putative ABC1 protein At2g40090	KOG1235
P52901	AT1G59900	Pyruvate dehydrogenase E1 component subunit alpha-1, mitochondrial	KOG0225
Q84J75	AT5G12210	RAB geranylgeranyl transferase beta subunit 1	KOG0366
Q541X0	AT3G07880	Rho GDP-dissociation inhibitor 1	KOG3205
Q9ZVT7	AT1G03360	ribosomal RNA processing 4	KOG3013
F4JRD3	AT4G12600	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein	KOG3387
Q9FNH2	AT5G06360	Ribosomal protein S8e family protein	KOG3163
F4IH25	AT2G40360	Ribosome biogenesis protein BOP1 homolog	KOG0650
Q9M1A8	AT3G59600	RNA polymerase Rpb8	KOG3400
Q9M9Y6	AT1G06790	RNA polymerase Rpb7 N-terminal domain-containing protein	KOG3297
Q9LVD0	AT5G57280	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	KOG1541
Q84XU2	AT2G42810	Serine/threonine-protein phosphatase 5	KOG0376
P37106	AT1G15310	Signal recognition particle 54 kDa protein 1	KOG0780
F4JVC5	AT4G39280	Similarity:Belongs to the class-II aminoacyl-tRNA synthetase family.	KOG2784
Q8W4G0	AT1G72340	Similarity:Belongs to the eIF-2B alpha/beta/delta subunits family.	KOG1466

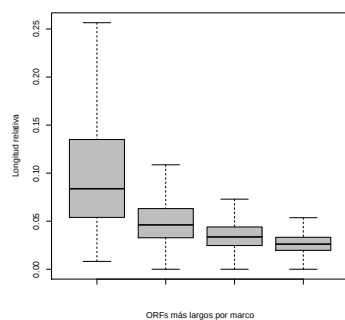
continúa en la siguiente página

Genes CEG conservados (continuación)

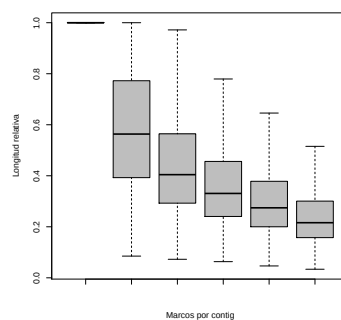
Proteína	Locus	Descripción	CEG
Q944L8	AT1G11930	Similarity:Belongs to the UPF0001 family.	KOG3157
F4KDI1	AT5G58060	Similarity:Contains longin domain.	KOG0861
F4JPK5	AT4G30220	Sm protein F	KOG3482
Q8LAK5	AT4G30330	Small nuclear ribonucleoprotein family protein	KOG1774
Q8RUH0	AT2G47640	Small nuclear ribonucleoprotein family protein	KOG3459
Q9ZUG4	AT2G05830	S-methyl-5-thioribose-1-phosphate isomerase	KOG1468
Q53YT9	AT3G10330	Transcription initiation factor IIB-2	KOG1597
Q9ZUN8	AT2G19540	Transducin family protein / WD-40 repeat family protein	KOG0302
Q9LPP3	AT3G10530	Transducin/WD40 repeat-like superfamily protein	KOG1272
Q67ZW2	AT4G11160	Translation initiation factor 2, small GTP-binding protein	KOG1145
Q8VZK2	AT5G11900	Translation machinery-associated protein 22	KOG3239
Q9FFT6	AT5G54940	Translation initiation factor SUI1 family protein	KOG1770
P31265	AT3G16640	Translationally-controlled tumor protein 1	KOG1727
A8MQD8	AT5G10780	Transmembrane:Helical	KOG3318
F4HQE6	AT1G79990	Transmembrane:Helical	KOG0276
F4K1U5	AT5G54750	Transport protein particle (TRAPP) component	KOG3330
P48491	AT3G55440	Triosephosphate isomerase, cytosolic	KOG1643
Q1H595	AT1G03330	U6 snRNA-associated Sm-like protein LSM2	KOG3448
Q9LVC8	AT5G57300	Ubiquinone biosynthesis methyltransferase COQ5	KOG1540
Q9SVK0	AT4G38930	Ubiquitin fusion degradation UFD1 family protein	KOG1816
P55852	AT4G26840	Ubiquitin-like protein SMT3	KOG1769
Q9M9P3	AT3G03250	UDP-glucose pyrophosphorylase 2	KOG2638
Q9XIK4	AT1G10490	UPF0202 protein At1g10490	KOG2036
O49048	AT1G77140	Vacuolar protein sorting-associated protein 45 homolog	KOG1299
Q9SLA2	AT2G25610	Vacuolar H(+)-ATPase subunit c'' isoform 2	KOG0233
Q9ZQX4	AT4G02620	Vacuolar H(+)-ATPase subunit F	KOG3432
C0SUV0	AT1G12840	V-type proton ATPase subunit C	KOG2909
Q0WWD3	AT3G58730	V-type proton ATPase subunit D	KOG1647
Q949W8	AT5G49650	xylulose kinase-2	KOG2531
Q940S5	AT1G60970	Zeta-1-coat protein	KOG3343
Q9ZQ18	AT2G24500	Zinc finger protein 622	KOG2785
F4K9W6	AT5G22480	ZPR1 zinc-finger domain protein	KOG2703

Apéndice B

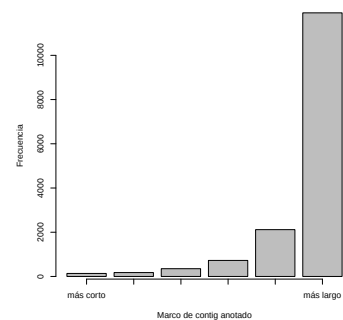
Figuras suplementarias



(a) ORFs dentro de cada marco



(b) Marcos dentro de cada contig



(c) ORFs más largos de cada frame en la anotación

Figura B.1: Justificación del uso del ORF más largo por marco y por contig para la anotación (ensamblado ZT60_20C)

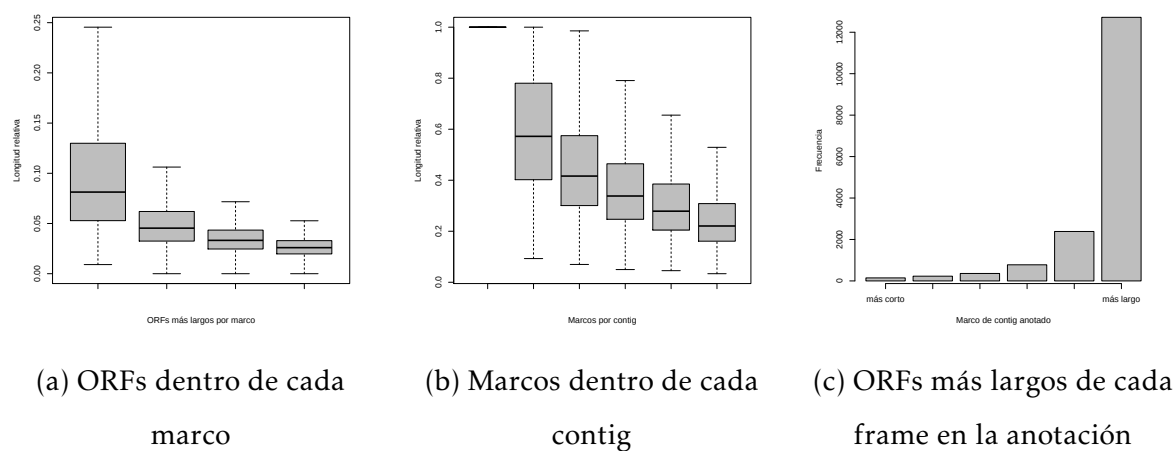


Figura B.2: Justificación del uso del ORF más largo por marco y por contig para la anotación (ensamblado ZT48_34C)

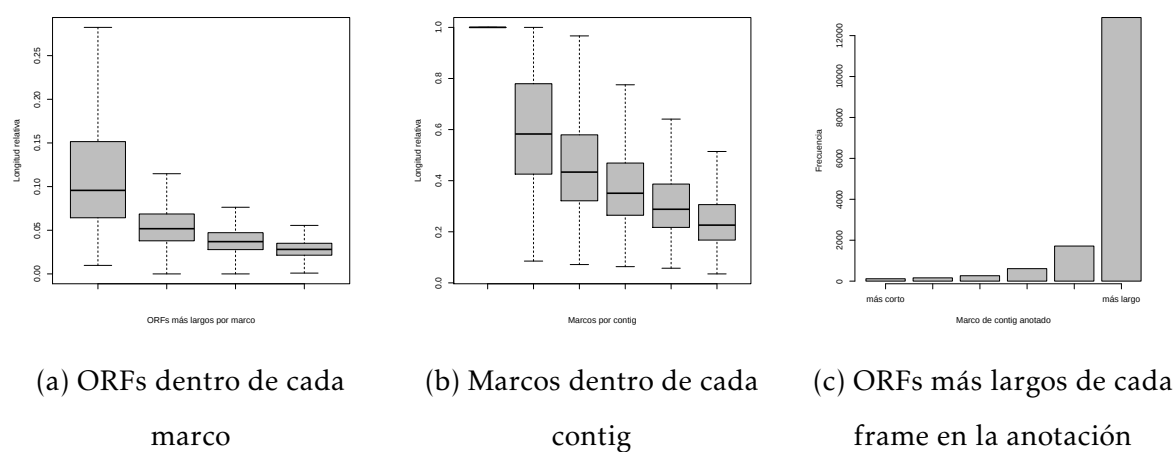


Figura B.3: Justificación del uso del ORF más largo por marco y por contig para la anotación (ensamblado ZT60_34C)

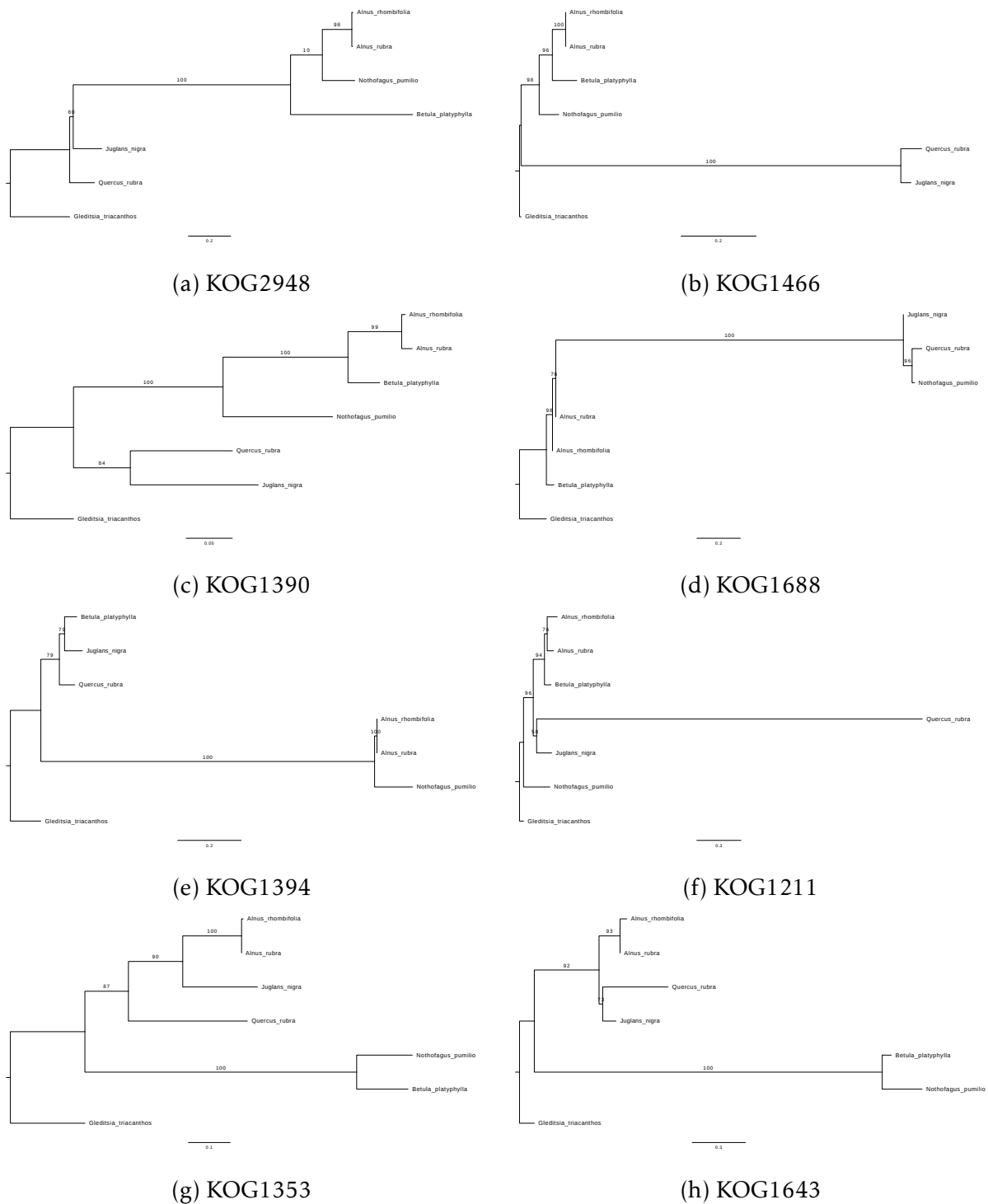


Figura B.4: Árboles filogenéticos de máxima verosimilitud de genes *outliers* que apoyan a la hipótesis T1 (valores positivos de ΔGLS). Se indica el soporte de *bootstrap* de cada nodo

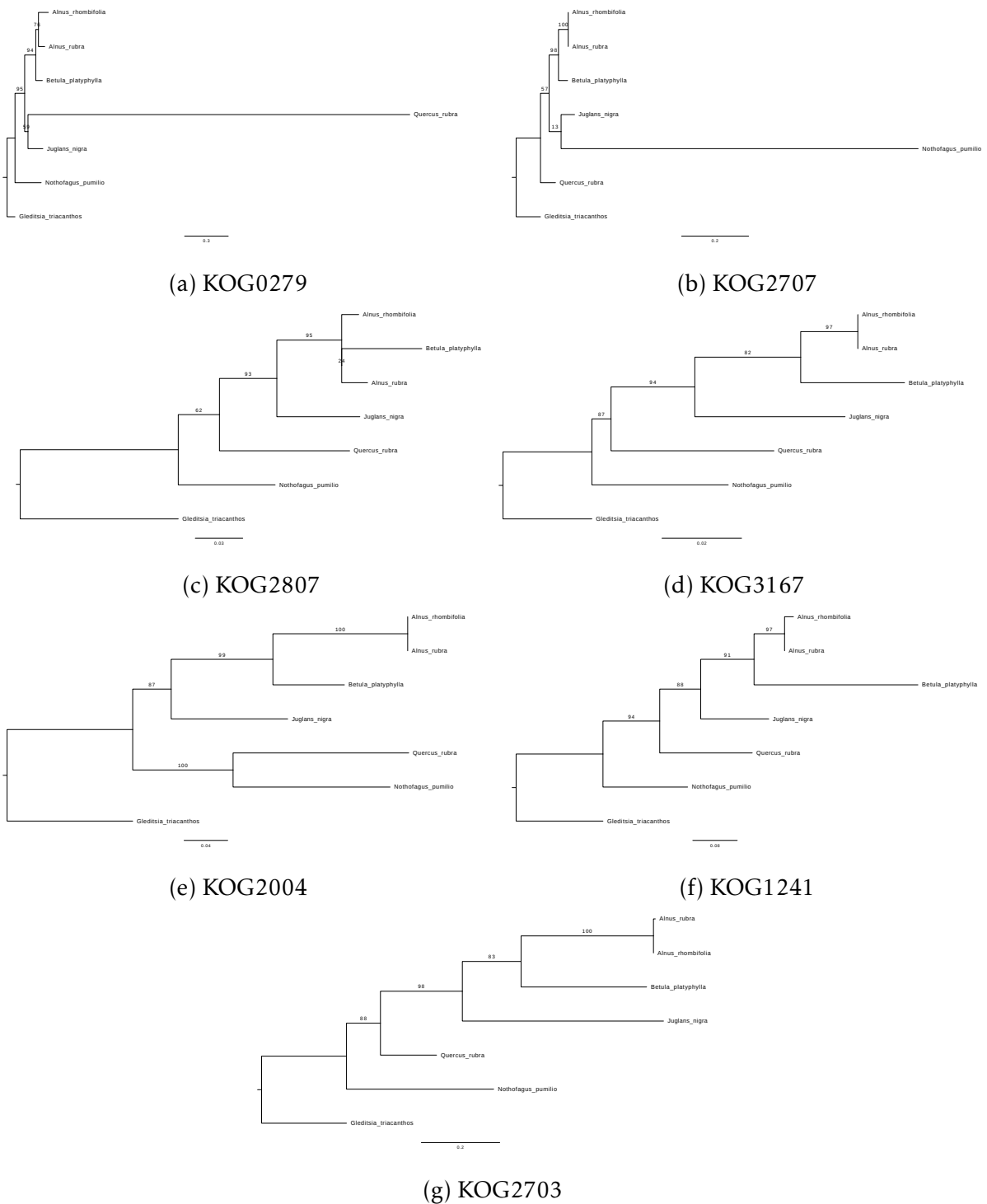


Figura B.5: Árboles filogenéticos de máxima verosimilitud de genes *outliers* que apoyan a la hipótesis T2 (valores negativos de ΔGLS). Se indica el soporte de *bootstrap* de cada nodo

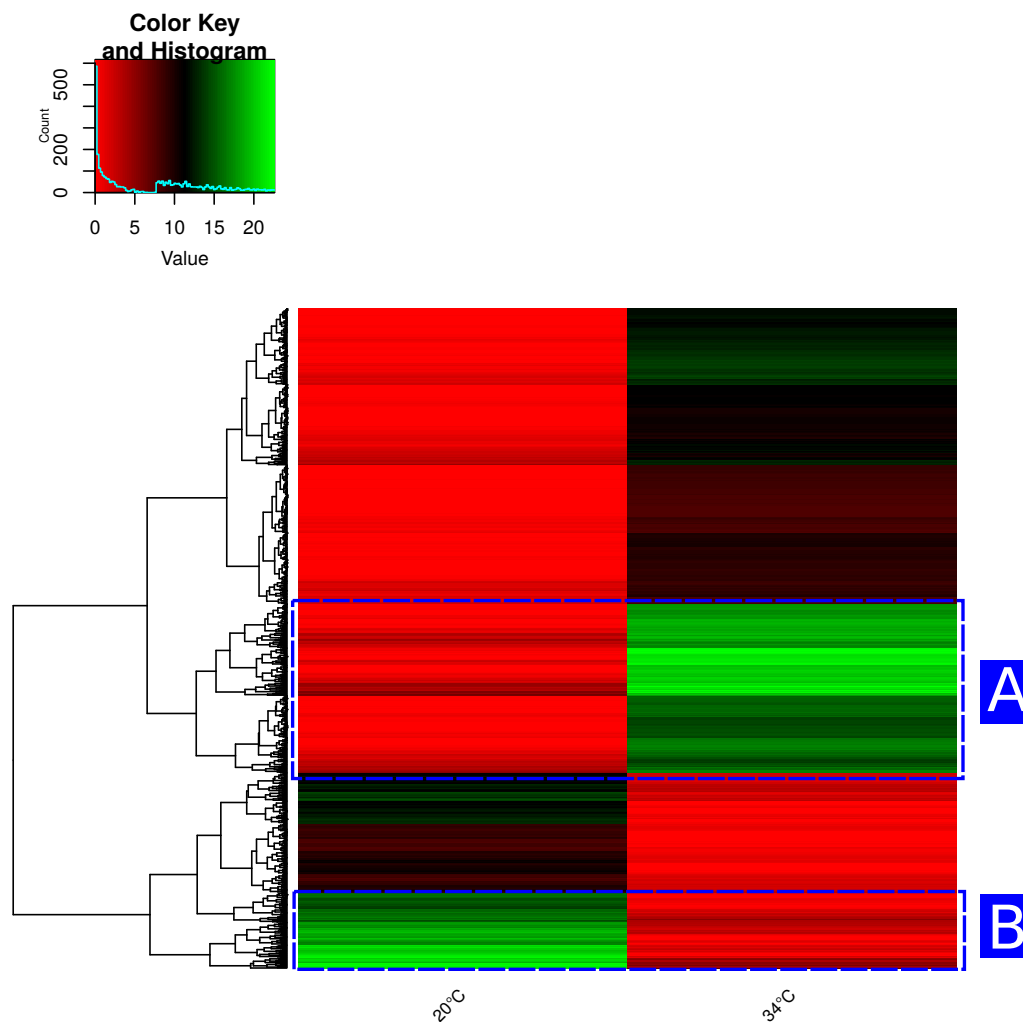


Figura B.6: *Heat-map* para genes diferencialmente expresados según DESeq en condiciones contrastantes de temperatura. Cada columna representa una condición, cada fila un contig, y el color la expresión (clave arriba a la izquierda). Se representan los contigs con un p-valor < 0.05 según DESeq y expresión intermedia (ver Sección 4.3.5)

Apéndice C

Tablas suplementarias

Tabla C.1: Términos de Gene Ontology sobrerrepresentados en genes más expresados a 20°C que a 34°C. Cada flecha representa un nivel de jerarquía superior

Nombre	p-valor
Proceso biológico	
reductive pentose-phosphate cycle	6.85×10^{-4}
⇒⇒ carbon fixation	2.93×10^{-3}
⇒⇒ photosynthesis, dark reaction	1.04×10^{-3}
⇒⇒⇒ photosynthesis	2.47×10^{-9}
glycine metabolic process	1.30×10^{-2}
⇒⇒⇒⇒⇒⇒⇒ small molecule metabolic process	1.71×10^{-2}
glycolytic process	3.72×10^{-2}
⇒⇒ ATP generation from ADP	3.72×10^{-2}
⇒⇒⇒ ADP metabolic process	4.46×10^{-2}
⇒⇒⇒⇒ purine ribonucleoside diphosphate metabolic process	4.46×10^{-2}
⇒⇒⇒⇒⇒ purine nucleoside diphosphate metabolic process	4.46×10^{-2}
⇒⇒⇒⇒⇒ ribonucleoside diphosphate metabolic process	4.46×10^{-2}
⇒⇒⇒ generation of precursor metabolites and energy	1.18×10^{-5}
⇒⇒ pyruvate metabolic process	9.24×10^{-3}
⇒⇒ nicotinamide nucleotide metabolic process	2.62×10^{-2}
⇒⇒⇒ pyridine nucleotide metabolic process	2.99×10^{-2}

continúa en la siguiente página

Términos de GO (*continuación*)

Nombre	p-valor
⇒⇒⇒⇒ pyridine-containing compound metabolic process	4.11×10^{-2}
⇒⇒⇒⇒ oxidoreduction coenzyme metabolic process	4.93×10^{-2}
⇒⇒⇒⇒⇒⇒ cofactor metabolic process	2.79×10^{-3}
hexose metabolic process	4.46×10^{-2}
photosynthesis, light reaction	1.48×10^{-2}
Función molecular	
catalytic activity	2.40×10^{-3}
Componente celular	
chloroplast photosystem I	3.30×10^{-2}
⇒⇒ photosystem I	4.07×10^{-4}
⇒⇒⇒ photosystem	7.46×10^{-5}
⇒⇒⇒⇒ photosynthetic membrane	3.84×10^{-14}
⇒⇒⇒⇒⇒ thylakoid part	2.58×10^{-14}
⇒⇒⇒⇒⇒⇒ thylakoid	6.83×10^{-19}
⇒⇒⇒ chloroplast thylakoid membrane	1.34×10^{-14}
⇒⇒⇒⇒ chloroplast thylakoid	1.05×10^{-16}
⇒⇒⇒⇒⇒ chloroplast part	8.29×10^{-24}
⇒⇒⇒⇒⇒⇒ chloroplast	1.47×10^{-15}
⇒⇒⇒⇒⇒⇒⇒ plastid	3.14×10^{-15}
⇒⇒⇒⇒⇒⇒⇒⇒ cytoplasmic part	1.52×10^{-4}
⇒⇒⇒⇒⇒⇒⇒⇒⇒ cytoplasm	4.96×10^{-2}
⇒⇒⇒⇒⇒⇒⇒ plastid part	1.75×10^{-23}
⇒⇒⇒⇒⇒⇒⇒⇒ intracellular organelle part	6.13×10^{-12}
⇒⇒⇒⇒⇒⇒⇒⇒⇒ organelle part	6.78×10^{-12}
⇒⇒⇒⇒⇒⇒ plastid thylakoid	1.12×10^{-16}
⇒⇒⇒⇒⇒⇒⇒ organelle subcompartment	4.33×10^{-7}
⇒⇒⇒⇒⇒ plastid thylakoid membrane	1.44×10^{-14}
⇒⇒⇒⇒⇒ thylakoid membrane	3.59×10^{-14}
stromule	1.56×10^{-4}

continúa en la siguiente página

Términos de GO (continuación)

Nombre	p-valor
⇒⇒ plastid envelope	8.84×10^{-20}
⇒⇒⇒ organelle envelope	7.58×10^{-18}
⇒⇒⇒⇒ envelope	7.58×10^{-18}
plastoglobule	1.46×10^{-2}
⇒⇒ chloroplast stroma	7.23×10^{-19}
⇒⇒⇒ plastid stroma	1.32×10^{-19}
apoplast	1.16×10^{-7}
chloroplast membrane	4.94×10^{-2}
⇒⇒ chloroplast envelope	4.14×10^{-20}
nucleus	1.08×10^{-5}

Tabla C.2: Términos de Gene Ontology sobrerrepresentados en genes más expresados a 34°C que a 20°C. Cada flecha representa un nivel de jerarquía superior

Nombre	p-valor
Proceso biológico	
response to hydrogen peroxide	6.32×10^{-3}
⇒⇒ response to inorganic substance	4.80×10^{-6}
⇒⇒⇒ response to chemical	1.64×10^{-4}
⇒⇒⇒⇒ response to stimulus	1.20×10^{-3}
⇒⇒ response to reactive oxygen species	1.84×10^{-3}
⇒⇒⇒ response to oxygen-containing compound	1.14×10^{-2}
⇒⇒⇒ response to oxidative stress	5.60×10^{-3}
⇒⇒⇒⇒ response to stress	6.30×10^{-6}
response to heat	3.98×10^{-4}
⇒⇒ response to temperature stimulus	2.89×10^{-2}
⇒⇒⇒ response to abiotic stimulus	1.28×10^{-2}
response to wounding	2.77×10^{-4}

continúa en la siguiente página

Términos de GO (continuación)	
Nombre	p-valor
secondary metabolite biosynthetic process	3.68×10^{-2}
⇒⇒ secondary metabolic process	3.92×10^{-4}
⇒⇒⇒ metabolic process	1.53×10^{-3}
oxidation-reduction process	1.12×10^{-5}
Función molecular	
oxidoreductase activity	2.76×10^{-4}
Componente celular	
cytosolic small ribosomal subunit	1.87×10^{-2}
⇒⇒ small ribosomal subunit	3.28×10^{-2}
⇒⇒⇒⇒ intracellular organelle part	9.41×10^{-4}
⇒⇒⇒⇒⇒ organelle part	1.00×10^{-3}
⇒⇒⇒⇒ cytosol	4.06×10^{-2}

Tabla C.3: Términos de Gene Ontology sobrerrepresentados en genes del grupo A del criterio *ad hoc* de significancia. Cada flecha representa un nivel de jerarquía superior

Nombre	p-valor
Proceso biológico	
-	-
Función molecular	
-	-
Componente celular	
apoplast	7.42×10^{-4}
chloroplast part	3.66×10^{-2}
⇒⇒ chloroplast	2.27×10^{-4}
⇒⇒⇒ plastid	9.19×10^{-5}
⇒⇒ plastid part	4.40×10^{-2}

Tabla C.4: Términos de Gene Ontology sobrerrepresentados en genes del grupo **B** del criterio *ad hoc* de significancia. Cada flecha representa un nivel de jerarquía superior

Nombre	p-valor
Proceso biológico	
porphyrin-containing compound biosynthetic process	1.44×10^{-2}
⇒⇒ porphyrin-containing compound metabolic process	9.92×10^{-4}
⇒⇒⇒ cofactor metabolic process	6.75×10^{-4}
⇒⇒⇒⇒ cellular metabolic process	5.98×10^{-9}
⇒⇒⇒⇒⇒ cellular process	8.56×10^{-9}
⇒⇒⇒⇒⇒ metabolic process	2.17×10^{-9}
⇒⇒⇒ tetrapyrrole metabolic process	1.28×10^{-3}
⇒⇒⇒⇒ organonitrogen compound metabolic process	3.23×10^{-11}
⇒⇒⇒⇒⇒ organic substance metabolic process	1.83×10^{-5}
⇒⇒⇒⇒⇒ nitrogen compound metabolic process	1.19×10^{-5}
⇒⇒ cofactor biosynthetic process	7.27×10^{-3}
⇒⇒ tetrapyrrole biosynthetic process	2.52×10^{-2}
⇒⇒⇒ organonitrogen compound biosynthetic process	2.01×10^{-3}
chlorophyll metabolic process	3.56×10^{-3}
photosynthesis, light reaction	7.45×10^{-3}
⇒⇒ photosynthesis	1.82×10^{-10}
cellular protein complex assembly	2.03×10^{-2}
⇒⇒⇒⇒⇒ cellular component biogenesis	1.09×10^{-2}
⇒⇒ protein complex assembly	4.31×10^{-2}
cellular amino acid metabolic process	1.34×10^{-2}
⇒⇒ primary metabolic process	3.57×10^{-4}
⇒⇒⇒⇒⇒ small molecule metabolic process	3.51×10^{-4}
response to light stimulus	2.83×10^{-2}
⇒⇒ response to radiation	4.50×10^{-2}
⇒⇒⇒ response to abiotic stimulus	7.86×10^{-4}

continúa en la siguiente página

Términos de GO (*continuación*)

Nombre	p-valor
⇒⇒⇒⇒ response to stimulus	2.29×10^{-3}
phosphate-containing compound metabolic process	2.22×10^{-2}
cellular protein metabolic process	1.29×10^{-2}
⇒⇒ protein metabolic process	2.62×10^{-3}
response to stress	3.14×10^{-2}

Función molecular

ATP binding	2.09×10^{-5}
⇒⇒ adenylyl ribonucleotide binding	7.61×10^{-5}
⇒⇒⇒ adenylyl nucleotide binding	7.74×10^{-5}
⇒⇒⇒⇒ purine nucleotide binding	2.40×10^{-5}
⇒⇒⇒⇒⇒ nucleotide binding	1.31×10^{-4}
⇒⇒⇒⇒⇒⇒ nucleoside phosphate binding	1.31×10^{-4}
⇒⇒⇒⇒⇒⇒⇒ heterocyclic compound binding	1.38×10^{-2}
⇒⇒⇒⇒⇒⇒⇒⇒ binding	1.99×10^{-5}
⇒⇒⇒⇒⇒⇒⇒⇒ organic cyclic compound binding	1.57×10^{-2}
⇒⇒⇒⇒⇒⇒⇒ small molecule binding	1.61×10^{-4}
⇒⇒⇒ purine ribonucleotide binding	2.36×10^{-5}
⇒⇒⇒⇒ ribonucleotide binding	3.04×10^{-5}
⇒⇒⇒⇒⇒ carbohydrate derivative binding	4.03×10^{-5}
⇒⇒ purine ribonucleoside triphosphate binding	6.68×10^{-6}
⇒⇒⇒ anion binding	1.09×10^{-5}
⇒⇒⇒⇒ ion binding	1.60×10^{-6}
protein binding	3.73×10^{-2}
catalytic activity	1.23×10^{-7}

Componente celular

chloroplast photosystem II	2.09×10^{-3}
⇒⇒ chloroplast thylakoid membrane protein complex	5.82×10^{-5}
⇒⇒⇒ thylakoid part	2.15×10^{-18}
⇒⇒⇒⇒ thylakoid	3.80×10^{-21}

continúa en la siguiente página

Términos de GO (*continuación*)

Nombre	p-valor
⇒⇒⇒ membrane protein complex	9.88×10^{-4}
⇒⇒⇒⇒⇒ membrane	1.99×10^{-2}
⇒⇒⇒⇒ protein complex	1.80×10^{-2}
⇒⇒⇒⇒⇒ macromolecular complex	5.95×10^{-5}
⇒⇒⇒ chloroplast thylakoid membrane	1.74×10^{-19}
⇒⇒⇒⇒ chloroplast thylakoid	6.46×10^{-21}
⇒⇒⇒⇒⇒ chloroplast part	2.78×10^{-34}
⇒⇒⇒⇒⇒⇒ chloroplast	9.70×10^{-21}
⇒⇒⇒⇒⇒⇒⇒ plastid	8.50×10^{-21}
⇒⇒⇒⇒⇒⇒⇒⇒ cytoplasmic part	6.23×10^{-8}
⇒⇒⇒⇒⇒⇒⇒⇒⇒ cytoplasm	1.59×10^{-7}
⇒⇒⇒⇒⇒⇒⇒ plastid part	9.16×10^{-34}
⇒⇒⇒⇒⇒⇒⇒⇒ intracellular organelle part	1.64×10^{-16}
⇒⇒⇒⇒⇒⇒⇒⇒⇒ organelle part	1.92×10^{-16}
⇒⇒⇒⇒⇒⇒⇒ plastid thylakoid	7.09×10^{-21}
⇒⇒⇒⇒⇒⇒⇒⇒ organelle subcompartment	9.45×10^{-14}
⇒⇒⇒⇒⇒⇒⇒ plastid thylakoid membrane	1.93×10^{-19}
⇒⇒⇒⇒⇒⇒⇒ thylakoid membrane	7.51×10^{-19}
⇒⇒⇒⇒⇒⇒⇒⇒ photosynthetic membrane	8.31×10^{-19}
⇒⇒ photosystem II	8.79×10^{-4}
⇒⇒⇒ photosystem	1.30×10^{-8}
photosystem I	2.46×10^{-5}
plastoglobule	2.17×10^{-10}
⇒⇒ chloroplast stroma	3.54×10^{-23}
⇒⇒⇒ plastid stroma	1.09×10^{-22}
stromule	4.07×10^{-3}
⇒⇒ plastid envelope	1.85×10^{-25}
⇒⇒⇒ organelle envelope	1.72×10^{-17}
⇒⇒⇒⇒ envelope	1.72×10^{-17}

continúa en la siguiente página

Términos de GO (*continuación*)

Nombre	p-valor
chloroplast thylakoid lumen	4.60×10^{-3}
⇒⇒ plastid thylakoid lumen	4.60×10^{-3}
⇒⇒⇒ thylakoid lumen	9.48×10^{-5}
chloroplast membrane	1.10×10^{-4}
⇒⇒ chloroplast envelope	6.05×10^{-26}
⇒⇒ plastid membrane	1.60×10^{-4}
apoplast	8.36×10^{-7}

Tabla C.5: Términos de Gene Ontology sobrerrepresentados en genes del grupo C del criterio *ad hoc* de significancia. Cada flecha representa un nivel de jerarquía superior

Nombre	p-valor
Proceso biológico	
response to jasmonic acid	1.66×10^{-2}
⇒⇒ response to acid chemical	3.78×10^{-4}
⇒⇒⇒ response to chemical	3.56×10^{-5}
⇒⇒⇒⇒ response to stimulus	1.73×10^{-5}
⇒⇒ response to oxygen-containing compound	3.14×10^{-4}
⇒⇒⇒ response to organic substance	2.20×10^{-2}
cellular amide metabolic process	1.80×10^{-2}
response to inorganic substance	1.98×10^{-2}
response to other organism	4.42×10^{-3}
⇒⇒ response to external biotic stimulus	4.55×10^{-3}
⇒⇒⇒ response to external stimulus	4.75×10^{-2}
⇒⇒⇒ response to biotic stimulus	7.94×10^{-3}
⇒⇒ multi-organism process	2.73×10^{-3}
response to abiotic stimulus	7.59×10^{-3}
response to stress	1.91×10^{-2}

continúa en la siguiente página

Términos de GO (<i>continuación</i>)	
Nombre	p-valor
organonitrogen compound metabolic process	1.01×10^{-2}
Función molecular	
-	-
Componente celular	
cytosolic large ribosomal subunit	1.50×10^{-2}
⇒⇒⇒ cytosol	5.39×10^{-6}
⇒⇒⇒⇒ cytoplasmic part	1.03×10^{-3}
⇒⇒⇒⇒⇒ cytoplasm	4.60×10^{-5}
⇒⇒ large ribosomal subunit	4.44×10^{-2}
⇒⇒⇒ ribosomal subunit	1.87×10^{-2}
⇒⇒⇒⇒ intracellular organelle part	3.04×10^{-2}
⇒⇒⇒⇒⇒ organelle part	3.18×10^{-2}
⇒⇒ cytosolic ribosome	4.49×10^{-2}
vacuole	2.88×10^{-2}

Tabla C.6: Términos de Gene Ontology sobrerrepresentados en genes del grupo D del criterio *ad hoc* de significancia. Cada flecha representa un nivel de jerarquía superior

Nombre	p-valor
Proceso biológico	
photosynthesis, light reaction	1.98×10^{-3}
⇒⇒ photosynthesis	2.85×10^{-5}
⇒⇒⇒ cellular metabolic process	4.66×10^{-3}
⇒⇒⇒⇒ cellular process	1.73×10^{-3}
⇒⇒⇒⇒ metabolic process	9.09×10^{-4}
⇒⇒ generation of precursor metabolites and energy	4.02×10^{-4}
plastid organization	1.56×10^{-2}
cofactor metabolic process	8.78×10^{-3}

continúa en la siguiente página

Términos de GO (*continuación*)

Nombre	p-valor
catabolic process	1.40×10^{-2}
Función molecular	
glutathione transferase activity	4.36×10^{-2}
Componente celular	
chloroplast photosystem I	4.17×10^{-2}
⇒⇒ photosystem I	2.00×10^{-2}
⇒⇒⇒ photosystem	1.66×10^{-4}
⇒⇒⇒⇒⇒ membrane	1.51×10^{-2}
⇒⇒⇒⇒ photosynthetic membrane	5.37×10^{-12}
⇒⇒⇒⇒⇒ thylakoid part	2.68×10^{-14}
⇒⇒⇒⇒⇒ thylakoid	2.63×10^{-15}
⇒⇒⇒ chloroplast thylakoid membrane	2.85×10^{-10}
⇒⇒⇒⇒ chloroplast thylakoid	2.92×10^{-12}
⇒⇒⇒⇒⇒ chloroplast part	2.52×10^{-24}
⇒⇒⇒⇒⇒ chloroplast	1.55×10^{-14}
⇒⇒⇒⇒⇒⇒ plastid	3.34×10^{-14}
⇒⇒⇒⇒⇒⇒⇒ cytoplasmic part	3.53×10^{-10}
⇒⇒⇒⇒⇒⇒⇒ cytoplasm	2.93×10^{-7}
⇒⇒⇒⇒⇒⇒ plastid part	5.55×10^{-24}
⇒⇒⇒⇒⇒⇒⇒ intracellular organelle part	4.19×10^{-12}
⇒⇒⇒⇒⇒⇒⇒ organelle part	4.65×10^{-12}
⇒⇒⇒⇒⇒ plastid thylakoid	3.09×10^{-12}
⇒⇒⇒⇒⇒ organelle subcompartment	2.32×10^{-5}
⇒⇒⇒⇒ plastid thylakoid membrane	3.03×10^{-10}
⇒⇒⇒⇒⇒ thylakoid membrane	5.88×10^{-11}
chloroplast thylakoid lumen	3.22×10^{-2}
⇒⇒ plastid thylakoid lumen	3.22×10^{-2}
⇒⇒⇒ thylakoid lumen	4.58×10^{-3}
chloroplast stroma	2.63×10^{-14}

continúa en la siguiente página

Términos de GO (*continuación*)

Nombre	p-valor
⇒⇒ plastid stroma	5.68×10^{-15}
chloroplast membrane	1.44×10^{-2}
⇒⇒ chloroplast envelope	2.12×10^{-15}
⇒⇒⇒ plastid envelope	4.13×10^{-15}
⇒⇒⇒⇒ organelle envelope	3.44×10^{-10}
⇒⇒⇒⇒⇒ envelope	3.44×10^{-10}
⇒⇒ plastid membrane	1.81×10^{-2}
apoplast	1.27×10^{-3}
vacuole	5.08×10^{-3}
whole membrane	1.73×10^{-2}
cytosol	2.67×10^{-3}

Tabla C.7: Términos de Gene Ontology sobrerrepresentados en genes del grupo A de la Figura B.6. Cada flecha representa un nivel de jerarquía superior

Nombre	p-valor
Proceso biológico	
response to salt stress	1.19×10^{-2}
⇒⇒response to osmotic stress	8.44×10^{-3}
⇒⇒⇒response to stress	5.20×10^{-6}
⇒⇒⇒⇒response to stimulus	2.58×10^{-3}
⇒⇒⇒response to abiotic stimulus	2.88×10^{-2}
response to drug	1.25×10^{-2}
⇒⇒response to chemical	2.55×10^{-2}
metabolic process	4.07×10^{-3}
cellular process	3.51×10^{-2}
Función molecular	

continúa en la siguiente página

Términos de GO (continuación)	
Nombre	p-valor
cofactor binding	2.70×10^{-2}
oxidoreductase activity	3.06×10^{-2}
⇒⇒catalytic activity	4.24×10^{-2}
Componente celular	
cell wall	8.79×10^{-5}
⇒⇒external encapsulating structure	1.63×10^{-5}
⇒⇒⇒cell periphery	3.25×10^{-3}
vacuolar membrane	5.53×10^{-3}
⇒⇒whole membrane	7.81×10^{-3}
⇒⇒vacuolar part	5.81×10^{-3}
⇒⇒⇒intracellular organelle part	5.72×10^{-4}
⇒⇒⇒⇒organelle part	5.81×10^{-4}
⇒⇒⇒vacuole	2.64×10^{-5}
⇒⇒⇒⇒cytoplasmic part	9.28×10^{-6}
⇒⇒⇒⇒⇒cytoplasm	6.62×10^{-5}
⇒⇒bounding membrane of organelle	1.49×10^{-2}
cytosol	3.75×10^{-3}

Tabla C.8: Términos de Gene Ontology sobrerrepresentados en genes del grupo **B** de la Figura B.6. Cada flecha representa un nivel de jerarquía superior

Nombre	p-valor
Proceso biológico	
regulation of photosynthesis, light reaction	2.94×10^{-2}
photosynthesis, light harvesting in photosystem I	2.94×10^{-2}
⇒⇒⇒generation of precursor metabolites and energy	2.56×10^{-3}
⇒⇒⇒photosynthesis, light reaction	1.55×10^{-4}
⇒⇒⇒⇒photosynthesis	7.93×10^{-5}

continúa en la siguiente página

Términos de GO (*continuación*)

Nombre	p-valor
Función molecular	
-	-
Componente celular	
photosystem I	5.24×10^{-4}
⇒⇒photosystem	1.21×10^{-7}
⇒⇒⇒membrane protein complex	6.89×10^{-4}
⇒⇒⇒⇒membrane	1.32×10^{-4}
⇒⇒⇒photosynthetic membrane	1.90×10^{-9}
⇒⇒⇒⇒thylakoid part	5.54×10^{-9}
⇒⇒⇒⇒⇒thylakoid	9.36×10^{-12}
chloroplast thylakoid membrane protein complex	1.85×10^{-2}
⇒⇒chloroplast thylakoid membrane	1.07×10^{-9}
⇒⇒⇒chloroplast thylakoid	7.83×10^{-10}
⇒⇒⇒⇒chloroplast part	1.47×10^{-9}
⇒⇒⇒⇒⇒chloroplast	1.14×10^{-5}
⇒⇒⇒⇒⇒⇒plastid	1.36×10^{-5}
⇒⇒⇒⇒⇒⇒⇒cytoplasmic part	2.81×10^{-3}
⇒⇒⇒⇒⇒⇒⇒⇒cytoplasm	3.82×10^{-3}
⇒⇒⇒⇒⇒⇒plastid part	1.98×10^{-9}
⇒⇒⇒⇒⇒⇒⇒intracellular organelle part	3.55×10^{-3}
⇒⇒⇒⇒⇒⇒⇒⇒organelle part	3.61×10^{-3}
⇒⇒⇒⇒⇒plastid thylakoid	8.01×10^{-10}
⇒⇒⇒⇒⇒organelle subcompartment	3.16×10^{-6}
⇒⇒⇒⇒plastid thylakoid membrane	1.10×10^{-9}
⇒⇒⇒⇒thylakoid membrane	1.85×10^{-9}
photosystem II	2.96×10^{-3}
plastoglobule	6.19×10^{-3}
⇒⇒chloroplast stroma	1.43×10^{-5}
⇒⇒⇒plastid stroma	1.89×10^{-5}

continúa en la siguiente página

Términos de GO (*continuación*)

Nombre	p-valor
chloroplast envelope	7.41×10^{-4}
$\Rightarrow \Rightarrow$ plastid envelope	9.29×10^{-4}
nucleus	3.72×10^{-2}

Apéndice D

Descargos de derechos de autor de Figuras

Figura 1.1: Reimpresión con permiso de la AAAS para fines científicos no comerciales.

Figura 1.2: Copyright Centre for Development and Environment (CDE), Swiss Agency for Development and Cooperation (SDC) and Geographica Bernensia, 2014. Reimpresión con fines científicos no comerciales.

Figura 1.3: Licencia Creative Commons BY-NC-SA. Reimpresión con fines científicos no comerciales.

Figura 1.4: Copyright John Wiley & Sons, Ltd, 2014. Reimpresión con fines científicos no comerciales.

Figura 1.5: Copyright John Wiley & Sons, Ltd, 2014. Reimpresión con fines científicos no comerciales.

Figura 1.6: Licencia Creative Commons BY-NC-SA. Reimpresión con fines científicos no comerciales.

Figura 1.7: Licencia Creative Commons BY-NC-SA. Reimpresión con fines científicos no comerciales.

Figura 1.8: Licencia Creative Commons BY-NC-SA. Reimpresión con fines científicos no comerciales.

Figura 1.9: Copyright Elsevier, Ltd, 2013. Reimpresión con fines científicos no comerciales.

Figura 2.1: Reimpresión con permiso de la AAAS para fines científicos no comerciales.

les.

Figura 2.2: Licencia Creative Commons BY. Reimpresión con fines científicos no comerciales.

Figura 4.27, panel 4.27a: Copyright Springer Science+Business Media B.V., 2010. Reimpresión con fines científicos no comerciales.

Figura 4.27, panel 4.27b: Reimpresión con permiso de la Oxford University Press y la Japanese Society of Plant Physiologists para fines científicos no comerciales.

Figura 4.27, panel 4.27c: Reimpresión con permiso de Springer Nature Publishing AG para fines científicos no comerciales.

Figura 4.28, panel 4.28a: Copyright American Society of Plant Biologists, 2006. Reimpresión con fines científicos no comerciales.

Figura 4.28, panel 4.28b: Copyright National Academy of Sciences, 2017. Reimpresión con fines científicos no comerciales.

Figura 4.28, panel 4.28c: Copyright National Academy of Sciences, 2005. Reimpresión con fines científicos no comerciales.