

El olivo es una planta de gran relevancia tanto social como económica: no solo es una de las especies productoras de aceite más importantes del área mediterránea, sino que también tiene trascendencia en la salud pública por sus propiedades saludables y su implicación en las alergias. A pesar de ello, se la ha estudiado poco desde los puntos de vista genético y genómico, por lo que seguimos sin conocer muchos aspectos de su biología, sobre todo en el ámbito reproductivo. En el presente trabajo se ha generado un transcriptoma reproductivo del olivo que se ofrece a la comunidad como la base de datos ReprOlive, de gran facilidad de uso. Gracias a ello, hemos obtenido nuevos genes marcadores para análisis de expresión génica en estos tejidos, hemos profundizado en la capacidad alergénica de su polen y hemos evaluado la propensión de las proteínas derivadas de estos transcritos a sufrir determinadas modificaciones postraduccionales. Confiamos en que toda esta información permita desarrollar futuros proyectos de investigación donde se diseñen nuevas estrategias que mejoren tanto la producción de aceitunas y aceite, como su calidad. Así mismo, sabemos que está permitiendo conocer mejor la contribución del olivo a las alergias y mejorar tanto el diagnóstico como el tratamiento de esta extendida enfermedad.

Rosario María  
Carmona MuñozEstudio bioinformático del transcriptoma reproductivo  
de olivo (*Olea europaea* L.) y aplicacionesTESIS DOCTORAL  
2021

# Estudio bioinformático del transcriptoma reproductivo de olivo (*Olea europaea* L.) y aplicaciones

*Rosario María Carmona Muñoz*

Dirigida por:

*Dr. M. Gonzalo Claros Díaz*  
*Dr. Juan de Dios Alché Ramírez*

Programa de Doctorado de Biología Celular y Molecular  
Facultad de Ciencias  
UNIVERSIDAD DE MÁLAGA

UNIVERSIDAD  
DE MÁLAGA

TESIS DOCTORAL

# Estudio bioinformático del transcriptoma reproductivo de olivo (*Olea europaea* L.) y aplicaciones



UNIVERSIDAD DE MÁLAGA



**CSIC**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

ROSARIO M<sup>a</sup> CARMONA MUÑOZ

DIRIGIDA POR:

Dr. M. Gonzalo Claros Díaz

Dr. Juan de Dios Alché Ramírez

Programa de Doctorado de Biología Celular y Molecular  
Facultad de Ciencias

UNIVERSIDAD DE MÁLAGA

Plataforma Andaluza de Bioinformática  
Departamento de Bioquímica, Biología Celular y Molecular de Plantas  
de la Estación Experimental del Zaidín  
Consejo Superior de Investigaciones Científicas

Málaga, Mayo de 2021





UNIVERSIDAD  
DE MÁLAGA

AUTOR: Rosario María Carmona Muñoz

 <https://orcid.org/0000-0002-4235-2886>

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga



Esta obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional:

<http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

Cualquier parte de esta obra se puede reproducir sin autorización  
pero con el reconocimiento y atribución de los autores.

No se puede hacer uso comercial de la obra y no se puede alterar, transformar o hacer obras derivadas.

Esta Tesis Doctoral está depositada en el Repositorio Institucional de la Universidad de Málaga (RIUMA): [riuma.uma.es](http://riuma.uma.es)



UNIVERSIDAD  
DE MÁLAGA



Escuela de Doctorado

## DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

Dña. ROSARIO MARÍA CARMONA MUÑOZ, estudiante del programa de doctorado del Programa de Biología Celular y Molecular de la Universidad de Málaga, autora de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada:

### **"ESTUDIO BIOINFORMÁTICO DEL TRANSCRIPTOMA REPRODUCTIVO DE OLIVO (*Olea europaea* L.) Y APLICACIONES"**

Realizada bajo la tutorización de D. Manuel Gonzalo Claros Díaz y dirección de D. Manuel Gonzalo Claros Díaz y D. Juan de Dios Alché Ramírez.

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 21 de Mayo de 2021

|                               |                  |
|-------------------------------|------------------|
| Fdo.:<br>Doctorando/a         | Fdo.:<br>Tutor/a |
| Fdo.:<br>Director/es de tesis |                  |



# Estudio bioinformático del transcriptoma reproductivo de olivo (*Olea europaea* L.) y aplicaciones

Memoria presentada por:

**Rosario M<sup>a</sup> Carmona Muñoz**

Para optar al grado de Doctora por la Universidad de Málaga

Tesis realizada dentro del Programa de Doctorado de Biología Celular y Molecular de la Facultad de Ciencias de la Universidad de Málaga, bajo la dirección del Dr. M. Gonzalo Claros Díaz y el Dr. Juan de Dios Alché Ramírez en la Plataforma Andaluza de Bioinformática y el grupo de Biología Reproductiva y Microscopía Avanzada de Plantas del Departamento de Bioquímica, Biología Celular y Molecular de Plantas de la Estación Experimental del Zaidín (CSIC)

Fdo. Rosario M<sup>a</sup> Carmona Muñoz

V<sup>o</sup>.B<sup>o</sup>. CODIRECTORES DE LA TESIS DOCTORAL:

Fdo. M. Gonzalo Claros Díaz

Fdo. Juan de Dios Alché Ramírez

Málaga, 21 de Mayo de 2021



D. M. GONZALO CLAROS DÍAZ, catedrático del departamento de Biología Molecular y Bioquímica de la Universidad de Málaga.

INFORMA:

Que doña ROSARIO M.<sup>a</sup> CARMONA MUÑOZ, licenciada en biología, ha realizado bajo mi dirección, en la Plataforma Andaluza de Bioinformática, el trabajo de investigación recogido en la presente memoria de tesis doctoral que lleva por título: «Estudio bioinformático del transcriptoma reproductivo de olivo (*Olea europaea* L.) y aplicaciones».

Tras la revisión de la presente memoria, he estimado oportuna su presentación ante la Comisión de Evaluación correspondiente, por lo que autorizo su exposición y defensa para optar al grado de doctora.

Y para que así conste, en cumplimiento de las disposiciones legales vigentes, firmo el presente informe.

Málaga, 21 de Mayo de 2021

Codirector de la tesis,

Dr. M. Gonzalo Claros Díaz

D. JUAN DE DIOS ALCHÉ RAMÍREZ, investigador científico del CSIC de la Estación Experimental del Zaidín (CSIC).

INFORMA:

Que doña ROSARIO M.<sup>a</sup> CARMONA MUÑOZ, licenciada en biología, ha realizado bajo mi dirección, en el grupo de Biología Reproductiva y Microscopía Avanzada de Plantas, el trabajo de investigación recogido en la presente memoria de tesis doctoral que lleva por título: «Estudio bioinformático del transcriptoma reproductivo de olivo (*Olea europaea* L.) y aplicaciones».

Tras la revisión de la presente memoria, he estimado oportuna su presentación ante la Comisión de Evaluación correspondiente, por lo que autorizo su exposición y defensa para optar al grado de doctora.

Y para que así conste, en cumplimiento de las disposiciones legales vigentes, firmo el presente informe.

Granada, 21 de Mayo de 2021

Codirector de la tesis,

Dr. Juan de Dios Alché Ramírez





# Proyectos de investigación

- Optimización de la polinización, cuajado y desarrollo del fruto en sandías triploides (2013 a 2015; Proyecto RECUPERA 2020 - 3.1.4. Ministerio de Economía y Competitividad. VII Plan Estatal de Investigación Científica y Técnica y de Innovación. Programa operativo Fondo Tecnológico) IP: Juan de Dios Alché Ramírez.
- Papel de las especies de oxígeno reactivo (ROS) y el NO en la biología reproductiva del olivo (2012 a 2014; Plan Nacional, BFU2011-22779; Cofinanciación fondos Feder) IP: Juan de Dios Alché Ramírez.
- Proteínas y Enzimas de harinas optimizadas de semilla de olivo para uso en alimentación de organismos marinos (PROTEOLIMAR) (2015 a 2018; Programa Retos-Colaboración 2015. Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, RTC-2015-4181-2) IP: Juan de Dios Alché Ramírez.
- Caracterización de proteínas de almacenamiento en la semilla del olivo y en subproductos de la elaboración del aceite (2013 a 2017; Proyecto de Excelencia, Junta de Andalucía, P2011-CVI-7487; Cofinanciación fondos Feder). IP: Juan de Dios Alché Ramírez.
- Obtención y caracterización de proteínas y enzimas de interés industrial procedentes de pulpa de aceituna deshuesada y deshidratada (PROENADD) (2016 a 2017; Plan Estatal, Retos Colaboración, RTC-2016-4824-2; Cofinanciación fondos Feder). IP: Juan de Dios Alché Ramírez.
- Caracterización y análisis funcional de los enzimas responsables del metabolismo del glutatión en la biología reproductiva del olivo (*Olea europaea* L.) (2017 a 2019; Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Subprograma Estatal de generación del conocimiento 2016, BFU-2016-77243-P) IP: Juan de Dios Alché y José Carlos Jiménez López.
- Germinados de semillas de olivo como fuente alimentaria, agronómica y ornamental (GERMOLIVA) (2018 a 2021; Programa Retos-Colaboración 2017. Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, RTC-2017-6654-2) IP: Juan de Dios Alché Ramírez.
- Estancia colaborativa para la extensión y comercialización de resultados obtenidos a través de la realización de una Tesis en Empresa (2020 a 2021; convocatoria 2017 de ayudas

a actividades de transferencia de conocimiento entre los agentes del sistema andaluz del conocimiento y el tejido productivo Junta de Andalucía. Referencia CA8313, AT17\_5247) IP: Juan de Dios Alché Ramírez.

- Identificación y caracterización de componentes de interés agroalimentario y carácter saludable en la semilla del olivo (2020 a 2023; convocatoria 2018 de ayudas a proyectos de I+D+i Plan Andaluz de Investigación, Desarrollo e Innovación, PAIDI 2020, Junta de Andalucía. Ref. P18-RT-1577) IP: Juan de Dios Alché Ramírez y Antonio Jesús Castro López.

# Publicaciones

## Artículos y capítulos de libro

- Carmona R., Zafra A., Seoane P., Castro A. J., Guerrero-Fernández D., Castillo-Castillo T., Medina-García A., Cánovas F. M., Aldana-Montes J., Navas-Delgado I., Alché J.D., Claros M. G. (2015). ReprOlive: a Database with Linked-Data for the Olive Tree (*Olea europaea* L.) Reproductive Transcriptome. Front Plant Sci 6:625. <http://dx.doi.org/10.3389/fpls.2015.00625> (PUBLICACIÓN AVAL, apéndice F.1)
- Seoane P., Carmona R., Bautista R., Guerrero-Fernández D., Claros M. G. (2014). AutoFlow: an easy way to build workflows. Proceedings IWBBIO 2014 International Work-Conference on Bioinformatics and Biomedical Engineering. Copicentro Granada S.L. Volumen 1, 342-349. (ISBN 978-84-15814-84-9 GR 738/2014, apéndice F.2)
- Jiménez-Quesada M. J., Traverso J. A., Zafra A., Jiménez-López J. C., Carmona R., Claros M. G., Alché J. D. (2015). Identification and *in silico* analysis of NADPH oxidase homologues involved in allergy from an olive pollen transcriptome. Bioinformatics and Biomedical Engineering Third International Conference, IWBBIO 2015 Granada, Spain, April 15–17, 2015 Proceedings, Part I. LNBI 9043. pp. 450-459. Springer (ISBN 978-3-319-16482-3) [http://dx.doi.org/10.1007/978-3-319-16483-0\\_44](http://dx.doi.org/10.1007/978-3-319-16483-0_44) (apéndice F.3)
- Zafra A., Carmona R., Jiménez-López J. C., Pulido. A., Claros M. G., Alché J. D. (2015). Identification of distinctive variants of the olive pollen allergen *Ole e 5* (Cu,Zn Superoxide Dismutase) throughout the analysis of the olive pollen transcriptome. Bioinformatics and Biomedical Engineering Third International Conference, IWBBIO 2015 Granada, Spain, April 15–17, 2015 Proceedings, Part I. LNBI 9043. pp. 460-470. Springer (ISBN 978-3-319-16482-3) [http://dx.doi.org/10.1007/978-3-319-16483-0\\_45](http://dx.doi.org/10.1007/978-3-319-16483-0_45) (apéndice F.4)
- Carmona R., Zafra A., Jiménez-López J. C., Seoane P., Alché J. D., Claros M. G. (2015). In the search of non described olive tree allergens by transcriptome-mining. Conference: 44th Annual Meeting of the European-Histamine-Research-Society (EHRS), Inflammation Research, Volume: 64 (Suppl 1):S1–S50. <https://doi.org/10.1007/s00011-015-0842-9> (apéndice F.5)

- Zafra A., Jiménez-López J. C., Carmona R., Claros M. G., Alché J. D. (2016). Transcriptome-based identification of a seed olive legumin (11S Globulin). Characterization of subunits, 3D modelling and molecular assessment of allergenicity. Bioinformatics and Biomedical Engineering Fourth International Conference, IWBBIO 2016 Granada, Spain, April 20–22, 2016 Proceedings, LNBI 9656. pp. 59-70. Springer (ISBN 978-3-319-31743-4) [http://dx.doi.org/10.1007/978-3-319-31744-1\\_6](http://dx.doi.org/10.1007/978-3-319-31744-1_6) (apéndice F.6)
- Carmona R., Seoane P., Zafra A., Jiménez-Quesada M. J., Alché J. D., Claros M. G. (2016). Automatic workflow for the identification of constitutively-expressed genes based on mapped NGS reads. Bioinformatics and Biomedical Engineering Fourth International Conference, IWBBIO 2016 Granada, Spain, April 20–22, 2016 Proceedings, LNBI 9656. pp. 403-414. Springer (ISBN 978-3-319-31743-4) [http://dx.doi.org/10.1007/978-3-319-31744-1\\_36](http://dx.doi.org/10.1007/978-3-319-31744-1_36) (apéndice F.7)
- Seoane P., Ocaña S., Carmona R., Bautista R., Madrid E., Torres A. M., Claros M. G. (2016). AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised de novo Transcriptome for a Non-Model Species, such as Faba Bean (*Vicia faba*). Current Bioinformatics, Volume 11(4): 440-450. <http://dx.doi.org/10.2174/1574893611666160212235117> (apéndice F.8)
- García-Quirós E., Carmona R., Zafra A., Claros M. G., Alché J. D. (2017). Identification and *in silico* Analysis of Glutathione Reductase Transcripts Expressed in Olive (*Olea europaea* L.) Pollen and Pistil. Bioinformatics and Biomedical Engineering Fifth International Conference, IWBBIO 2017 Granada, Spain, April 26–28, 2017 Proceedings, Part II, LNBI 10209. pp. 185-195. Springer (ISBN 978-3-319-56153-0) [http://dx.doi.org/10.1007/978-3-319-56154-7\\_18](http://dx.doi.org/10.1007/978-3-319-56154-7_18) (apéndice F.9)
- Carmona R., Arroyo M., Jiménez-Quesada M. J., Seoane P., Zafra A., Larrosa R., Alché J. D., Claros M. G. (2017). Automated identification of reference genes based on RNA-seq data. BioMedical Engineering OnLine, 16(Suppl 1):65. <http://dx.doi.org/10.1186/s12938-017-0356-5> (apéndice F.10)
- Zafra A., Carmona R., Traverso J. A., Hancock J. T., Goldman M. H. S., Claros M. G., Hiscock S. J., Alché J. D. (2017). Identification and Functional Annotation of Genes Differentially Expressed in the Reproductive Tissues of the Olive Tree (*Olea europaea* L.) through the Generation of Subtractive Libraries. Front. Plant Sci. 8:1576. <http://dx.doi.org/10.3389/fpls.2017.01576> (apéndice F.11)
- Jiménez-Quesada M. J., Carmona R., Lima-Cabello E., Traverso J. A., Castro A. J., Claros M. G., Alché J. D. (2017). Generation of nitric oxide by olive (*Olea europaea* L.) pollen during in vitro germination and assessment of the S-nitroso- and nitro-proteomes by computational predictive methods. Nitric Oxide, 68, 23-37. <http://dx.doi.org/10.1016/j.niox.2017.06.005> (apéndice F.12)
- Carmona R., Jiménez-Quesada M. J., Lima-Cabello E., Traverso J. A., Castro A. J., Claros M. G., Alché J. D. (2017). S-nitroso- and nitro- proteomes in the olive (*Olea europaea*

L.) pollen. Predictive versus experimental data by nano-LC-MS. Data in Brief, Dec; 15: 474–477. <https://doi.org/10.1016/j.dib.2017.09.058> (apéndice F.13)

- Carmona R., Claros M. G., Alché J. D. (2018). Bioinformatic prediction of *S*-nitrosylation sites in large protein datasets. Methods in molecular biology (Clifton, N.J.). 1747:241-250. [http://dx.doi.org/10.1007/978-1-4939-7695-9\\_19](http://dx.doi.org/10.1007/978-1-4939-7695-9_19) (apéndice F.14)
- Seoane P., Espigares M., Carmona R., Polonio A., Quintana J., Cretazzo E., Bota J., Pérez-García A., Alché J. D., Gómez L., Claros M. G. (2018). TransFlow: A modular framework for assembling and assessing accurate de novo transcriptomes in non-model organisms. BMC Bioinformatics, 19 (Suppl 14) :416. <https://doi.org/10.1186/s12859-018-2384-y> (apéndice F.15)
- Lima-Cabello E., Martínez-Beas I., García-Quirós E., Carmona R., Claros M. G., Jiménez-López J. C., Alché J. D. (2019). Structural and functional features of glutathione reductase transcripts from olive (*Olea europaea* L.) seeds. Lecture Notes in Computer Science. LNCS, volume 11465: 178-191. [https://doi.org/10.1007/978-3-030-17938-0\\_17](https://doi.org/10.1007/978-3-030-17938-0_17) (apéndice F.16)

### Comunicaciones a congresos

- S. Morales, A. J. Castro, J. C. Jiménez-López, A. Zafra, M. J. Jiménez-Quesada, J. F. Florido, R. Carmona, M. I. Rodríguez-García, J. D. Alché. Desarrollo y aplicaciones de un sistema multiplex para el análisis y la expresión de alérgenos y productos génicos de interés en el polen del olivo. Segundas jornadas del máster universitario en biotecnología y biomedicina de la Universidad de Jaén. Jaén, diciembre 2012 (comunicación oral).
- A. Zafra, R. Carmona, M. J. Jiménez-Quesada, J. Traverso, A. Castro, M. I. Rodríguez-García, R. Bautista, M. G. Claros, J. D. Alché. Transcriptome analysis of enzyme activities regulating ROS metabolism in olive (*Olea europaea* L.) reproductive tissues. 11th International POG conference Reactive Oxygen and Nitrogen Species in Plants. Polonia, julio de 2013 (póster).
- A. Zafra, R. Carmona, E. Domingo, M. J. Jiménez-Quesada, A. J. Castro, R. Bautista, M. G. Claros, J. D. Alché. De novo assembly of the olive (*Olea europaea* L.) seed transcriptome during germination and seedling growth. 2nd Plant Genomics Congress: Europe. Londres, mayo 2014 (póster).
- R. Carmona, A. Zafra, M. J. Jiménez-Quesada, E. García-Quirós, J. A. Traverso, A. J. Castro, R. Bautista, M. G. Claros, J. D. Alché. Identification of transcripts corresponding to enzyme activities regulating ROS metabolism in the olive (*Olea europaea* L.) reproductive tissues through NGS approaches. 2nd Plant Genomics Congress: Europe. Londres, mayo 2014 (póster).
- A. Zafra, R. Carmona, M. J. Jiménez-Quesada, J. A. Traverso, A. J. Castro, M. I. Rodríguez-García, R. Bautista, M. G. Claros, J. D. Alché. De novo assembly, functional annotation

and applications of the olive tree (*Olea europaea* L.) reproductive transcriptome. 2nd Plant Genomics Congress: Europe. Londres, mayo 2014 (póster).

- P. Seoane, R. Carmona, R. Bautista, D. Guerrero-Fernández, M. G. Claros. Using Auto-flow, a workframe to resolve workflows, to build a de novo plant transcriptome. 2nd Plant Genomics Congress: Europe. Londres, mayo 2014 (póster).
- R. Carmona, A. Zafra, P. Seoane, A. J. Castro, J. D. Alché, M. G. Claros. From annotated transcriptome to a browseable database in the seek of olive tree allergens. XII Symposium on Bioinformatics. Sevilla, septiembre 2014 (póster).
- H. Benzekri, P. Seoane, R. Carmona, R. Bautista, D. Guerrero-Fernández, N. Fernández-Pozo, M. G. Claros. A workflow with dedicated tools for preparing reference transcriptomes from non-model organisms has evidenced important biological information. XII Symposium on Bioinformatics. Sevilla, septiembre 2014 (póster).
- M. J. Jiménez-Quesada, J. A. Traverso, A. Zafra, J. C. Jiménez-Lopez, R. Carmona, M. G. Claros, J. D. Alché. Identification and *in silico* analysis of NADPH oxidase homologues involved in allergy from an olive pollen transcriptome. 3rd International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2015). Granada, abril 2015 (póster).
- A. Zafra, R. Carmona, J. C. Jiménez-Lopez, A. Pulido, M. G. Claros, J. D. Alché. Identification of distinctive variants of the olive pollen allergen *Ole e 5* (Cu,Zn Superoxide Dismutase) throughout the analysis of the olive pollen transcriptome. 3rd International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2015). Granada, abril 2015 (póster).
- P. Seoane, R. Carmona, A. Zafra, A. J. Castro, D. Guerrero-Fernández, T. Castillo-Castillo, A. Medina-García, J. F. Aldana-Montes, I. Navas-Delgado, J. D. Alché, M. G. Claros. Automated construction of transcriptome databases for uncharacterised-genome plants, such as olive tree. The 3rd Plant Genomic Congress: Europe. Londres, mayo 2015 (póster).
- M. J. Jiménez-Quesada, A. Zafra, R. Carmona, J. A. Traverso, A. J. Castro, M. G. Claros, J. D. Alché. Identification and *in silico* analysis of enzymes involved in the generation and dismutation of the superoxide anion in olive (*Olea europaea* L.) pollen throughout the screening of the transcriptome. The 3rd Plant Genomic Congress: Europe. Londres, mayo 2015 (póster).
- R. Carmona, P. Seoane, A. Zafra, A. J. Castro, J. C. Jiménez-López, M. G. Claros, J. D. Alché. Transcriptome-mining approach in the seek of non-described olive tree allergens. The 3rd Plant Genomic Congress: Europe. Londres, mayo 2015 (póster).
- R. Carmona, A. Zafra, J. C. Jiménez-López, P. Seoane, J. D. Alché, M. G. Claros. In the search of non described olive tree allergens by transcriptome-mining. European Histamine Research Society (EHRS) 44th Annual Meeting. Málaga, mayo 2015 (comunicación oral).

- R. Carmona, P. Seoane, A. Zafra, M. J. Jiménez-Quesada, J. D. Alché, M. G. Claros. Automatic workflow for the identification of constitutively-expressed genes based on NGS reads mapping. 4th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2016). Granada, abril 2016 (comunicación oral).
- A. Zafra, J.C. Jiménez-López, R. Carmona, M. G. Claros, J. D. Alché. Transcriptome-based identification of a seed olive legumin (11S globulin). Characterization of subunits, 3D modelling and molecular assessment of allergenicity. 4th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2016). Granada, abril 2016 (comunicación oral).
- R. Carmona, M. J. Jiménez-Quesada, A. Zafra, M. G. Claros, J. D. Alché. Transcriptomic approaches to the *in silico* identification of the S-nitrosylome in the olive tree reproductive tissues. 6th Plant Nitric Oxide International Meeting. Granada, septiembre 2016 (comunicación oral).
- R. Carmona. De novo assembly, functional annotation and applications of the olive tree reproductive transcriptome. Young Researchers' Science Symposium. Estación Experimental del Zaidín (CSIC). Granada, diciembre 2016 (comunicación oral).
- R. Carmona, M. J. Jiménez-Quesada, M. G. Claros, J. D. Alché, A. J. Castro. Transcriptomic analysis of storage lipid mobilization in the olive tree pollen during germination. XII Reunión de la Sociedad Española de Fisiología Vegetal. Barcelona, junio 2017 (póster).
- R. Carmona. Avances en la obtención y aplicaciones del transcriptoma reproductivo de olivo. Second Young Researchers' Science Symposium. Estación Experimental del Zaidín (CSIC). Granada, diciembre 2017 (comunicación oral).
- E. García-Quirós, R. Carmona, A. Zafra, M. G. Claros, J. D. Alché. Identification and *in silico* Analysis of Glutathione Reductase transcripts expressed in olive (*Olea europaea* L.) pollen and pistil. 5th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2017). Granada, abril 2017 (comunicación oral).
- E. García-Quirós, R. Carmona, M. G. Claros, J. D. Alché. An overview of genes and transcripts corresponding to enzymes regulating glutathione metabolism in the olive (*Olea europaea* L.) reproductive tissues. The 5th Plant Genomics & Gene Editing Congress: Europe. Ámsterdam, marzo 2017 (póster).
- R. Carmona, M. J. Jiménez-Quesada, A. Zafra, A. J. Castro, M. G. Claros, J. D. Alché. Predictive S-nitrosylation and Tyr-nitration in the olive tree (*Olea europaea* L.) reproductive tissues transcriptomes: an *in silico* approach. The 5th Plant Genomics & Gene Editing Congress: Europe. Ámsterdam, marzo 2017 (póster).
- M. Arroyo, R. Bautista, R. Carmona, R. Larrosa, J. L. De La Cruz, M. Cobo, M. G. Claros. Identification of reference genes in lung cancer from RNA-seq data. European Respiratory Society International Congress 2017. Milán, septiembre 2017 (póster).

- J. C. Jiménez-López, E. Lima-Cabello, E. García-Quirós, R. Carmona, A. Zafra, A. J. Castro, J. D. Alché. Characterization and functional analysis of enzymes involved in glutathione metabolism in olive reproductive biology. 6th International Conference on the Olive Tree and Olive Products (Olive Biotech'18). Sevilla, octubre 2018 (comunicación oral).
- E. Lima-Cabello, M. J. Jiménez-Quesada, R. Carmona, J. C. Jiménez-López, A. J. Castro, M. G. Claros, J. D. Alché. Immunoprecipitation and nanoLC-MS analysis of the *S*-nitroso- and Tyr-nitroproteomes of the olive (*Olea europaea* L.) pollen. 7th Plant nitric oxide international meeting. Niza (Francia), octubre 2018 (póster).
- E. Lima-Cabello, I. Martínez-Beas, E. García-Quirós, R. Carmona, M. G. Claros, J.C. Jiménez-López, J.D. Alché. Structural and Functional Features of Glutathione Reductase Transcripts from Olive (*Olea europaea* L.) seeds. 7th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2019). Granada, mayo 2019 (comunicación oral).
- J. D. Alché, R. Carmona, A. J. Castro, M. G. Claros, J. C. Jiménez-López, M. J. Jiménez-Quesada, E. Lima-Cabello, A. Zafra. Desarrollo y aplicaciones de la transcriptómica reproductiva del olivo. Congreso en red de Olivicultura, Citricultura y Fruticultura de la Sociedad Española de Ciencias Hortícolas (SECH). Marzo 2021 (póster).



*A mis padres Ramón y Teresa*

# Agradecimientos

Esta tesis es fruto del esfuerzo directo o indirecto de muchas personas, a las que me gustaría dar por ello las gracias.

En primer lugar me gustaría mencionar a Enrique y Ana, a los que considero mis mentores, por animarme tanto en momentos duros y ser los primeros que confiaron en mí y me brindaron una oportunidad en este mundo de la bioinformática. Esas cosas nunca se olvidan.

Por supuesto, dar especialmente las gracias a mis directores de tesis, Gonzalo y Juande, por darme la oportunidad de llevar a cabo este trabajo, por su apoyo, su guía, sus consejos y sus enseñanzas durante todo este tiempo.

Además este trabajo tampoco habría sido posible sin la infraestructura y el personal de la Plataforma Andaluza de Bioinformática (PAB), tanto de la parte de supercomputación (Rafa y Darío), como el servicio de secuenciación (Pepi).

A Rocío tengo que agradecerle mis primeros pasos en el campo de la bioinformática, en el que con tanto entusiasmo me introdujo. A Pedro le agradezco su paciencia y toda la sabiduría que ha compartido conmigo, siempre dispuesto a ayudar.

A mi amigo Diego, compañero de carrera, de trabajo y de otras muchas aventuras. Por ser mi mayor apoyo moral durante mi estancia en la plataforma, con muchas risas y alguna que otra lágrima por el camino.

Al resto de mis compañeros de la plataforma de bioinformática: Hicham, Isa, David, Noé... Y otros muchos que han pasado por aquí en todo este tiempo y que me han aportado su ayuda, además de buenos momentos y experiencias.

A todos mis compañeros de la Estación Experimental del Zaidín por su importantísima colaboración en los resultados de esta tesis y por acogerme tan bien cada vez que he tenido que ir a Granada.

A mis amigos Pepe y Silveria, y mis cuñados Antonio y Vanesa, por sus palabras de ánimo, su confianza y por ayudarme a desconectar de vez en cuando, algo no menos importante.

A mi familia, el motor de mi vida: mis padres, Ramón y Teresa, a los que les debo todo, que son mi mejor ejemplo a seguir y que me han enseñado que con sacrificio, esfuerzo y humildad, todo es posible, y a mis hermanos, M.<sup>a</sup> Teresa y Ramón, por su amor y apoyo incondicional. Ellos cuatro han sido, son y serán mi mayor impulso en la vida.

A Carlos, que apareció en mi vida en el momento más inesperado, y que me dio el empujón que necesitaba para escribir por fin esta tesis, además de ayudarme muchísimo en la escritura y proporcionarme un inimaginable apoyo, tanto emocional como logístico. Espero poder por fin compensarle todo

este tiempo de sacrificio a mi lado.

En definitiva, a todas las personas, familiares o amigos, que de alguna u otra forma han hecho posible que este proyecto llegue a su fin. Por desgracia algunas de ellas ya no están conmigo, aunque estarán siempre presentes en mi corazón. Va por ellos...

Muchas gracias a todos.

# Resumen y contextualización

Comencé mi doctorado en el año 2013, dentro del grupo de Biología Reproductiva de Plantas del Departamento de Bioquímica, Biología Celular y Molecular de Plantas de la Estación Experimental del Zaidín (EEZ) de Granada, perteneciente al Consejo Superior de Investigaciones Científicas (CSIC). No obstante, dado el carácter bioinformático de la tesis, he llevado a cabo mi trabajo en la Plataforma Andaluza de Bioinformática (PAB), ubicada en el edificio de Supercomputación y Bioinnovación de la Universidad de Málaga, situado en el Parque Tecnológico de Andalucía. Esta plataforma proporciona recursos de supercomputación y ultrasecuenciación (inicialmente un equipo Titanium GS-FLX+454 de Roche y la adquisición posterior de un equipo NextSeq550 de Illumina) tanto a la comunidad universitaria, como a instituciones públicas y privadas.

La investigación que ha dado lugar a esta tesis parte del interés del grupo de Biología Reproductiva de llevar a cabo un estudio transcriptómico de los tejidos reproductivos del olivo, en un momento en el que aún no se había publicado el genoma de esta planta. Esta tesis es por tanto el fruto de la colaboración de la Plataforma Andaluza de Bioinformática con el grupo de Biología Reproductiva de la EEZ, ante su necesidad de apoyo bioinformático en su área de investigación. En este sentido me gustaría recalcar la ventaja que supone para un grupo de bioinformática colaborar con un potente grupo de investigación experimental, capaz de comprobar y validar en el laboratorio los resultados bioinformáticos obtenidos.

El trabajo de mi tesis se ha centrado en el procesamiento de los datos generados mediante la secuenciación en la PAB del material biológico aportado por el grupo de investigación de la Estación Experimental del Zaidín, su puesta a punto para hacerlos accesibles a dicha comunidad científica y el desarrollo de aplicaciones derivadas del análisis de dichos datos con el fin de ampliar los conocimientos biológicos.

En primer lugar generamos un transcriptoma reproductivo del olivo que ha sido incorporado a ReprOlive, una base de datos pública y de fácil uso. El acceso a esta información permitirá a los científicos desarrollar futuros proyectos de investigación y diseñar estrategias que mejoren tanto la producción como la calidad de estos cultivos.

A continuación exploramos el transcriptoma reproductivo generado mediante distintas aproximaciones, con el fin de ampliar el conocimiento acerca de esta planta. Así, hemos obtenido nuevos genes marcadores para análisis de expresión génica en estos tejidos, hemos profundizado en su capacidad alergénica y hemos evaluado la propensión de las proteínas derivadas de estos transcritos a sufrir determinadas modificaciones postraduccionales. Parte de los resultados obtenidos en estas aproximaciones han sido validados en el laboratorio por el grupo de investigación de la EEZ.

Todos los estudios mencionados se han llevado a cabo mediante el diseño de flujos de trabajo automáticos, que garantizan la reproducibilidad de los resultados, y que son fácilmente ejecutables y adaptables a nuevos programas o versiones de los mismos. Además, pueden aplicarse en otras especies.

Debemos destacar que en los últimos años se han publicado varios genomas del olivo, pertenecientes a distintas variedades, que suponen sin duda la mejor herramienta para profundizar en la biología de esta planta. No obstante, toda la información específica generada en esta tesis fue anterior a estas publicaciones y ha supuesto un importante punto de referencia para la comunidad científica y continúa siendo útil para corroborar los nuevos hallazgos.

# Acrónimos y Abreviaturas

|                               |                                                                                                                                                  |                |                                                                                                                                                                              |
|-------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>AP<sub>non-olive</sub></b> | proteínas alergénicas de especies distintas a olivo                                                                                              | <b>ddNTP</b>   | didesoxinucleótidos trifosfato ( <i>d</i> ideoxy <u>N</u> ucleotides <u>T</u> ri <u>P</u> hosphate)                                                                          |
| <b>ATP</b>                    | trifosfato de adenosina ( <u>A</u> denosine <u>T</u> ri <u>P</u> hosphate)                                                                       | <b>DDT</b>     | <u>D</u> icloro <u>D</u> ifenil <u>T</u> ricloroetano                                                                                                                        |
| <b>BAC</b>                    | cromosoma artificial bacteriano ( <u>B</u> acterial <u>A</u> rtificial <u>C</u> hromosome)                                                       | <b>DNA</b>     | ácido desoxirribonucleico ( <u>D</u> eoxy <u>r</u> ibo <u>N</u> ucleic <u>A</u> cid)                                                                                         |
| <b>BAM</b>                    | mapa binario de alineamiento ( <u>B</u> inary <u>A</u> lignment <u>M</u> ap)                                                                     | <b>DNA-seq</b> | secuenciación de DNA                                                                                                                                                         |
| <b>BLAST</b>                  | herramienta de búsqueda de alineamiento de secuencias de tipo local ( <u>B</u> asic <u>L</u> ocal <u>A</u> lignment <u>S</u> earch <u>T</u> ool) | <b>dNTP</b>    | desoxinucleótidos trifosfato ( <i>d</i> eoxy <u>N</u> ucleotides <u>T</u> ri <u>P</u> hosphate)                                                                              |
| <b>BSAS</b>                   | secuenciación de amplicones con bisulfito ( <u>B</u> i <u>S</u> ulfite <u>A</u> mplicon <u>S</u> equencing)                                      | <b>EBI</b>     | Instituto Europeo de Bioinformática ( <u>E</u> uropean <u>B</u> ioinformatics <u>I</u> nstitute)                                                                             |
| <b>cDNA</b>                   | ácido desoxirribonucleico complementario ( <i>c</i> omplementary <u>D</u> eoxy <u>r</u> ibo <u>N</u> ucleic <u>A</u> cid)                        | <b>EC</b>      | clasificación numérica de enzimas ( <u>E</u> nzyme <u>C</u> ommission)                                                                                                       |
| <b>ChIP-seq</b>               | secuenciación de la cromatina inmunoprecipitada ( <u>C</u> hromatin <u>I</u> mmuno <u>P</u> recipitation <u>s</u> equencing)                     | <b>EEZ</b>     | <u>E</u> stación <u>E</u> xperimental del <u>Z</u> aidín                                                                                                                     |
| <b>CNAG-CRG</b>               | <u>C</u> entro <u>N</u> acional de <u>A</u> nálisis <u>G</u> enómico y <u>C</u> entro de <u>R</u> egulación <u>G</u> enómica de Barcelona        | <b>ENA</b>     | archivo europeo de secuencias nucleotídicas ( <u>E</u> uropean <u>N</u> ucleotide <u>A</u> rchive)                                                                           |
| <b>CNV</b>                    | variación en el número de copias ( <u>C</u> opy <u>N</u> umber <u>V</u> ariation)                                                                | <b>EST</b>     | etiquetas de secuencias expresadas ( <u>E</u> xpressed <u>S</u> equences <u>T</u> ag)                                                                                        |
| <b>CPU</b>                    | unidad central de procesamiento ( <u>C</u> entral <u>P</u> rocessing <u>U</u> nity)                                                              | <b>FAO</b>     | Organización de las Naciones Unidas para la Alimentación y la Agricultura ( <u>F</u> ood and <u>A</u> griculture <u>O</u> rganization of the <u>U</u> nited <u>N</u> ations) |
| <b>CSIC</b>                   | <u>C</u> onsejo <u>S</u> uperior de <u>I</u> nvestigaciones <u>C</u> ientíficas                                                                  | <b>FDR</b>     | <u>F</u> alse <u>D</u> iscovery <u>R</u> ate                                                                                                                                 |
| <b>CV</b>                     | coeficiente de variación                                                                                                                         | <b>FPKM</b>    | fragmentos por kilobase de transcrito, por millón de lecturas mapeadas ( <u>F</u> ragments <u>P</u> er <u>K</u> ilobase of transcript, per <u>M</u> illion mapped reads)     |
| <b>Cys</b>                    | cisteína (aminoácido)                                                                                                                            | <b>GO</b>      | <u>G</u> ene <u>O</u> ntology                                                                                                                                                |
| <b>DDJB</b>                   | base de datos de DNA de Japón ( <u>D</u> NA <u>D</u> ata <u>B</u> ank of <u>J</u> apan)                                                          | <b>GSH</b>     | glutación                                                                                                                                                                    |
|                               |                                                                                                                                                  | <b>GSNO</b>    | <i>S</i> -nitrosoglutación                                                                                                                                                   |
|                               |                                                                                                                                                  | <b>GSSG</b>    | glutación oxidado                                                                                                                                                            |

|                         |                                                                                                                   |                |                                                                                                                                            |
|-------------------------|-------------------------------------------------------------------------------------------------------------------|----------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| <b>GTP</b>              | trifosfato de guanosina ( <i>Guanosine TriPhosphate</i> )                                                         | <b>nt</b>      | nucleótido                                                                                                                                 |
| <b>HCPC</b>             | <i>Hierarchical Clustering on Principal Components</i>                                                            | <b>OLC</b>     | <i>Overlap-Layout-Consensus</i>                                                                                                            |
| <b>IgE</b>              | Inmunoglobulina E                                                                                                 | <b>ORF</b>     | marco abierto de lectura ( <i>Open Reading Frame</i> )                                                                                     |
| <b>indel</b>            | inserciones y deleciones                                                                                          | <b>PAB</b>     | Plataforma Andaluza de Bioinformática                                                                                                      |
| <b>IOGC</b>             | <i>International Olive Genome Consortium</i>                                                                      | <b>pb</b>      | pares de bases                                                                                                                             |
| <b>IUBMB</b>            | <i>International Union of Biochemistry and Molecular Biology</i>                                                  | <b>PCA</b>     | análisis de componentes principales ( <i>Principal Component Analysis</i> )                                                                |
| <b>IUIS</b>             | <i>International Union of Immunological Societies</i>                                                             | <b>PCR</b>     | Reacción en cadena de la polimerasa ( <i>Polymerase Chain Reaction</i> )                                                                   |
| <b>KEGG</b>             | <i>Kyoto Encyclopedia of Genes and Genomes</i>                                                                    | <b>PDB</b>     | <i>Protein Data Bank</i>                                                                                                                   |
| <b>LC-MS</b>            | cromatografía de líquidos acoplada a espectrometría de masas ( <i>Liquid Chromatography - Mass Spectrometry</i> ) | <b>PTM</b>     | modificaciones postraduccionales ( <i>Post-Translational Modifications</i> )                                                               |
| <b>lncRNA</b>           | RNA largo no codificante ( <i>long non-coding RNA</i> )                                                           | <b>qPCR</b>    | Reacción en cadena de la polimerasa en tiempo real ( <i>quantitative Polymerase Chain Reaction</i> )                                       |
| <b>LTP</b>              | Proteína de transferencia de lípidos ( <i>Lipid Transfer Protein</i> )                                            | <b>RAM</b>     | memoria de acceso aleatorio ( <i>Random Access Memory</i> )                                                                                |
| <b>mRNA</b>             | ácido ribonucleico mensajero ( <i>messenger RiboNucleic Acid</i> )                                                | <b>RDF</b>     | marco de descripción de recursos ( <i>Resource Description Framework</i> )                                                                 |
| <b>miRNA</b>            | microARN ( <i>micro RiboNucleic Acid</i> )                                                                        | <b>RG</b>      | genes de referencia ( <i>Reference Genes</i> )                                                                                             |
| <b>NAD<sup>+</sup></b>  | Dinucleótido de nicotinamida y adenina en su forma oxidada                                                        | <b>RNA</b>     | ácido ribonucleico ( <i>RiboNucleic Acid</i> )                                                                                             |
| <b>NADP<sup>+</sup></b> | Dinucleótido fosfato de nicotinamida y adenina en su forma oxidada                                                | <b>RNA-seq</b> | secuenciación de RNA                                                                                                                       |
| <b>NADPH</b>            | Dinucleótido fosfato de nicotinamida y adenina en su forma reducida                                               | <b>RNS</b>     | especies reactivas de nitrógeno ( <i>Reactive Nitrogen Species</i> )                                                                       |
| <b>NCBI</b>             | Centro Nacional para la Información Biotecnológica ( <i>National Center for Biotechnology Information</i> )       | <b>rRNA</b>    | ácido ribonucleico ribosómico ( <i>ribosomal RiboNucleic Acid</i> )                                                                        |
| <b>ncRNA</b>            | RNA no codificante ( <i>non-coding RNA</i> )                                                                      | <b>ROS</b>     | especies de oxígeno reactivo ( <i>Reactive Oxygen Species</i> )                                                                            |
| <b>NGS</b>              | secuenciación de nueva generación ( <i>Next Generation Sequencing</i> )                                           | <b>RPKM</b>    | lecturas por kilobase de transcrito, por millón de lecturas mapeadas ( <i>Reads Per Kilobase of transcript, per Million mapped reads</i> ) |
| <b>NO</b>               | óxido nítrico ( <i>Nitric Oxide</i> )                                                                             | <b>RPMM</b>    | lecturas por millón de lecturas mapeadas ( <i>Reads Per Mapped Million</i> )                                                               |
| <b>nsLTP</b>            | Proteína de transferencia de lípidos no específica ( <i>non-specific Lipid Transfer Protein</i> )                 | <b>R-SNO</b>   | S-nitrosotioles, R denota cualquier grupo orgánico                                                                                         |

|                                  |                                                                                                                              |                                    |                                                                                                                          |
|----------------------------------|------------------------------------------------------------------------------------------------------------------------------|------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| <b>RT-PCR</b>                    | Reacción en cadena de la polimerasa con transcripción inversa ( <i>Reverse Transcription Polymerase Chain Reaction</i> )     | <b>TT<sup>newcand pollen</sup></b> | transcritos tentativos del transcriptoma de polen de ReprOlive considerados candidatos a alérgenos nuevos                |
| <b>SAM</b>                       | mapa del alineamiento de secuencias ( <i>Sequence Alignment Map</i> )                                                        | <b>TT<sup>new pollen</sup></b>     | transcritos tentativos del transcriptoma de polen de ReprOlive considerados posibles nuevos alérgenos de olivo           |
| <b>SCBI</b>                      | Centro de Supercomputación y Bioinformática de la UMA                                                                        | <b>TT<sub>pollen</sub></b>         | transcritos tentativos del transcriptoma de polen de ReprOlive                                                           |
| <b>scRNA-seq</b>                 | secuenciación de RNA de célula única ( <i>single cell RNA-seq</i> )                                                          | <b>TT<sup>source pollen</sup></b>  | transcritos tentativos del transcriptoma de polen de ReprOlive considerados fuente de posibles nuevos alérgenos de olivo |
| <b>SDS-PAGE</b>                  | <i>Sodium Dodecyl Sulfate polyacrylamide gel electrophoresis</i>                                                             | <b>Tyr</b>                         | tirosina (aminoácido)                                                                                                    |
| <b>siRNA</b>                     | ácido ribonucleico pequeño nucleolar ( <i>small nucleolar RiboNucleic Acid</i> )                                             | <b>UMA</b>                         | Universidad de Málaga                                                                                                    |
| <b>snoRNA</b>                    | ácido ribonucleico pequeño de interferencia ( <i>small interfering RiboNucleic Acid</i> )                                    | <b>UTR</b>                         | región no traducida ( <i>UnTranslated Region</i> )                                                                       |
| <b>SNP</b>                       | polimorfismo de nucleótido simple ( <i>Single Nucleotide Polymorphism</i> )                                                  | <b>WHO</b>                         | Organización Mundial para la Salud ( <i>World Health Organization</i> )                                                  |
| <b>SRA</b>                       | <i>Sequence Read Archive</i>                                                                                                 | <b>WGS</b>                         | secuenciación del genoma completo ( <i>Whole Genome Sequencing</i> )                                                     |
| <b>SSR</b>                       | repeticiones de secuencia simple ( <i>Single Sequence Repeat</i> )                                                           | <b>ZMW</b>                         | onda guía de modo cero ( <i>Zero Mode Waveguide</i> )                                                                    |
| <b>STR</b>                       | repetición corta en tándem ( <i>Short Tandem Repeat</i> )                                                                    |                                    |                                                                                                                          |
| <b>SV</b>                        | variación estructural ( <i>Structural Variation</i> )                                                                        |                                    |                                                                                                                          |
| <b>TLP</b>                       | proteína de tipo taumatina ( <i>Thaumatococcus-like Protein</i> )                                                            |                                    |                                                                                                                          |
| <b>tRNA</b>                      | ácido ribonucleico de transferencia ( <i>transfer RiboNucleic Acid</i> )                                                     |                                    |                                                                                                                          |
| <b>TT</b>                        | transcritos tentativos ( <i>Tentative Transcripts</i> )                                                                      |                                    |                                                                                                                          |
| <b>TT<sup>iso pollen</sup></b>   | transcritos tentativos del transcriptoma de polen de ReprOlive considerados isoformas de los alérgenos de olivo ya conocidos |                                    |                                                                                                                          |
| <b>TT<sup>known pollen</sup></b> | transcritos tentativos del transcriptoma de polen de ReprOlive que son alérgenos de olivo ya conocidos                       |                                    |                                                                                                                          |



# Índice general

|                                                                   |           |
|-------------------------------------------------------------------|-----------|
| <b>Proyectos de investigación</b>                                 | <b>11</b> |
| <b>Publicaciones</b>                                              | <b>13</b> |
| <b>Agradecimientos</b>                                            | <b>21</b> |
| <b>Resumen y contextualización</b>                                | <b>23</b> |
| <b>Acrónimos y Abreviaturas</b>                                   | <b>25</b> |
| <b>1. INTRODUCCIÓN</b>                                            | <b>37</b> |
| 1.1. Bioinformática . . . . .                                     | 37        |
| 1.1.1. Las ómicas . . . . .                                       | 37        |
| 1.1.2. Tecnologías de secuenciación . . . . .                     | 38        |
| ¿En qué consiste la secuenciación? . . . . .                      | 38        |
| La primera generación de la secuenciación . . . . .               | 38        |
| Segunda generación: la NGS . . . . .                              | 40        |
| Pirosecuenciación (454/FLX de Roche) . . . . .                    | 41        |
| Solexa/Illumina . . . . .                                         | 43        |
| SOLiD . . . . .                                                   | 44        |
| Ion Torrent . . . . .                                             | 45        |
| Tercera generación: secuenciación de una única molécula . . . . . | 45        |
| Helicos, PacBio y Nanopore . . . . .                              | 46        |
| Tipos de lecturas . . . . .                                       | 46        |
| Aplicaciones de las tecnologías de secuenciación . . . . .        | 49        |

|                                                                                                |    |
|------------------------------------------------------------------------------------------------|----|
| Secuenciación de genomas <i>de novo</i> (DNA-seq) . . . . .                                    | 49 |
| Resecuenciación (DNA-seq) . . . . .                                                            | 50 |
| Epigenómica (DNA-seq) . . . . .                                                                | 50 |
| Metagenómica (DNA-seq) . . . . .                                                               | 51 |
| Secuenciación de transcriptomas <i>de novo</i> (RNA-seq) . . . . .                             | 51 |
| Análisis de expresión génica (RNA-seq) . . . . .                                               | 52 |
| 1.1.3. Obtención de un transcriptoma . . . . .                                                 | 52 |
| Preprocesamiento . . . . .                                                                     | 54 |
| Ensamblaje . . . . .                                                                           | 55 |
| Filtrado de las lecturas . . . . .                                                             | 55 |
| Construcción de grafos . . . . .                                                               | 55 |
| Simplificación de grafos . . . . .                                                             | 59 |
| Posprocesamiento . . . . .                                                                     | 59 |
| Mapeo . . . . .                                                                                | 59 |
| Anotación . . . . .                                                                            | 60 |
| Enriquecimiento de las anotaciones . . . . .                                                   | 61 |
| Incorporación a bases de datos . . . . .                                                       | 62 |
| 1.1.4. Flujos de trabajo . . . . .                                                             | 63 |
| 1.2. El olivo como modelo . . . . .                                                            | 64 |
| 1.2.1. Visión general . . . . .                                                                | 64 |
| 1.2.2. La reproducción del olivo . . . . .                                                     | 64 |
| 1.2.3. El papel de las ROS y el NO en la interacción polen-pistilo y la señalización . . . . . | 66 |
| Las ROS en la reproducción del olivo . . . . .                                                 | 66 |
| El NO y las modificaciones postraduccionales . . . . .                                         | 66 |
| ¿Por qué centrarnos en el NO? . . . . .                                                        | 66 |
| <i>S</i> -Nitrosilación . . . . .                                                              | 67 |
| Tyr-Nitración . . . . .                                                                        | 67 |
| 1.2.4. El olivo y la alergia . . . . .                                                         | 67 |
| 1.2.5. Ómicas en el olivo . . . . .                                                            | 70 |
| 1.2.6. Desafíos futuros . . . . .                                                              | 74 |

## 2. OBJETIVOS

75

|                                                             |           |
|-------------------------------------------------------------|-----------|
| <b>3. MATERIAL Y MÉTODOS</b>                                | <b>77</b> |
| 3.1. Infraestructura bioinformática . . . . .               | 77        |
| 3.2. Programas informáticos . . . . .                       | 77        |
| 3.2.1. Para manipulación . . . . .                          | 78        |
| SeqTrimNext . . . . .                                       | 78        |
| Bowtie2 . . . . .                                           | 78        |
| Samtools . . . . .                                          | 78        |
| CD-HIT EST . . . . .                                        | 78        |
| SeaView . . . . .                                           | 78        |
| 3.2.2. Para ensamblaje . . . . .                            | 78        |
| MIRA3 y MIRA4 . . . . .                                     | 78        |
| Euler-SR . . . . .                                          | 79        |
| CAP3 . . . . .                                              | 79        |
| OASES . . . . .                                             | 79        |
| SOAPdenovo-Trans . . . . .                                  | 79        |
| RAY . . . . .                                               | 79        |
| Minimus2 . . . . .                                          | 79        |
| 3.2.3. Para anotación . . . . .                             | 80        |
| Sma3s . . . . .                                             | 80        |
| Full-LengtherNext . . . . .                                 | 80        |
| MREPS . . . . .                                             | 81        |
| BUSCO . . . . .                                             | 81        |
| 3.2.4. Para enriquecimiento biológico . . . . .             | 81        |
| GOrilla . . . . .                                           | 81        |
| GeneCodis . . . . .                                         | 81        |
| 3.2.5. Para predicción de <i>S</i> -nitrosilación . . . . . | 81        |
| iSNO-AAPair . . . . .                                       | 81        |
| GPS-SNO . . . . .                                           | 82        |
| SNOSite . . . . .                                           | 82        |
| 3.2.6. Para predicción de Tyr-nitración . . . . .           | 82        |
| GPS-YNO2 . . . . .                                          | 82        |

|                                                                                          |            |
|------------------------------------------------------------------------------------------|------------|
| 3.2.7. Para diseño de flujos de trabajo . . . . .                                        | 82         |
| AutoFlow . . . . .                                                                       | 82         |
| TransFlow . . . . .                                                                      | 85         |
| 3.3. Datos biológicos . . . . .                                                          | 90         |
| 3.3.1. Secuencias para los transcriptomas reproductivo y vegetativo de olivo (ReprOlive) | 90         |
| 3.3.2. Secuencias para el transcriptoma mejorado de polen maduro . . . . .               | 90         |
| 3.3.3. Secuencias para el desarrollo de marcadores de RT-PCR . . . . .                   | 91         |
| 3.3.4. Secuencias para la identificación de <i>Ole e 7</i> . . . . .                     | 91         |
| 3.4. Obtención de transcriptomas de ReprOlive . . . . .                                  | 92         |
| 3.4.1. Preprocesamiento de las lecturas brutas . . . . .                                 | 93         |
| 3.4.2. Ensamblaje de los transcriptomas . . . . .                                        | 94         |
| 3.4.3. Anotación de los transcriptomas . . . . .                                         | 94         |
| 3.4.4. Datos de expresión . . . . .                                                      | 95         |
| 3.4.5. Construcción de la base de datos ReprOlive . . . . .                              | 95         |
| 3.5. Obtención del transcriptoma mejorado de polen maduro . . . . .                      | 97         |
| 3.6. Detección de genes de referencia para RT-PCR a partir de datos de RNA-seq . . . . . | 97         |
| 3.6.1. Flujo de trabajo para la detección de genes de referencia . . . . .               | 97         |
| Obtención de la tabla de recuentos de lecturas . . . . .                                 | 97         |
| Detección de genes de referencia . . . . .                                               | 98         |
| 3.6.2. Detección de genes de referencia en el olivo . . . . .                            | 99         |
| Validación experimental de genes de referencia en el olivo . . . . .                     | 99         |
| 3.6.3. Detección de genes de referencia en <i>Arabidopsis</i> . . . . .                  | 100        |
| 3.7. Detección de posibles nuevos alérgenos de olivo . . . . .                           | 100        |
| 3.7.1. Bases de datos derivadas de Allergome . . . . .                                   | 101        |
| 3.7.2. Flujo de trabajo para la identificación de alérgenos . . . . .                    | 101        |
| 3.7.3. Identificación de <i>Ole e 7</i> en transcriptomas y genomas de olivo . . . . .   | 102        |
| 3.8. Predicción de modificaciones postraduccionales . . . . .                            | 104        |
| <b>4. RESULTADOS</b>                                                                     | <b>105</b> |
| 4.1. Transcriptoma reproductivo del olivo . . . . .                                      | 105        |
| 4.1.1. El transcriptoma reproductivo del olivo . . . . .                                 | 105        |
| 4.1.2. La base de datos ReprOlive . . . . .                                              | 106        |

|                                                                                                                                                                    |            |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| Ensamblajes y transcritos tentativos . . . . .                                                                                                                     | 107        |
| SSR, Descriptions, GO, EC e InterPro . . . . .                                                                                                                     | 108        |
| Rutas KEGG . . . . .                                                                                                                                               | 110        |
| Datos de expresión . . . . .                                                                                                                                       | 110        |
| Búsquedas por secuencia, anotación y semántica . . . . .                                                                                                           | 112        |
| 4.2. Transcriptoma mejorado de polen maduro . . . . .                                                                                                              | 117        |
| 4.2.1. Ensamblaje mejorado . . . . .                                                                                                                               | 117        |
| 4.2.2. Comparación del transcriptoma mejorado con el de ReprOlive . . . . .                                                                                        | 117        |
| 4.3. Marcadores para la expresión génica por RT-PCR de los tejidos reproductivos de olivo . . . . .                                                                | 123        |
| 4.3.1. Tiempos de ejecución del flujo de trabajo . . . . .                                                                                                         | 123        |
| 4.3.2. Candidatos a genes de referencia en el olivo . . . . .                                                                                                      | 123        |
| Validación experimental de los candidatos a RG en el olivo . . . . .                                                                                               | 126        |
| 4.3.3. Candidatos a genes de referencia en <i>Arabidopsis</i> . . . . .                                                                                            | 129        |
| 4.4. Detección bioinformática de nuevos alérgenos . . . . .                                                                                                        | 131        |
| 4.4.1. Alérgenos de polen de olivo ya conocidos . . . . .                                                                                                          | 131        |
| 4.4.2. Reactividad cruzada con otras especies . . . . .                                                                                                            | 132        |
| 4.4.3. Posibles alérgenos nuevos y nuevas isoformas de los alérgenos ya conocidos . . . . .                                                                        | 137        |
| 4.5. Predicción <i>in silico</i> de posibles modificaciones postraduccionales ( <i>S</i> -nitrosilación y Tyr-nitración) en el polen de olivo . . . . .            | 145        |
| 4.5.1. Predicción <i>in silico</i> del <i>S</i> -nitrosoproteoma . . . . .                                                                                         | 145        |
| 4.5.2. Predicción <i>in silico</i> del Tyr-nitroproteoma . . . . .                                                                                                 | 145        |
| 4.5.3. Proteínas propensas a ambos tipos de modificación . . . . .                                                                                                 | 145        |
| 4.5.4. Funciones celulares afectadas por las PTM . . . . .                                                                                                         | 146        |
| 4.5.5. Comparación entre resultados predictivos y validación experimental . . . . .                                                                                | 150        |
| <b>5. DISCUSIÓN</b>                                                                                                                                                | <b>153</b> |
| 5.1. Transcriptoma reproductivo del olivo . . . . .                                                                                                                | 153        |
| 5.2. Transcriptoma mejorado de polen maduro . . . . .                                                                                                              | 155        |
| 5.3. Idoneidad del algoritmo de detección de genes de referencia . . . . .                                                                                         | 157        |
| 5.4. El análisis bioinformático ha profundizado en la heterogeneidad alérgica del polen de olivo . . . . .                                                         | 160        |
| 5.5. Predicción <i>in silico</i> de posibles modificaciones postraduccionales ( <i>S</i> -nitrosilación y Tyr-nitración) sobre el transcriptoma de polen . . . . . | 165        |

|                        |            |
|------------------------|------------|
| <b>6. CONCLUSIONES</b> | <b>169</b> |
|------------------------|------------|

|                  |            |
|------------------|------------|
| <b>APÉNDICES</b> | <b>173</b> |
|------------------|------------|

|                                                                                                         |            |
|---------------------------------------------------------------------------------------------------------|------------|
| <b>A. Flujo de trabajo basado en AutoFlow para la obtención del transcriptoma reproductivo de olivo</b> | <b>173</b> |
|---------------------------------------------------------------------------------------------------------|------------|

|                                                                                    |            |
|------------------------------------------------------------------------------------|------------|
| <b>B. Script de lanzamiento y plantillas de los distintos módulos de TransFlow</b> | <b>181</b> |
|------------------------------------------------------------------------------------|------------|

|                                                                                                                      |            |
|----------------------------------------------------------------------------------------------------------------------|------------|
| <b>C. Script de lanzamiento y plantilla de AutoFlow de <i>findRGs</i> para el desarrollo de marcadores de RT-PCR</b> | <b>207</b> |
|----------------------------------------------------------------------------------------------------------------------|------------|

|                                                                                                                                |            |
|--------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>D. Script de lanzamiento y plantilla de AutoFlow para la detección bioinformática de posibles nuevos alérgenos de olivo</b> | <b>213</b> |
|--------------------------------------------------------------------------------------------------------------------------------|------------|

|                                                                                                            |            |
|------------------------------------------------------------------------------------------------------------|------------|
| <b>E. Script de parseo e intersección de las salidas de los programas de predicción de S-nitrosilación</b> | <b>219</b> |
|------------------------------------------------------------------------------------------------------------|------------|

|                                               |            |
|-----------------------------------------------|------------|
| <b>F. Publicaciones derivadas de la tesis</b> | <b>225</b> |
|-----------------------------------------------|------------|

|                                                                                                                                         |     |
|-----------------------------------------------------------------------------------------------------------------------------------------|-----|
| F.1. ReprOlive: a database with linked data for the olive tree ( <i>Olea europaea</i> L.) reproductive transcriptome (publicación aval) | 225 |
|-----------------------------------------------------------------------------------------------------------------------------------------|-----|

|                                               |     |
|-----------------------------------------------|-----|
| F.2. AutoFlow: an easy way to build workflows | 227 |
|-----------------------------------------------|-----|

|                                                                                                                                      |     |
|--------------------------------------------------------------------------------------------------------------------------------------|-----|
| F.3. Identification and <i>in silico</i> analysis of NADPH oxidase homologues involved in allergy from an olive pollen transcriptome | 229 |
|--------------------------------------------------------------------------------------------------------------------------------------|-----|

|                                                                                                                                                                                |     |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| F.4. Identification of distinctive variants of the olive pollen allergen <i>Ole e 5</i> (Cu,Zn Superoxide Dismutase) throughout the analysis of the olive pollen transcriptome | 231 |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|

|                                                                                  |     |
|----------------------------------------------------------------------------------|-----|
| F.5. In the search of non described olive tree allergens by transcriptome-mining | 233 |
|----------------------------------------------------------------------------------|-----|

|                                                                                                                                                                      |     |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| F.6. Transcriptome-based identification of a seed olive legumin (11S Globulin). Characterization of subunits, 3D modelling and molecular assessment of allergenicity | 235 |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|

|                                                                                                            |     |
|------------------------------------------------------------------------------------------------------------|-----|
| F.7. Automatic workflow for the identification of constitutively-expressed genes based on mapped NGS reads | 237 |
|------------------------------------------------------------------------------------------------------------|-----|

|                                                                                                                                                                          |     |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| F.8. AutoFlow, a versatile workflow engine illustrated by assembling an optimised de novo transcriptome for a non-model species, such as faba bean ( <i>Vicia faba</i> ) | 239 |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|

|                                                                                                                                                 |     |
|-------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| F.9. Identification and in silico Analysis of Glutathione Reductase Transcripts Expressed in Olive ( <i>Olea europaea</i> L.) Pollen and Pistil | 241 |
|-------------------------------------------------------------------------------------------------------------------------------------------------|-----|

|                                                                         |     |
|-------------------------------------------------------------------------|-----|
| F.10. Automated identification of reference genes based on RNA-seq data | 243 |
|-------------------------------------------------------------------------|-----|

|                                                                                                                                                                                                           |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| F.11. Identification and functional annotation of genes differentially expressed in the reproductive tissues of the olive tree ( <i>Olea europaea</i> L.) through the generation of subtractive libraries | 245 |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|

|                                                                                                                                                                                                         |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| F.12. Generation of nitric oxide by olive ( <i>Olea europaea</i> L.) pollen during in vitro germination and assessment of the <i>S</i> -nitroso and nitro-proteomes by computational predictive methods | 247 |
| F.13. <i>S</i> -nitroso- and nitro- proteomes in the olive ( <i>Olea europaea</i> L.) pollen. Predictive versus experimental data by nano-LC-MS . . . . .                                               | 249 |
| F.14. Bioinformatic prediction of <i>S</i> -nitrosylation sites in large protein datasets . . . . .                                                                                                     | 251 |
| F.15. TransFlow: A modular framework for assembling and assessing accurate de novo transcriptomes in non-model organisms . . . . .                                                                      | 253 |
| F.16. Structural and functional features of glutathione reductase transcripts from olive ( <i>Olea europaea</i> L.) seeds . . . . .                                                                     | 255 |



# Capítulo 1

## INTRODUCCIÓN

### 1.1. Bioinformática

#### 1.1.1. Las ómicas

El avance de las tecnologías de secuenciación en las últimas décadas ha posibilitado el desarrollo de las denominadas disciplinas ómicas [1]. Las más destacadas se comentan a continuación:

- **Genómica:** Consiste en el estudio del genoma, que se define como la secuencia completa de nucleótidos (A, C, T y G en el caso de genomas de DNA) de un individuo. Se organiza en cromosomas y constituye el material genético de un organismo. Incluye tanto las regiones codificantes (genes), como el DNA no codificante y también el DNA mitocondrial y cloroplastídico. Un subconjunto del genoma sería el exoma, formado por los exones o regiones codificantes de los genes, que constituirán el RNA mensajero y que al ser traducido dará lugar a las proteínas. En humanos el exoma constituye alrededor de un 1 % del total del genoma [2].
- **Transcriptómica:** Se centra en el estudio del transcriptoma, definido como el conjunto de moléculas de RNA presentes en una célula o en un grupo de células. Aunque hace referencia a todos los tipos de RNA, tanto codificante como no codificante (ncRNA), suele usarse refiriéndose únicamente al conjunto de RNA mensajeros o transcritos. Es un reflejo de los genes que se están expresando en un momento dado y bajo unas condiciones determinadas.
- **Proteómica:** Se encarga del estudio del proteoma, que hace referencia al conjunto de proteínas que se expresan o pueden expresarse en una célula, un tejido o un organismo completo en un momento determinado y bajo unas condiciones específicas.
- **Metabolómica:** Estudia el metaboloma, que se define como el conjunto de moléculas pequeñas o metabolitos que se pueden encontrar en una muestra biológica bajo unas condiciones determinadas. Estos metabolitos pueden ser moléculas producidas por el organismo (aminoácidos, ácidos nucleicos, azúcares...) o bien compuestos exógenos (drogas, toxinas, antibióticos...).

Adviértase que, al contrario que el genoma, que es algo estático en una célula (exceptuando el proceso de mutación), tanto el transcriptoma, como el proteoma y el metaboloma son conceptos dinámicos y

cambian en función del estado y de las condiciones biológicas de la célula, tejido u organismo en ese momento [3].

Mientras que las tecnologías tradicionales tenían un enfoque reduccionista, estas nuevas disciplinas ómicas suponen una aproximación holística, en la que el funcionamiento de un sistema biológico puede ser más fácilmente entendible si se considera como un todo. Así, en un sistema biológico, la explicación del fenotipo resultante se busca en las complejas interacciones entre genes, proteínas y metabolitos.

Las tecnologías ómicas tienen un amplio campo de aplicación, como por ejemplo entender el funcionamiento de los procesos fisiológicos normales, la causa de las enfermedades, el descubrimiento de marcadores y de nuevos fármacos, entre muchas otras.

Cualquier experimento en alguna de las disciplinas ómicas requiere un diseño y preparación de las muestras que involucra varias tecnologías y genera a su vez un importante volumen de datos cuyo análisis resulta complejo. Por ello, la implicación de bioinformáticos y bioestadísticos durante todo el proceso es esencial.

### 1.1.2. Tecnologías de secuenciación

#### ¿En qué consiste la secuenciación?

Secuenciar una molécula de DNA (o RNA) consiste en determinar el orden exacto de los nucleótidos dentro de dicha molécula. Esta secuencia constituye la información genética heredable de los seres vivos y conocerla supone la base de la investigación tanto básica como aplicada de los procesos biológicos fundamentales. Dentro de una misma especie la mayoría de los nucleótidos son idénticos entre los individuos, pero la secuenciación de muchos de ellos nos permite entender los procesos responsables de la diversidad genética.

En los últimos años, el desarrollo de las distintas ómicas ha sido imparable de la mano del desarrollo de las tecnologías de secuenciación, porque ha supuesto una simplificación, un aumento de rendimiento y un abaratamiento considerable de los costes de secuenciación (figura 1.1). A día de hoy es factible secuenciar y analizar cualquier genoma o transcriptoma, tanto de especies modelo como no modelo, de forma que las bases de datos van acumulando secuencias de animales, plantas y microorganismos de manera exponencial. Un ejemplo de la evolución de las capacidades de estas tecnologías fue el Proyecto Genoma Humano [4, 5], un proyecto a gran escala, impensable de abordar hace no tantos años. Tras este logro, el avance gracias a la mejora de estas tecnologías ha sido imparable y desde entonces se han llevado a cabo o bien están en marcha varios proyectos cuyo objetivo es secuenciar miles [6], cientos de miles<sup>1,2</sup> y hasta un millón de genomas humanos [7]. En cuanto a plantas, también hay en marcha una iniciativa para la secuenciación de 10 000 genomas<sup>3</sup>.

Existen tres generaciones de tecnologías de secuenciación que comentamos a continuación.

#### La primera generación de la secuenciación

En 1977 se publicaron simultáneamente el método de secuenciación química de Maxam y Gilbert [8] y el método de secuenciación enzimática de Sanger [9], lo que les valió el Premio Nobel de Química

<sup>1</sup><https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/>

<sup>2</sup><https://genomeasia100k.org/>

<sup>3</sup><https://db.cngb.org/10kp/>

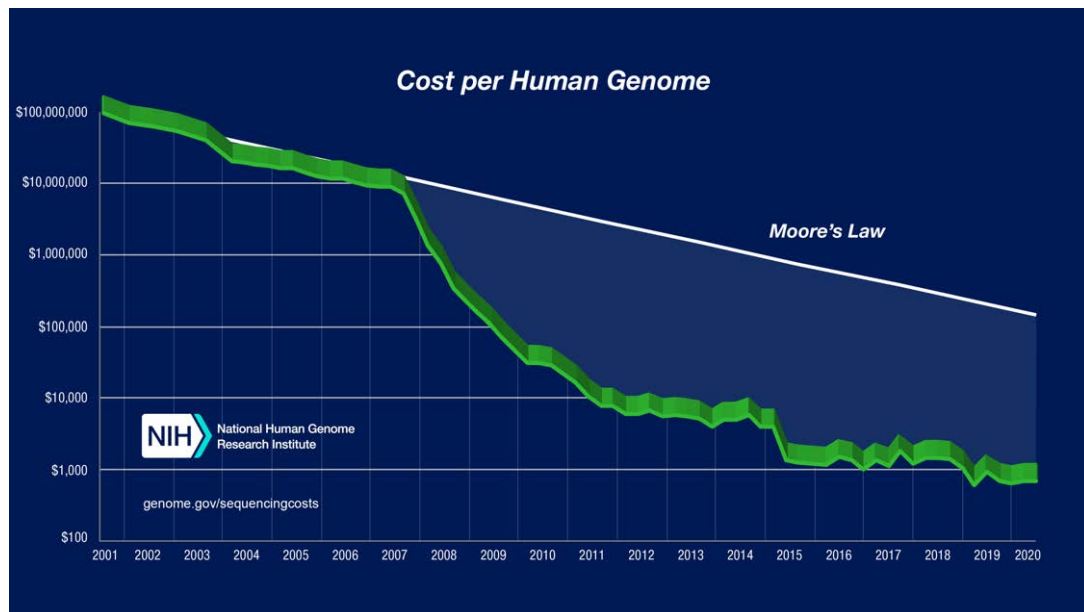


**Figura 1.1:** Evolución del coste de secuenciar un genoma humano. La rebaja de costes de 2003 se debió a la entrada de la pirosecuenciación (la primera tecnología comercial de NGS), la que comenzó en 2007 fue por la difusión de los secuenciadores de Illumina y la de 2015 por la aparición del Illumina X Ten. Todavía no se aprecia el efecto que va a tener la plataforma NovaSeq (Tomada de: <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>)

en el año 1980.

En el método de secuenciación de Maxam y Gilbert, la molécula de DNA se marca radiactivamente en uno de sus extremos con  $^{32}\text{P}$ . La técnica consiste en fragmentar estas moléculas marcadas utilizando agentes químicos específicos para cada uno o dos nucleótidos en cuatro reacciones independientes (G, A+G, C, C+T). Los productos de estas reacciones se separan por tamaño en una electroforesis en gel de poliacrilamida, donde el patrón de bandas radiactivas obtenido permite leer la secuencia (figura 1.2). Esta técnica, cuya principal ventaja es que se secuencia el DNA de la célula y no un clon ni un amplicón, empleaba productos químicos muy tóxicos, además de una gran dosis de radioactividad, por lo que acabó perdiendo la carrera ante el método de Sanger.

El método de secuenciación de Sanger precisa de una hebra molde de DNA de cadena simple, un cebador de DNA, una DNA polimerasa, nucleótidos normales marcados radiactivamente y unos nucleótidos modificados, los didesoxinucleótidos trifosfato (ddNTP), que producirán una terminación en la elongación de la cadena, por lo que el método también recibe el nombre de terminación de la cadena. Se llevan a cabo cuatro reacciones de secuenciación independientes, cada una solo con uno de los cuatro ddNTP (ddATP, ddGTP, ddCTP, ddTTP). Estos ddNTP carecen de un grupo 3'-OH necesario para la formación del enlace fosfodiéster entre dos nucleótidos durante la elongación de la cadena, por lo que su incorporación a la cadena naciente trunca su elongación. Dado que la concentración a la que se añaden estos didesoxinucleótidos es baja, se generarán todas las posibilidades de fragmentos. Los fragmentos resultantes de cada reacción por separado se desnaturalizan y se separan por tamaño mediante electroforesis en pocillos independientes y se visualizan por autorradiografía o luz ultravioleta.

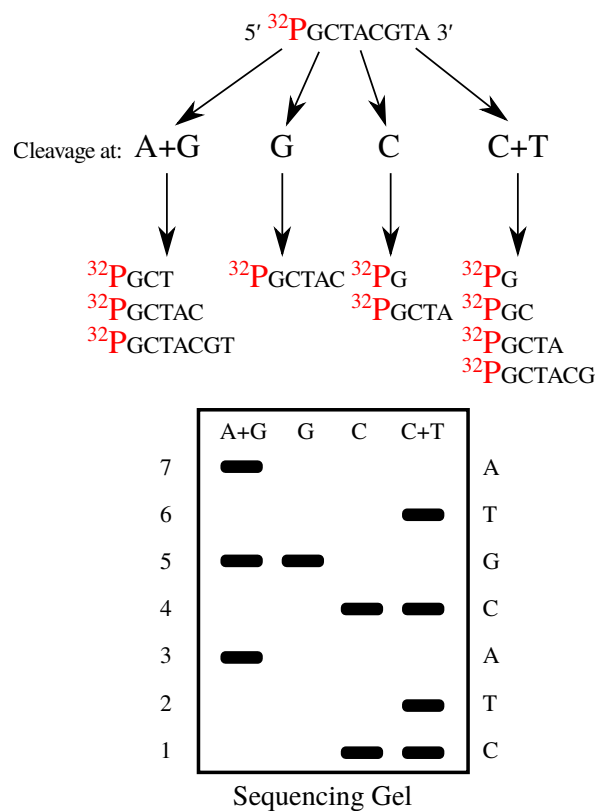


**Figura 1.2:** Método de secuenciación química de Maxam y Gilbert. Detallada en el texto. (Tomada de: [https://en.wikipedia.org/wiki/Maxam-Gilbert\\_sequencing](https://en.wikipedia.org/wiki/Maxam-Gilbert_sequencing))

La secuencia de DNA puede leerse directamente de la autorradiografía.

La puesta en marcha del Proyecto del Genoma Humano impulsó una serie de mejoras en la técnica, entre las que está la sustitución del marcaje radiactivo por un marcaje con cuatro fluorocromos distintos, de forma que la reacción se podía llevar a cabo en un único tubo. Además se sustituyeron los geles planos por electroforesis capilar. Inicialmente se llevaban a cabo en un solo capilar, y más adelante, las máquinas consiguieron incorporar varios capilares y paralelizar el proceso. Se producía un cromatograma o electroferograma, del cual se podía deducir la secuencia en un ordenador (figura 1.3). La técnica permite obtener secuencias de unas 1000 pb y un ligero abaratamiento de costes (figura 1.1 hasta 2003).

## Segunda generación: la NGS

La obtención del primer borrador del genoma humano [4, 5] costó unos 3 000 millones de dólares para leer 3 000 millones de nucleótidos (1 \$ por nucleótido). Estos costes resultaban desorbitados para cualquier laboratorio, por lo que surgió la necesidad de encontrar soluciones más baratas, dando lugar al desarrollo de las tecnologías de secuenciación de nueva generación o *Next Generation Sequencing* (NGS). Pertenecen a esta generación los primeros secuenciadores de alto rendimiento (*high-throughput*), que gracias a la inmovilización de las reacciones de secuenciación en una superficie sólida, son capaces de producir miles

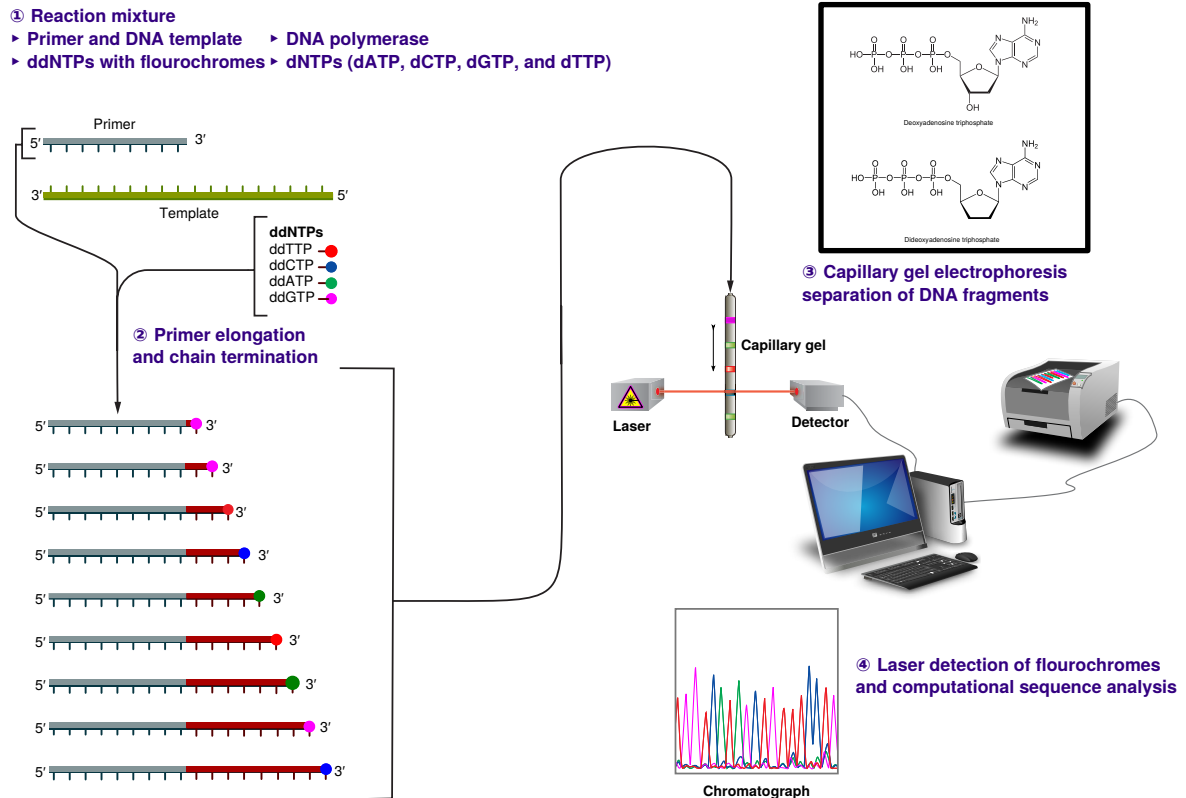


**Figura 1.3:** Método de secuenciación por didesoxinucleótidos de Sanger. Explicada en el texto. (Tomada de: [https://en.wikipedia.org/wiki/Sanger\\_sequencing](https://en.wikipedia.org/wiki/Sanger_sequencing))

de reacciones en paralelo, con el consiguiente ahorro de reactivos y abaratamiento del coste por base leída. Tienen además la ventaja de que son capaces de secuenciar el DNA sin necesidad de clonarlo, evitando así todo lo que ello conlleva (creación de genotecas, replicación, purificación, conservación...). Podríamos decir que se juntan las ventajas de los métodos de Sanger y de Maxam-Gilbert, con alto rendimiento y menos riesgos para la salud. No obstante, cabe destacar que, aunque las tecnologías NGS desplazaron a las de primera generación, la secuenciación de Sanger se sigue usando todavía a día de hoy y es la referencia (*gold standard*) para la secuenciación de alta calidad de fragmentos cortos con múltiples aplicaciones.

**Pirosecuenciación (454/FLX de Roche)** Fue la primera técnica NGS introducida. Es un método basado en el principio de secuenciación por síntesis, en el cual, una hebra simple de DNA es secuenciada mientras se sintetiza enzimáticamente la hebra complementaria. Consiste en la detección de luz en una reacción en la que se libera pirofosfato cada vez que la DNA polimerasa incorpora un nucleótido, de ahí el nombre de pirosecuenciación [10]. La hebra molde está inmovilizada e hibridada a un cebador y se van añadiendo y quitando secuencialmente cada uno de los cuatro nucleótidos. Solo se produce luz cuando se añade el nucleótido complementario (figura 1.4). La luz producida se detecta en una cámara y se genera un flujograma o pirograma que será analizado por un programa que deducirá la secuencia.

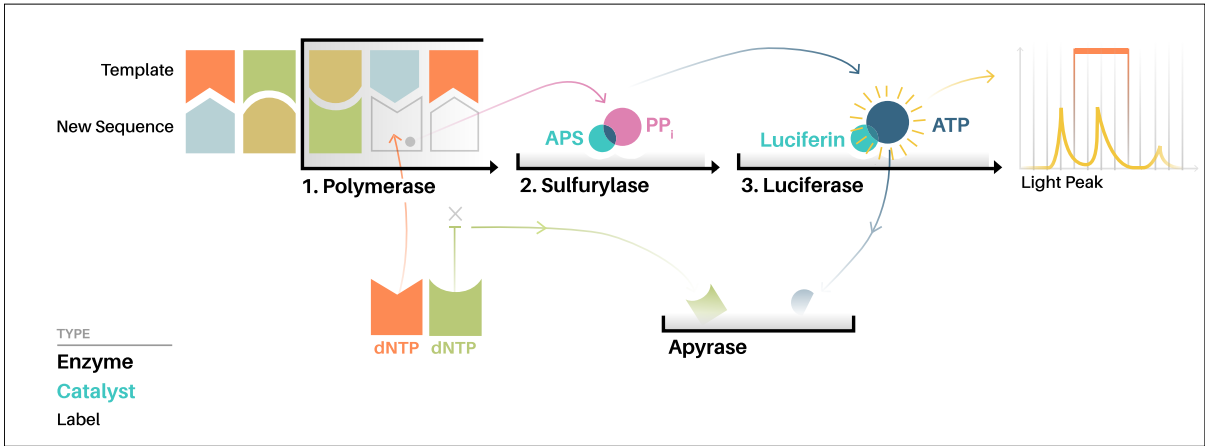


**Figura 1.4:** Método de secuenciación por pirosecuenciación. Tras la amplificación por PCR de la molécula a secuenciar, cada molécula simple de DNA se hibrida con un cebador y se incuba con las enzimas DNA polimerasa, ATP sulfurilasa, luciferasa y apirasa y los sustratos adenosina 5'-fosfosulfato (APS) y luciferina. (1) Si el nucleótido añadido es el que necesita la polimerasa, se incorpora a la cadena en elongación y se libera pirofosfato. (2) La ATP sulfurilasa transforma el pirofosfato (PPi) en ATP que actúa como sustrato de (3) la luciferasa, que producirá oxiluciferina y una señal luminosa. La apirasa degradará los nucleótidos que no hayan sido incorporados. Una cámara captura la señal lumínica que se produce cada vez que se incorpora un nucleótido y se traducirá en un pico en un pirograma. La secuencia será deducida a partir de los picos obtenidos. (Tomada de: <https://en.wikipedia.org/wiki/Pyrosequencing>)

**Tabla 1.1:** Datos técnicos de las tecnologías de secuenciación más destacadas de segunda y tercera generación. Información extraída de [11, 12, 13, 14] y de las webs oficiales de las distintas casas comerciales. ND: no disponible.

| Generación | Compañía                    | Secuenciador     | Longitud de lectura | Nº de lecturas por carrera | Total de nucleótidos | Tiempo de carrera | Coste por Gb |
|------------|-----------------------------|------------------|---------------------|----------------------------|----------------------|-------------------|--------------|
| 2ª         | Roche 454                   | 454 GS-FLX       | 400 – 700 pb        | 1 M                        | 0,7 Gb               | 24 h              | 9 500\$      |
|            | Illumina                    | MiSeq            | 2 x 300 pb          | 25 M                       | 15 Gb                | 4 – 55 h          | 110\$        |
|            | Illumina                    | NextSeq          | 2 x 150 pb          | 400 M                      | 120 Gb               | 12 – 30 h         | 33\$         |
|            | Illumina                    | HiSeq X          | 2 x 150 pb          | 6 000 M                    | 1 800 Gb             | 3 días            | 7\$          |
|            | Illumina                    | NovaSeq          | 2 x 250 pb          | 20 000 M                   | 6 000 Gb             | 13 – 44 h         | 18\$         |
|            | Life Technologies (SOLiD)   | 5500xl           | 35 – 75 pb          | 400 M                      | 30 Gb                | 2 – 7 días        | 70\$         |
|            | Thermo Fisher (Ion Torrent) | Ion Proton       | 200 pb              | 60 – 80 M                  | 10 Gb                | 2 – 4 h           | 80\$         |
| 3ª         | Helicos BioScience          | Heliscope        | 25 – 45 pb          | 555 M                      | 25 Gb                | 8 días            | ND           |
|            | Pacific Biosciences         | Sequel II System | 30 000 – 100 000 pb | 0.5 – 0.66 M               | 20 – 50 Gb           | 30 h              | 80\$         |
|            | Oxford Nanopore             | MinION           | 10 000 pb           | 3 M                        | 10 – 30 Gb           | 2 días            | 750\$        |

Los secuenciadores basados en la pirosecuenciación supusieron el primer sistema de secuenciación verdaderamente escalable y paralelizable, con mucha más capacidad y eficiencia que la secuenciación Sanger. Pueden llegar a generar hasta un millón de lecturas que originariamente eran de 25 nt, y que más adelante alcanzaron los 400 a 700 pb en 24 horas (tabla 1.1). Estas lecturas largas resultan muy útiles cuando se trabaja en el ensamblaje *de novo* de genomas complejos de organismos con numerosas y extensas regiones repetitivas, como es el caso de plantas [15, 16]. Cuando esto ocurre, el ensamblaje se complica considerablemente, pues, al fragmentar la región repetitiva para secuenciarla, muchas de las lecturas resultantes son iguales o muy similares, lo cual puede ocasionar que las secuencias de la misma repetición colapsen en una única repetición, impidiendo resolver correctamente la repetición,

o bien podrían acabar conectando dos fragmentos realmente distantes en el genoma, formándose las denominadas quimeras [17].

Sin embargo, este método de secuenciación tiene el inconveniente de que comete errores en las secuencias homopoliméricas, ya que la incorporación de varios nucleótidos del mismo tipo consecutivos genera un destello de tal intensidad que impide al sistema de detección determinar con precisión el número exacto de nucleótidos iguales que se han incorporado [18].

La tecnología fue inicialmente desarrollada por la empresa 454 Life Sciences<sup>4</sup>, que posteriormente, en 2007, fue absorbida por Roche. Debido a que el resto de tecnologías NGS siguió evolucionando más exitosamente, Roche canceló en 2016 el soporte de toda la gama de plataformas FLX por dejar de ser competitiva.

**Solexa/Illumina** Esta técnica también se basa en el principio de secuenciación por síntesis, pero con la particularidad del uso de unos dNTPs terminadores reversibles y DNA polimerasas modificadas [19]. Fue inicialmente desarrollada por la empresa Solexa, creada en 1998, que ocho años después adquirió Illumina.

En esta tecnología, el DNA se fragmenta mediante el uso de transposasas o nebulización y se le añaden una serie de cebadores, índices (para diferenciar distintas muestras) y adaptadores que la técnica requiere (figura 1.5A). Las moléculas de DNA y los cebadores se fijan sobre una superficie sólida o chip especializado (*flow cells*) donde tendrán lugar tanto la amplificación como la secuenciación. Los fragmentos anclados los amplifica una polimerasa mediante una PCR de tipo puente cuyo objetivo es crear cientos de fragmentos idénticos o *clusters* de DNA que se secuenciarán posteriormente (figura 1.5B). Para determinar la secuencia se añaden desoxinucleótidos terminadores reversibles marcados con fluorescencia, de forma que cuando uno de ellos se incorpora a la secuencia, la reacción se detiene y una cámara toma imágenes con las que determinar qué nucleótido se ha unido en esa posición, en función de la longitud de onda y de la intensidad de la emisión (figura 1.5C). Los nucleótidos que no se incorporen son retirados de la reacción. Una vez identificado el nucleótido, se elimina el agente bloqueante del extremo 3' que impedía proseguir la síntesis y puede comenzar el siguiente ciclo, con lo que solo se añade un nucleótido por ronda. Con esta tecnología se pueden secuenciar miles de posiciones en paralelo de forma masiva.

Illumina ha comercializado diversos tipos de plataformas de secuenciación aumentando mucho sus capacidades, y sigue en constante evolución, siendo a día de hoy la NGS más usada. La mayoría de ellas producen lecturas pareadas (apartado 1.1.2) de pequeño tamaño, de  $2 \times 150$  pb, aunque puede llegar a  $2 \times 250$  pb en el secuenciador NovaSeq, e incluso  $2 \times 300$  pb en MiSeq (tabla 1.1). Pero, en lo que más ha evolucionado la tecnología es en el número total de lecturas capaz de producir en una sola reacción, que puede ir desde 4 millones de lecturas (iSeq 100 System) hasta los 20 000 millones (NovaSeq 6000 System)<sup>5</sup>, lo cual se traduce en altos valores de profundidad de secuenciación o cobertura (esto es, el número medio de veces que se secuencia cada nucleótido de la secuencia) y un considerable abaratamiento del coste de la secuenciación. Además, resuelve mejor las regiones homopoliméricas y la tasa de error por lectura es inferior a la obtenida con otras tecnologías como 454 o Ion Torrent [20, 12]. Por todo esto, es la tecnología más empleada a la hora de secuenciar genomas completos a bajo coste, o en estudios de expresión génica, o para búsqueda de variantes en estudios de población.

Sin embargo, presenta problemas en el ensamblaje *de novo* de genomas completos, pues, al no ser secuencias muy largas, éste se hace costoso, además de la dificultad que supone la resolución de regiones

<sup>4</sup>[https://en.wikipedia.org/wiki/454\\_Life\\_Sciences](https://en.wikipedia.org/wiki/454_Life_Sciences)

<sup>5</sup><https://www.illumina.com/systems/sequencing-platforms.html>

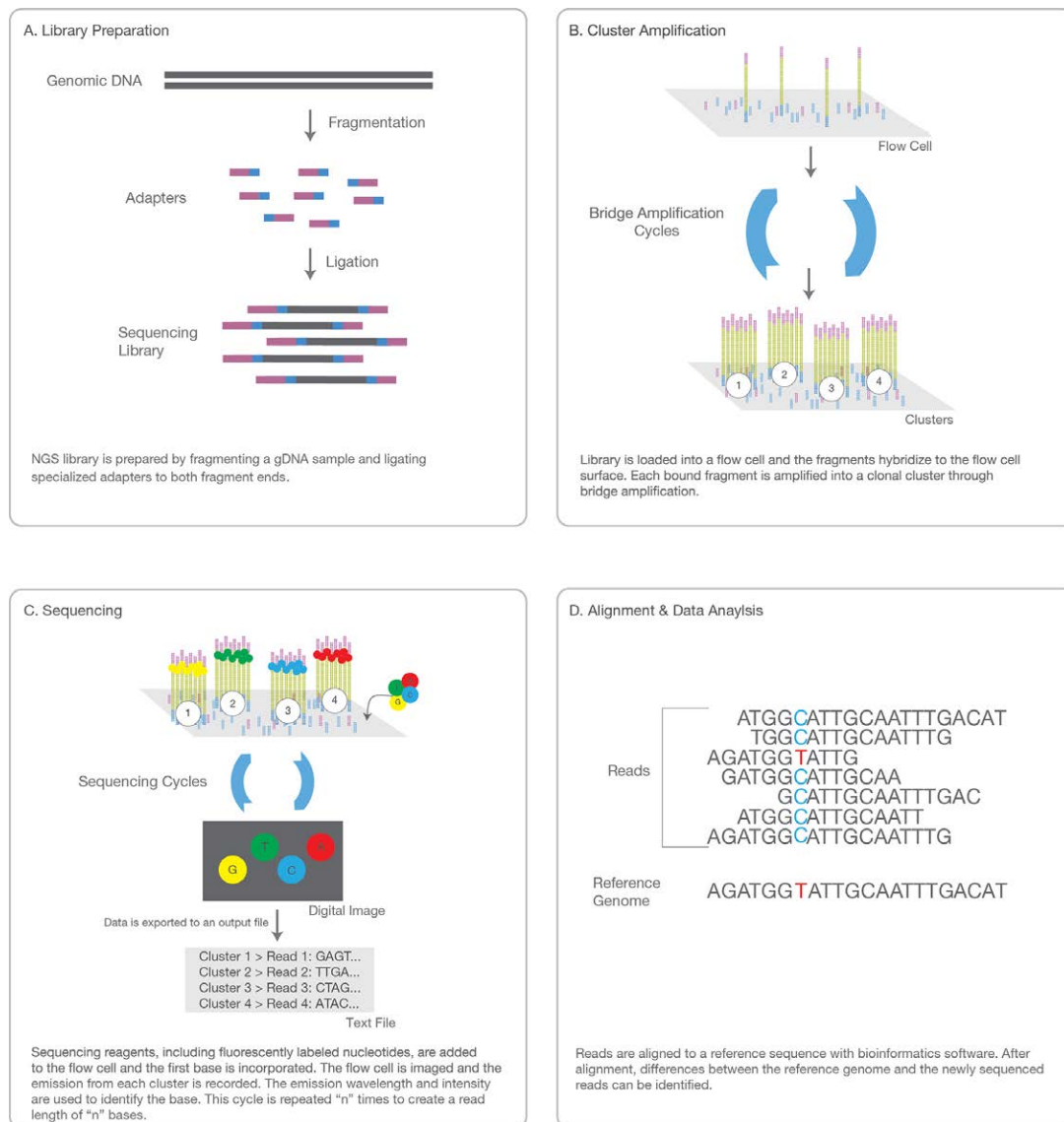


**Figura 1.5:** Método de secuenciación de la tecnología Illumina. Más detalles en el texto. (Tomada de: <https://emea.illumina.com/company/news-center/multimedia-images.html>)

repetitivas como se comentaba anteriormente. Otro inconveniente es la tasa de error en las lecturas, debidos a la pérdida de calidad en los extremos 3' de las lecturas, por imprecisiones de las polimerasas y sesgos en la amplificación por PCR entre otros [21, 22].

**SOLiD** La tecnología de secuenciación SOLiD (del inglés *Sequencing by Oligonucleotide Ligation and Detection*) [23], al contrario que los métodos anteriores basados en polimerización, utiliza un método enzimático denominado secuenciación por ligación. En esta técnica, mediante sonicación o disrupción física se prepara una librería de fragmentos de la muestra de DNA que queremos secuenciar y se disponen

sobre perlas magnéticas. Se lleva a cabo una PCR por emulsión para amplificar la muestra y a continuación se depositan sobre una superficie sólida para llevar a cabo el proceso de secuenciación. Para ello se dispone de una colección de octámeros marcados, con todas las combinaciones de secuencia posibles, que se irán ligando por complementariedad mediante una DNA ligasa, con la posterior detección de señal fluorescente emitida tras cada ligación.

Al igual que Illumina, resuelve mejor las regiones homopoliméricas que 454. Sin embargo, tiene problemas a la hora de secuenciar palíndromos [24]. Además, al generar lecturas cortas (75 pb a lo sumo), no son las más indicadas para ensamblajes *de novo*. La tecnología fue rápidamente superada por Illumina en cuanto a longitud de lectura y volumen de datos generado, y por ende, también en cuanto a precio (tabla 1.1).

La tecnología fue desarrollada por la empresa Life Technologies y adquirida por ThermoFisher en 2014, aunque justo al año siguiente comenzaron a dejar de dar soporte a este tipo de secuenciadores, recomendando el uso de la tecnología Ion Torrent como alternativa<sup>6</sup>.

**Ion Torrent** La tecnología de Ion Torrent consiste también en un método de secuenciación por síntesis, pero basado en la detección de iones hidrógeno (variaciones de pH) que se liberan durante la polimerización del DNA [25]. La molécula de DNA a secuenciar se introduce en un micropocillo y se inunda con un único tipo de nucleótido. Si el nucleótido en cuestión es el que corresponde añadir a la molécula de DNA, se incorpora en ella, liberándose un ión hidrógeno que será detectado por un sensor nanotecnológico. Si el nucleótido no es el complementario que correspondería añadir, no se incorpora y no tiene lugar ninguna reacción. Después de cada ciclo, se lavan los nucleótidos para añadir otro diferente y repetir el proceso. Si hay dos o más nucleótidos iguales consecutivos, se incorporan todos en un único ciclo, con la liberación de tantos iones hidrógeno como nucleótidos se hayan adicionado y produciendo una señal mayor.

El hecho de no emplear nucleótidos modificados y que el método de detección sea eléctrico y no óptico, abarató mucho los costes en esta tecnología e hizo el proceso de secuenciación mucho más rápido y escalable. En cuanto a limitaciones de la técnica, comparte el problema de los homopolímeros de la pirosecuenciación. De nuevo, al ser secuencias cortas (de hasta 200 pb en el secuenciador Ion Proton<sup>7</sup>; tabla 1.1), no están aconsejadas para ensamblajes *de novo*.

Esta tecnología fue desarrollada por la empresa Ion Torrent Systems Inc., y puesta en el mercado en 2010. La empresa fue posteriormente adquirida por Life Technologies.

### Tercera generación: secuenciación de una única molécula

En la búsqueda por abaratar aún más los costes de secuenciación y aumentar la fiabilidad de las secuencias obtenidas, surgen las plataformas de secuenciación de tercera generación. El denominador común de estas tecnologías es la secuenciación de una única molécula de DNA (*single molecule real time sequencing*), pues no requieren PCR previa a la secuenciación, lo cual acorta el tiempo de reacción, además de evitar los artefactos generados por la PCR y el sesgo en GC [13]. La señal se captura en directo.

<sup>6</sup><https://www.thermofisher.com/order/catalog/product/4460730>

<sup>7</sup><https://www.thermofisher.com/order/catalog/product/4476610>

**Helicos, PacBio y Nanopore** La compañía Helicos BioSciences fue una de las primeras en poner en el mercado un ejemplo de tecnología de tercera generación. En su metodología [26] la muestra de DNA original, sin amplificar, se fragmenta y se le añaden adaptadores poli-A con los que se fija a una superficie sólida o *flow cell*. La secuenciación se produce mediante lavados cíclicos cada vez con un único nucleótido marcado fluorescentemente. Las lecturas se van generando en un secuenciador Heliscope. Se generan lecturas cortas de 35 pb (tabla 1.1), y, a pesar de que se publicó un genoma humano secuenciado con esta tecnología [27], la compañía quebró en 2012.

La empresa Pacific Biosciences por su parte desarrolló una metodología en la cual lo que está fijo en el fondo del pocillo es la DNA polimerasa sin modificar, con una única hebra molde de DNA. A esta unidad se le denomina ZMW (onda guía de modo cero), y hace que la luz pueda penetrar y crear un canal de visualización que permite monitorizar la actividad de la DNA polimerasa a nivel molecular. Cada pocillo presenta muchas de esas unidades en el fondo. El método de secuenciación es también por síntesis, empleando nucleótidos modificados que liberan un tipo de fluorescencia distinto dependiendo de la base, con lo cual no es necesario realizar lavados y el proceso se agiliza. Desde la puesta en el mercado de su primer secuenciador en 2010, la empresa ha ido evolucionando con el lanzamiento de distintos equipos (figura 1.6), con un aumento progresivo del número de ZMWs por pocillo y con versiones de fungibles químicos cada vez de mayor capacidad. Lo más destacable de esta metodología es la longitud de sus lecturas, que supera con creces las de las tecnologías anteriores de secuenciación. Así, a finales de 2018, la empresa lanzó una versión de reactivos para su secuenciador Sequel II System capaz de generar lecturas de 30 hasta 100 kpb (tabla 1.1). Esta longitud de lectura hace que la tecnología sea ideal para ensamblajes *de novo* de grandes genomas, combinándolo con otros tipos de lecturas con las que corregir su elevada tasa de error, pues su longitud permite resolver mucho mejor las regiones repetitivas, o también para terminar de resolver genomas previamente secuenciados [28]. La tecnología permite además detectar modificaciones en los nucleótidos como la metilación [29] (también abordable con tecnologías como Illumina), además de isoformas completas y variantes de ajuste (*splicing*) [30]. Su gran inconveniente sigue siendo su enorme tasa de error que se sitúa en el 10 %, por más que la compañía asegure que no supera el 3 %, por lo que es claramente más alta que la de sus predecesoras [31, 32], si bien es cierto que si se realizan varios pases de secuenciación sobre la misma molécula, es posible obtener una secuencia consenso de alta calidad.

La secuenciación basada en nanoporos [33] es el método más novedoso de las tecnologías secuenciación de tercera generación. De hecho, hay quienes lo clasifican como un método no de tercera, sino de cuarta generación, categoría que combinaría la secuenciación de molécula única con miniaturización (el secuenciador cabe en la palma de la mano y se enchufa al puerto USB de un portátil) [34]. Consiste en hacer pasar la hebra de DNA a través de un nanoporo y, en función de los cambios electrofisiológicos observados, debido a los iones que lo atraviesan, identificar qué nucleótidos están pasando [35], pues cada tipo de nucleótido bloquea en canal iónico durante un periodo de tiempo diferente. Esta aproximación fue desarrollada por la compañía Oxford Nanopore que tiene en el mercado por ejemplo su secuenciador MinION, muy económico y minimalista (ver figura 1.7), capaz de generar lecturas de una longitud media de 10 kpb (tabla 1.1). El método no requiere nucleótidos modificados y la preparación de la muestra es simple. En cuanto a limitaciones, y aunque está en camino de mejora, cabe mencionar que la tasa de error en la identificación de los nucleótidos sigue siendo alta, si bien es cierto que secuenciar varias veces la misma molécula puede mejorar la calidad de los datos [32].

## Tipos de lecturas

En los proyectos de secuenciación basados en las técnicas de NGS se usan dos tipos generales de lectura: individuales y pareadas. Las lecturas individuales, también llamadas simples (*single*), se generan



**Figura 1.6:** Secuenciadores Sequel System de Pacific Biosciences. (Tomada de: <https://www.pacb.com>)

secuenciando del extremo de un fragmento de DNA, tantos nucleótidos como permita la tecnología en cuestión. Este tipo de lectura no contiene más información que la secuencia en sí. Las lecturas pareadas, por su parte, se producen secuenciando ambos extremos de un fragmento de DNA o inserto de tamaño conocido, por lo que las dos lecturas generadas, al pertenecer al mismo fragmento, están relacionadas entre sí y además se conoce, como información adicional, la distancia aproximada que hay entre ellas, y por lo tanto la posición relativa que una debe tener respecto a la otra. Según el sentido en el que se lea el DNA se generan distintos tipos de lecturas pareadas: si es en sentido contrario y de forma convergente hablamos de lecturas pareadas de Illumina (*paired-end* según Illumina), si es en sentido contrario, pero de forma divergente, tendríamos lecturas compañeras (*mate-pair* según Illumina), mientras que si se leen en el mismo sentido tenemos lecturas pareadas 454 (*paired-end* según 454/Roche) (figura 1.8).

Es esencial conocer que existen distintos tipos de librerías en función de si se conserva o no la información sobre la hebra que se secuencia. En las librerías sin direccionalidad conocida (llamadas *unstranded*), la información sobre la hebra no se conserva, se pierde durante el proceso de amplificación de los fragmentos, mientras que en las librerías *stranded* sí se mantiene dicha información. Las librerías *stranded* pueden ser *first strand* (la segunda lectura corresponde a la hebra original, mientras que la primera es de la hebra complementaria), o bien *second strand* (es la primera lectura la que corresponde a la hebra original).

El concepto de la direccionalidad es importante sobre todo cuando se llevan a cabo análisis de expresión por RNA-seq (ver apartado 1.1.2), pues si este parámetro se usa erróneamente, se pueden asignar las lecturas a la hebra equivocada, que puede, o bien no contener genes, o contener un gen solapante. En ambos casos, la estimación sería errónea.



**Figura 1.7:** Secuenciador MinION USB, de Oxford Nanopore. (Tomada de: <https://www.extremetech.com>)

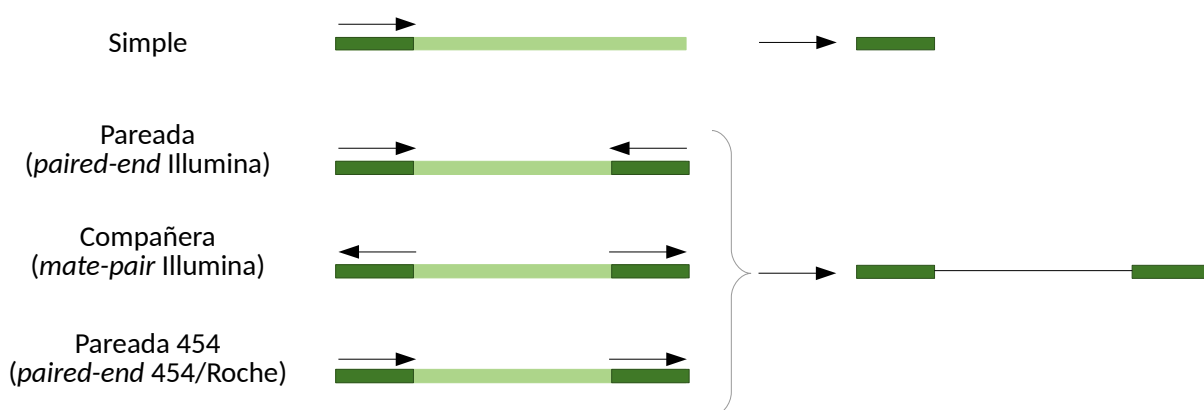


**Figura 1.8:** Tipos de lectura empleados en la secuenciación por NGS. Más detalles en el texto.

El uso de lecturas pareadas ayuda a resolver o al menos mejorar muchos problemas en bioinformática. En los ensamblajes de transcriptomas o genomas permiten conectar *contigs* entre sí para generar *scaffolds* (ver figura 1.9). Por otra parte, como ya se comentó previamente en el apartado 1.1.2, el ensamblaje de las regiones repetitivas del genoma supone un problema, ya que las lecturas tienden a colapsarse sobre una única repetición. Sin embargo, con el uso de lecturas pareadas, habrá casos en los que al menos un extremo de la pareada corresponderá al extremo flanqueante de la repetición, ayudando a mejorar su resolución. Incluso puede ocurrir que la repetición sea más pequeña que el tamaño del inserto, en cuyo caso se puede deducir el número de repeticiones en la región. La tendencia de las casas comerciales es aumentar cada vez más el tamaño del inserto para poder resolver regiones repetitivas mayores. El uso de lecturas pareadas también es muy útil a la hora de detectar distintas isoformas de un transcrito [36], variantes estructurales [37] o fusiones génicas quiméricas [38].

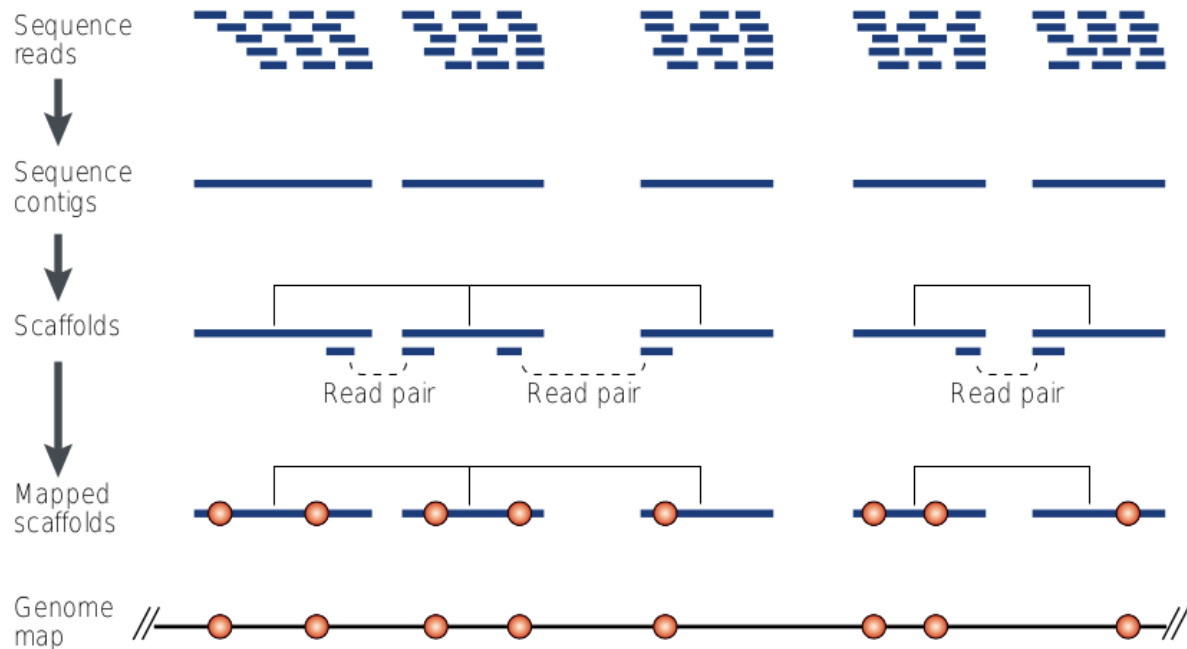


**Figura 1.9:** Representación de una estrategia de ensamblaje tradicional con lecturas simples y pareadas. Las lecturas simples se ensamblan entre sí para formar contigs. Los contigs se conectan entre sí para formar scaffolds gracias a las lecturas pareadas. Si se dispone de un mapa genético u otra referencia, los scaffolds se pueden ordenar entre sí. (Tomada de [39])

## Aplicaciones de las tecnologías de secuenciación

En función de si el tipo de molécula en el que se centra un proyecto de secuenciación es DNA (DNA-seq) o RNA (RNA-seq), las aplicaciones serán diferentes.

**Secuenciación de genomas *de novo* (DNA-seq)** La obtención por primera vez de la secuencia de nucleótidos del genoma de una especie (*de novo sequencing*) supone poder disponer de su genoma de referencia, gracias al cual es posible conocer no solo qué genes lo componen, sino además su estructura, función, localización y regulación dentro del mismo.

Para secuenciar un genoma *de novo* existen dos estrategias generales, en función del tamaño y la complejidad del genoma a ensamblar. La primera de ellas y más tradicional se denomina secuenciación jerárquica y se basa en digerir el genoma en fragmentos grandes solapantes que se clonan y se ordenan con respecto al genoma. Luego se secuencian por separado y finalmente se ensamblan para reconstruir el genoma. Con este método se simplifica el problema de las regiones repetitivas que comentábamos en el apartado 1.1.2, pero tiene el inconveniente de que la clonación y el ordenamiento de los clones es un proceso caro, lento y tedioso.

La segunda estrategia consiste en la secuenciación aleatoria del genoma completo (*whole genome shotgun sequencing*, WGS) por el cual el genoma se trocea en fragmentos de un tamaño concreto, que se secuenciarán y se ensamblarán computacionalmente para dar lugar a una secuencia consenso [40]. Entre sus inconvenientes está que no existe información previa que dé una idea de la localización aproximada

de las lecturas, con lo cual genera ensamblajes de peor calidad, llenos de huecos (*gaps* en inglés). Se suma además el inconveniente de las regiones repetitivas, aunque algunas de las tecnologías lo solventaban en parte con la generación de lecturas o bien pareadas, o bien de mayor longitud. Otro problema que puede surgir de la estrategia WGS es que, al ser una secuenciación aleatoria, algunas regiones del genoma pueden quedar sin secuenciar o pobremente representadas. Esto se resuelve secuenciando con una cobertura suficiente, algo a día de hoy asequible gracias a los reducidos precios de la mayoría de las NGS. Este menor coste y su mayor rapidez hicieron que el método WGS sustituyera al método jerárquico en las tecnologías de secuenciación de segunda y tercera generación, permitiendo desde secuenciar cromosomas completos hasta obtener el primer borrador del genoma humano [4].

Una de las múltiples aplicaciones de la obtención de un genoma de referencia en combinación con el uso de las NGS es la búsqueda de marcadores moleculares como SNP o SSR [41], muy útiles para estudiar la variación genética dentro de una población, o como indicadores de enfermedad, o para la selección de características fenotípicas de interés en agricultura.

En las especies no modelo, al hecho de no poseer un genoma de referencia se une el problema de la escasez de herramientas disponibles para el análisis de las secuencias en estos organismos. No obstante en los últimos años se está llevando a cabo un notable esfuerzo por desarrollar estrategias que permitan abordar el estudio de estos organismos no modelo [42, 43, 16].

**Resecuenciación (DNA-seq)** Cuando se dispone de un genoma de referencia, a ser posible de la misma especie en estudio, o, en su defecto, de alguna especie muy cercana, es posible llevar a cabo una resecuenciación, que consiste en secuenciar el genoma problema y alinear (o mapear, del inglés *mapping*) las lecturas sobre el genoma de referencia, para detectar identidades y diferencias entre ambos [44, 45, 46]. Posibilita por tanto identificar variaciones del tipo SNP, indel, CNV o SV de individuos o poblaciones de una especie, en definitiva, ampliar el conocimiento de la variación genética entre individuos.

La resecuenciación tiene la ventaja de que los algoritmos de mapeo son más precisos y rápidos que los de ensamblaje, y además no se necesita tanta profundidad de secuenciación o cobertura como en un ensamblaje *de novo*, con lo cual, disminuyen el coste económico, el tiempo y la dedicación.

A veces, en vez de resecuenciar el genoma completo, puede interesar al investigador secuenciar solo una región o regiones concretas del genoma, esto es, realizar una secuenciación dirigida (*targeted sequencing*) [47]. Así por ejemplo se pueden capturar solo determinados genes (panel de genes), lo cual puede interesar por ejemplo en el estudio de enfermedades, o bien el exoma completo. Esta simplificación permite a los investigadores secuenciar esas zonas con mayor profundidad a menor coste.

La resecuenciación favorece el aumento de la representatividad de cada especie en las bases de datos.

**Epigenómica (DNA-seq)** La epigenómica se encarga del estudio de los mecanismos que regulan la expresión génica y que no implican una modificación de la secuencia de DNA. Estos procesos incluyen, entre otros, la metilación del DNA, las modificaciones de las histonas y el papel de los ncRNA sobre la transcripción o la traducción. El acoplamiento de las tecnologías de NGS a la epigenómica también ha dado sus frutos.

Uno de los primeros métodos que se usaron para estudiar las interacciones DNA-proteína claves en cualquier proceso celular fue ChIP-on-chip, que combinaba las micromatrices con la inmunoprecipitación de cromatina. Posteriormente se aplicó NGS a los fragmentos purificados de DNA que se obtenían, dando lugar a la tecnología ChIP-seq, que genera mapas de interacción DNA-proteína con menos ruido de fondo y mayor resolución. Se ha usado, por ejemplo para localizar sitios de unión de factores de transcripción

en genes de proliferación y diferenciación celular [48].

En los organismos eucariotas, la metilación de las citosinas forma parte del proceso de regulación de la replicación y la transcripción del DNA. Para determinar los patrones de metilación (metiloma) se trata el DNA con bisulfito de sodio, que convierte las citosinas en uracilos, de forma que cuando se secuencian, las citosinas se secuencian como timinas, mientras que las metilcitosinas se siguen secuenciando como citosinas. La combinación de este tipo de secuenciación con NGS dio lugar a la técnica denominada secuenciación de amplicones con bisulfito (BSAS), que permite determinar de forma rápida y eficiente el metiloma de un genoma o región de interés [49].

**Metagenómica (DNA-seq)** La metagenómica consiste en el estudio genético de comunidades completas de organismos, mediante la secuenciación de todo el material genético presente en una muestra. Este enfoque proporciona una idea general de la constitución, complejidad y evolución de ecosistemas en todo su conjunto, y facilita el estudio de las causas y las consecuencias de los cambios ambientales [50]. El desarrollo de esta disciplina también se ha visto disparado con los avances y el abaratamiento de las NGS.

Un estudio metagenómico se puede abordar según dos aproximaciones distintas: secuenciando todo el material genético encontrado, o solo determinadas secuencias específicas uniformemente conservadas (como por ejemplo rRNA). Este segundo enfoque tiene varias ventajas. Por una parte, la complejidad de los datos es mucho menor y además permite un análisis de carácter cuantitativo, que resultaría muy complicado si se parte de todas las secuencias mezcladas de muchos organismos con tamaños de genoma muy dispares [51].

El enfoque de la metagenómica ha sido muy explotado en el estudio de comunidades de microorganismos. Por ejemplo, en casos de enfermedad, la comparación entre el DNA presente en una muestra control sana frente a una muestra de enfermedad puede servir para detectar DNA extra en la muestra problema que podría corresponder al agente infeccioso [52].

**Secuenciación de transcriptomas *de novo* (RNA-seq)** La obtención del transcriptoma proporciona información sobre los genes de un organismo a menor coste y esfuerzo que secuenciando su genoma, pues solo se secuenciarán las regiones que se están transcribiendo y además en estas regiones las secuencias repetitivas son escasas. Antiguamente, los trabajos de transcriptómica se basaban en secuenciación Sanger de secuencias etiquetadas por su expresión o EST, con la que solo se podían detectar los transcritos más abundantes. Por el contrario, las nuevas tecnologías con sus altos valores de profundidad de secuenciación permiten detectar hasta los transcritos más raros [53]. El desarrollo de la NGS ha permitido así obtener un transcriptoma de forma rápida y barata, lo que ha disparado la publicación de análisis transcriptómicos. En las especies no modelo, cuando no existe un genoma disponible, la obtención de su transcriptoma suele ser la primera opción a llevar a cabo [54, 55, 56, 57, 58].

Disponer de un transcriptoma permite catalogar todos los transcritos de una especie, no solo los mRNA, sino también los ncRNA [59], así como determinar la estructura transcripcional de los genes en términos de sitio de inicio de la transcripción, extremos 5' y 3', límites exón-intrón, patrones de ajuste (*splicing* en inglés) y otras modificaciones postranscripcionales [60]. Si además se dispone del genoma de la especie que se está analizando, es posible identificar nuevos transcritos, corroborar y mejorar supuestos transcritos ya descritos e identificar polimorfismos mononucleotídicos (SNP), variantes estructurales y nuevas formas de ajuste.

**Análisis de expresión génica (RNA-seq)** La transcriptómica también permite estudiar el nivel de expresión de los genes que se están transcribiendo en una célula o tejido durante el desarrollo, o en diferentes condiciones fisiológicas o patológicas.

Tradicionalmente, los análisis de expresión se llevaban a cabo mediante las llamadas micromatrices, *microarrays* o chips de DNA. Consisten en láminas de vidrio sobre las cuales se imprimen sondas que pueden ser oligonucleótidos o DNA complementario (cDNA). Las moléculas de mRNA que se quieren estudiar se convierten a DNA complementario mediante PCR inversa (RT-PCR). Posteriormente, se marcan con un fluoróforo y se ponen en contacto con la superficie del chip donde se hibridarán con las sondas por complementariedad. La intensidad de fluorescencia en cada punto del *microarray* será proporcional al nivel de expresión del gen en cuestión. Esta forma indirecta de cuantificar la expresión tiene además la limitación de que solo permite estudiar genes conocidos. En los últimos años está siendo casi totalmente sustituida por la secuenciación de RNA mediante tecnologías de nueva generación o RNA-seq, cuyos fundamentos técnicos se describen en el siguiente párrafo [61, 62].

En un análisis de expresión por RNA-seq se extraen en primer lugar las moléculas de mRNA de una célula o tejido en una condición determinada y se convierten en cDNA mediante RT-PCR, que posteriormente es amplificado y secuenciado. Las secuencias de las lecturas obtenidas se alinean sobre un genoma o transcriptoma y se cuantifican para saber a qué gen pertenecen y cuántas corresponden a cada gen. La clave de la técnica radica en que el número de secuencias que alinean sobre un gen es directamente proporcional al número de moléculas de mRNA que proceden de ese gen, y por tanto refleja el nivel de expresión de dicho gen en la condición dada [63]. Esta técnica permite detectar tanto genes conocidos como desconocidos y cuantificar su expresión de forma directa [64]. Cabe señalar que, cuando el objetivo es descubrir qué genes se están expresando, es necesario que la muestra de mRNA esté normalizada, para así aumentar la representación de los genes pocos expresados y disminuir la de los sobreexpresados [65]. Es necesario también tener en cuenta que los niveles de expresión de mRNA estimados pueden no corresponder exactamente a la cantidad de proteína presente en la célula, ya que pueden existir pequeñas alteraciones durante el proceso de traducción que den lugar a grandes cambios en los niveles finales de proteína.

Las tecnologías de secuenciación de tercera generación (apartado 1.1.2) van un paso más allá y permiten abordar la expresión de los genes a nivel de una sola célula mediante la secuenciación de RNA de células sueltas (scRNA-seq). Entre las múltiples aplicaciones de esta aproximación, hace posible la identificación de poblaciones celulares raras, como las que surgen en el desarrollo de un proceso tumoral y que no serían detectables de otra forma [66]. Puede ayudar también a dilucidar los procesos de diferenciación celular que tienen lugar durante el desarrollo embrionario y que parten de una única célula.

### 1.1.3. Obtención de un transcriptoma

A la hora de llevar a cabo un análisis bioinformático, hay una serie de etapas o consideraciones generales que son comunes tanto si la molécula de partida es DNA como RNA. Luego existen algunos aspectos específicos de cada tipo de análisis, en función de la finalidad del mismo (por ejemplo si es construir un genoma o un transcriptoma) y de otros muchos factores (como la disponibilidad o no de un genoma de referencia). Puesto que una de las tareas fundamentales de la presente tesis es llevar a cabo un análisis bioinformático del transcriptoma del olivo, esta sección se orientará hacia ese tipo de análisis, teniendo en cuenta las peculiaridades concretas de las plantas.

Los pasos a seguir en el ensamblaje de un transcriptoma están esquematizados en la figura 1.10. A partir de una muestra del tejido de interés se extrae el RNA mensajero o total y se purifica. Este RNA

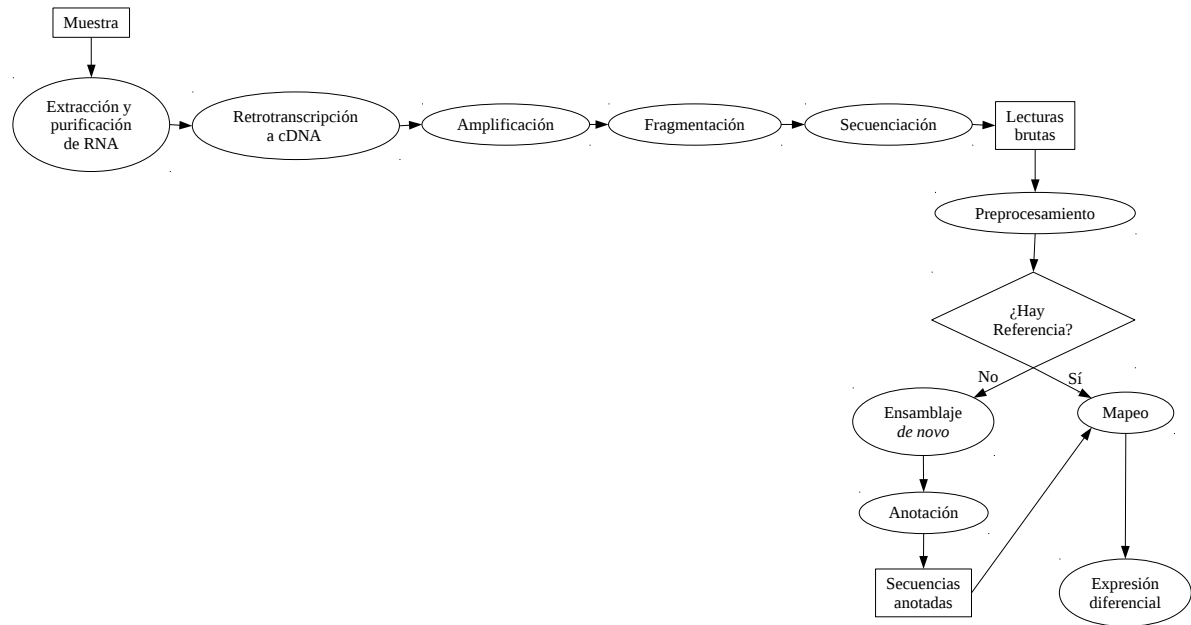


**Figura 1.10:** Etapas de la obtención y anotación de un transcriptoma. Los rectángulos indican entradas o salidas del flujo de trabajo, las elipses representan tareas a llevar a cabo y los rombos son puntos que implican una pregunta o decisión en el flujo. Más detalles en el texto.

se convierte a cDNA mediante RT-PCR. Si la cantidad de cDNA obtenida es muy escasa, se puede llevar a cabo un paso de amplificación por PCR. El cDNA se fragmenta en función del tipo de secuenciación que se elija y que proporcionará una gran cantidad de lecturas en bruto, normalmente en formato de texto FASTQ, que incluye tanto la secuencia de los nucleótidos como su calidad. Para evitar errores en las siguientes etapas del análisis, se lleva a cabo un preprocesamiento de las lecturas en bruto, que consiste en evaluar su calidad y retirar aquellos fragmentos de baja calidad, así como adaptadores de secuenciación y elementos contaminantes. En el caso de disponer de un transcriptoma o un genoma de referencia, las lecturas preprocesadas se mapean sobre él para identificar los transcritos que se están expresando en la muestra y las diferencias con la referencia. Si no se dispone de una referencia, las lecturas preprocesadas se ensamblan *de novo* para reconstruir las secuencias de los transcritos originales. Los transcritos obtenidos se anotan mediante comparación con bases de datos públicas. La información generada puede incluirse a su vez en una base de datos para hacerla fácilmente accesible. En ambos casos es posible llevar a cabo un análisis de expresión para determinar qué genes se están transcribiendo y cuantificar su expresión.

Para organizar la toma de muestras, no debemos olvidar que, como comentamos en el apartado 1.1.1, el transcriptoma es algo dinámico, de forma que el conjunto de genes que se transcribe varía en función del tipo celular o tejido de que se trate y a su vez de en qué estadio de desarrollo o bajo qué condiciones ambientales se encuentren esas células, con lo cual, si el objetivo final es conseguir el transcriptoma completo de un organismo, sería necesario muestrear cuantos más tipos de tejido y bajo más condiciones ambientales variadas o etapas de desarrollo distintas, mejor, para tener la certeza de que todos los genes están representados.

El interés por los ncRNA ha aumentado en los últimos años, ya que cada vez hay más pruebas de la importante función reguladora que llevan a cabo [67], por lo que también se están desarrollando protocolos específicos para su extracción, purificación y secuenciación. Para su detección, es imprescindible que se utilicen librerías *stranded* (apartado 1.1.2), puesto que muchos son complementarios de transcritos reales.

Aunque tradicionalmente las tecnologías de secuenciación más usadas en transcriptómica han sido 454/Roche e Illumina, cada vez son más los transcriptomas que se obtienen exclusivamente o en combinación con las tecnologías PacBio y Nanopore [68, 69, 70]. A la hora de elegir una u otra se debe tener en cuenta cuál es el objetivo final. En el caso de querer ensamblar un transcriptoma *de novo*, aunque la tendencia más habitual es utilizar lecturas pareadas de Illumina por su bajo coste, es recomendable el empleo de lecturas largas como las de 454, PacBio o Nanopore, que facilitan el ensamblaje. Si por el contrario, se dispone de un genoma o un transcriptoma de referencia, la secuenciación con Illumina es una buena elección para obtener lecturas que mapear sobre dicha referencia, con una alta cobertura. Cabe destacar también que el gran volumen de lecturas cortas que genera la tecnología Illumina resulta extremadamente potente para, entre otras cosas, corregir la secuencia consenso que se genera tras el ensamblaje, detectar formas de ajuste alternativo [71, 72], realizar análisis de expresión génica y para corregir las secuencias de molécula única. La tecnología probablemente más adecuada, aunque bastante más cara, para obtener transcritos completos y estudiar de forma precisa el ajuste alternativo es PacBio Iso-seq<sup>8</sup>.

En relación a las plantas, en las últimas décadas se han publicado numerosos estudios transcriptómicos de muchas especies para las que no existe genoma de referencia (especies no modelo), pero que han permitido generar información bastante útil [53]. No obstante, existen determinadas características de las plantas que pueden complicar este tipo de análisis. Los distintos alelos de un gen y las duplicaciones génicas pueden complicar el ensamblaje de un transcrito, pues resulta difícil distinguir entre errores de secuenciación, heterocigosidad, genes duplicados o distintas isoformas generadas por ajuste alternativo. En las especies de plantas modelo, este problema se ve reducido ya que suele tratarse de líneas que se autofecundan, por lo que presentan un alto grado de homocigosidad. En las especies de plantas no modelo, sin embargo, suele darse la polinización cruzada, lo que genera altos niveles de heterocigosidad, que complican la distinción entre verdaderos SNP y errores de secuenciación [53], si bien, unos buenos valores de los parámetros de calidad y de frecuencia esperable para esos SNP pueden ayudar a solventar el problema. Los grandes eventos de duplicación a lo largo de la evolución típicos de las angiospermas y responsables de su poliploidía [73] y de la presencia de genes parálogos añade dificultad al ensamblaje. Cuando el evento de duplicación ha sido reciente, y las secuencias de los genes parálogos no han divergido lo suficiente, puede haber problemas a la hora de reconstruir los transcritos, pues es difícil diferenciar qué lecturas corresponden a cada parálogo [53].

A continuación se explican con más detalle algunas de las etapas del análisis de un transcriptoma.

## Preprocesamiento

El preprocesamiento de las lecturas en bruto consiste en evaluar la calidad y fiabilidad de las mismas, con el objetivo de detectar y eliminar:

- Fragmentos de vectores, adaptadores y etiquetas propios de la técnica de secuenciación.
- Fragmentos o lecturas completas por baja calidad o complejidad.
- DNA procedente de orgánulos (cloroplastos y mitocondrias) y rRNA.

<sup>8</sup><https://www.pacb.com/applications/rna-sequencing/>

- Colas poliA/poliT.
- Secuencias contaminantes. Las más habituales suelen proceder de hongos y bacterias.

La finalidad es conservar la parte informativa más larga posible de las lecturas y con ello evitar posibles errores posteriores en el ensamblaje o en el alineamiento de secuencias [74]. Hay ocasiones en las que el número final de lecturas útiles tras el preprocesamiento es tan bajo que resulta conveniente plantearse incluso repetir la secuenciación de la muestra. Esto puede ocurrir por múltiples razones, como pueden ser una contaminación de la muestra, o errores graves durante el proceso de secuenciación y detectarlas antes de proseguir con el análisis puede suponer mucho ahorro de tiempo y dinero al investigador. Para ello, se han descrito en la bibliografía muchos programas para llevar a cabo esta tarea, algunos específicos de cada tecnología de secuenciación [75, 76, 77] y otros adaptables a todo tipo de tecnologías, como es el caso de SeqTrimNext [74], desarrollado en nuestro laboratorio.

## Ensamblaje

El propósito del ensamblaje es reconstruir la secuencia original de nucleótidos de los transcritos a partir de las lecturas útiles que quedan tras el preprocesamiento. Los programas de ensamblaje organizan el trabajo en las siguientes fases [78]:

**Filtrado de las lecturas** El objetivo de esta etapa es corregir o eliminar lecturas con errores de secuenciación antes de comenzar el ensamblaje, lo cual facilitará el proceso y evitará posteriores ensamblajes erróneos. El proceso de corrección puede ir desde recortar la lectura en función de unos valores de calidad, hasta un proceso más complejo de corrección basado en la frecuencia de las lecturas erróneas, concretamente en el hecho de que las lecturas poco frecuentes suelen originarse por errores de secuenciación. Por ello, hay algoritmos de corrección que identifican estas lecturas poco frecuentes y las comparan con las más frecuentes para decidir si se corrigen o por el contrario se eliminan. Esta decisión depende del tipo de ensamblador.

**Construcción de grafos** En este paso, las lecturas que solapan totalmente, o bien parcialmente, por sus extremos se combinan entre sí para formar lecturas contiguas más largas denominadas *contigs* (apócope de *contiguous sequences* [79]). La conexión entre las lecturas da lugar a un grafo, que describe las relaciones de similitud que hay entre ellas. Desde el punto de vista matemático, un grafo está compuesto por nodos o vértices y aristas. En un ensamblaje, los nodos del grafo representan las lecturas o fragmentos de lectura, mientras que las aristas serían las conexiones existentes entre esas lecturas. El objetivo de esta etapa consiste en comparar todas las lecturas entre sí para construir un grafo. En este paso es posible que se corrijan más errores de secuenciación. Hay distintas estrategias para construir los grafos que se comentan a continuación:

- **OLC**

La estrategia solapamiento-composición-consenso (del inglés *Overlap-Layout-Consensus*) fue la primera que se usó por ser muy intuitiva [80]. En primer lugar se comparan todas las lecturas entre sí para determinar todos los solapamientos posibles. Luego, esta información se organiza en un grafo donde los nodos corresponden a las lecturas completas y la aristas a los solapamientos entre ellas. El paso final consiste en determinar el camino más corto que recorre el grafo visitando cada nodo una sola vez (camino hamiltoniano), que da lugar a la secuencia consensuada. Un problema de esta

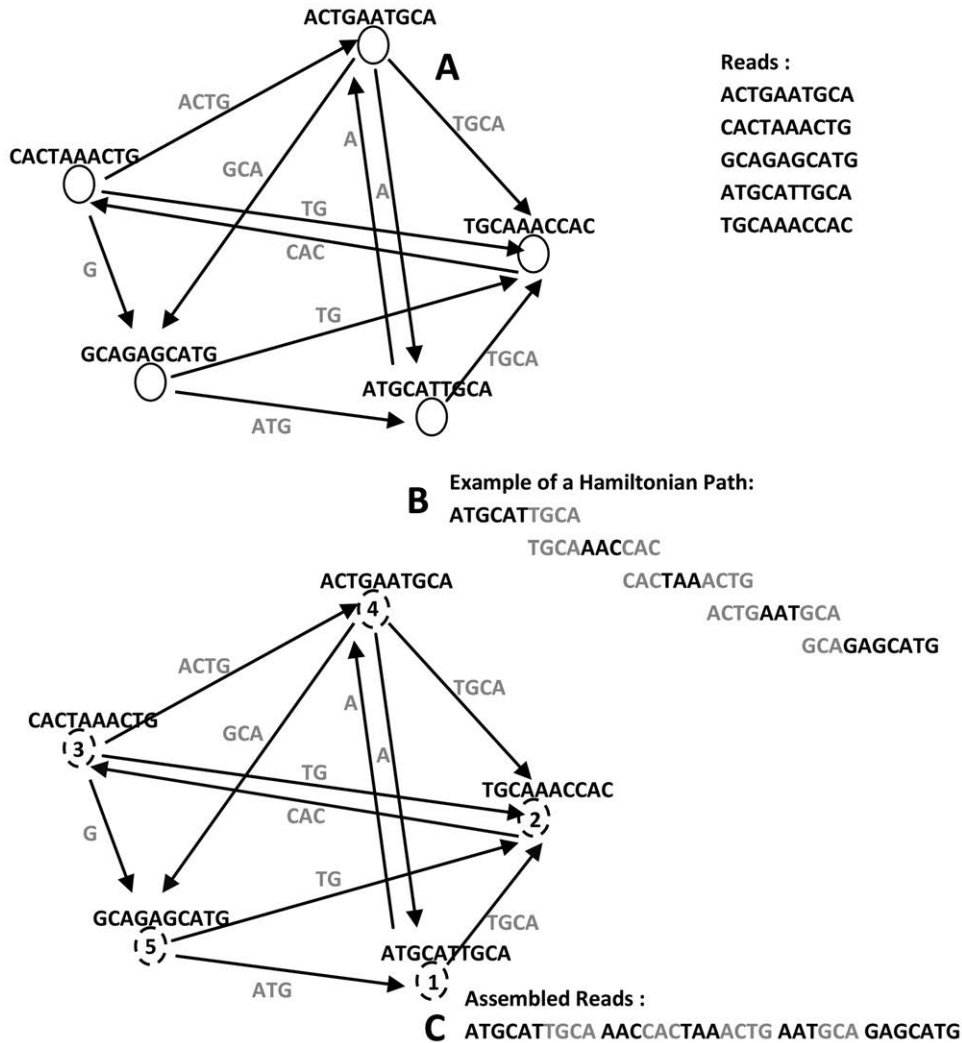


**Figura 1.11: Representación de un grafo generado por un ensamblador de tipo OLC (solapamiento-composición-consenso).** (A) Grafo donde los nodos son las lecturas completas y las aristas los solapamientos entre ellas. (B) Camino más corto que recorre todos los nodos visitando cada nodo solo una vez (camino hamiltoniano). (C) Ensamblaje resultante del recorrido de los nodos descrito en B. Los números representan el orden en el que se visitan los nodos. (Tomada de [78])

estrategia es que, conforme aumenta la cantidad de lecturas, el número de comparaciones lo hace exponencialmente, lo cual requiere mucha memoria RAM además de mucho más tiempo.

Esta estrategia comenzó usándose para ensamblaje de lecturas Sanger y luego fue adaptada para NGS. Suele emplearse en ensamblajes *de novo*. Algunos ensambladores que siguen esta estrategia son CAP3 [81], MIRA3 [82] y Newbler [83].

#### ■ Espectro de $k$ -meros

Esta aproximación parte las lecturas en todos los fragmentos (o  $k$ -meros) solapantes de una determinada longitud fija, de forma que cada fragmento está desplazado un nucleótido con respecto al

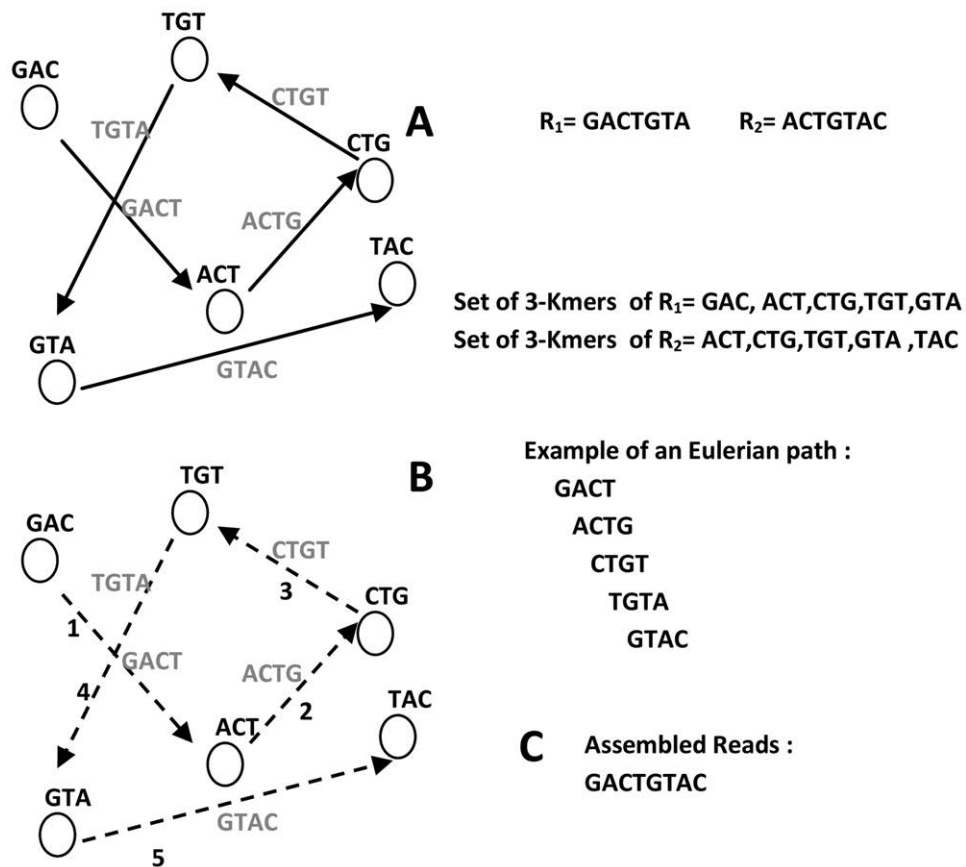


**Figura 1.12: Representación de un grafo generado por un ensamblador de tipo de Brujin.** (A) Grafo de de Brujin donde los nodos son los  $k$ -meros y las aristas son los solapamientos de longitud  $k - 1$  entre los nodos adyacentes. (B) Ejemplo de camino euleriano que recorre cada arista del grafo únicamente una vez (los números de las flechas discontinuas representan el orden del recorrido). (C) Resultado de las lecturas ensambladas tras recorrer el grafo. (Tomada de [78])

anterior según la lectura original. Esta fragmentación de las lecturas evita tener que comparar todas las secuencias entre sí. En este caso cada nodo del grafo corresponde a un  $k$ -mero y cada arista representa un solapamiento de longitud  $k - 1$  entre nodos. Una adaptación de esta estrategia utiliza grafos de de Brujin modificados [84], en los que las aristas corresponden a los  $k$ -meros y los nodos a la secuencia de solapamiento entre dos  $k$ -meros adyacentes. La secuencia original se reconstruye mediante la búsqueda de caminos eulerianos que recorren cada arista solo una vez (figura 1.12).

Partir las lecturas en  $k$ -meros conlleva la necesidad de almacenamiento y algoritmos potentes, aunque no necesitan tanta RAM como los grafos de tipo OLC. Además, este tipo de grafo es muy sensible al valor del parámetro  $k$ . La longitud del  $k$ -mero debería ser lo suficientemente grande como para que las secuencias solapen, pero no demasiado pequeño como para que se encuentren solapamientos por azar. Para escoger su valor se tienen en cuenta el tamaño y la cobertura de las lecturas y el tipo de errores de secuenciación esperables.

Muchos ensambladores de lecturas cortas de Illumina siguen esta estrategia. Algunos de ellos son EULER-SR [84], Velvet [85], ABySS [86], SOAPdenovo [87] y Trinity [88].

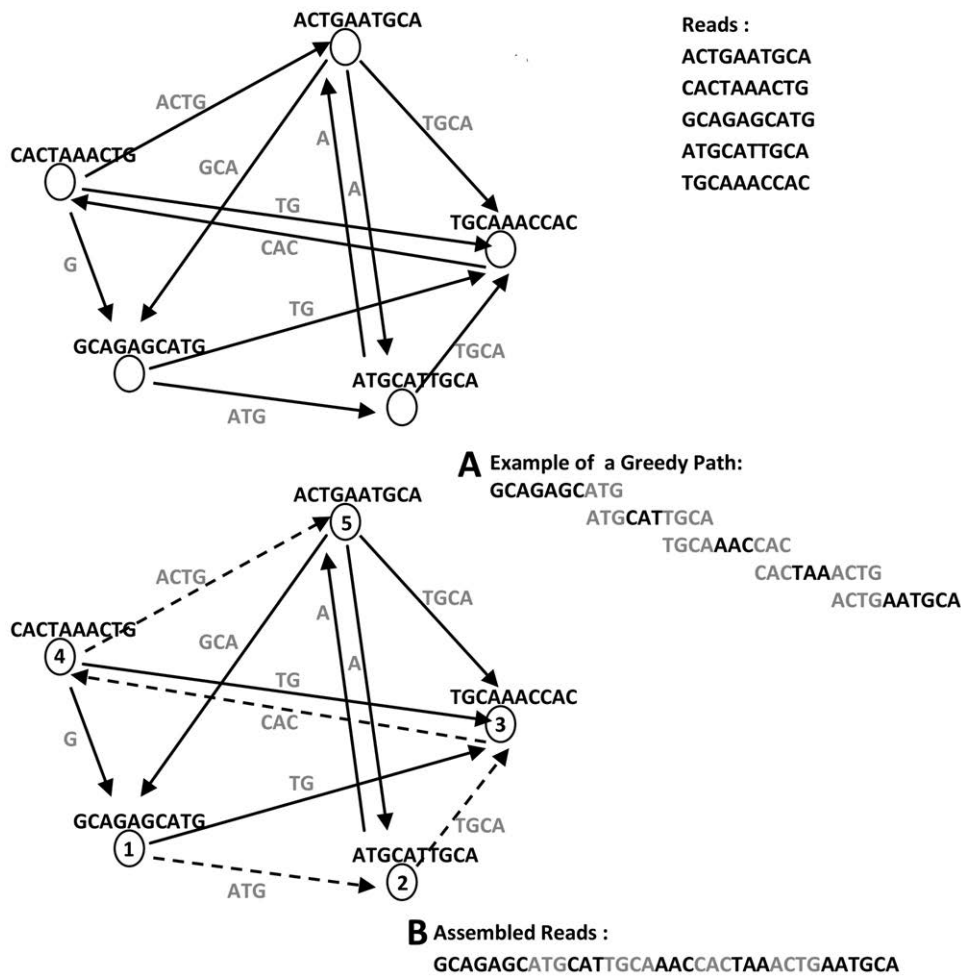


**Figura 1.13: Representación de un grafo generado por un algoritmo voraz.** (A) Los nodos del grafo corresponden a las lecturas y las aristas al solapamiento entre ellas. (B) A partir de un nodo inicial elegido aleatoriamente (el nodo 1 en el ejemplo) el algoritmo evalúa el solapamiento con los nodos vecinos y se decanta por el que tenga mayor solapamiento para ir trazando el camino de Hamilton (flechas discontinuas) y generar la secuencia final resultante del ensamblaje. (Tomada de [78])

#### ■ Algoritmos voraces

Es una aproximación heurística que surgió como alternativa a los grafos OLC. En este tipo de grafo, los nodos representan las lecturas y las aristas el solapamiento entre ellas. Para reconstruir la secuencia, el algoritmo parte de cualquier nodo al azar y busca cuál de los nodos adyacentes tiene el mayor solapamiento con la lectura del nodo actual. De esta forma, va recorriendo el grafo para construir el camino de Hamilton hasta volver al nodo de partida (figura 1.13). Tiene varios problemas por los que ya casi no se usa: comprime las lecturas de las zonas repetitivas y además no tiene en cuenta ninguna información global que puedan aportar las relaciones entre las lecturas, por ejemplo lecturas pareadas. No obstante, son apropiados para genomas de pequeño tamaño. Algunos ejemplos de ensambladores que usan este tipo de algoritmo son VCAKE [89] y SSAKE [90].

**Simplificación de grafos** Los grafos generados contienen gran cantidad de nodos, aristas, caminos y subgrafos, sobre todo los generados a partir de ultrasecuenciación de lecturas cortas. Para contrarrestar las limitaciones de memoria y reducir el tiempo de computación el grafo generado se simplifica. Las lecturas con errores de secuenciación que no fueron detectados durante el filtrado inicial pueden complicar el grafo y el proceso de ensamblaje, y crear estructuras erróneas, por lo que hay que detectarlas y eliminarlas.

Para simplificar el nodo se llevan a cabo varios procesos: los nodos conectados entre sí de forma lineal y con un único recorrido posible se fusionan en uno solo; los caminos sin salida o ramificaciones cortas de un camino principal bien definido suelen deberse a errores de secuenciación o baja cobertura en las lecturas, por lo que se suelen eliminar en función de un umbral de profundidad o de longitud; se evalúan las burbujas causadas por variaciones en la secuencia biológica (debidas por ejemplo a SNP) y se eliminan los nodos con menor cobertura; los nudos o estructuras con forma de X se deben a secuencias repetidas en distintos puntos del genoma y se pueden resolver mediante el uso de lecturas pareadas que mapeen en los extremos, siempre que el tamaño de la repetición sea menor que el de la lectura.

**Posprocesamiento** Una vez construido el grafo, debe ser recorrido para construir lecturas más largas o *contigs*, que a su vez se conectarán entre sí para formar *scaffolds* (figura 1.9). Este proceso conlleva una serie de pasos previos de filtrado y simplificación tanto para la construcción de *contigs*, como la de *scaffolds*, para evitar así ensamblajes erróneos. Las lecturas pareadas se usan como guía para ordenar y orientar los *contigs* durante el *scaffolding*: dos *contigs* se unen o no en función de la posición que ocupan los extremos de las pareadas en ellos, su orientación y el tamaño de inserto esperado. Si el extremo de una lectura pareada forma parte de un *contig*, y el otro extremo forma parte de otro *contig* distinto, ambos *contigs* se pueden conectar en un *scaffold*. Como se conoce el tamaño del inserto, se puede saber la separación entre ambos *contigs* y rellenarla con indeterminaciones (caracteres N que indican nucleótidos desconocidos), en el caso de que el hueco no pueda ser rellenado con otros *contigs*. Las lecturas pareadas y su frecuencia son útiles también para detectar *contigs* quiméricos en los que se han ensamblado en el mismo *contig* regiones de diferentes localizaciones en el genoma. Algunos ensambladores que presentan un módulo de *scaffolding* son EULER-SR [84], Velvet [85] y SOAPdenovo [87].

Como hemos visto, la variedad de algoritmos que pueden usar los ensambladores es bastante amplia. Cabe también mencionar el desarrollo de ensambladores especializados en lecturas largas procedentes de tecnologías de tercera generación (PacBio y Nanopore), como por ejemplo HGAP [91], Falcon [92], Canu [93] o Hinge [94]. Cada ensamblador tiende a especializarse en un tipo de análisis o situación concretos, y por ello es habitual el uso simultáneo de distintos ensambladores con el fin de aprovechar los puntos de fuertes de cada uno y fusionar los resultados para obtener un ensamblaje final más completo (es lo que se denomina un ensamblaje híbrido o reconciliador). Esto va asociado al uso en un mismo ensamblaje de distintos tipos de tecnología de secuenciación y de lecturas generadas, lo que permite aprovechar las ventajas de cada una de ellas.

## Mapeo

Si se dispone de un genoma o transcriptoma de referencia, en vez de llevar a cabo un ensamblaje *de novo* es posible realizar un mapeo de las lecturas sobre dicha referencia, es decir, se comparan las lecturas con la referencia y se posicionan allí donde la identidad es perfecta o casi perfecta (en función del margen de error permitido). La finalidad del mapeo es determinar qué transcritos se están expresando en nuestra muestra, y qué diferencias existen entre ambas secuencias (SNP, indel...). Cuando las lecturas del mapeo proceden de transcritos, dependiendo de si la referencia es un genoma o un transcriptoma, se presentan

distintos problemas. Si la referencia de la que disponemos es un genoma puede haber problemas con la presencia de los intrones, ya que pueden interferir a la hora de mapear las lecturas de los transcritos que proceden del límite entre exones. Para tratar con este problema existen hoy en día programas capaces de mapear las lecturas sobre los exones, evitando los intrones, como son TopHat2 [95], HISAT2 [96], STAR [97] o GSNAP [98]. Si por el contrario la referencia es un transcriptoma, puede ocurrir que no estén representadas todas las variantes de ajuste alternativo posibles. La solución suele ser la utilización de herramientas capaces de realizar alineamientos con huecos y evaluar posteriormente las variantes de ajuste [53].

Existen muchas herramientas para llevar a cabo un mapeo [99], aunque hay dos grandes grupos de mapeadores en función de la estrategia que utilizan. Un primer grupo realiza un mapeo por indexado, en el cual la referencia se parte en  $k$ -meros y se ordenan y se usan como índice. Del mismo modo, la lectura se fragmenta también en  $k$ -meros y se compara con el índice para posicionarla sobre la referencia. Un ejemplo de este tipo de mapeadores es Stampy [100]. El otro grupo de mapeadores lleva a cabo un mapeo por búsqueda inversa mediante una transformación de Burrows-Wheeler para poder posicionar la lectura sobre la referencia. Se trata de un algoritmo bastante rápido y eficiente, y es la base de programas como BWA [101] o Bowtie2 [102]. Los programas de mapeo en general tienen el inconveniente de que para tener un alto rendimiento de trabajo pueden permitir muy pocos errores.

## Anotación

El siguiente paso tras haber obtenido los posibles transcritos de un transcriptoma (ensamblaje) o determinado las diferencias con respecto a una referencia (mapeo) es anotar dichas secuencias, lo cual consiste en compararlas con otras ya conocidas, y en función de esta similitud, poder comprender y extraer información biológica del estudio que estemos llevando a cabo.

Es necesario distinguir entre dos tipos de anotaciones: estructural y funcional. El primer tipo identifica cuáles son los elementos que hay en la secuencia y cuál es su estructura. En el caso de un transcriptoma, la anotación estructural consistiría en delimitar las regiones no traducidas (UTR) y el marco abierto de lectura (ORF) del transcrito. Esta información permite determinar si el transcrito está completo o no [103]. La anotación funcional, por su parte, consiste en asignar funciones a los genes o transcritos, una vez que ya se conoce su estructura. Para ello, se deduce la secuencia codificada y se compara con aquellas secuencias proteicas o nucleotídicas de función conocida contenidas en bases de datos.

Existen herramientas bioinformáticas para realizar la anotación de forma automática. La herramienta clásica de anotación es BLAST [104], utilizada para encontrar regiones de similitud entre secuencias. El programa compara secuencias de nucleótidos o aminoácidos con secuencias de bases de datos y calcula la significancia estadística de las coincidencias. BLAST permite distintas formas de búsqueda en función del tipo de secuencia (nucleotídica o aminoacídica) tanto de la entrada, como de la base de datos que se quiera consultar. Así, en el caso de secuencias nucleotídicas obtenidas en el ensamblaje de un transcriptoma, suele emplearse Blastx, que las traduce a posibles secuencias aminoacídicas para buscar en bases de datos de proteínas, tales como RefSeq [105] de GenBank [106], Swiss-Prot de UniProtKB [107], cuyas anotaciones están curadas a mano, o TrEMBL, también de UniProtKB y que emplea anotaciones automáticas que proceden de referencias cruzadas. Algunos programas especializados en anotación funcional son Blast2Go [108] y Sma3s [109, 110].

La anotación por similitud tiene el inconveniente de que la información contenida en estos repositorios corresponde principalmente a organismos modelo, por lo que, cuando se trabaja con especies no modelo, el número de anotaciones obtenidas es bajo y suelen encontrarse gran cantidad de genes ortólogos de función desconocida. Además, la anotación se suele basar en el mejor *hit* que devuelve BLAST,

y se sabe que muchas veces no es la secuencia más parecida (P. Seoane y M.G. Claros, comunicación personal). En nuestro laboratorio se ha desarrollado el algoritmo de Full-LengtherNext para evitar estos problemas y conseguir anotaciones fiables tanto estructurales como funcionales.

**Enriquecimiento de las anotaciones** Además de la información obtenida de bases de datos como RefSeq, Swiss-Prot o TrEMBL, la búsqueda por similitud también permite enriquecer la anotación con otros datos como son:

- **Gene Ontology:** El objetivo de GO<sup>9</sup> fue producir un vocabulario estructurado, controlado y dinámico, aplicable a todos los organismos que permita entender mejor el papel de los genes y las proteínas en la célula [111]. GO engloba tres ontologías independientes: procesos biológicos, funciones moleculares y componentes celulares, que proporcionan un set común de términos y definiciones que facilitan el intercambio, análisis y manejo de datos biológicos. Cada elemento de la ontología consiste en un código numérico y una descripción (por ejemplo GO:0046863 *ribulose-1,5-bisphosphate carboxylase/oxygenase activator activity*). Existe una jerarquía dentro de los términos GO que posibilita la anotación a distintos niveles.
- **Rutas KEGG:** La enciclopedia de genes y genomas de Kioto es una base de datos que integra información genómica, química y funcional. Los principales esfuerzos de KEGG<sup>10</sup> [112] se han centrado en el desarrollo de mapas de rutas que son diagramas gráficos que representan las redes de interacciones moleculares que tienen lugar dentro de la célula, a nivel de metabolismo, procesamiento de información genética y ambiental, procesos celulares, sistemas orgánicos, enfermedades y desarrollo de fármacos. Es considerada como una representación informática de un sistema biológico y puede ser utilizada para la modelización y extracción de datos.
- **Códigos EC:** Los números EC son un sistema de clasificación numérica para las enzimas basado en las reacciones químicas que catalizan. Estos números no especifican enzimas, sino reacciones catalizadas por ellas. De esta forma, las enzimas de organismos diferentes que catalicen la misma reacción química, recibirán el mismo identificador EC con el fin de evitar que a una misma enzima se le asignaran nombres diferentes [113]. Resulta interesante, por tanto, completar la anotación con este código en el caso de proteínas con función enzimática. Se puede consultar toda la información de los códigos EC en una base de datos alojada en el portal de recursos bioinformáticos ExPASy<sup>11</sup>.
- **InterPro:** Esta plataforma<sup>12</sup> lleva a cabo un análisis de las proteínas clasificándolas en familias y prediciendo dominios, sitios funcionales y regiones de interés [114]. Para ello usa modelos predictivos proporcionados por diferentes bases de datos que componen el consorcio InterPro, combinándolos en un solo recurso, reduciendo la redundancia y facilitando así la interpretación y el análisis de los resultados.
- **Ortólogos en especies modelo:** Cuando se trabaja con especies no modelo (aquellas que carecen de genoma de referencia) resulta muy útil determinar genes ortólogos a los nuestros en organismos modelo, cuyos genes están bien caracterizados y permiten extrapolar el conocimiento que ya se tiene a la especie problema. En plantas la especie modelo por excelencia es *Arabidopsis thaliana*, que a pesar de ser una mala hierba presenta una serie de características que la hacen idónea para la investigación: es pequeña y de fácil manejo, tiene un tiempo de generación corto, su fecundación es autógama, produce numerosas semillas y tiene un número reducido de cromosomas y un genoma de

---

<sup>9</sup><http://www.geneontology.org>

<sup>10</sup><https://www.kegg.jp>

<sup>11</sup><https://enzyme.expasy.org/>

<sup>12</sup><https://www.ebi.ac.uk/interpro>

pequeño tamaño. Otra especie modelo muy utilizada en plantas para determinar genes ortólogos en el caso de especies leñosas es el chopo (*Populus thichocarpa*). Los genes de estos organismos modelo pueden obtenerse de bases de datos generales como RefSeq [105] o Ensembl [115], o bien de bases de datos más específicas como por ejemplo TAIR10 [116], centrada en *Arabidopsis thaliana*.

Y también se puede complementar la información con anotaciones no funcionales, con elementos como:

- **SSR:** Los microsatélites o repeticiones de secuencia simple (SSR), también llamados repeticiones cortas en tándem (STR), son secuencias de DNA en las que un fragmento de pocas pares de bases (de dos a seis) se repite consecutivamente y la variación en el número de estas repeticiones constituyen alelos diferentes. En genética de plantas resultan muy útiles como marcadores para la selección de características fenotípicas interesantes desde el punto de vista agrícola [41].
- **SNP:** Los polimorfismos de un solo nucleótido son variaciones en la secuencia de DNA con una frecuencia en la población de al menos un 1 %, si no se considera mutación puntual. Suelen heredarse de forma estable por lo que son útiles en estudios de poblaciones. Existen numerosas bases de datos bioinformáticas de SNP en humano, como dbSNP<sup>13</sup> [117] del NCBI y también en plantas, como la de arroz RiceVarMap<sup>14</sup> [117].
- **ncRNA:** Las moléculas de RNA no codificantes son moléculas de RNA funcional que no se traducen a proteínas. Existen varios tipos, algunos con una funcionalidad ya muy bien descrita, como RNA de transferencia (tRNA) o el ribosómico (rRNA) y otros de más reciente descubrimiento como los micro-rna (miRNA), capaces de regular la expresión de otros genes, los RNA pequeños de interferencia (siRNA), que reducen la expresión de genes específicos, los RNA largos no codificantes (lncRNA), involucrados en la regulación de la transcripción, o los RNA pequeños nucleolares (snoRNA) que intervienen en la modificación química de otras moléculas de RNA, entre otros. No obstante, todavía se desconoce la función de muchos de ellos y se sospecha que en algunos casos podrían no ser funcionales [118]. El portal ncRNA.org<sup>15</sup> contiene herramientas bioinformáticas y bases de datos especializadas en estos ncRNA, como es el caso de fRNAdb [118].

## Incorporación a bases de datos

Tradicionalmente las secuencias tanto nucleotídicas como aminoácídicas de cualquier organismo se han ido depositando en bases de datos genéricas como por ejemplo GenBank<sup>16</sup> [106], del NCBI o UniProtKB<sup>17</sup> [107].

Así mismo, los datos generados a partir de los análisis de NGS deben ser accesibles al resto de la comunidad científica. De hecho, cualquier publicación científica que implique este tipo de análisis requiere el depósito de las secuencias brutas en bases de datos públicas como son SRA<sup>18</sup> (*Sequence Read Archive*), del NCBI; ENA<sup>19</sup> (*European Nucleotide Archive*), del EBI, o el archivo de secuencias de lecturas del DDJB<sup>20</sup> (*DNA Data Bank of Japan*), su equivalente en Japón.

<sup>13</sup><https://www.ncbi.nlm.nih.gov/snp/>

<sup>14</sup><http://ricevarmap.ncpgr.cn>

<sup>15</sup><https://www.ncrna.org/>

<sup>16</sup><https://www.ncbi.nlm.nih.gov/genbank/>

<sup>17</sup><https://www.uniprot.org/>

<sup>18</sup><https://www.ncbi.nlm.nih.gov/sra/>

<sup>19</sup><https://www.ebi.ac.uk/ena/>

<sup>20</sup><https://www.ddbj.nig.ac.jp/dra/>

Con respecto a plantas existen bases de datos muy importantes con acceso a información de numerosas especies, como es el caso de Phytozome<sup>21</sup>; PLAZA<sup>22</sup> y Gramene<sup>23</sup>, ambos para estudios de genómica comparativa; y EnsemblPlants<sup>24</sup>, entre otras. También es posible encontrar bases de datos especializadas en familias concretas de plantas, como es el caso por ejemplo de Sol Genomics Network<sup>25</sup>, para solanáceas, o CuGenDB<sup>26</sup>, para cucurbitáceas. De igual modo hay disponibles bases de datos exclusivas de una sola especie, como por ejemplo la base de datos TAIR<sup>27</sup> [116], para la especie modelo *Arabidopsis thaliana*, y otras específicas de organismos no modelo como el caso que nos ocupa, el olivo, tales como OliveTreeDB<sup>28</sup> o la del consorcio internacional del genoma de olivo IOGC<sup>29</sup>.

Cabe destacar también que la revista Plant and Cell Physiology<sup>30</sup> tiene un apartado donde recoge exclusivamente bases de datos públicas de plantas. Algunas de sus últimas incorporaciones son PeachVarDB [119], Melonet-DB [120] o TOMATOMICS [121].

En nuestro grupo de investigación se han desarrollado bases de datos públicas que contienen información de organismos no modelo como son EuroPineDB [55] y su versión más moderna SustainPineDB [54], que albergan el transcriptoma del pino marítimo (*Pinus pinaster*) y SoleaDB que contiene información sobre el transcriptoma del lenguado común (*Solea solea*) y del lenguado senegalés (*Solea senegalensis*) [56].

#### 1.1.4. Flujos de trabajo

Como hemos visto en el apartado 1.1.3, la obtención de un transcriptoma emplea muchos y muy diferentes programas bioinformáticos. El uso de múltiples algoritmos y la necesidad de ejecutarlos con distintos parámetros hace útil y necesario el uso de programas que permitan el diseño y la gestión de flujos de trabajo, que se definen como una serie de tareas concatenadas, de manera que los resultados producidos por un programa en una tarea se use como entrada para otro programa o tarea. La automatización de los flujos de trabajo disminuye la cantidad de tiempo y la probabilidad de error que supondría llevarlo a cabo de forma manual por el usuario, además, aumenta la reproducibilidad del trabajo.

Los programas gestores de flujos de trabajo tienen que lidiar con varios problemas, entre ellos la escasez en bioinformática de formatos de archivo estándar, lo cual exige introducir tareas que conviertan unos formatos en otros. Existe también mucha diversidad en cuanto a los lenguajes de programación utilizados para desarrollar los algoritmos, lo cual conlleva requisitos diferentes a la hora de integrar esos programas en un flujo de trabajo [43]. Para bregar con estos y otros muchos inconvenientes, se han desarrollado gestores de flujo de trabajo como Kepler [122], Taverna [123] o Galaxy [124]. En este ámbito, Pedro Seoane ha desarrollado en nuestro laboratorio dos herramientas muy útiles para el presente trabajo: AutoFlow [125] para el diseño y la gestión de flujos de trabajo y TransFlow [126], un sistema modular o *framework* basado en AutoFlow y diseñado para el ensamblaje de transcriptomas *de novo* en organismos no modelo y su posterior evaluación.

<sup>21</sup><https://phytozome.jgi.doe.gov/pz/portal.html>

<sup>22</sup><https://bioinformatics.psb.ugent.be/plaza/>

<sup>23</sup><https://www.gramene.org/>

<sup>24</sup><https://plants.ensembl.org/>

<sup>25</sup><https://solgenomics.net/>

<sup>26</sup><http://cucurbitgenomics.org/>

<sup>27</sup><https://www.arabidopsis.org/>

<sup>28</sup><https://genomaolivar.dipujaen.es/db/>

<sup>29</sup><http://olivegenome.org>

<sup>30</sup><http://pcp.oxfordjournals.org/>

Estos programas de gestión están muy enfocados hacia el trabajo en supercomputadores que cuentan con un sistema de colas. Este sistema actúa como puente entre el usuario y la máquina, y su papel consiste en recibir las peticiones de recursos de los distintos usuarios, asignar dichos recursos, decidir cuándo y dónde se ejecuta cada trabajo y vigilar su desarrollo.

En esta tesis se ha optado por realizar todos los análisis bioinformáticos en forma de flujos de trabajo para potenciar su repetitividad y su reusabilidad.

## 1.2. El olivo como modelo

### 1.2.1. Visión general

El olivo (*Olea europaea* L., subespecie *europaea*) es una planta leñosa de hoja perenne de la familia *Oleaceae*, de gran importancia tanto a nivel social como económico en el área mediterránea. A pesar de ello, y por no tratarse de una especie modelo, esta planta ha sido escasamente caracterizada a nivel genético y genómico y a día de hoy todavía se desconocen muchos aspectos de su biología.

El olivo parece ser originario de Asia Menor, desde donde se expandió por toda la cuenca mediterránea, y en los últimos tiempos se ha extendido a zonas donde no era cultivo tradicional, como son Australia, Estados Unidos, Sudáfrica, países latinoamericanos como Chile y Argentina y lugares exóticos como Hawái. A día de hoy, todavía existe debate acerca de sus orígenes y las bases genéticas de los cambios fenotípicos asociados a su domesticación y evolución [127, 128, 129].

Es una de las especies productoras de aceite más importantes del mundo. España es líder mundial en cuanto a superficie de cultivo de esta planta (2,6 millones de hectáreas), producción de aceite de oliva (60 % de la producción a nivel mundial) y producción de aceituna de mesa (22 % de la producción a nivel mundial) [130]. Existen más de 2600 variedades de olivo (FAO, 2010) en todo el mundo (alrededor de 250 solo en España), con muchas variedades locales y ecotipos.

A nivel genómico, se trata de una especie diploide ( $n = 23$ ), con un genoma de alrededor de 1,4 Gb [131], con gran cantidad de regiones repetitivas [132], fundamentalmente repeticiones en tándem (31 % del genoma completo) y elementos transponibles [133, 134], lo cual ha dificultado enormemente la reciente secuenciación y ensamblaje de su genoma [135].

El olivo tiene también un papel muy importante en el ámbito de la salud pública, ya que por una parte su polen es uno de los principales causantes de casos de alergia en los países donde se produce, y por otra parte, tanto el fruto como el aceite poseen características saludables cada vez mejor caracterizadas. Se ha visto que el principal ácido graso monoinsaturado presente en el aceite de oliva, el ácido oleico, tiene un efecto protector y modula la actividad de una gran variedad de tipos celulares. Está ampliamente demostrado que su consumo como componente clave de la dieta mediterránea tiene beneficios para el sistema cardiovascular, así como propiedades antimicrobianas, anticancerígenas y antioxidantes, y está relacionado con un aumento de la longevidad [136, 137].

### 1.2.2. La reproducción del olivo

En nuestra latitud, el olivo florece durante los meses de mayo y junio, como respuesta a una combinación de fotoperiodo y temperatura. Sus flores tienen simetría radial, y habitualmente son hermafroditas, aunque algunas veces pueden ser unisexuales, concretamente masculinas. Se agrupan en panículas, o



**Figura 1.14:** Flores de olivo (fotografía de Manuel Díaz Zurita).

inflorescencias racimosas que contienen de 10 a 35 flores, en función de la variedad, de las cuales solo una o unas pocas darán lugar a fruto. La corola está formada por cuatro pétalos blancos. El androceo está constituido por dos estambres, mientras que el gineceo consiste en un único pistilo con un estigma frecuentemente bilobular, un estilo sólido y un ovario con dos lóculos o cavidades, cada uno de los cuales contiene dos sacos embrionarios (figura 1.14).

La polinización del olivo es mayoritariamente anemófila (por medio del viento). Las anteras producen una gran cantidad de polen y, una vez abiertas (dehiscencia), el grano de polen es transportado por el viento hasta el estigma receptor, por lo que es capaz de viajar enormes distancias. Una vez ahí, si las condiciones son las adecuadas, el grano de polen se hidrata y germina para producir un tubo polínico que invade los tejidos del estigma, penetra el estilo y deposita finalmente los núcleos espermáticos en el saco embrionario, donde tiene lugar la doble fertilización característica de las plantas superiores con la que se genera el cigoto y el endospermo. Se requiere la fertilización de uno de los cuatro sacos embrionarios para que se desarrolle el fruto. La fertilización de esta planta es principalmente alógama, es decir, se da la polinización cruzada, y la fecundación tiene lugar entre individuos genéticamente diferentes. Para ello, las distintas especies de olivo presentan grados variables de autoincompatibilidad que pueden impedir la autofecundación y favorecen el intercambio genético [138, 139, 140].

El fruto es una drupa y está compuesto por tres tipos de tejido: endocarpo, mesocarpo y exocarpo, de los cuales, los dos últimos conforman el pericarpo. El desarrollo del fruto conlleva entre cuatro y cinco meses y consta de las siguientes fases: cuajado del fruto tras la fertilización, desarrollo de la semilla, endurecimiento o esclerificación del hueso, desarrollo del mesocarpo (que corresponde a la pulpa del fruto y es donde se produce el almacenamiento de aceite), y, por último, maduración del fruto, tras el consecuente cambio de color o envero.

En la reproducción de esta planta es relativamente frecuente la alternancia de años de abundante floración y producción de fruto con otros de cosechas escasas (vecería), lo cual puede estar determinado tanto por condiciones ambientales y fisiológicas de la planta, como por un componente genético [141].

### 1.2.3. El papel de las ROS y el NO en la interacción polen-pistilo y la señalización

#### Las ROS en la reproducción del olivo

El grupo de investigación de Biología Reproductiva y Microscopía Avanzada de Plantas (BRMAP) de la EEZ realiza análisis destinados a conocer las complejas cascadas de señalización en las que participan las ROS y RNS en los tejidos reproductivos de las Angiospermas [142], especialmente en el caso del olivo [143]. Además analizan diversas especies de plantas con un amplio panel de sistemas de incompatibilidad y pertenecientes a diferentes familias de Angiospermas [144]. Las principales conclusiones de estos estudios indican que las ROS son detectables sobre la superficie del estigma en aquellos estadios considerados “receptivos” a la interacción con el polen, produciéndose una disminución de sus niveles cuando el polen comienza a interactuar con el estigma. En estadios anteriores a la recepción polínica, las ROS pueden considerarse moléculas protectoras frente al ataque por diversos agentes como bacterias, hongos y otros patógenos. Gran parte de las ROS provienen de la acción enzimática de NADPH oxidasas de la familia Rboh (OeRbohH en el caso del olivo) generadoras de superóxido en los sistemas de membrana [145, 146]. Este enzima se puede considerar un vórtice de regulación a través de numerosos factores que influyen en su actividad y que incluyen activación por  $\text{Ca}^{2+}$  y fosforilación. Las pequeñas GTPasas de la familia Rop y el ácido fosfatídico son también capaces de estimular la producción de ROS por estas enzimas [147, 146]. El ión superóxido es considerado clave en el control redox celular, ya que es la fuente principal de ROS a partir del cual derivan las restantes especies de oxígeno reactivo a través de la acción de sistemas enzimáticos como las peroxidasas, catalasas y superóxido dismutasas, enzimas también sujeto de estudio por parte del grupo en el olivo [148, 149, 150, 151].

El metabolismo de las ROS está además estrechamente imbricado en el metabolismo del óxido nítrico, así como en el de antioxidantes como el glutatión, a través por ejemplo de intermediarios como el *S*-nitrosoglutatión (GSNO) [152]. El glutatión forma una parte muy importante del sistema regulador de los niveles de diferentes ROS a través de su presencia celular, especialmente en la forma reducida (GSH), y su estudio en los sistemas reproductivos está confirmando un importante papel en el mantenimiento de la fisiología del polen y el estigma [153, 154].

#### El NO y las modificaciones postraduccionales

**¿Por qué centrarnos en el NO?** Las modificaciones postraduccionales (PTM) son cambios que sufren las proteínas una vez sintetizadas por los ribosomas y que suelen dar lugar a la proteína madura, o bien modificar su actividad. Son muy variadas. Pueden consistir en la modificación de un grupo funcional existente, la adición de uno nuevo (como por ejemplo fosforilación, metilación, glucosilación...), la eliminación de determinados aminoácidos del extremo amino de la proteína, cortes en la cadena peptídica, etc. De todos los tipos de PTM, nos vamos a centrar en las que son dependientes del óxido nítrico (NO) [155], entre las que se incluyen la *S*-nitrosilación y la Tyr-nitración.

El NO es una importante molécula de señalización implicada en numerosos procesos en todos los seres vivos. Se trata de un radical libre, una molécula gaseosa, soluble en disolución acuosa y capaz de difundir a través de las membranas biológicas [156]. En las plantas, el NO está involucrado en numerosos procesos fisiológicos como el crecimiento, el desarrollo, la inmunidad, la respuesta a interacciones de tipo abiótico [157] y también la reproducción sexual [158, 143]. Su acción modula la expresión génica o interactúa con segundos mensajeros, canales iónicos, fitohormonas, ROS o proteína cinasas. Junto con las especies reactivas de oxígeno (ROS), el NO ha sido propuesto como mecanismo de señalización celular durante el desarrollo de los tejidos reproductivos de las plantas, la interacción entre el polen y el pistilo, el

direccionamiento del tubo polínico y la doble fertilización [159]. La implicación del NO en la reproducción sexual de las plantas se demostró inicialmente en especies modelo como *Arabidopsis* o *Lilium*, y ha sido recientemente estudiada en el olivo [160], donde se ha evaluado su papel en la viabilidad del polen, su capacidad para germinar y la elongación del tubo polínico. Aún así, todavía quedan cuestiones que resolver sobre la función del NO en la reproducción sexual de las plantas.

**S-Nitrosilación** La *S*-nitrosilación es una PTM involucrada en señalización celular en las plantas. Tiene lugar cuando el NO reacciona con grupos tiol reducidos de un residuo de cisteína (Cys), formando *S*-nitrosotioles (R-SNO). La unión de NO al residuo de Cys afecta a las proteínas, al modular características fundamentales como su actividad, su localización o su estabilidad [161]. Por otra parte, el NO reacciona con el glutatión reducido (GSH) para formar *S*-nitrosoglutatión (GSNO), que se considera un reservorio móvil capaz de mediar reacciones de nitrosilación. La *S*-nitrosilación es una modificación reversible, de forma que puede actuar como un mecanismo de control de encendido y apagado en la señalización celular, la regulación transcripcional, la remodelación de la cromatina y la apoptosis, entre otros procesos.

Se ha visto que las proteínas que se *S*-nitrosilan pertenecen a grupos muy diversos, como enzimas metabólicas, proteínas de señalización, relacionadas con mecanismos redox, con respuesta a estrés, enzimas de la fotosíntesis o proteínas del citoesqueleto. En el polen concretamente, parece ser una modificación relevante que afecta a una serie de productos génicos clave que incluyen alérgenos, NADPH oxidasa [162], actina [163] y moléculas implicadas en mecanismos de autoincompatibilidad, entre otros.

**Tyr-Nitración** La nitración del anillo de tirosina es una PTM mediada por el peroxinitrito ( $\text{ONOO}^-$ ) generado por la reacción de NO con superóxido



que, al reaccionar con la tirosina, forma nitrotirosina.

Esta PTM afecta a la función de la proteína, en la mayoría de los casos de manera irreversible, y contribuye así a la regulación del funcionamiento de la proteína, al metabolismo celular y a las rutas de señalización mediadas por NO [155]. Recientemente se ha revisado la función de esta PTM en el desarrollo y la respuesta a estrés (a)biótico en las plantas [164].

#### 1.2.4. El olivo y la alergia

En este apartado se describen los aspectos más generales de la implicación del olivo en la alergia [165].

##### ¿En qué consiste la alergia?

La alergia consiste en una respuesta desproporcionada del sistema inmunitario cuando identifica como dañinas determinadas moléculas (alérgenos) de carácter inocuo. La fuente alérgénica puede ser de naturaleza muy diversa (pólenes, alimentos, fármacos, frutos secos, ácaros, látex, etc). El contacto con el alérgeno dispara la producción de anticuerpos como la inmunoglobulina E (IgE), que a su vez inducen la liberación de sustancias como la histamina, causantes de los síntomas indeseados: rinitis, conjuntivitis, asma, dermatitis, y, en el caso más extremo, la anafilaxia, que puede llegar a ser mortal.

Además del tratamiento clásico con antihistamínicos, es frecuente el uso de la vacunación o inmunoterapia específica, consistente en administrar los alérgenos cada cierto tiempo para que disminuya la hipersensibilidad a estas moléculas, lo cual puede conseguir aliviar, e incluso eliminar los síntomas.

Las alergias afectan a más de un 25 % de la población en países desarrollados, por lo que suponen un grave problema de salud pública. Aún así, siguen desconociéndose numerosos detalles sobre los alérgenos implicados, su variabilidad, el mecanismo de interacción con el sistema inmune y los mediadores generados, y esto se traduce en carencias en su diagnóstico y tratamiento.

### Alergia al polen del olivo

Las grandes extensiones de terreno dedicadas a su cultivo y el elevado potencial alergogénico de este polen hacen que sea la principal causa de sensibilización por polen en Andalucía y Castilla la Mancha, y la segunda más importante en España tras la alergia al polen de gramíneas.

Es un polen fuertemente alergénico y relativamente pequeño (alrededor de 30  $\mu\text{m}$ ), capaz de recorrer largas distancias y permanecer mucho tiempo en suspensión en la atmósfera, donde alcanza concentraciones de hasta 5 000 granos/ $\text{m}^3$ . Causa una alergia de carácter estacional coincidente con el periodo de floración de la planta, cuyos síntomas más comunes son rinoconjuntivitis y asma. Cuando los granos de polen entran en contacto con la mucosa respiratoria, los alérgenos se liberan en unos minutos, lo cual explica la rápida aparición de los síntomas.

### Alérgenos del olivo

Hasta la fecha se han descrito 14 alérgenos distintos en el polen de olivo y uno en el fruto (tabla 1.2). Sin embargo, los alergogramas de extractos de polen de olivo muestran más de 20 bandas correspondientes a proteínas que se unen a IgE [166] y además se sospecha la presencia de alérgenos enmascarados coincidentes en punto isoelectrico y masa molecular con otros alérgenos altamente prevalentes y abundantes [167]. Todo esto hace pensar que quedan aún alérgenos por identificar y se necesitan más estudios para completar el alergograma de polen de olivo.

Los alérgenos de olivo conocidos son proteínas con una elevada solubilidad, bajo peso molecular, alto grado de polimorfismo y con funciones biológicas claves en el grano de polen. La mayoría de ellos han sido ya clonados y producidos como proteínas recombinantes, lo cual ha permitido analizar parcial o totalmente su estructura tridimensional y determinar los correspondientes epítomos (regiones de la secuencia del alérgeno reconocidas por el anticuerpo).

Los alérgenos con una elevada similitud de secuencia en las regiones de los epítomos pueden llegar a ser reconocidos por los mismos anticuerpos y desencadenar reacciones alérgicas cruzadas. Esto suele ocurrir entre especies taxonómicamente relacionadas. Como ejemplo, el alérgeno *Ole e 1* de olivo presenta reactividad cruzada con otras especies de la familia *Oleaceae*, y los pacientes alérgicos a *Ole e 1* de olivo, también suelen serlo al fresno, jazmín y aligustre, entre otros.

Cuando los alérgenos son proteínas con funciones vitales básicas, muy conservadas evolutivamente y están presentes en fuentes biológicas muy diversas, pueden provocar reactividad cruzada entre especies no relacionadas. Estos alérgenos se denominan panalérgenos (del griego 'pan-: todo', por estar ampliamente distribuidos). Existe un síndrome de alergia a polen-látex-fruta debido al gran parecido entre determinados panalérgenos que están presentes a la vez en el polen, los alimentos vegetales y el látex, por lo que una persona sensible a estos alérgenos puede desencadenar una respuesta alérgica no solo por

**Tabla 1.2: Alérgenos identificados en el polen y fruto de olivo.** La prevalencia hace referencia a la proporción de la población de individuos alérgicos al olivo que lo son al alérgeno en cuestión. Cuando la prevalencia es superior al 50 %, se habla de alérgeno mayor, mientras que si es inferior, se trata de un alérgeno menor. ND: no determinado.

| Alérgeno        | Tejido | Familia                                    | Prevalencia (%) | Reactividad cruzada                     |
|-----------------|--------|--------------------------------------------|-----------------|-----------------------------------------|
| <i>Ole e 1</i>  | Polen  | Ole e 1-like                               | 55-90           | Oleaceae                                |
| <i>Ole e 2</i>  | Polen  | Profilina                                  | 24              | Pólenes; alimentos; látex (Panalérgeno) |
| <i>Ole e 3</i>  | Polen  | Polcalcina                                 | 20-30           | Pólenes (Panalérgeno)                   |
| <i>Ole e 5</i>  | Polen  | Cu/Zn superóxido dismutasa                 | 35              | ND                                      |
| <i>Ole e 6</i>  | Polen  | ND                                         | 10-55           | Oleaceae                                |
| <i>Ole e 7</i>  | Polen  | Proteína de transferencia de lípidos (LTP) | 47              | Oleaceae                                |
| <i>Ole e 8</i>  | Polen  | Proteína de unión a calcio                 | 5               | Oleaceae                                |
| <i>Ole e 9</i>  | Polen  | 1,3- $\beta$ -glucanasa                    | 65              | Pólenes; alimentos; látex               |
| <i>Ole e 10</i> | Polen  | Proteína de unión a carbohidratos (CBM-43) | 55              | Pólenes; alimentos; látex (Panalérgeno) |
| <i>Ole e 11</i> | Polen  | Pectina metilesterasa                      | 55-75           | ND                                      |
| <i>Ole e 12</i> | Polen  | Isoflavona reductasa-like                  | ND              | ND                                      |
| <i>Ole e 13</i> | Fruto  | Taumatina-like (TLP)                       | ND              | ND                                      |
| <i>Ole e 14</i> | Polen  | Poligalacturonasa                          | 5-19            | Pólenes                                 |
| <i>Ole e 15</i> | Polen  | Ciclofilina                                | 12-14           | Pólenes; alimentos; animales            |

exposición al polen, sino también al tomar determinados alimentos o entrar en contacto con productos derivados del látex.

A continuación se comentan las características principales de los alérgenos de olivo identificados hasta la fecha:

- ***Ole e 1*** es el alérgeno de olivo mejor caracterizado y fue el primero en ser purificado [168, 169]. Se trata de una proteína glucosilada, específica de polen, de la que todavía se desconoce su función biológica, aunque es muy similar a proteínas de otras especies implicadas en la hidratación y germinación del polen [170, 171]. Ha dado lugar a la “familia de proteínas similares a *Ole e 1*”, descrita en la base de datos Pfam<sup>31</sup>.
- ***Ole e 2*** pertenece a la familia de las profilinas [172], proteínas ubicuas, de bajo peso molecular e implicadas en la regulación de la polimerización de la actina [173, 174, 175].
- ***Ole e 3*** pertenece a la familia de las polcalcinas (proteínas de unión a calcio) y se cree que su función es la modulación del flujo de calcio durante el crecimiento del tubo polínico [176, 177].
- ***Ole e 4*** parece ser un producto de degradación de *Ole e 9*, pues presenta una gran similitud de secuencia con su extremo amino [178, 166].
- ***Ole e 5*** se ha identificado como una Cu/Zn superóxido dismutasa [178, 179]. Estas proteínas intervienen en la señalización celular, en la protección frente al estrés oxidativo y en la interacción entre el polen y el estigma durante la reproducción [180, 151].
- ***Ole e 6*** es una proteína pequeña rica en residuos de cisteína de función desconocida, que no presenta similitud significativa con otras proteínas [181, 182].
- ***Ole e 7*** es un homólogo de las proteínas de transferencia inespecífica de lípidos (LTP) [183, 184].
- ***Ole e 8*** es una proteína de unión al calcio muy poco abundante en el polen. Es posible que funcione como regulador de las rutas de transducción de señales [185, 186]. Guarda similitud de secuencia con *Ole e 3*, si bien solo en los dominios de unión a calcio.

<sup>31</sup><https://pfam.xfam.org>

- **Ole e 9** es una 1,3- $\beta$ -glucanasa glucosilada que posee en su extremo carboxilo un dominio de unión a los glúcidos, mientras que su extremo amino es el responsable de la actividad catalítica [187].
- **Ole e 10** es otra proteína de unión a los glúcidos [188, 189]. Posee gran similitud con el extremo carboxilo de *Ole e 9* y se cree que interviene en la formación de la pared celular del tubo polínico durante el proceso de germinación.
- **Ole e 11** es una pectina metilesterasa. Se trata de una enzima ubicua, implicada en la modificación de la pared celular y participa en muchos procesos biológicos, como la maduración y el desarrollo del tubo polínico durante la reproducción [190, 191].
- **Ole e 12** es un alérgeno menor, correspondiente a una isoflavona reductasa. Estas enzimas parecen participar en procesos como la defensa de la planta, la germinación del polen, el crecimiento del tubo polínico y, probablemente, en el reconocimiento entre el polen y el estigma [192, 193].
- **Ole e 13** es una proteína de tipo taumatina (TLP), involucrada en mecanismos de defensa frente a patógenos y en respuesta a estrés biótico o abiótico y es el único alérgeno identificado en el fruto de olivo hasta la fecha [194, 195, 196].
- **Ole e 14** es una poligalacturonasa involucrada en elongación celular, abscisión de órganos, liberación de microsporas y crecimiento del tubo polínico [167].
- **Ole e 15** es una ciclofilina y ha sido el último alérgeno identificado hasta la fecha [197].

## El problema de la variabilidad alérgica

Pueden existir múltiples formas moleculares de un mismo alérgeno llamadas isoalérgenos, con un porcentaje de identidad de secuencia aminoacídica entre ellos superior al 67 %, y se nombran añadiendo al nombre del alérgeno un punto seguido de dos cifras (por ejemplo *Ole e 1.01*). A su vez, un isoalérgeno puede presentar distintas isoformas o variantes polimórficas con un porcentaje de identidad de secuencia superior al 90 % y que se nombran añadiendo dos cifras más al nombre del isoalérgeno (por ejemplo *Ole e 1.0101*, *Ole e 1.0102*...) [198].

La mayoría de los alérgenos son altamente polimórficos. En nuestro laboratorio también se ha visto que las distintas variedades de olivo muestran diferencias tanto en los niveles de expresión de los diferentes alérgenos, como en el número y el grado de polimorfismo de las isoformas alérgicas expresadas [180, 199, 174]. Además, los pacientes alérgicos suelen ser sensibles a diferentes combinaciones de alérgenos, lo cual dificulta mucho el diagnóstico y su tratamiento, pues los extractos de polen usados para diagnóstico y diseño de vacunas deben imitar lo máximo posible la composición alérgica a la que el paciente está expuesto en su entorno. Una vacuna o un diagnóstico poco precisos o demasiado generales pueden conducir a tratamientos poco eficaces, reacciones adversas, o en el peor de los casos, a una sensibilización secundaria a algún componente de la vacuna.

Es necesaria, por tanto, una caracterización precisa de los alérgenos ya descritos, así como de los posibles alérgenos nuevos del olivo y de todas las especies con las que pueda existir reactividad cruzada, además de una identificación y cuantificación adecuada de las distintas isoformas de alérgenos, con el fin de mejorar tanto el diagnóstico como el tratamiento de esta extendida enfermedad.

### 1.2.5. Ómicas en el olivo

A pesar de su importancia económica, hasta hace relativamente pocos años, la información disponible sobre el olivo a nivel genómico y transcriptómico era bastante escasa, si bien en los últimos años,

y gracias al desarrollo de las tecnologías NGS, se ha producido un gran avance al respecto.

**Genómica** Varios grupos han tenido como objetivo la secuenciación del genoma completo de esta planta. El proyecto italiano OLEA [200] surgió con la idea de identificar y caracterizar los genes responsables de su fenotipo reproductivo y vegetativo. Se centraron en la variedad Lecchino y combinaron las tecnologías Illumina y 454 hasta conseguir una cobertura media 40×. Por su parte Barguini et al. [133], partiendo también de la misma variedad, generaron un ensamblaje *de novo* que combina lecturas de 454 e Illumina, pero con una cobertura más baja, observando gran cantidad de regiones repetitivas y elementos transponibles del genoma.

El primer borrador público completo del genoma de olivo corresponde a un individuo con más de 1 200 años de *Olea europaea* L., subespecie *europaea*, variedad Farga, de escaso interés agronómico. Fue liberado por un equipo de investigadores del Centro de Regulación Genómica (CRG) de Barcelona, del Real Jardín Botánico de Madrid-CSIC y del Centro Nacional de Análisis Genómico de Barcelona (CNAG-CRG) [135] y puede ser consultado en <http://denovo.cnag.cat/olive>. La tecnología empleada fue Illumina y combinaron secuenciación por *shotgun* con genotecas de fósidos, para generar un genoma final con un N50<sup>32</sup> de *scaffold* de 443 kb, una longitud total de 1,31 Gb, que supone el 95 % de la longitud estimada del genoma (1,38 Gb) y un 63 % de secuencias repetitivas. La anotación se llevó a cabo con datos de RNA-seq de diversos tejidos y reveló 56 349 secuencias codificantes únicas.

Poco después, en 2017, el consorcio internacional del genoma de olivo IOGC (<http://olivegenome.org>) publicó el genoma y transcriptoma del olivo silvestre o acebuche (*Olea europaea* L., subespecie *europaea*, variedad *sylvestris*) [201], también llamado oleaster. La estrategia de secuenciación elegida fue *shotgun*, empleando Illumina, para obtener un genoma final de 1,48 Gb, con un N50 de *scaffold* de 228 Kb, un 51 % de secuencias repetitivas y en la anotación llevada a cabo fueron predichas 50 684 secuencias codificantes (47 124 respaldadas por RNA-seq). Este grupo ha conseguido datar dos grandes duplicaciones del genoma completo a lo largo de la evolución. En nuestro laboratorio se ha comprobado que este genoma produce mejores resultados cuando se usa como referencia para los análisis de expresión génica (J.D. Alché y M.G. Claros, comunicación personal).

Muy recientemente, en febrero de 2020 se ha hecho público el genoma de la variedad Picual, muy extendida en España y Andalucía [129]. El trabajo ha sido liderado por el grupo de Francisco Luque, de la Universidad de Jaén. Se trata de un genoma de 1,68 Gb que parece tener 79 667 modelos de gen que codifican 78 079 proteínas y 1 588 tRNA. Presenta un 59 % de secuencias repetitivas. A partir de estos genomas se ha propuesto que la presión selectiva debida a la domesticación y la aclimatación del olivo hizo que algunas familias génicas se expandieran o contrajeran, seguramente de la mano de los transposones [129].

La disponibilidad del genoma de olivo marca sin duda un antes y un después en el estudio de esta planta, y permitirá con seguridad elucidar muchos de los aspectos relacionados con su biología.

**Transcriptómica** Antes de la publicación de los primeros borradores del genoma de olivo [135, 201] se han llevado a cabo numerosos estudios transcriptómicos que han permitido ampliar el conocimiento acerca de esta planta. Los más destacados quedan resumidos en la tabla 1.3.

<sup>32</sup>N50: longitud del *contig* o *scaffold* a partir del cual está contenido el 50 % de los nucleótidos ensamblados.

**Tabla 1.3:** Resumen de los principales estudios transcriptómicos llevados a cabo en el olivo.

| Objetivo o Proceso biológico            | Variedad                                                 | Tejido                           | Tecnología de secuenciación o tipo de análisis | Publicación                       |
|-----------------------------------------|----------------------------------------------------------|----------------------------------|------------------------------------------------|-----------------------------------|
| Búsqueda de nuevos genes                | Gemlik                                                   | Hoja; fruto                      | Sanger                                         | Ozgenturk et al. 2010 [202]       |
| Desarrollo y arquitectura de la planta  | Arbequina; Picual x Arbequina                            | Plantón; yema                    | Microarray                                     | García-López et al. 2014 [203]    |
|                                         | Picual; Arbequina; Picual x Arbequina                    | Esqueje; plantón                 | Microarray                                     | Jiménez-Ruiz et al. 2015 [204]    |
|                                         | Arbequina; Picual; Chiquitita; Arbosana                  | Yema                             | Microarray                                     | González-Plaza et al. 2016 [205]  |
| Desarrollo de flor y fruto              | Arbequina                                                | Semilla; embrión; plantón        | Illumina                                       | Jiménez-Ruiz et al. 2018 [206]    |
|                                         | Leccino                                                  | Fruto                            | Sanger                                         | Galla et al. 2009 [207]           |
|                                         | Coratina; Tendellone                                     | Fruto                            | Pirosecuenciación                              | Alagna et al. 2009 [208]          |
|                                         | Lechin de Sevilla; Picual; Arbequina; Picual x Arbequina | Fruto; yema; raíz; semilla; hoja | Sanger; Pirosecuenciación                      | Munoz-Merida et al. 2013 [209]    |
|                                         | Picual                                                   | Polen; pistilo; hoja; raíz       | Sanger; Pirosecuenciación                      | Carmona et al. 2015 [57]          |
|                                         | Leccino; Dolce Agogia; Frantoio                          | Flor                             | Pirosecuenciación                              | Alagna et al. 2016 [210]          |
|                                         | Nocellara del Belice; Nera di Gonnos                     | Polen; pistilo                   | Illumina                                       | Iaria et al. 2016 [211]           |
| Abscisión del fruto                     | Leucocarpa; Cassanese                                    | Fruto                            | Illumina                                       | Iaria et al. 2016 [212]           |
|                                         | Picual                                                   | Polen; pistilo; hoja             | Sanger                                         | Zafra et al. 2017 [213]           |
| Respuesta a estrés biótico              | Picual                                                   | Fruto                            | Pirosecuenciación                              | Gil-Amado et al. 2013 [214]       |
|                                         | Picual                                                   | Fruto                            | Pirosecuenciación                              | Parra et al. 2013 [215]           |
| Respuesta a estrés abiótico             | Picual                                                   | Raíz                             | Illumina                                       | Jiménez-Ruiz et al. 2017 [216]    |
|                                         | Kalamon; Chondrolia Chalkidikis                          | Raíz                             | Sanger                                         | Bazakos et al. 2012 [217]         |
|                                         | Kalamon                                                  | Hoja; Raíz                       | Pirosecuenciación                              | Bazakos et al. 2015 [218]         |
| Búsqueda de marcadores                  | Leccino                                                  | Hoja                             | Illumina                                       | Guerra et al. 2015 [219]          |
|                                         | Picual                                                   | Hoja                             | Illumina                                       | Leyva-Pérez et al. 2015 [220]     |
| Análisis de miRNA                       | Gemlik x Edincik Su                                      | Hoja                             | Illumina                                       | İpek et al. 2017 [221]            |
|                                         | Picual; Arbequina                                        | Yema                             | Pirosecuenciación                              | Donaire et al. 2011 [222]         |
|                                         | Ayvalik                                                  | Fruto; hoja                      | Illumina                                       | Yanik et al. 2013 [223]           |
| Transcriptómica comparativa y evolutiva | Leucocarpa; Cassanese                                    | Fruto                            | Illumina                                       | Carbone et al. 2013 [224]         |
|                                         | Arbequina; subsp. cuspidata                              | Fruto; hoja; flor                | Illumina                                       | Sarah et al. 2017 [225]           |
|                                         | Arbequina; Acebuche; subsp. cuspidata                    | Hoja; flor                       | Illumina                                       | Gros-Balthazard et al. 2019 [128] |

Muchos de estos trabajos han permitido encontrar nuevos genes a partir de la secuenciación, ensamblaje y anotación de genotecas de cDNA de numerosos tejidos de olivo [209, 202]. El enriquecimiento de las bases de datos con nuevas secuencias y el desarrollo de la transcriptómica han facilitado, por ejemplo, la búsqueda masiva de marcadores moleculares de tipo SNP para la selección de características fenotípicas de interés [221].

A nivel reproductivo, se han estudiado, entre otros, los cambios fisiológicos que tienen lugar durante el desarrollo de la flor [210], se han identificado genes involucrados en el desarrollo del tubo polínico [211] y se han empleado genotecas sustractivas para identificar y anotar genes diferencialmente expresados en el polen y en el pistilo [213]. También en este ámbito se ha desarrollado la base de datos ReprOlive<sup>33</sup> [57], que contiene un transcriptoma reproductivo de polen y pistilo, y que constituye el eje principal de esta tesis, por lo que se expone en detalle en los siguientes capítulos. Su publicación [57] avala la presente tesis y puede consultarse en detalle en el apéndice F.1.

En cuanto al desarrollo del fruto, existen estudios sobre la acumulación de determinados transcritos [208], la influencia del aporte de luz [207], y el metabolismo de antocianinas y flavonoides [212] durante dicho proceso. La abscisión del fruto y sus mecanismos de control también han suscitado bastante interés [214, 215].

Igualmente con transcriptómica se ha abordado la respuesta del olivo a estrés de tipo biótico, como la infección por hongos patógenos [216], y abiótico, como la tolerancia a sal [217, 218] o la respuesta de la planta al frío [219, 220].

El desarrollo de la planta desde su estado embrionario hasta la etapa adulta también ha sido analizado desde un punto de vista transcriptómico [206], así como las características arquitectónicas tales como la altura, forma de la copa o densidad de las ramas, de especial interés en las nuevas prácticas de cultivo de alta densidad [205].

El creciente interés por los miRNA también en esta planta ha dado lugar a la publicación de un inventario con una alta diversidad de este tipo de moléculas en el olivo [222]. El fenómeno de la vecería también ha sido caracterizado a nivel de miRNA [223, 224].

El olivo también ha sido incluido en estudios evolutivos de transcriptómica comparativa entre especies cultivadas y especies silvestres filogenéticamente próximas [225]. Además, esta rama de la transcriptómica evolutiva ha permitido concluir que la domesticación de esta planta a lo largo de la historia ha producido principalmente cambios en la expresión génica, y no tanto cambios a nivel de genoma [128].

**Proteómica y metabolómica** Los trabajos de proteómica en el olivo se han centrado sobre todo en la calidad y el desarrollo del fruto. Así, por ejemplo, mediante la obtención de perfiles proteicos por espectrometría de masas y su comparación con transcriptómica y bases de datos, se han identificado numerosas proteínas involucradas en los procesos metabólicos que tienen lugar en el fruto [226, 227]. Otros análisis proteómicos realizados en el olivo han permitido caracterizar el denominado “secretoma” del estigma [228], el contenido de nanovesículas secretadas por el tubo polínico [229] o el contenido proteico del “pollen coat” [230].

Por otra parte, la caracterización proteómica de las células especializadas de los tricomas ha permitido identificar las enzimas involucradas en la adaptación de la planta a las condiciones de estrés de tipo biótico y abiótico [231]. También se han identificado los genes diferencialmente expresados en la aceituna en respuesta a la infección por larvas de la mosca *Bactrocera oleae* [232, 233], el principal agente biótico que afecta a los cultivos de olivo.

<sup>33</sup><http://reprolive.eez.csic.es>

En cuanto a análisis metabolómicos en el olivo, la mayoría de los escasos estudios existentes se han centrado en la caracterización de metabolitos presentes en el aceite de oliva [234], aunque algunos análisis se han centrado en el fruto [235], así como en diversos productos derivados de su procesamiento [236, 237].

Aunque en general las aportaciones al campo de la proteómica y la metabolómica de olivo son todavía escasas a día de hoy, ya se están llevando a cabo análisis proteómicos a mayor escala, aprovechando la reciente publicación del genoma de la planta, y que está sirviendo por ejemplo para identificar nuevos alérgenos de olivo [197].

### 1.2.6. Desafíos futuros

Todavía existen numerosas cuestiones por resolver respecto a la biología reproductiva del olivo, entre ellas la definición de criterios para la mejora de la planta en relación a la selección de genotipos, las condiciones de cultivo para prevenir la vecería, una fase juvenil excesivamente larga en algunas de las variedades y la presencia de genotipos autoincompatibles. También se sabe poco sobre las interacciones entre el polen y el pistilo durante la fertilización y los procesos de señalización que tienen lugar en la germinación del grano de polen.

En cuanto a los mecanismos de autoincompatibilidad, recientemente se ha confirmado que el olivo tiene un sistema de autoincompatibilidad de tipo esporofítico [138, 139, 140], aunque todavía están por identificar los genes involucrados en este sistema y quedan cuestiones por esclarecer, como el fenómeno de la autofertilización en algunas variedades.

A nivel clínico, los avances de las ómicas en el olivo están permitiendo una mejor caracterización de los alérgenos ya conocidos [166], así como la identificación de otros nuevos [167, 197], con el fin mejorar el diagnóstico y la terapia de la alergia a esta planta [238].

La disponibilidad de un genoma de olivo supone una valiosa herramienta para generar conocimiento enfocado a mejoras en la producción de esta planta mediante la identificación de genes importantes responsables de características de interés agronómico, la búsqueda de nuevos marcadores moleculares, selección genómica, desarrollo de nuevas variedades, aceleración en el crecimiento de la planta..., entre otros muchos aspectos. Así mismo, un mayor conocimiento de su biología ayudará a combatir las amenazas a las que está expuesto este cultivo, como por ejemplo, el impacto del cambio climático o el caso de la infección por el patógeno *Xylella fastidiosa*, que ha generado una gran preocupación, puesto que su propagación produciría graves pérdidas económicas en el sector [239].

## Capítulo 2

# OBJETIVOS

A pesar de ser el olivo una planta de cultivo muy extendida y de gran interés, hay todavía muchos aspectos de su biología, especialmente del aspecto reproductivo, que se desconocen.

El objetivo principal de este trabajo consiste en la obtención de un transcriptoma reproductivo de olivo que nos permita ampliar los conocimientos que existen sobre esta planta. Este objetivo se abordará mediante los siguientes objetivos específicos:

- Ensamblaje y anotación de un transcriptoma reproductivo de olivo y otro transcriptoma vegetativo como control.
- Obtención de un transcriptoma de polen maduro de olivo más completo y comparación con otros transcriptomas.
- Desarrollo de nuevos marcadores o genes constitutivos de RT-PCR para análisis de expresión génica en estos tejidos.
- Detección bioinformática de posibles alérgenos nuevos de olivo.
- Predicción *in silico* de posibles modificaciones postraduccionales (*S*-nitrosilación y Tyr-nitración) a partir de los transcritos de polen.

Con este trabajo se pretende ayudar a la comunidad científica a desarrollar futuras investigaciones y diseñar estrategias de mejora, tanto en la producción como en la calidad de los cultivos de olivo a nivel agronómico. Así mismo, se espera que sus aproximaciones a nivel clínico deriven en un aumento del conocimiento acerca de los putativos alérgenos presentes en el polen de olivo.



## Capítulo 3

# MATERIAL Y MÉTODOS

### 3.1. Infraestructura bioinformática

Los programas necesarios para la obtención de un transcriptoma demandan por lo general una infraestructura informática con gran capacidad de almacenamiento en disco, cómputo y memoria RAM, pues generan una gran cantidad de datos que necesitan ser almacenados y su procesamiento suele ser costoso a nivel de recursos computacionales. Además, la puesta a punto de estos programas y algoritmos bioinformáticos requiere mucho tiempo y personal especializado. Por todo esto, a la hora de abordar experimentos de NGS en general, es recomendable disponer de una plataforma bioinformática especializada, pues de no ser así la capacidad del investigador quedará muy limitada.

Los análisis bioinformáticos llevados a cabo en esta tesis se realizaron en Picasso, supercomputador perteneciente a la Red Española de Supercomputación, que se encuentra en el Centro de Supercomputación y Bioinnovación de la Universidad de Málaga. Está formado por 223 nodos (un nodo es equivalente a un ordenador personal) con procesadores Intel E5-2670 a 2,6 GHz e Intel E7-4870 a 2,4 GHz que suman un total de 4016 núcleos y 22,4 TB de RAM. Estos nodos están conectados por una red Infiniband (54/40 Gbps) que permite la transmisión de la información entre todos los nodos. Como sistema operativo usa SUSE Linux Enterprise Server 11SP2 con un sistema de colas SLURM que gestiona las tareas de computación. Se pueden consultar características más específicas en <http://www.scbi.uma.es/site/scbi/hardware>.

### 3.2. Programas informáticos

A continuación se listan todos los programas informáticos que se han empleado en la presente tesis. Algunos de ellos, por su relevancia en este trabajo, se describen con más nivel de detalle, como AutoFlow, con el que se han diseñado los flujos de trabajo, o Full-LengtherNext, en el que se basa principalmente la anotación.

### 3.2.1. Para manipulación

#### SeqTrimNext

El programa SeqtrimNext, basado en SeqTrim [74], fue desarrollado en nuestro laboratorio para el preprocesamiento de las lecturas brutas. Permite eliminar de las lecturas fragmentos de baja calidad, regiones de baja complejidad, vectores, adaptadores, DNA procedente de los orgánulos, colas poliA/poliT y secuencias contaminantes, con el fin de conservar la parte informativa más larga posible de la lectura. Se trata de un programa distribuido aplicable a cualquier tecnología de secuenciación gracias a un sistema de plantillas de configuración y plugins, capaz de procesar gran cantidad de lecturas brutas en poco tiempo. Posee además un sistema de detección de contaminantes basado en BLAST que permite al usuario indicar sus propios contaminantes. En este trabajo se ha usado la versión 2.0.59 de SeqTrimNext.

#### Bowtie2

Bowtie2 [102] es una herramienta para el alineamiento de lecturas sobre secuencias largas de referencia, ya sean de genoma, o de transcriptoma. En este trabajo se ha usado la versión 2.0.0-beta7 de Bowtie2.

#### Samtools

Samtools [240] es un paquete de programas informáticos y librerías que permite parsear y manipular alineamientos en formato SAM y BAM. En este trabajo se han empleado las herramientas Bio-samtools [241] adaptadas a BioRuby [242], concretamente la versión 0.1.16 (r963:234) de Samtools.

#### CD-HIT EST

CD-HIT EST [243] es un programa que sirve para eliminar redundancia. Se puede utilizar tanto para secuencias nucleotídicas como aminoacídicas. Las agrupa en función de un porcentaje de identidad indicado por el usuario y devuelve una secuencia consenso para cada conjunto. En este trabajo se ha usado la versión 4.5.4 de CD-HIT EST.

#### SeaView

SeaView [244] es una interfaz gráfica de usuario multiplataforma que permite realizar alineamientos múltiples de secuencias y análisis filogenéticos. En este trabajo se ha utilizado la versión 5.0.4 de SeaView.

### 3.2.2. Para ensamblaje

#### MIRA3 y MIRA4

MIRA3 [82] es un programa de ensamblaje basado en el solapamiento de los extremos de las secuencias con el algoritmo OLC (ver apartado 1.1.3). En este trabajo se ha usado la versión V3.2.0 de MIRA3

para el ensamblaje de ReprOlive y la versión 4.0.2 de MIRA4 para la construcción del transcriptoma mejorado de polen maduro.

### Euler-SR

Euler-SR [84] es un programa de ensamblaje basado en  $k$ -meros que utiliza grafos de Brujin (apartado 1.1.3). En este trabajo se ha usado la versión 120408 de Euler-SR.

### CAP3

CAP3 [81] es un programa de ensamblaje basado en el solapamiento de los extremos de las secuencias OLC (ver apartado 1.1.3). Es uno de los ensambladores más usados con secuencias tipo Sanger, y también para reconciliar ensamblajes distintos. En este trabajo se ha usado la versión de CAP3 vigente a fecha 10/15/07.

### OASES

Oases [245] es un programa de ensamblaje basado en  $k$ -meros que utiliza grafos de Brujin. Es una extensión del ensamblador Velvet [85] destinada a ensamblar lecturas de transcriptómica especialmente cuando no hay genoma de referencia disponible. Aprovecha las lecturas largas y las lecturas pareadas para reconstruir isoformas distintas y distintas formas de ajuste (*splicing*). En este trabajo se ha usado la versión 0.2.08.

### SOAPdenovo-Trans

SOAPdenovo-Trans [246] es un ensamblador *de novo* de transcriptomas basado en SOAPdenovo [87] que utiliza un algoritmo de ensamblaje paralelizable basado en grafos de Brujin. Es capaz de lidiar con formas de ajuste alternativo y con distintos niveles de expresión de los transcritos. En este trabajo se ha usado la versión 1.03.

### RAY

RAY [247] es un ensamblador basado en grafos de Brujin que utiliza memoria distribuida en vez de compartida. Es capaz de realizar ensamblajes combinando distintas plataformas de secuenciación. En este trabajo se ha usado la versión 2.3.1.

### Minimus2

Minimus2 es un programa de ensamblaje de la *suite* AMOS [248] que hemos utilizado para reconciliar ensamblajes. En este trabajo se ha usado la versión 3.1.0 de la *suite* AMOS.

### 3.2.3. Para anotación

#### Sma3s

Sma3s [109, 110] es un programa de anotación funcional, y que lleva a cabo BLAST contra la base de datos UniProtKB. Es paralelizable, por lo que resulta muy eficiente para anotar grupos grandes de unigenes ya que reduce bastante el tiempo de ejecución. En este trabajo se ha usado la versión 2.1 de Sma3s.

#### Full-LengtherNext

Full-LengtherNext<sup>1</sup> ([43]; Seoane et al., en preparación) es una herramienta desarrollada en nuestro laboratorio para anotar tanto estructural como funcionalmente, analizar y corregir transcriptomas *de novo*, ideal para organismos no modelo.

Las principales capacidades de Full-LengtherNext son las siguientes: clasifica el set de transcritos tentativos (TT) resultantes de un ensamblaje en función de si codifican proteínas completas o incompletas, construye un ORF ajustando las posibles roturas del marco abierto de lectura de los TT, descubre posibles transcritos específicos de especie, extrae ncRNA tentativos, construye un transcriptoma mínimo representativo o de referencia, que minimiza la redundancia, maximiza la longitud de secuencia completa y proporciona una visión general para la selección del mejor transcriptoma *de novo*.

Full-LengtherNext utiliza para anotar Swiss-Prot, TrEMBL y la base de datos de ncRNA<sup>2</sup>, aunque también admite una base de datos proporcionada por el usuario. Para aquellos TT sin similitud con ninguna secuencia de las bases de datos anteriores Full-LengtherNext utiliza TransDecoder<sup>3</sup> [249] para detectar si son secuencias codificantes [250].

Para la construcción del transcriptoma mínimo de referencia, Full-LengtherNext toma todos los transcritos que tengan similitud con alguna proteína y los agrupa usando el identificador de su ortólogo en UniProtKB. De cada grupo conserva solo la secuencia más larga y completa posible. Como algunos pueden pertenecer a la misma familia de proteínas, para no sobreestimar en número de TT diferentes, y como paso opcional, Full-LengtherNext los agrupa en función de su identificador en Pfam (base de datos de familias de proteínas) y vuelve a seleccionar solo el transcrito más largo y completo. A los transcritos ya seleccionados se añaden también para completar el transcriptoma representativo aquellos predichos por TransDecoder como codificantes, que sean más largos de 500 nt después de reducir la redundancia de secuencia usando CD-HIT EST al 95 % de identidad.

Además Full-LengtherNext calcula un índice empírico que representa la cantidad total de secuencias informativas que hay en un ensamblaje, de forma que permite evaluar de manera automática varios ensamblajes y decidir cuál de ellos es el mejor [125, 251].

En este trabajo se ha usado la versión 0.4.4 de Full-LengtherNext.

---

<sup>1</sup><http://www.scbi.uma.es/fulllengthernext>

<sup>2</sup><https://www.ncrna.org/>

<sup>3</sup><https://transdecoder.github.io/>

## MREPS

MREPS<sup>4</sup> [252] permite detectar microsatélites (SSR) o repeticiones de secuencia simple a partir de un fichero en formato FASTA. En este trabajo se ha usado la versión 2.5 de MREPS.

## BUSCO

BUSCO [253] es un programa diseñado para evaluar cuantitativamente el nivel de compleción de un ensamblaje en función de la presencia o no de un conjunto de genes ortólogos conservados de copia única que se supone que deben estar presentes. En este trabajo se ha usado la versión 2.0.

### 3.2.4. Para enriquecimiento biológico

#### GORilla

GORilla<sup>5</sup> [254] permite identificar y visualizar términos GO enriquecidos en listas no redundantes de genes ordenados según algún criterio biológico (por ejemplo, el nivel de expresión). El programa busca los términos GO que están enriquecidos en la parte alta de la lista comparados con los del resto de la lista. En este trabajo se ha utilizado la interfaz web de la herramienta, en la versión liberada a fecha 8 de marzo de 2013 y con los parámetros definidos por defecto.

#### GeneCodis

GeneCodis<sup>6</sup> [255, 256, 257] es una herramienta web para el análisis ontológico de listas de genes. Integra distintas fuentes de información para buscar anotaciones que frecuentemente concurren en un conjunto de genes y las ordena con significancia estadística para su interpretación biológica. En este trabajo se ha utilizado la versión 3 de GeneCodis.

### 3.2.5. Para predicción de *S*-nitrosilación

Se han utilizado las siguientes herramientas para predecir los residuos de Cys propensos a *S*-nitrosilación a partir de un archivo de secuencias proteicas en formato FASTA:

#### iSNO-AAPair

ISNO-AAPair [258] proporciona un servidor web de libre acceso<sup>7</sup> y fácil de usar, soporta archivos de entrada grandes y la salida es un archivo HTML. En este trabajo se ha utilizado la interfaz web de la herramienta descrita en su publicación, de octubre de 2013.

---

<sup>4</sup><http://mreps.univ-mlv.fr>

<sup>5</sup><http://cbl-gorilla.cs.technion.ac.il>

<sup>6</sup><http://genecodis.cnb.csic.es>

<sup>7</sup><http://app.aporc.org/iSNO-AAPair>

## GPS-SNO

GPS-SNO [259] cuenta con un servidor web de libre acceso<sup>8</sup>, aunque los autores recomiendan descargar la versión local del programa para archivos de entrada con más de 1 000 secuencias. La salida es un archivo de texto simple. Permite elegir entre tres umbrales distintos en función del nivel de precisión, sensibilidad y especificidad. En este trabajo se ha descargado e instalado en una máquina local la versión 1.0 del programa, y se ha elegido el umbral más restrictivo en su ejecución.

## SNOSite

SNOSite [260] está disponible a través de una interfaz web de libre acceso<sup>9</sup>. No acepta archivos FASTA de tamaño superior a 2 Mb y la salida es una tabla en formato HTML. Permite elegir entre tres niveles de especificidad: alto (95 %), medio (90 %) o bajo (85 %). En este trabajo se utilizó la versión 1.0 de su interfaz web, con el nivel de especificidad más alto.

### 3.2.6. Para predicción de Tyr-nitración

## GPS-YNO2

GPS-YNO2 [261] predice sitios propensos a Tyr-nitración en proteínas a partir de un archivo de secuencias proteicas en formato FASTA. Cuenta con un servidor web de libre acceso<sup>10</sup>, si bien sus autores recomiendan descargar la versión local del programa para archivos de entrada con más de 1 000 secuencias. La salida es un archivo de texto simple. Permite elegir entre tres umbrales distintos en función del nivel de precisión, sensibilidad y especificidad. En este trabajo se ha descargado e instalado en una máquina local la versión 1.0.1 del programa. Se han usado el umbral recomendado por defecto (medio) y el más restrictivo (alto), con fines comparativos.

### 3.2.7. Para diseño de flujos de trabajo

## AutoFlow

AutoFlow [262, 125] es un programa basado en Ruby creado para diseñar y gestionar de flujos de trabajo dinámicos, que permite integrar y combinar cualquier tipo de herramienta informática (independientemente del lenguaje de programación en que esté escrito). Está diseñado para definir tareas independientes usando sintaxis de bash, que se lanzarán automáticamente a un sistema de cola.

Los flujos de trabajo grandes pueden superar el tiempo máximo de ejecución permitido si se realizara una única petición de tiempo para todo el flujo. AutoFlow solventa este problema permitiendo asignar a cada tarea de forma individual un tiempo máximo de ejecución. Permite la iteración de tareas y además el programa trabaja dentro de un sistema de variables de entorno que facilita la persistencia de ciertos datos a lo largo de todas las tareas y la toma de decisiones sobre los datos mientras el flujo se está ejecutando, aportando dinamismo al sistema. Esto facilita también la selección o reconciliación de los

---

<sup>8</sup><http://sno.biocuckoo.org>

<sup>9</sup><http://csb.cse.yzu.edu.tw/SNOSite>

<sup>10</sup><http://yno2.biocuckoo.org>

distintos resultados cuando sea necesario. Incluye herramientas para monitorizar el estado de las tareas, controlar los tiempos de ejecución, buscar archivos y realizar representaciones gráficas de los flujos de trabajo. Una vez diseñado un flujo, cabe destacar la reproducibilidad y la posibilidad de poderlo reutilizar con diferentes datos de entrada, así como compartirlo entre distintos grupos de investigación.

Existen otras plataformas para construir y ejecutar flujos de trabajo como son Taverna [123] o Galaxy [124] que tienen una interfaz web en la que el usuario puede diseñar flujos de trabajo personalizados. Están enfocadas a usuarios que carecen de habilidades informáticas, pero esta usabilidad resta flexibilidad, ya que no permite el diseño de flujos de trabajo complejos, ni la inclusión de nuevos programas. El usuario no tiene control sobre los recursos usados y esto puede disminuir el rendimiento del flujo de trabajo. AutoFlow evita estas limitaciones, al carecer de interfaz gráfica y de repositorio de programas. Funciona por línea de comandos, por lo que el usuario debe tener habilidades informáticas y conocer la herramienta que va a emplear, aunque no necesita tener experiencia en el diseño de flujos de trabajo.

AutoFlow usa un archivo de texto simple que se denomina plantilla de flujo en la que se describe cada tarea con todos sus atributos, para ser lanzada posteriormente a un sistema de cola. Un ejemplo de la estructura general de cada tarea sería el siguiente:

```
preprocessing_454){
  module load seqtrimnext
  ?
  seqtrimnext -t transcriptomics_plants.txt -Q 454_reads.fastq >
    preprocessed_454_reads.txt
}
```

En la primera línea se asigna la etiqueta o identificador a la tarea y se termina con el carácter “)”. Esta etiqueta debe ser única, pues será utilizada para hacer referencia a ella posteriormente y para representaciones gráficas. A continuación entre llaves (“{ }”) se define la tarea. Esta definición consta de dos partes separadas por el carácter “?”. En la primera parte se inicializa el entorno, mientras que en la segunda parte se indican los comandos que deben ser ejecutados.

Para iterar una tarea basta con añadir al final del nombre de la etiqueta y entre corchetes (“[ ]”) los distintos parámetros necesarios para crear cada tarea, separados por punto y coma, como por ejemplo:

```
preprocessing_[Sanger;Illumina;454]{
  module load seqtrimnext
  ?
  seqtrimnext -t transcriptomics_plants.txt -Q (*)_reads.fastq >
    preprocessed_(*)_reads.txt
}
```

Reemplazando la etiqueta “(\*)” con los distintos parámetros (Sanger, Illumina y 454) se creará una tarea independiente para cada uno de ellos. Esta capacidad del programa resulta muy útil para buscar los mejores parámetros en el uso de un programa con unos datos determinados.

Cuando una tarea depende de otra anterior, no será lanzada hasta que la ejecución de la tarea de la que depende no haya finalizado. La llamada a la tarea previa se hace de la siguiente forma:

```
preprocessing_454){
  module load seqtrimnext
  ?
```

```

    seqtrimnext -t transcriptomics_plants.txt -Q 454_reads.fastq >
    preprocessed_454_reads.txt
}

assembly){
    module load euler
    ?
    Assemble.pl preprocessing_454)/preprocessed_454_reads.txt 25
}

```

Cuando hay tareas repetitivas con dependencias, la sintaxis de AutoFlow evita tener que escribir una nueva tarea para cada dependencia de la iteración, haciendo uso del carácter “!”. A este respecto nos puede interesar o bien que se genere un trabajo por cada dependencia, o bien que se genere un solo trabajo que depende de varios. La sintaxis de AutoFlow en cada escenario sería la siguiente:

```

# Supongamos una tarea iterativa en la que listamos varias carpetas, y
# despues queremos mostrar el contenido de cada lista

List_[folderA;folderB;folderC]{
    # No se requiere inicializacion
    ?
    ls (*) > dir
}

# La siguiente tarea genera un trabajo por cada dependencia

Show_[folderA;folderB;folderC]{
    # No se requiere inicializacion
    ?
    cat !List_*/dir
}

# En este caso se genera un solo trabajo que depende de varios

Show_result){
    # No se requiere inicializacion
    ?
    cat !List_*/dir
}

```

AutoFlow también permite el uso de variables estáticas y dinámicas, que pueden ser declaradas tanto dentro de la plantilla como desde la línea de comandos al lanzar el flujo de trabajo, si bien, esta segunda opción es la recomendable, pues permite al usuario cambiar el flujo de trabajo sin tener que modificar la plantilla. Las variables estáticas son aquellas que no cambiarán a lo largo del flujo de trabajo y se declaran precedidas del carácter “\$”, mientras que las variables dinámicas pueden ser modificadas durante el flujo y se preceden del carácter “@”. ESTAS últimas pueden ser transferidas de una tarea a otra, cambiando su valor, por lo que son muy útiles para la toma de decisiones sobre la marcha. El uso de variables también permite concatenar plantillas de flujo de trabajo compatibles.

Para lanzar el flujo de trabajo, la línea a ejecutar es la siguiente:

```
Autoflow -w template
```

donde *-w* indica la plantilla de texto que debe ejecutarse y es el único parámetro obligatorio. El comando opcional *-graph* genera una representación gráfica del flujo de trabajo (véase como ejemplo la figura 3.2), cuyo uso antes de lanzar el flujo resulta muy útil para comprobar si hay errores en el diseño y poder resolverlos.

Una vez lanzada una plantilla de flujo, AutoFlow la lee, parsea la descripción del flujo de trabajo a tareas y genera en el sistema de cola un trabajo (*job* en inglés) por cada tarea. Se crea una carpeta para cada tarea y dentro de cada una un archivo bash con toda la información necesaria para el sistema de cola. Cada uno de estos archivos bash es enviado al sistema de cola y las tareas se irán ejecutando de forma ordenada en función de las dependencias descritas por el usuario. En este trabajo se ha usado la versión 0.5.1 de AutoFlow.

## TransFlow

TransFlow [126] es una sistema modular o *framework* diseñado también en nuestro laboratorio por Pedro Seoane para el ensamblaje de transcriptomas *de novo* en organismos no modelo y su posterior evaluación. Esta herramienta está basada en AutoFlow (apartado 3.2.7) y permite combinar lecturas obtenidas con distintas tecnologías de secuenciación (en principio Illumina o 454/Roche, aunque es adaptable a otras plataformas de secuenciación), y ensamblarlas con diferentes programas. Los ensamblajes obtenidos de las distintas combinaciones de lecturas y ensambladores se evalúan y puntúan según una serie de parámetros de calidad, y los valores se comparan con los obtenidos para otros transcriptomas de referencia de gran fiabilidad para establecer un *ranking* y seleccionar cuál de ellos es el mejor.

TransFlow consta de cinco módulos independientes (figura 3.1), basados en un sistema de cinco plantillas que AutoFlow conecta entre sí (entre paréntesis se indica el código o etiqueta que se le añade al nombre de cada ensamblaje en función de la estrategia seguida):

- **Módulo 1: Ensamblaje con Illumina y *scaffolding***

Antes del ensamblaje las lecturas de Illumina, se preprocesan con la herramienta SeqTrimNext. Para el ensamblaje se utilizan los ensambladores OASES, SOAPdenovo-Trans y RAY, cada uno con dos *k*-meros diferentes (25 y 35). Así mismo, con cada ensamblador se lleva a cabo: (1) un ensamblaje primario, en el que solo se ensamblan *contigs* (se añade la etiqueta ‘ct’, el *k*-mero y el nombre del ensamblador al nombre del ensamblaje) y (2) un ensamblaje con *scaffolding*, basado en el ensamblaje primario previo, donde se aprovechan las lecturas pareadas para construir *scaffolds* con los *contigs* (‘sc’) (véase figura 1.9).

A continuación se emplea CD-HIT para eliminar la redundancia de los ensamblajes, con un porcentaje de identidad del 100 % (‘cd’) y los ensamblajes no redundantes se reconcilian usando Minimus2, con los parámetros por defecto, para proporcionar un conjunto de *contigs* más grandes para cada par de *k*-meros (‘rc’).

- **Módulo 2: Ensamblaje con 454/Roche y reconciliación**

Como en el caso anterior, este módulo comienza con la limpieza de las lecturas de 454/Roche crudas con la herramienta SeqTrimNext. Para el ensamblaje de las lecturas útiles se emplean los

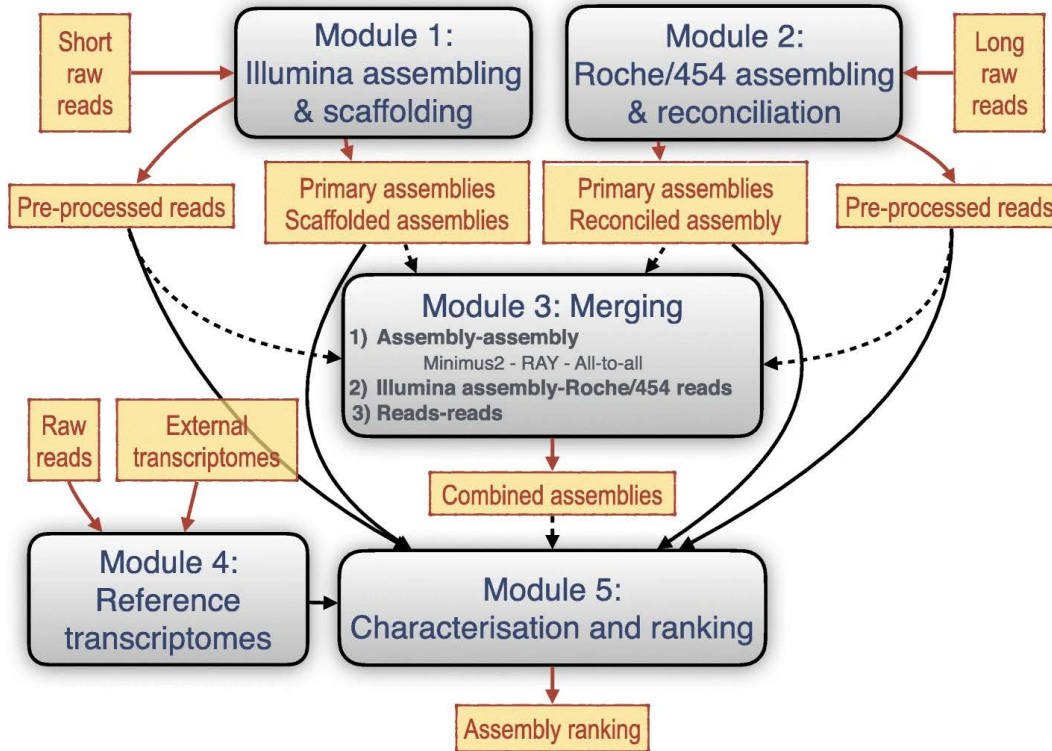


**Figura 3.1:** Diagrama que representa los distintos módulos de TransFlow y las dependencias entre ellos. Las flechas continuas hacen referencia a dependencias obligatorias si el módulo parental está presente. Las flechas discontinuas indican dependencias opcionales, incluso si el módulo parental está presente. Más detalles en el texto. Tomada de [126]

ensambladores MIRA4 y Euler-SR, este último con  $k$ -mero 29. Se produce un ensamblaje primario (se añade ‘ct’ y el nombre del ensamblador al nombre del ensamblaje). Los *contigs* que resultan de ambos ensambladores se reconcilian con CAP3 (‘rc’), para generar un ensamblaje que se espera que sea mejor que los ensamblajes primarios [57, 56].

#### ■ Módulo 3: Combinación de los ensamblajes de los módulos 1 y 2

En este módulo se fusionan los datos de ambas plataformas de secuenciación para producir ensamblajes combinados, que en principio deberían ser mejores que los ensamblajes resultantes de una sola tecnología. Este módulo es opcional y solo se activa si los módulos 1 y 2 se han activado previamente (obviamente también es posible partir de lecturas de un solo tipo). Se siguen tres estrategias diferentes: (1) combinación ensamblaje-ensamblaje de los generados en los módulos 1 y 2 respectivamente (‘aa’, de *assembling-assembling*), (2) combinación de ensamblaje con Illumina y lecturas 454/Roche (‘ar’, de *assembling-reads*) y (3) combinación de las lecturas de ambas plataformas (‘rr’, de *reads-reads*). A continuación se detalla cada una de las estrategias.

**Ensamblaje - ensamblaje (‘aa’):** Esta aproximación sirve para fusionar cada ensamblaje primario con *scaffolding* de Illumina del módulo 1 con los ensamblajes primarios y reconciliados del módulo 2. Se producen dos combinaciones diferentes, una usando Minimus2 con los parámetros por defecto, y otra usando RAY con  $k$ -meros 25 y 35. Por otra parte, se obtiene otro ensamblaje combinado tras eliminar con CD-HIT EST la redundancia de todos los ensamblajes primarios de

Illumina y mezclándolos con el ensamblaje reconciliado de 454/Roche usando Minimus 2. Por último, se genera otro ensamblaje más repitiendo lo anterior, pero usando ensamblajes de *scaffolding* en vez de ensamblajes primarios.

**Ensamblaje de Illumina - lecturas 454/Roche ('ar'):** Cada ensamblaje primario con *scaffolding* de Illumina del módulo 1 se fusiona con las lecturas 454/Roche preprocesadas del módulo 2 usando MIRA4 (considerando los *contigs* como si fuesen lecturas Sanger) y también usando RAY con *k*-meros 25 y 35. Como resultado se obtiene un ensamblaje combinado para cada ensamblaje primario con *scaffolding* de Illumina.

**Lecturas - lecturas ('rr'):** En este caso, las lecturas preprocesadas de ambos módulos 1 y 2 se ensamblan juntas directamente usando RAY con *k*-meros 25 y 35, puesto que es el único ensamblador capaz de tratar con tal volumen de lecturas.

#### ■ Módulo 4: Transcriptomas de referencia

Para poder evaluar los ensamblajes generados, se necesita compararlos con transcriptomas de referencia bien caracterizados, extraídos de bases de datos públicas, de especies próximas a aquella con la que trabajemos, como son *Arabidopsis thaliana* y *Populus trichocarpa* en el caso de plantas. Igualmente se necesitarán lecturas brutas de RNA-seq de estas especies, del mismo tipo y características que las que se empleen en los módulos 1 y 2. Cada transcriptoma de referencia se analiza con Full-LengtherNext para calcular una serie de parámetros que se emplearán en la evaluación y que se describen en la tabla 3.1.

#### ■ Módulo 5: Caracterización y *ranking*

En este módulo se evalúan todos los ensamblajes obtenidos para seleccionar cuál de ellos es el mejor. Para ello se necesita una referencia con la que comparar, que es la que se obtiene en el módulo 4. Para cada ensamblaje se calculan la misma serie de parámetros de evaluación que se obtuvieron en el módulo 4 para los transcriptomas de referencia, y que se describen con detalle en la tabla 3.1. Un conjunto de esos parámetros se calcula con Full-LengtherNext parametrizado con la base de datos SwissProt, mientras que el resto se obtiene con el programa BUSCO, que evalúa lo completo que está el ensamblaje, para lo cual extraen de la base de datos OrthoDB [263] los ortólogos universales de copia única para un taxón determinado, o lo que es lo mismo, qué genes se supone que deben estar siempre presentes en un determinado taxón. En el caso de plantas, se selecciona embriófitas, donde se ha visto que hay unos 1 400 genes que siempre deberían estar presentes. Cuantos más de estos genes estén contenidos en el transcriptoma en evaluación, tanto más completo estará.

En cuanto al análisis estadístico, cada ensamblaje queda caracterizado con una serie de factores cualitativos referidos a: (1) el programa usado en el último paso de ensamblaje, (2) la tarea llevada a cabo, o lo que es lo mismo, el objetivo buscado con la ejecución del programa, (3) el *k*-mero empleado en el ensamblaje y (4) la plataforma o tecnología de secuenciación. Se construye una tabla resumen que contiene para cada ensamblaje las características anteriores junto con los valores de los parámetros de evaluación, donde cada factor especifica la categoría particular que se ha usado. A partir de esta tabla se lleva a cabo un análisis de componentes principales (PCA) usando el paquete de R FactoMineR [264], que permite añadir datos suplementarios de individuos para facilitar la interpretación del PCA. Cada uno de los ensamblajes de la tabla se considera un individuo, mientras que los transcriptomas de referencia del módulo 4 se consideran individuos suplementarios que proporcionan una referencia para la evaluación de los ensamblajes. El PCA y los análisis estadísticos posteriores se llevan a cabo conservando las tres primeras componentes que puedan explicar hasta el 85 % de la varianza observada entre individuos.

FactoMineR permite hacer análisis de correlación con todas las coordenadas de los ensamblajes en el espacio PCA. Se calcula la correlación entre cada parámetro de evaluación y las coordenadas del individuo en cada una de las componentes del PCA, para obtener el coeficiente de correlación (*R*) y

**Tabla 3.1: Parámetros de evaluación utilizados en TransFlow calculados para cada uno de los ensamblajes.** Se indica su nombre, el programa que lo calcula (FLN: Full-LengtherNext), una breve descripción, y la tendencia esperada para cada uno de ellos. Adaptada de [126].

| Parámetro      | Software | Descripción                                                                                                                                                                                                   | Tendencia (*) |
|----------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| AllTransSize   | FLN      | La suma de la longitud de todos los transcritos en nucleótidos                                                                                                                                                | ↓             |
| N50            | FLN      | La longitud más corta del <i>contig</i> o <i>scaffold</i> (en nucleótidos) en el transcriptoma necesaria para cubrir el 50 % de AllTransSize                                                                  | ↑             |
| N90            | FLN      | La longitud más corta del <i>contig</i> o <i>scaffold</i> (en nucleótidos) en el transcriptoma necesaria para cubrir el 90 % de AllTransSize                                                                  | ↑             |
| Contigs        | FLN      | Número de <i>contigs</i> en los que mapea al menos un par de lecturas                                                                                                                                         | ↓             |
| Contigs500     | FLN      | Igual que el parámetro previo pero teniendo en cuenta solo <i>contigs</i> > 500 nt                                                                                                                            | ↑             |
| MeanContigLen  | FLN      | Media de la longitud de la secuencia (en nucleótidos) de todos los <i>contigs</i> o <i>scaffolds</i> útiles                                                                                                   | ↑             |
| Ns             | FLN      | Número de indeterminaciones (N) en los <i>contigs</i> o <i>scaffolds</i>                                                                                                                                      | ↓             |
| MeanGapLen     | FLN      | Media de la longitud de la indeterminación (en nucleótidos), donde 1 indica que los <i>gaps</i> se distribuyen aleatoriamente, mientras que valores mayores indican <i>gaps</i> reales                        | ↓             |
| DiffProts      | FLN      | Número de proteínas diferentes únicas                                                                                                                                                                         | ↑             |
| DiffComplProts | FLN      | Igual que el parámetro anterior, pero considerando solo las proteínas que parecen estar completas                                                                                                             | ↑             |
| MissAssembl    | FLN      | Porcentaje de <i>contigs</i> donde la proteína anotada encuentra similitud en ambas hebras ( <i>plus</i> y <i>minus</i> )                                                                                     | ↓             |
| MeanContigCov  | FLN      | Fracción de la longitud del <i>contig</i> (expresada como porcentaje) cubierta por lecturas que mapean. Primero se calcula para cada <i>contig</i> y luego se calcula el promedio para el ensamblaje completo | ↑             |
| ComplOrtho     | BUSCO    | Porcentaje de ortólogos de OrthoDB de un taxón identificados completos en un único <i>contig</i>                                                                                                              | ↑             |
| FragOrtho      | BUSCO    | Porcentaje de ortólogos de OrthoDB de un taxón que están fragmentados en varios <i>contigs</i>                                                                                                                | ↑             |
| DuplOrtho      | BUSCO    | Porcentaje de ortólogos de OrthoDB de un taxón que están repetidos en varios <i>contigs</i>                                                                                                                   | ↓             |

(\*) ↑ indica que cuanto mayor sea el valor de este parámetro, mejor será el transcriptoma; ↓ indica que cuanto menor sea el valor de este parámetro, mejor será el transcriptoma

la significación estadística ( $P$ ) para cada par parámetro de evaluación-componente PCA. Se descartan

los pares con  $P > 0,05$ . El paquete también ofrece la posibilidad de realizar agrupaciones o *clustering* jerárquico sobre las componentes principales (HCPC), generando el número óptimo de *clusters* en función de la diferencia de varianza entre los *clusters*, para explorar las similitudes entre los individuos. Se aplica la función HCPC de FactoMineR con los valores por defecto y se obtiene una matriz de diferencias que sirve para establecer un *ranking* objetivo basado en las distancias medias a los transcriptomas de referencia. Esta matriz permite calcular la distancia euclidiana de cada ensamblaje a cada transcriptoma de referencia. Posteriormente se calcula la distancia media (MD) a lo largo de los transcriptomas de referencia para cada ensamblaje. Se asume que cuanto más cercanos están los parámetros del ensamblaje a los de las referencias (esto es, cuanto más próxima a 0 esté MD), mejor es la calidad del ensamblaje.

TransFlow se lanza a partir de un archivo de bash (véase como ejemplo el apéndice B.1), donde el usuario puede editar y personalizar una serie de variables:

1. *TEMPLATES* (para indicar cuáles plantillas, y por tanto cuáles módulos se van a ejecutar),
2. *reference* (ruta a la carpeta que contiene el archivo/s FASTA para el/los transcriptomas de referencia),
3. *reads* (ruta a la carpeta que contiene los archivos FASTQ con las lecturas que se mapearán sobre los transcriptomas de referencia),
4. *read\_454* (ruta del archivo con las lecturas 454/Roche),
5. *ill\_type* (tipo de lecturas de Illumina: *single* o *paired*),
6. *read\_illumina\_pair\_1* - *read\_illumina\_pair\_2* / *single\_illumina* (ruta a los archivos de lecturas *paired/single* de Illumina, respectivamente),
7. *BUSCO\_DB* (taxón específico para BUSCO),
8. *FLN\_DB* (nombre de la base de datos para Full-LengtherNext),
9. *k-meros*,
10. *key\_organisms* (identificadores de la tabla resumen de ensamblajes que se usarán como transcriptomas de referencia).

Una vez personalizado, se lanza el archivo y se ejecuta AutoFlow, que gestionará los distintos módulos de TransFlow. Los resultados generados se empaquetan en un informe HTML interactivo por el que se puede navegar para elegir el mejor ensamblaje (véase como ejemplo el archivo adicional 2 de [126]). Entre los numerosos resultados que produce destacan una tabla con el *ranking* de los ensamblajes, las tablas con los parámetros de evaluación, factores y sus pesos para las dos primeras dimensiones del PCA y un *heatmap* de los parámetros de evaluación que permite ver su influencia en cada ensamblaje, así como los *clusters* que se forman. Los resultados del PCA y el HCPC se muestran embebidos en un PDF.

Al estar basado en AutoFlow, TransFlow proporciona flujos de trabajo automáticos y reproducibles que se pueden reciclar fácilmente. En este trabajo se ha usado la versión de TransFlow vigente a fecha 18/07/2017.

### 3.3. Datos biológicos

#### 3.3.1. Secuencias para los transcriptomas reproductivo y vegetativo de olivo (ReprOlive)

Se han utilizado lecturas largas de 454/Roche y lecturas Sanger de distintos tejidos de la variedad de olivo Picual.

**Lecturas de 454/Roche:** Se obtuvieron a partir de RNA y mRNA de granos de polen maduro, polen germinado *in vitro* a dos tiempos diferentes tras la hidratación, (1 y 5 horas) y de pistilos en estadio de desarrollo 2 (capullo blanco), 3 (flor recién abierta) y 4 (antera dehiscente) como queda descrito en [143]. Para el transcriptoma vegetativo también se obtuvieron y secuenciaron librerías de hoja, raíz y radícula de olivo. La secuenciación se llevó a cabo con el sistema Roche GS-FLX Titanium+ y las lecturas brutas obtenidas fueron depositadas en la base de datos SRA con identificador de BioProject ID PRJNA287107<sup>11</sup>.

**Lecturas Sanger:** Se añadieron también tres librerías sustractivas secuenciadas por el método Sanger, obtenidas de la comparación de polen, pistilos en estadio de desarrollo 4 y hojas [213]. Las lecturas brutas de estas librerías están disponibles públicamente en la base de datos ENA bajo el identificador de acceso PRJEB13716<sup>12</sup>.

Las librerías utilizadas y el número total de lecturas brutas de cada una de ellas se resume en la tabla 3.2.

**Tabla 3.2: Librerías utilizadas para la obtención de los transcriptomas reproductivo y vegetativo de olivo.** ‘-Subs’: librería sustractiva. Tomada de [57].

| Gene library | Tissue | Developmental stage | Sequencing method | Raw reads | Useful reads |
|--------------|--------|---------------------|-------------------|-----------|--------------|
| PM-Subs      | Pollen | Mature              | Sanger            | 666       | 518          |
| PM           | Pollen | Mature              | Pyrosequencing    | 216 497   | 111 242      |
| PG1          | Pollen | 1 h germination     | Pyrosequencing    | 258 167   | 141 232      |
| PG5          | Pollen | 5 h germination     | Pyrosequencing    | 233 921   | 120 276      |
| S2           | Pistil | Stage2              | Pyrosequencing    | 257 813   | 138 077      |
| S3           | Pistil | Stage3              | Pyrosequencing    | 247 401   | 141 903      |
| S4           | Pistil | Stage4              | Pyrosequencing    | 262 269   | 149 929      |
| S4-Subs      | Pistil | Stage4              | Sanger            | 480       | 256          |
| L            | Leaf   | Mature              | Pyrosequencing    | 223 399   | 41 178       |
| L-Subs       | Leaf   | Mature              | Sanger            | 403       | 251          |
| R1           | Root   | Mature              | Pyrosequencing    | 231 237   | 25 899       |
| R2           | Root   | Radicle             | Pyrosequencing    | 145 204   | 22 075       |

#### 3.3.2. Secuencias para el transcriptoma mejorado de polen maduro

Se ha combinado la librería previa de polen maduro de lecturas 454/Roche descritas en tabla 3.2 con las lecturas de Illumina (NextSeq 550) 2 × 75 nt de polen maduro descritas en la tabla 3.3. Están

<sup>11</sup><http://www.ncbi.nlm.nih.gov/bioproject/287107>

<sup>12</sup><https://www.ebi.ac.uk/ena/data/view/PRJEB13716>

disponibles pero protegidas en la base de datos SRA con identificador de BioProject ID 392587.

**Tabla 3.3:** Librerías utilizadas para la obtención de un transcriptoma de polen maduro más completo.

| Librería | Estadio de desarrollo | Método de secuenciación | Lecturas brutas | Lecturas útiles     |
|----------|-----------------------|-------------------------|-----------------|---------------------|
| PM       | Polen maduro          | Pirosecuenciación       | 216 497         | 111 242 (51,4 %)    |
| Po-Pi    | Polen maduro          | Illumina                | 40 488 002      | 32 529 229 (80,3 %) |

Para la evaluación de los transcriptomas de referencia usados por TransFlow se han empleado los transcriptomas de *Arabidopsis thaliana* y *Populus trichocarpa* obtenidos de Phytozome 12.1<sup>13</sup>, así como lecturas brutas RNA-seq de  $2 \times 100$  nt obtenidas con secuenciadores Illumina HiSeq detalladas en la tabla 3.4.

**Tabla 3.4:** Información sobre los transcriptomas de Phytozome (versión y número de secuencias codificantes) y los datos de RNA-seq (identificador de la librería y número de lecturas brutas) descargados para evaluar el mejor transcriptoma de referencia en plantas. Adaptada de [126].

| Source                                | Transcriptome version | Protein-coding transcripts | Library ID | Raw reads  |
|---------------------------------------|-----------------------|----------------------------|------------|------------|
| <i>Arabidopsis thaliana</i> seedlings | TAIR 10               | 35 386                     | SRR4897845 | 78 742 616 |
| <i>Populus trichocarpa</i> leaves     | 3.1                   | 63 498                     | SRR1030352 | 65 442 430 |

### 3.3.3. Secuencias para el desarrollo de marcadores de RT-PCR

Para la búsqueda de marcadores en el olivo se han utilizado las librerías de tejido reproductivo (polen y pistilo) descritas en la tabla 3.2 y el transcriptoma reproductivo de ReprOlive (tabla 4.1).

Para la búsqueda de marcadores en *Arabidopsis thaliana* se han utilizado lecturas de Illumina correspondientes a un estudio que compara una variedad de floración tardía (Columbia) con la variedad de referencia Killeen para determinar los genes involucrados en floración temprana [265]. Para este estudio secuenciaron tres réplicas biológicas de plantas de 10 días (Illumina HiSeq 1000  $2 \times 100$  nt). Cada réplica consiste en una planta de la variedad Columbia, y una de la variedad Killeen. Las lecturas brutas se han descargado de la base de datos SRA (identificador de BioProject ID PRJEB9470<sup>14</sup>). El tamaño de las librerías se detalla en la tabla 3.5. Estas lecturas se han mapeado sobre el transcriptoma de referencia de *Arabidopsis* (35 386 transcritos) descrito en la tabla 3.4.

### 3.3.4. Secuencias para la identificación de *Ole e 7*

Para la identificación de *Ole e 7* en genoma se han utilizado tanto los transcriptomas como los proteomas correspondientes a estas tres variedades de olivo: Farga<sup>15</sup> [135], acebuche<sup>16</sup> [201] y Picual<sup>17</sup> [129]. Los archivos utilizados se detallan en la tabla 3.6. Para la detección de este alérgeno se ha empleado

<sup>13</sup><https://phytozome.jgi.doe.gov/pz/portal.html>

<sup>14</sup><http://www.ncbi.nlm.nih.gov/bioproject/PRJEB9470>

<sup>15</sup><https://denovo.cnag.cat/olive>

<sup>16</sup><http://olivegenome.org/>

<sup>17</sup><https://genomaolivar.dipujaen.es>

**Tabla 3.5:** *Librerías utilizadas para desarrollo de marcadores de RT-PCR en Arabidopsis thaliana. Adaptada de [266].*

| Variedad   | Réplica    | Lecturas brutas | Lecturas útiles      |
|------------|------------|-----------------|----------------------|
| Columbia-0 | Col-0-rep1 | 12 251 921      | 11 861 826 (96,82 %) |
| Columbia-0 | Col-0-rep2 | 12 297 628      | 11 912 459 (96,87 %) |
| Columbia-0 | Col-0-rep3 | 13 076 233      | 12 678 437 (96,96 %) |
| Killeen-0  | Kil-0-rep1 | 11 210 913      | 10 914 810 (97,36 %) |
| Killeen-0  | Kil-0-rep2 | 11 523 570      | 11 198 260 (97,18 %) |
| Killeen-0  | Kil-0-rep3 | 9 107 610       | 8 859 088 (97,27 %)  |

tanto la secuencia parcial de *Ole e 7* de 21 aminoácidos (P81430) [183] como la secuencia completa de este alérgeno de 95 aminoácidos descrita en [184].

**Tabla 3.6:** *Transcriptomas y proteomas derivados de los genomas de olivo de las variedades Farga [135], acebuche [201] y Picual [129].*

| Variedad | Versión  | Ómica         | Archivo                                            | N.º secuencias |
|----------|----------|---------------|----------------------------------------------------|----------------|
| Farga    | Oe6      | Transcriptoma | OE6A.transcripts.fa                                | 89 982         |
|          |          | Proteoma      | OE6A.pep.fa                                        | 79 910         |
| Acebuche | Oeu      | Transcriptoma | Olea_europaea.gene.cds.final.chr_and_chrUn_noTE.fa | 50 684         |
|          |          | Proteoma      | Olea_europaea.gene.pep.final.chr_and_chrUn_noTE.fa | 50 684         |
| Picual   | Oleur061 | Transcriptoma | Olea_europaea_CDS_v061.fasta                       | 79 867         |
|          |          | Proteoma      | Olea_europaea_proteins_v061.fasta                  | 79 867         |

### 3.4. Obtención de transcriptomas de ReprOlive

En esta sección se describen las distintas etapas que se han llevado a cabo para la obtención de los transcriptomas y demás información contenidos en la base de datos ReprOlive<sup>18</sup> [57]. Para el preprocesamiento de las lecturas brutas y el ensamblaje de los transcriptomas se ha seguido el mismo flujo de trabajo desarrollado previamente en nuestro laboratorio para los transcriptomas de pino y lengüado [54, 56]. El flujo de trabajo llevado a cabo ha sido automatizado usando la herramienta AutoFlow (ver apartado 3.2.7). La figura 3.2 muestra el diagrama del flujo, mientras que la plantilla con todas las tareas desde el preprocesamiento hasta la anotación y las líneas de comando de cada programa puede consultarse en el apéndice A.

<sup>18</sup><http://reprolive.eez.csic.es>

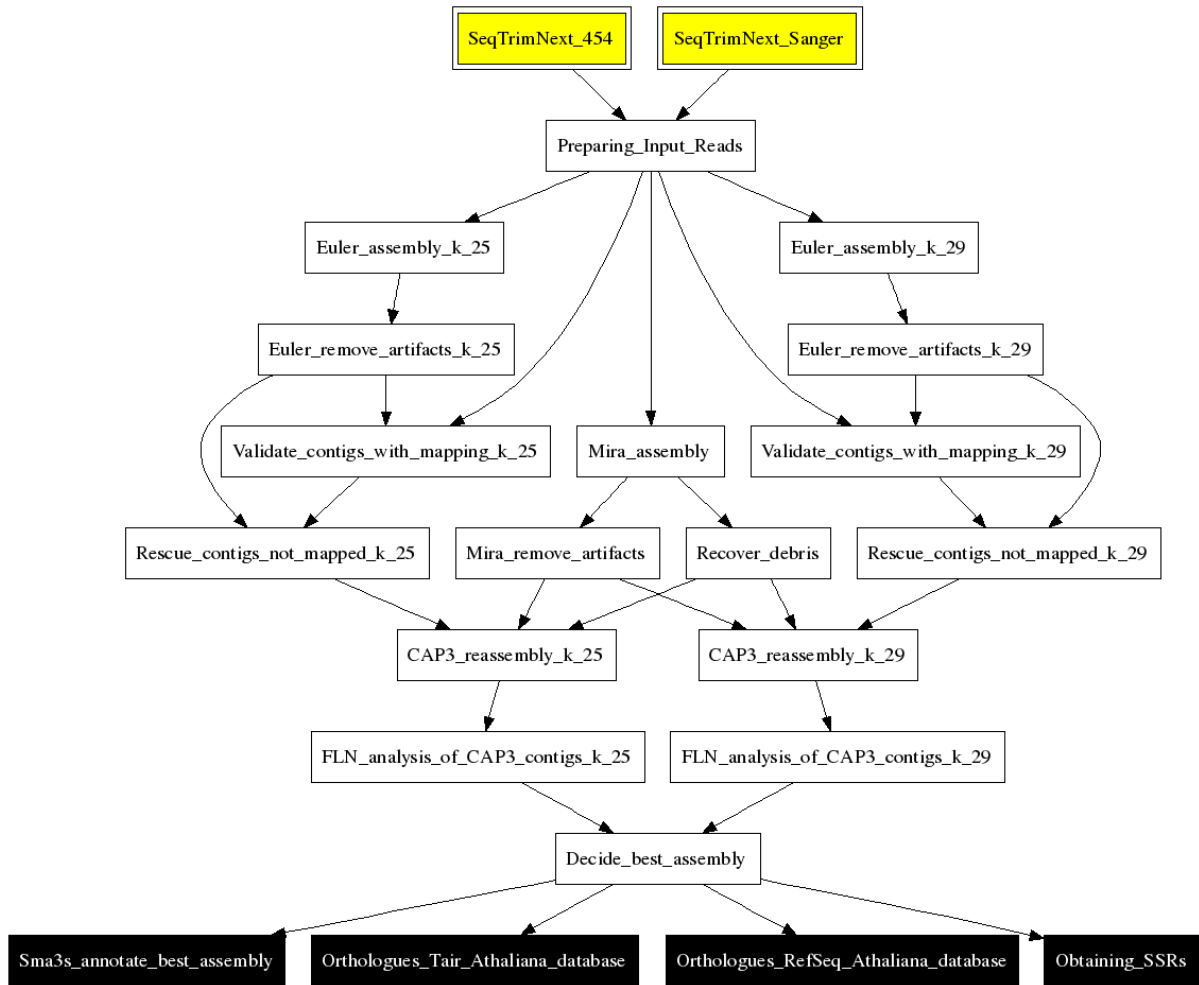


**Figura 3.2:** Diagrama de flujo de la estrategia para el preprocesamiento, ensamblaje y anotación de los transcriptomas, generado con AutoFlow. Las cajas amarillas indican las entradas, mientras que las negras proporcionan los resultados de salida. Tomada de [57].

### 3.4.1. Preprocesamiento de las lecturas brutas

Se ha llevado a cabo el preprocesamiento de las lecturas brutas utilizando SeqTrimNext. Aquellas secuencias de 454 de menos de 40 pb y las secuencias Sanger inferiores a 100 pb fueron descartadas del análisis. El número total de lecturas útiles tras el preprocesamiento figura en la tabla 3.2. En el caso del polen, un total de 709 251 lecturas produjeron 373 268 lecturas útiles (52,6 %), en el pistilo de 767 963 resultaron 430 165 lecturas útiles (56 %). Las lecturas rechazadas corresponden a secuencias contaminantes y secuencias de baja calidad. En cuanto a los tejidos vegetativos, de 600 243 lecturas, fueron útiles 89 403 (14,9 %). El bajo rendimiento en el caso de tejidos vegetativos se debió a problemas de contaminación con rRNA y hongos (*Saccharomyces cerevisiae* y *Aspergillus niger*).

### 3.4.2. Ensamblaje de los transcriptomas

Se siguió el protocolo que se explica en párrafos que vienen a continuación para construir cuatro transcriptomas: un transcriptoma reproductivo con las librerías de polen y pistilo (tabla 3.2), uno exclusivo de polen, otro exclusivo de pistilo y por último, un transcriptoma vegetativo, como control, a partir de las librerías de hoja y raíz. Los TT de estos transcriptomas de polen, pistilo, tejido reproductivo y vegetativo se nombraron con los prefijos “po11\_olive”, “st11\_olive” (de *stigma*), “rp11\_olive” y “vg11\_olive”, respectivamente.

Una vez preprocesadas las lecturas útiles fueron ensambladas combinando dos programas de algoritmos muy diferentes para aprovechar los puntos fuertes de cada uno de ellos, MIRA3 y Euler-SR. Tras en ensamblaje con MIRA3, los *contigs* obtenidos se evaluaron con Full-LengtherNext para eliminar artefactos (tarea *Mira\_remove\_artifacts* en figura 3.2). Además, las lecturas descartadas en el ensamblaje (llamados *debris*) también se evaluaron con Full-LengtherNext, para recuperar aquellas que sean codificantes (tarea *Recover\_debris* en figura 3.2).

Con Euler-SR se llevaron a cabo dos ensamblajes independientes, con *k*-meros 25 y 29 respectivamente. Los *contigs* obtenidos también se evaluaron con Full-LengtherNext para eliminar artefactos (tarea *Euler\_remove\_artifacts\_k\** en figura 3.2) y además se validaron mediante el mapeo de las lecturas con Bowtie2 sobre los *contigs* generados, rescatando aquellos sobre los que mapea al menos una lectura (tarea *Validate\_contigs\_with\_mapping\_k\** en figura 3.2). Como Euler-SR fragmenta las lecturas en *k*-meros, puede ocurrir que genere *contigs* más pequeños que el tamaño de la lectura, que no serán mapeados, pero que pueden ser codificantes y por tanto interesantes, por lo que se llevó a cabo otro análisis con Full-LengtherNext sobre ellos para rescatarlos (tarea *Rescue\_contigs\_not\_mapped\_k\** en figura 3.2).

Todos los *contigs* de interés obtenidos se reconciliaron con CAP3 usando un porcentaje mínimo de identidad del 95 % y una longitud mínima de solapamiento del 40 pb para obtener un conjunto final de transcritos tentativos (TT) con la secuencia más próxima a los transcritos originales [267, 55]. Se hicieron dos reconciliaciones diferentes: los *contigs* de MIRA3 junto con los de Euler-SR con *k*-mero 25 y por otra parte, los *contigs* de MIRA3 junto con los de Euler-SR con *k*-mero 29. Se generaron por tanto dos ensamblajes diferentes y se utilizó Full-LengtherNext para decidir cuál de ellos era el mejor.

### 3.4.3. Anotación de los transcriptomas

Una vez seleccionado el mejor ensamblaje en cada caso, se llevó a cabo una anotación funcional del conjunto de TT obtenido, indispensable para futuros estudios comparativos. Se utilizaron varios programas para anotar, combinando con ello distintas fuentes de información que se comentan a continuación.

La anotación con Sma3s incluyó el nombre del gen y su descripción (tomados del ortólogo más cercano de la división de plantas de UniProtKB), términos GO y códigos EC e InterPro. La división de plantas se volvió no redundante para eliminar anotaciones espurias. A partir de los códigos EC se recuperaron los mapas de rutas KEGG.

La anotación con Full-LengtherNext generó (1) otra descripción para el gen basada en el ortólogo en plantas más próximo (> 45 % de identidad), (2) los codones putativos de inicio y fin de la transcripción, (3) la secuencia aminoacídica deducida, (4) el estado del ORF (si es una proteína completa o incompleta), (5) los ncRNA putativos (excluidos los miRNA maduros), (6) los ortólogos en *Arabidopsis thaliana* de las bases de datos TAIR10 [116] y RefSeq [105] (en la versión de noviembre de 2012), (7) si se trata o no de una secuencia desconocida, pero codificante (según TransDecoder), y (8) un transcriptoma mínimo o de referencia.

Para detectar microsatélites o SSR como posibles marcadores genéticos, se utilizó el programa MREPS contando repeticiones cuya longitud fuera como mínimo de 2 nucleótidos, el periodo o número de nucleótidos que se repiten fuera como mínimo 2 y el número mínimo de veces que se repiten sea como mínimo 3.

La detección de SNP en poblaciones naturales requiere una gran cantidad de datos y su interpretación es difícil [56], por lo que su predicción no se ha abordado en este trabajo.

### 3.4.4. Datos de expresión

Al contrario que la tecnología de Illumina, la plataforma Roche FLX produce un número limitado de lecturas, por lo que no es apropiada para llevar a cabo análisis de expresión génica cuantitativos (véase el apartado 1.1.2). No obstante, en nuestro caso se ha llevado a cabo una aproximación que permite un análisis cualitativo de la expresión.

Las librerías descritas en la tabla 3.2 se combinaron entre sí para obtener un set de lecturas de polen (agrupando las librerías PM-Subs, PM, PG1 y PG2, 373 268 lecturas en total), uno de pistilo (librerías S2, S3, S4 y S4-Subs, 430 165 lecturas) y otro de tejidos vegetativos (librerías L, L-Subs, R1 y R2, 89 403 lecturas). Estas lecturas se mapearon sobre todos los transcriptomas de referencia incluidos en ReprOlive usando Bowtie2 y permitiendo que cada lectura mapee en cada TT complementario. Las lecturas mapeadas se contaron con Bio-samtools de BioRuby y se incluyeron en la base de datos tanto los recuentos brutos, como valores de RPKM calculados como sigue:

$$RPKM = \frac{n.^{\circ} \text{ de lecturas que mapean sobre el TT}}{\left(\frac{\text{longitud del TT}}{1\,000}\right) \cdot \left(\frac{\text{total de lecturas mapeadas sobre el transcriptoma}}{1\,000\,000}\right)} \quad (3.1)$$

Esta aproximación permite determinar qué genes se están expresando o no en un determinado tejido.

### 3.4.5. Construcción de la base de datos ReprOlive

Debido al gran volumen de datos generado tras la construcción y anotación funcional completa de los transcriptomas, en el laboratorio se creó, basado en Ruby-on-Rails, una base de datos específica para albergarlos y hacerlos disponibles a través de la web. El desarrollo corrió a cargo de Pedro Seoane [43], y sigue el modelo de otras bases de datos desarrolladas en nuestro laboratorio por Noé Fernández Pozo e Hicham Benzekri, basadas en trabajos previos de Darío Guerrero [54, 56, 55]. ReprOlive [57] sigue estando disponible en <http://reprolive.eez.csic.es> y consta de cuatro máquinas virtuales (figura 3.3) que impiden su bloqueo cuando hay peticiones concurrentes, y también facilitan su migración y el mantenimiento:

- **Máquina para base de datos:** Contiene una base de datos relacional MySQL que alberga toda la información producida por el flujo de trabajo.
- **Máquina para cálculo:** se ocupa de las tareas de cálculo intensivo, como por ejemplo la realización de BLAST a petición del usuario.
- **Máquina para interfaz web:** Contiene la aplicación web completa y también se encarga de la importación automática de los nuevos transcriptomas, cuando es necesario.

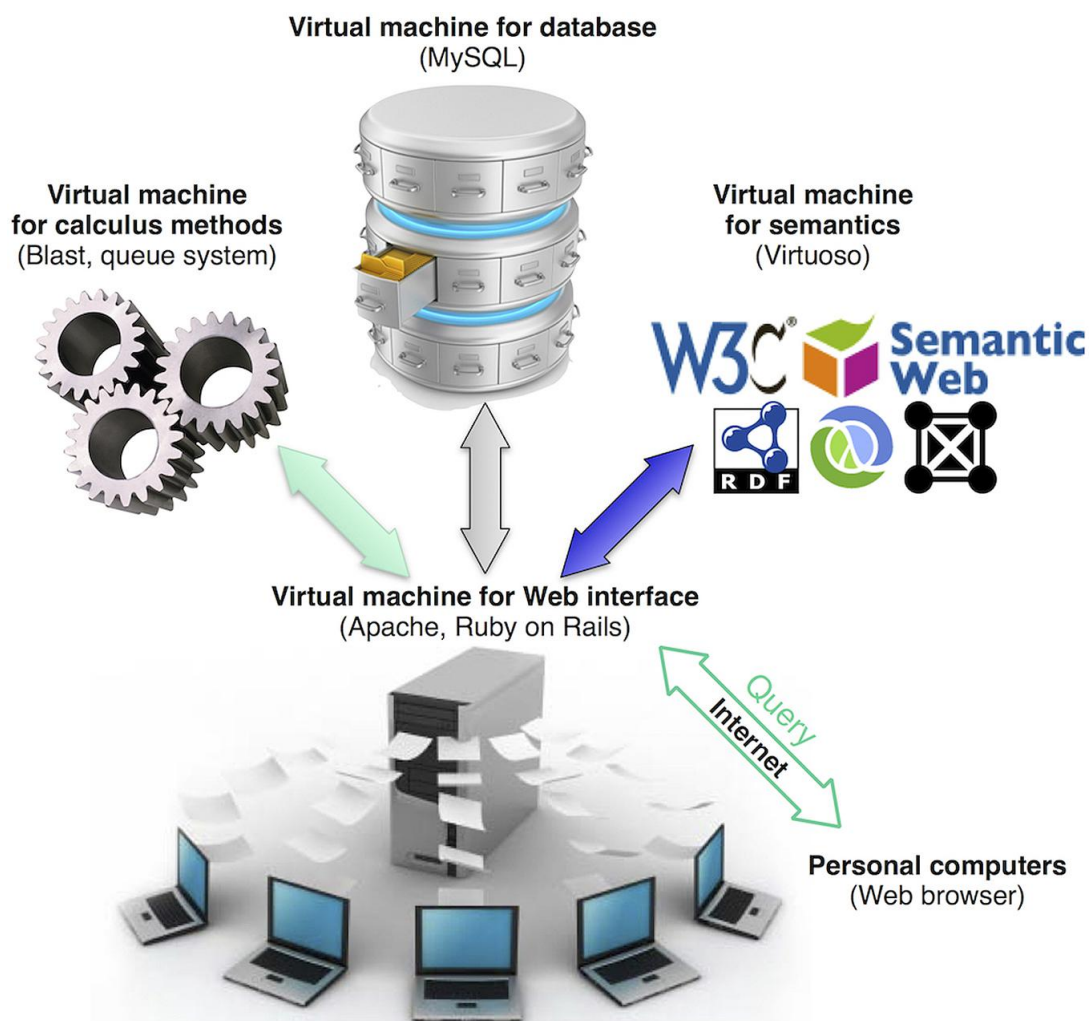


**Figura 3.3: Organización de la base de datos en máquinas virtuales.** A través de un navegador web el usuario puede recuperar información realizando búsquedas o navegando por la interfaz. El uso de las diferentes máquinas virtuales es transparente para el usuario, que solo ve las entradas y las salidas en el navegador de la web. Tomada de [57].

- **Máquina para semántica:** Permite realizar búsquedas semánticas en servidores de información por medio de un marco de descripción de recursos (RDF) que permite una búsqueda de datos vinculados. Esto permite al usuario formular consultas más intuitivas y ágiles en otras bases de datos, basándose en la utilización de conceptos en vez de la comparación de cadenas de texto, principio en el que se basa la web semántica [268]. Esta máquina supone una novedad con respecto al las otras bases de datos del grupo de investigación, y surgió de la colaboración con el grupo Khaos de Ismael Navas y José Aldana en la Universidad de Málaga.

### 3.5. Obtención del transcriptoma mejorado de polen maduro

Se ha utilizado la herramienta TransFlow (apartado 3.2.7) para obtener un transcriptoma de polen más completo que el generado en ReprOlive y centrado en el polen maduro. Para ello se han combinado las lecturas 454/Roche correspondientes a polen maduro que se utilizaron en ReprOlive (tabla 3.2) con lecturas pareadas de Illumina (tabla 3.3), por lo que se han ejecutado los cinco módulos de TransFlow. En el apéndice B se pueden consultar tanto el archivo .sh de lanzamiento de TransFlow (apéndice B.1), como las plantillas de los distintos módulos de TransFlow (apéndices B.2, B.3, B.4, B.5 y B.6).

Durante el preprocesamiento con SeqTrimNext de las lecturas de Illumina (módulo 1) se descartaron del análisis aquellas de menos de 60 nt, mientras que para las lecturas de 454/Roche (módulo 2) se eliminaron las menores de 40 pb.

Las especies que se usaron como transcriptomas de referencia para evaluar los distintos ensamblajes (módulo 4) fueron *Arabidopsis thaliana* y *Populus trichocarpa*, ya que han resultado ser los mejores tras la evaluación con la propia herramienta TransFlow entre varias especies de plantas bien caracterizadas [126]. Los transcriptomas de referencia y las lecturas utilizadas para estas especies se detallan en el apartado 3.3.2.

Los  $k$ -meros empleados en los ensamblajes que los requerían han sido 25 y 35 (módulo 2 y 3), y el taxón empleado por BUSCO ha sido embriófitas, concretamente la versión Odb9 (módulo 5).

### 3.6. Detección de genes de referencia para RT-PCR a partir de datos de RNA-seq

Se ha utilizado la herramienta AutoFlow (apartado 3.2.7) para diseñar un flujo de trabajo que permita establecer nuevos marcadores o genes de referencia (RG) estables y altamente expresados a lo largo de todas las condiciones experimentales de interés a partir de datos de RNA-seq.

#### 3.6.1. Flujo de trabajo para la detección de genes de referencia

El flujo de trabajo (figura 3.4) consta de un primer bloque para proporcionar una tabla con el recuento de lecturas por gen y condición, y a continuación otro bloque donde se analizan estos datos para detectar los genes cuya expresión es elevada y prácticamente constante en todas las condiciones. Los detalles se describen a continuación:

#### Obtención de la tabla de recuentos de lecturas

Este primer bloque (figura 3.4A) recibe como archivos de entrada el transcriptoma y las lecturas brutas, ambos en formato FASTA. Cada uno de los archivos de lecturas se considera una condición experimental. Estas lecturas se preprocesan con la herramienta SeqTrimNext para eliminar secuencias problemáticas y retener únicamente las lecturas fiables. Las lecturas útiles se mapean sobre el transcriptoma correspondiente usando Bowtie2 con los parámetros por defecto y la opción *-a* que permite que cada lectura mapee en todos los transcritos posibles. Para hacer el recuento de las lecturas mapeadas se usa Bio-samtools de BioRuby. En caso de tratarse de lecturas pareadas de Illumina se utiliza la opción *-f2* para contar únicamente aquellas lecturas donde ambos extremos de la pareada mapean en el mismo

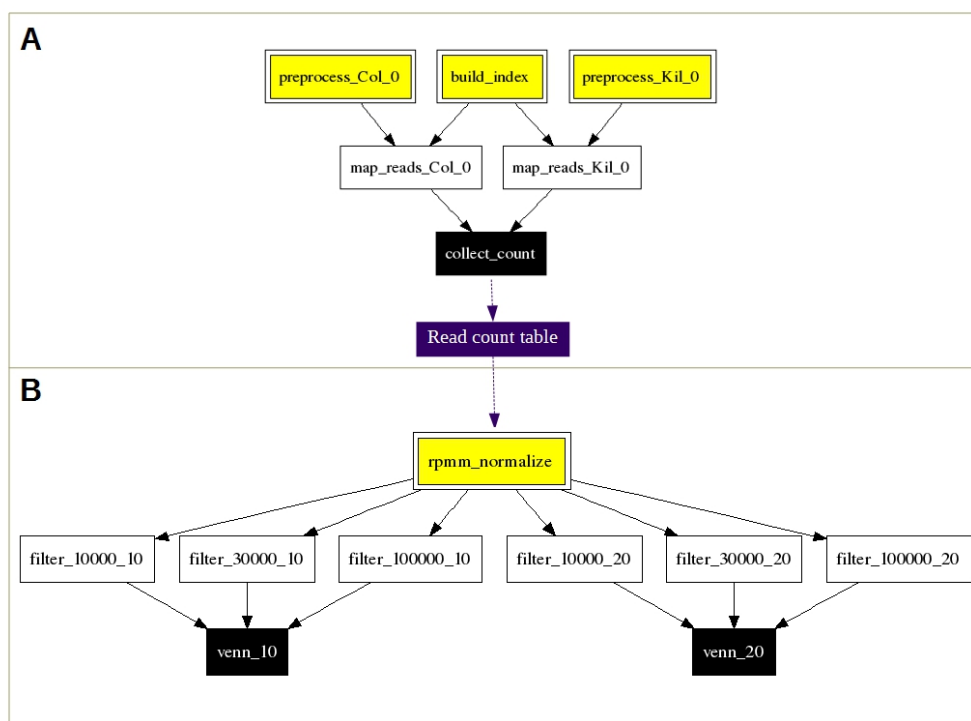


**Figura 3.4:** Diagrama de flujo generado con AutoFlow para la detección de genes de referencia (RG) en *Arabidopsis* usando los sets de datos de SRA (identificador de Bio-Project ID PRJEB9470). (A) El primer flujo prepara las lecturas, las mapea sobre el transcriptoma y proporciona la tabla de recuentos. (B) El segundo flujo (llamado *findRGs*) detecta candidatos a RG. En este ejemplo se probaron distintos valores de los parámetros de filtrado: 10 y 20 % de coeficiente de variación (CV) máximo, y 10 000, 30 000 y 100 000 para el mínimo de lecturas mapeadas por transcrito y condición experimental. Se obtiene un diagrama de Venn para cada uno de los umbrales de CV. Tomada de [266].

transcrito. Como resultado se obtiene un tabla en formato tabulado con los recuentos de lecturas, donde las filas representan los distintos transcritos, las columnas las condiciones experimentales y los valores son el número de lecturas que mapea sobre cada transcrito en cada condición.

### Detección de genes de referencia

Este segundo bloque del flujo de trabajo (llamado *findRGs*, figura 3.4 B) buscará los RG más estables y con mayor expresión a lo largo de todas las condiciones experimentales en estudio. Requiere como entrada un fichero tabulado de recuentos de lecturas como el que se genera en el bloque anterior, o bien puede ser aportado externamente por el usuario, siempre que presente el mismo formato. Opcionalmente se puede aportar un fichero de anotación tabulado con al menos dos columnas: una primera con el identificador de cada transcrito y una segunda con el identificador de referencia de un gen ortólogo. Se recomienda añadir además una tercera columna con una descripción del ortólogo. Los valores de la tabla de recuentos se normalizan como el número de lecturas mapeadas sobre un transcrito dividido entre

el número total de lecturas de la muestra correspondiente mapeadas sobre el transcriptoma (RPMM o lecturas por millón de lecturas mapeadas):

$$RPMM = \frac{n.^{\circ} \text{ de lecturas que mapean sobre el TT}}{\text{total de lecturas mapeadas sobre el transcriptoma}} \cdot 1\,000\,000 \quad (3.2)$$

A continuación se calcula para cada transcrito el coeficiente de variación (CV, cociente entre la desviación estándar y la media, expresado en porcentaje) de los valores de RPMM a lo largo de todas las condiciones experimentales. La tabla normalizada con los valores de RPMM y CV se puede filtrar en función de dos parámetros personalizables: (1) CV máximo (se aconseja un 10 %, aunque el intervalo puede ir de 0 hasta 20) para seleccionar genes cuya expresión sea lo más invariable posible a lo largo de todas las condiciones (a menor valor de CV, la expresión es más homogénea), y (2) número mínimo de lecturas por transcrito y condición (se recomienda un mínimo de 10, aunque se propone comenzar con el valor resultante de multiplicar 0,00003 por el tamaño de la librería más pequeña) para seleccionar los RG con los niveles más altos de expresión y así garantizar la correcta amplificación por PCR (a mayor valor de este parámetro, mayor nivel de expresión del transcrito). El flujo de trabajo permite combinar diferentes valores de estos dos parámetros en una sola ejecución. Si el transcriptoma de referencia que se usa para el mapeo sobreestima el número de transcritos, muchos de ellos harán referencia al mismo gen. En ese caso, la tabla de anotación opcional sirve para detectar transcritos que comparten el mismo ortólogo o identificador, y filtrarlos para conservar solo aquel con el mayor valor de RPMM. Por último se genera un diagrama de Venn que muestra el número de ortólogos específicos y comunes entre las diferentes combinaciones de los parámetros utilizados, para visualizar los resultados y comprobar la idoneidad de tales parámetros. En el apéndice C puede consultarse tanto el archivo .sh de lanzamiento (apéndice C.1) como la plantilla de AutoFlow para *findRGs* (apéndice C.2).

### 3.6.2. Detección de genes de referencia en el olivo

En el caso de olivo se llevaron a cabo tres ejecuciones del flujo previamente descrito: polen, pistilo y ambos tejidos reproductivos juntos. En cada ejecución las lecturas correspondientes a cada estadio de desarrollo por separado del tejido en cuestión descritas en el apartado 3.3.1 (longitud media de 385 nt) se mapearon por separado sobre el transcriptoma reproductivo de ReprOlive (tabla 4.1), pero solo sobre los transcritos que codifican proteínas completas (9157 transcritos). Los valores de los parámetros de filtrado de *findRGs* que se usaron en este caso fueron: CV máximo 10 % (defecto) y 20 % (no estricto), y 10, 50 y 100 (con fines comparativos) para el mínimo de lecturas mapeadas por transcrito y condición experimental.

### Validación experimental de genes de referencia en el olivo

Se llevó a cabo una validación experimental preliminar de algunos de los candidatos a RG por parte de María José Jiménez-Quesada mediante RT-PCR semi-cuantitativa a tiempo real, de acuerdo a Alché et al. [170]. Diversos transcritos de distintos tejidos de olivo (hoja, inflorescencia, polen maduro, pistilo y semilla, con fines comparativos) se sometieron a 25 ciclos de amplificación para asegurar una tasa de amplificación exponencial. Los genes evaluados fueron: 18S, ubicuitina, actina, proteína cinasa activada por mitógeno, transportador de prolina, homólogo de la L-ascorbato oxidasa, gliceraldehído-3-fosfato deshidrogenasa, proteína de unión a ácido salicílico y proenzima S-adenosilmetionina descarboxilasa.

### 3.6.3. Detección de genes de referencia en *Arabidopsis*

En el caso de *Arabidopsis* se llevaron a cabo tres ejecuciones del flujo previamente descrito, una por cada réplica. En cada ejecución, las lecturas de cada una de las variedades de *Arabidopsis thaliana* (apartado 3.3.3) se mapearon sobre el transcriptoma de referencia de *Arabidopsis* (35 386 transcritos) descrito en la tabla 3.4. En este caso, los valores de los parámetros de filtrado de *findRGs* usados fueron: CV máximo 10 % (defecto) y 20 % (no estricto), y 10 000, 30 000 y 100 000 (con fines comparativos) para el mínimo de lecturas mapeadas por transcrito y condición experimental.

En la tabla 3.7 se pueden consultar los porcentajes de lecturas útiles tras el preprocesamiento y el porcentaje de mapeo sobre los transcriptomas descritos, tanto para el caso de olivo como el de *Arabidopsis*.

**Tabla 3.7: Estadísticas sobre el preprocesamiento con SeqTrimNext y el mapeo con Bowtie2 de las librerías utilizadas para desarrollo de marcadores de RT-PCR en el olivo y *Arabidopsis thaliana*. Adaptada de [266].**

|                                    |            | Lecturas brutas | Lecturas útiles     | Lecturas mapeadas   |
|------------------------------------|------------|-----------------|---------------------|---------------------|
| <b>Olivo</b>                       |            |                 |                     |                     |
| Librería                           | Tejido     |                 |                     |                     |
| PM                                 | Polen      | 217 163         | 111 760 (51,5 %)    | 76 197 (68,2 %)     |
| PG1                                | Polen      | 258 167         | 141 232 (54,7 %)    | 96 746 (68,5 %)     |
| PG5                                | Polen      | 233 921         | 120 276 (51,4 %)    | 84 081 (69,9 %)     |
| S2                                 | Pistilo    | 257 813         | 138 077 (53,6 %)    | 63 030 (45,7 %)     |
| S3                                 | Pistilo    | 247 401         | 141 903 (57,4 %)    | 65 934 (46,5 %)     |
| S4                                 | Pistilo    | 262 749         | 150 185 (57,2 %)    | 71 044 (47,3 %)     |
| <b><i>Arabidopsis thaliana</i></b> |            |                 |                     |                     |
| Variedad                           | Réplica    |                 |                     |                     |
| Columbia-0                         | Col-0-rep1 | 12 251 921      | 11 861 826 (96,8 %) | 11 143 112 (93,9 %) |
| Columbia-0                         | Col-0-rep2 | 12 297 628      | 11 912 459 (96,9 %) | 11 178 500 (93,8 %) |
| Columbia-0                         | Col-0-rep3 | 13 076 233      | 12 678 437 (97,0 %) | 11 988 627 (94,6 %) |
| Killeen-0                          | Kil-0-rep1 | 11 210 913      | 10 914 810 (97,4 %) | 7 544 520 (69,1 %)  |
| Killeen-0                          | Kil-0-rep2 | 11 523 570      | 11 198 260 (97,2 %) | 9 214 640 (82,3 %)  |
| Killeen-0                          | Kil-0-rep3 | 9 107 610       | 8 859 088 (97,3 %)  | 6 945 942 (78,4 %)  |

## 3.7. Detección de posibles nuevos alérgenos de olivo

Se ha utilizado la herramienta AutoFlow (apartado 3.2.7) para diseñar un flujo de trabajo que permita detectar posibles alérgenos nuevos de polen de olivo, así como posibles nuevos casos de reactividad cruzada de olivo con otras especies, partiendo del transcriptoma de polen obtenido en ReprOlive (apartado 4.1.1).

### 3.7.1. Bases de datos derivadas de Allergome

Para el flujo de trabajo se crearon dos bases de datos derivadas de UniProt<sup>19</sup> [269] y Allergome<sup>20</sup> [270, 271] con fines comparativos. Estas bases de datos se obtuvieron haciendo uso del panel de búsqueda avanzada de UniProt, que entre múltiples opciones permite por ejemplo buscar y descargar en formato FASTA secuencias proteicas de especies o taxones concretos contenidas en determinadas bases de datos, como puede ser Allergome.

- ***Olive allergens DB***: Base de datos de alérgenos de olivo. Incluye todas las secuencias proteicas presentes en UniProt en el momento de la fecha de descarga (11 de julio de 2017) que corresponden a *Olea europaea* y que tienen al menos una referencia cruzada a la base de datos Allergome. Incluye 107 secuencias que corresponden a secuencias completas o parciales de los isoalérgenos e isoformas de los alérgenos de olivo identificados hasta ese momento (*Ole e 1* a *Ole e 13*).
- ***Non-olive allergens DB***: Base de datos de alérgenos que no son de olivo. Contiene 4161 secuencias que corresponden a todas las secuencias proteicas presentes en UniProt en el momento de la fecha de descarga (11 de julio de 2017) pertenecientes a cualquier organismo distinto de *Olea europaea*, ya sea animal, planta, hongo o bacteria, con al menos una referencia cruzada a la base de datos Allergome.

### 3.7.2. Flujo de trabajo para la identificación de alérgenos

El algoritmo está esquematizado en la figura 3.5 y consta de las siguientes etapas:

#### ■ Etapa A: Detección de alérgenos ya conocidos

El algoritmo comienza con la detección de los alérgenos de olivo ya conocidos que se han conseguido ensamblar en el transcriptoma de polen de ReprOlive, para poder descartarlos del análisis. Para ello se lleva a cabo una comparación con BLASTX (v2.2.29+) masiva y estricta entre los TT del transcriptoma de polen de ReprOlive ( $TT_{pollen}$ ) y *Olive allergens DB*, utilizando los siguientes filtros:  $E < 10^{-1}$ , identidad  $\geq 80\%$  y una cobertura de alineamiento  $\geq 80\%$  de la longitud del alérgeno de la base de datos. Estos valores tan restrictivos aseguran una asignación fiable entre los  $TT_{pollen}$  y el repertorio de alérgenos de olivo de la base de datos (figura 3.5A). De esta forma, los  $TT_{pollen}$  que pasen los filtros se consideran alérgenos de olivo ya conocidos ( $TT_{pollen}^{known}$ ), mientras que los otros suponen una fuente de posibles alérgenos nuevos ( $TT_{pollen}^{source}$ ).

#### ■ Etapa B: Detección de reactividad cruzada

Para obtener información acerca de la reactividad cruzada de los  $TT_{pollen}^{known}$  con alérgenos de otras especies se lleva a cabo una comparación con BLASTX con *Non-olive allergens DB*, de acuerdo a las directrices de la FAO/WHO sobre la evaluación de la alergia. Según estas, un alérgeno candidato a producir reactividad cruzada debe cumplir alguno de los siguientes requisitos: su identidad con la secuencia completa de la proteína alergénica debe ser  $\geq 50\%$ ; o bien  $\geq 35\%$  dentro de una ventana de secuencia de 80 aminoácidos; o bien ser 100 % idéntico en una ventana de 8 aminoácidos (figura 3.5B). Los alérgenos de otras especies ( $AP_{non-olive}$ , de *allergenic proteins*) que cumplan alguno de estos requisitos podrían producir reacciones cruzadas con olivo.

<sup>19</sup><https://www.uniprot.org>

<sup>20</sup><http://www.allergome.org>

### ■ Etapa C: Búsqueda de posibles alérgenos nuevos

Los  $TT_{pollen}^{source}$  se comparan con *Non-olive allergens DB* en condiciones menos estrictas de filtrado: identidad  $\geq 40\%$  y una cobertura de alineamiento  $\geq 40\%$  de la longitud del alérgeno de la base de datos (figura 3.5C). Los transcritos que cumplan con estos criterios mínimos se consideran candidatos a alérgenos nuevos de polen de olivo ( $TT_{pollen}^{newcand}$ ).

### ■ Etapa D: Distinción entre posibles alérgenos nuevos e isoformas distantes

Los  $TT_{pollen}^{newcand}$  se comparan de nuevo con *Olive allergens DB*, pero usando una combinación de parámetros de filtrado más laxa que en la etapa A. En este caso, las secuencias con una identidad  $\geq 60\%$  y una cobertura de alineamiento  $\geq 60\%$  de la longitud del alérgeno de la base de datos son clasificadas como nuevas isoformas más distantes de los alérgenos de olivo ya conocidos ( $TT_{pollen}^{iso}$ ), mientras que los  $TT_{pollen}^{newcand}$  que no coinciden con ningún alérgeno de olivo en esta comparación, se consideran posibles alérgenos nuevos de polen de olivo ( $TT_{pollen}^{new}$ ), no descritos previamente en la bibliografía (figura 3.5D).

En el apéndice D se pueden consultar tanto el archivo .sh de lanzamiento del algoritmo de búsqueda de alérgenos (D.1), como la plantilla de AutoFlow con las distintas etapas del algoritmo (D.2).

### 3.7.3. Identificación de *Ole e 7* en transcriptomas y genomas de olivo

Para profundizar en la identificación del alérgeno *Ole e 7* se han llevado a cabo comparaciones con BLASTX entre las dos secuencias publicadas de este alérgeno (apartado 3.3.4), el transcriptoma de polen de ReprOlive y el transcriptoma de polen mejorado. Así mismo, se ha analizado la presencia de la secuencia completa de *Ole e 7* tanto en los transcriptomas (BLASTX) como en los proteomas (BLASTP) correspondientes a los genomas de las variedades Farga, acebuche y Picual (apartado 3.3.4). En todas las comparaciones se ha utilizado la versión v2.2.29+ de BLAST, con  $E < 10^{-1}$ .

Las secuencias obtenidas en estas comparaciones se han alineado utilizando Clustal  $\Omega$  de la plataforma SeaView (apartado 3.2.1) para su visualización.

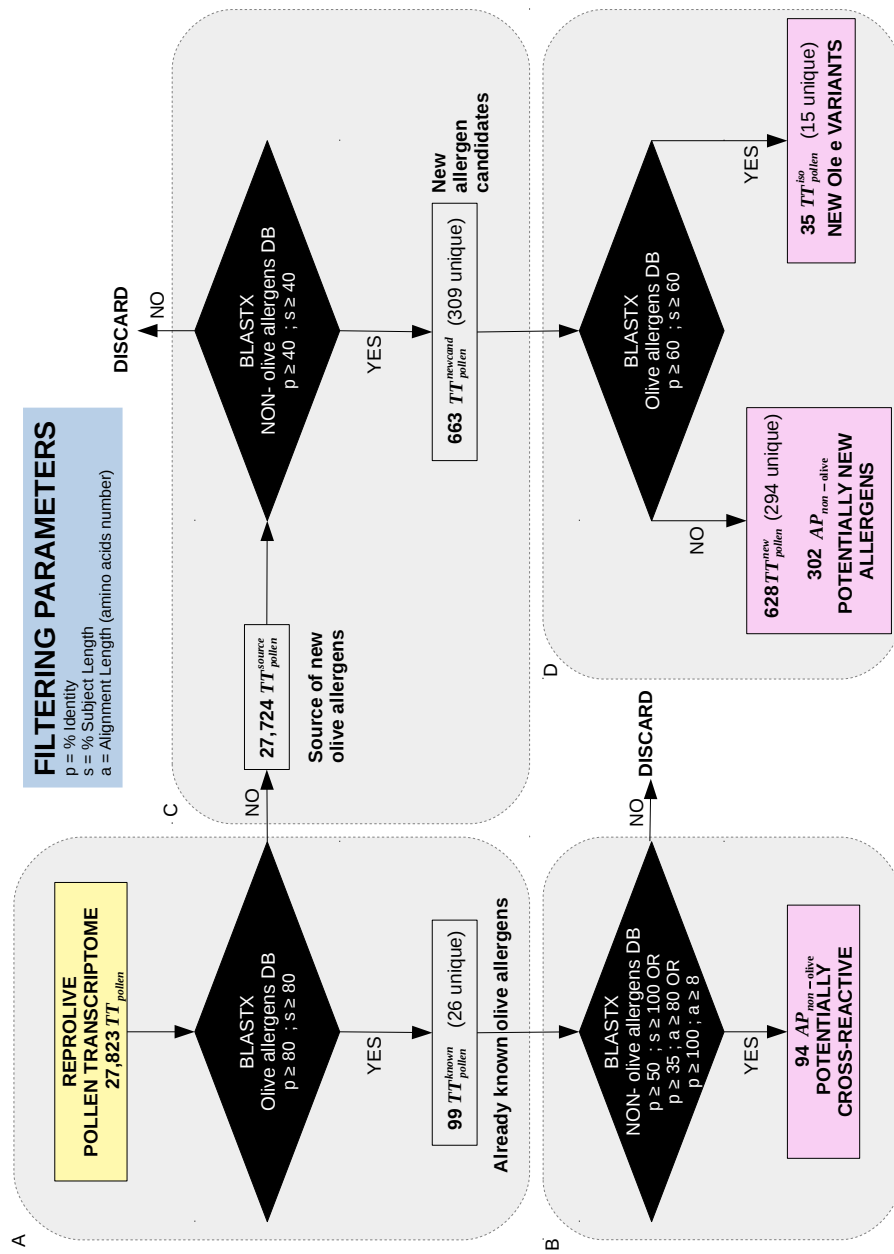


**Figura 3.5: Etapas del flujo de trabajo usado para clasificar los diferentes  $TT_{pollen}$  de ReprOlive con respecto a su posible alergia.** (A) Comparación del transcriptoma de polen con Olive allergens DB para detectar los alérgenos de polen de olivo ya conocidos ( $TT_{known pollen}$ ). (B) Comparación entre los  $TT_{known pollen}$  y Non-olive allergens DB para detectar reactividad cruzada de los alérgenos de polen de olivo con proteínas alérgicas de otras especies ( $AP_{non-olive}$ ). (C) Comparación de los transcritos considerados como fuente de alérgenos nuevos ( $TT_{source pollen}$ ) con Non-olive allergens DB para buscar candidatos a alérgenos nuevos de polen de olivo ( $TT_{new cand pollen}$ ). (D) Nueva comparación con Olive allergens DB de los  $TT_{new cand pollen}$ , en condiciones de filtrado menos estrictas, para distinguir entre posibles alérgenos nuevos de polen de olivo ( $TT_{pollen}^{new}$ ) e isoformas más distantes de los alérgenos de olivo ya conocidos ( $TT_{iso pollen}$ ). Esta figura procede del manuscrito en preparación.

### 3.8. Predicción de modificaciones postraduccionales

Con el fin de investigar el papel de la *S*-nitrosilación y la Tyr-nitración, dos modificaciones postraduccionales (PTM) mediadas por NO, se llevó a cabo una búsqueda bibliográfica de herramientas bioinformáticas libremente disponibles para la predicción de tales PTM, que fuesen capaces de trabajar con un gran número de secuencias. Tras evaluar y probar varias de estas herramientas, se seleccionaron: iSNO-AAPair, GPS-SNO y SNOSite para *S*-nitrosilación, y GPS-YNO2 para Tyr-nitración (véanse apartados 3.2.5 y 3.2.6).

El conjunto de las 21 607 proteínas de polen deducidas del transcriptoma obtenido en ReprOlive (tabla 4.1) se analizó con cada herramienta de forma independiente. En el caso de la herramienta SNOSite, que no admite archivos de tamaño superior a 2 Mb, el archivo FASTA con todas las proteínas de polen se fragmentó en varios archivos, que se analizaron de forma separada para finalmente integrar los resultados.

Los archivos de salida obtenidos en cada caso se procesaron para generar archivos de texto tabulado más fáciles de manejar y solo con la información de interés: el identificador de la proteína en ReprOlive y la posición/es propensas a modificación. Además, se añadieron los siguientes datos de anotación procedentes de ReprOlive y generados con Full-LengtherNext: el ortólogo en plantas más próximo, el estado del ORF (si es una proteína completa o incompleta), una descripción del gen y el identificador del ortólogo correspondiente en *Arabidopsis thaliana* de la base de datos TAIR10, para análisis funcionales posteriores.

En el caso de la *S*-nitrosilación, los resultados de las distintas herramientas se cruzaron entre sí para ver qué posiciones susceptibles de modificación eran predichas por todas a la vez. Para más detalle, en el apéndice E se puede consultar el programa en bash utilizado para procesar la salida de los distintos programas y llevar a cabo la intersección. El parseo de los resultados de la herramienta GPS-YNO2 de predicción de Tyr-nitración fue idéntico al de los resultados de la herramienta GPS-SNO, de predicción de *S*-nitrosilación, puesto que son herramientas desarrolladas por el mismo grupo de investigación con el mismo formato de interfaz gráfica y de ficheros de entrada y salida.

A partir de sus ortólogos en *Arabidopsis* se llevó a cabo un análisis de enriquecimiento biológico con la herramienta GeneCodis (véase apartado 3.2.4) de la lista de proteínas potencialmente *S*-nitrosilables según el programa iSNO-AAPair, y otro para las potencialmente Tyr-nitrables obtenidas con GPS-YNO2, con el umbral por defecto (medio). Se usó la siguiente combinación de parámetros en GeneCodis: el genoma de *Arabidopsis thaliana* como referencia, el nivel más bajo de anotaciones GO, 10 como número mínimo de genes anotados, test hipergeométrico para computar los valores de *P* y corrección por FDR. Se hicieron análisis de enriquecimiento de procesos biológicos, funciones moleculares, componentes celulares y rutas KEGG.

## Capítulo 4

# RESULTADOS

### 4.1. Transcriptoma reproductivo del olivo

#### 4.1.1. El transcriptoma reproductivo del olivo

En el momento de la realización de esta investigación, el genoma de olivo todavía no había sido publicado. Hasta entonces habían salido a la luz numerosos estudios transcriptómicos (véase el apartado 1.2.5) que abordaban distintos aspectos de la biología de esta planta. A sabiendas de que los transcriptomas reproductivos presentan características que los diferencian de los correspondientes a los transcriptomas vegetativos [272, 273], nuestro grupo decidió elaborar un transcriptoma reproductivo del olivo.

Con este objetivo se han construido un transcriptoma reproductivo del olivo con muestras de polen y pistilo en diferentes estadios de desarrollo y un transcriptoma vegetativo como control, a partir de hoja y raíz. Se partió de 2 075 908 lecturas crudas de 454/Roche y 1 549 lecturas Sanger (tabla 3.2). Estas secuencias se preprocesaron, ensamblaron, mapearon y anotaron con datos de expresión, descripciones, términos GO, códigos EC e InterPro, rutas KEGG, ORF y SSR, todo ello mediante un flujo de trabajo predefinido basado en herramientas de *software* libre. Los transcritos tentativos (TT) obtenidos se anotaron también con sus ortólogos correspondientes en *Arabidopsis thaliana* en las bases de datos TAIR y RefSeq. En la tabla 4.1 se muestran las estadísticas de la anotación con Full-LengtherNext para los distintos transcriptomas ensamblados.

Como resultado se obtuvo un transcriptoma reproductivo constituido por 72 846 *contigs* de los que el 87,8 % incluían al menos una anotación funcional y el 75,9 % tenían un ortólogo (tabla 4.1). Como era de esperar, el número de genes diferentes (ortólogos únicos) era menor: solo 23 568 TT diferentes, de los cuales 5 835 contenían un ORF completo. Este transcriptoma se redujo también a un transcriptoma mínimo representativo (obtenido por uno de los métodos implementados en Full-LengtherNext), constituido por 28 972 TT, muy útil para estudios de expresión génica. Se han ensamblado además un transcriptoma exclusivo de polen y otro de pistilo.

**Tabla 4.1: Principales características de los transcriptomas de ReprOlive basados en los análisis con Full-LengtherNext. Extraída de [57].**

| Feature                                                 | Pollen  | Stigma  | Vegetative | Reproductive |
|---------------------------------------------------------|---------|---------|------------|--------------|
| <b>Assembling statistics</b>                            |         |         |            |              |
| Number of useful reads                                  | 373 268 | 430 165 | 89 403     | 803 433      |
| Mean length (nt)                                        | 383     | 385     | 545        | 384          |
| Number of MIRA3 contigs                                 | 54 754  | 73 823  | 42 310     | 116 298      |
| Number of Euler-SR contigs                              | 4 807   | 15 216  | 490        | 16 211       |
| Number of Contigs after CAP3 reconciliation             | 28 094  | 60 964  | 39 425     | 73 589       |
| Number of TTs without chimeras and artefacts*           | 27 823  | 60 400  | 38 919     | 72 846       |
| Mean length (nt)                                        | 608     | 678     | 664        | 686          |
| N50 (nt)                                                | 661     | 780     | 683        | 798          |
| <b>Annotation statistics</b>                            |         |         |            |              |
| Longest TT (nt)                                         | 7 016   | 7 757   | 2 865      | 7 950        |
| Number of ncRNAs                                        | 31      | 17      | 265        | 45           |
| Number of TTs with annotation                           | 24 861  | 54 129  | 36 700     | 63 965       |
| Number of TTs with orthologue                           | 21 607  | 46 910  | 32 076     | 55 356       |
| With unique orthologue IDs                              | 11 672  | 21 326  | 15 003     | 23 568       |
| With orthologue from <i>Arabidopsis thaliana</i> RefSeq | 21 233  | 46 924  | 31 945     | 54 890       |
| Unique RefSeq IDs                                       | 9 769   | 16 565  | 12 489     | 17 612       |
| With orthologue from <i>Arabidopsis thaliana</i> TAIR10 | 21 312  | 47 038  | 31 980     | 55 067       |
| Unique TAIR10 IDs                                       | 8 922   | 14 656  | 11 247     | 15 503       |
| Number of TTs coding a complete protein                 | 2 809   | 7 137   | 3 559      | 9 157        |
| Unique, complete proteins                               | 1 976   | 4 822   | 2 220      | 5 835        |
| Number of TTs without orthologue                        | 6 185   | 13 473  | 6 578      | 17 445       |
| Likely coding for a complete protein                    | 170     | 446     | 242        | 628          |
| Likely coding for an incomplete protein                 | 2 610   | 5 312   | 2 523      | 6 486        |
| <b>Reference transcriptome</b>                          |         |         |            |              |
| Number of representative TTs                            | 13 589  | 25 720  | 17 340     | 28 972       |
| <i>Arabidopsis thaliana</i> RefSeq orthologues          | 10 878  | 20 612  | 14 576     | 22 565       |
| Unique RefSeq IDs                                       | 8 281   | 13 901  | 10 349     | 14 706       |
| <i>Arabidopsis thaliana</i> TAIR10 orthologues          | 10 900  | 20 658  | 14 581     | 22 638       |
| Unique TAIR10 IDs                                       | 7 842   | 12 883  | 9 756      | 13 584       |

(\*) Los artefactos se refieren a repeticiones internas, directas o inversas.

#### 4.1.2. La base de datos ReprOlive

Todos estos transcriptomas han sido incorporados a la base de datos ReprOlive, que dispone de un portal web de libre acceso (<http://reprolive.eez.csic.es>, figura 4.1), que puede ser consultada de forma anónima, con opciones de descarga de los resultados generados.



**Figura 4.1:** Captura de pantalla de la página de inicio de ReprOlive (<http://reprolive.eez.csic.es>). El panel de la izquierda contiene enlaces al historial de versiones de la base de datos y a la última versión del flujo de trabajo basado en AutoFlow. El panel de la derecha muestra información sobre los programas utilizados, la versión de la base de datos y los logos de las entidades responsables de su financiación y desarrollo.

## Ensamblajes y transcritos tentativos

Actualmente ReprOlive aloja las secuencias anotadas de los transcriptomas de polen, pistilo, tejido vegetativo y un transcriptoma reproductivo de olivo que combina polen y pistilo (tabla 4.1 y figura 4.2A). De cada uno de estos transcriptomas es posible descargar (i) el set completo de TT en formato FASTA, (ii) el transcriptoma de referencia con sus anotaciones, (iii) la secuencia y anotaciones para un conjunto de identificadores de TT definido por el usuario y (iv) los datos crudos de expresión. Al acceder al set de TT de cada uno de los transcriptomas (figura 4.2B) es posible filtrar por nombre o descripción del TT y se despliega para cada uno de ellos información sobre su longitud, descripción, estado del ORF y si forma parte o no del transcriptoma de referencia (etiqueta “REF” en la columna “Where” en figura 4.2B).

Se puede acceder a información más detallada de cada TT (figura 4.3) como: (1) su secuencia en formato FASTA; (2) una completa anotación que incluye las descripciones asignadas por Sma3s y Full-LengtherNext, y que aporta información sobre sus posibles funciones, además de términos GO, EC, rutas KEGG y términos InterPro, todos con sus correspondientes *E* asociados, para evaluar la calidad de la anotación, y también los ortólogos TAIR y RefSeq de *Arabidopsis thaliana*, que permiten análisis funcionales y de enriquecimiento génico; (3) una predicción del putativo ORF del TT, aportada por

**A**

| Assembly name                                     | Organism      | Tissue                 | Transcripts number | Description                                                    |
|---------------------------------------------------|---------------|------------------------|--------------------|----------------------------------------------------------------|
| pollen_transcriptome_v1.1                         | Olea europaea | pollen                 | 27823              | Complete transcriptome assembly of pollen.                     |
| reproductive_tissues_transcriptome_v1.1 (default) | Olea europaea | pollen and pistil      | 72846              | Complete transcriptome assembly of pollen and pistil together. |
| pistil_transcriptome_v1.1                         | Olea europaea | pistil                 | 60400              | Complete transcriptome assembly of pistil.                     |
| vegetative_tissues_transcriptome_v1.1             | Olea europaea | leaf, radicle and root | 38919              | Transcriptome assembly of vegetative tissues.                  |

**B**

Assembly Name: reproductive\_tissues\_transcriptome\_v1.1  
Listing 63979 Annotated Transcripts

All transcripts | Annotated transcripts | Download Annotated sequences

Complete | Putative Complete | N-terminus | Putative N-terminus | Internal | C-terminus | Putative C-terminus  
Coding | Putative Coding | Putative ncRNA | Unknown

Search: Name Containing In All transcripts Search

| Name              | Length | Software | Description                                                                                                               | E-value    | ORF Status          | Where |
|-------------------|--------|----------|---------------------------------------------------------------------------------------------------------------------------|------------|---------------------|-------|
| rp11_olive_026975 | 966    | FL-Next  | sp=adenylate kinase 5, chloroplastic Arabidopsis thaliana (Mouse-ear cress)                                               | 0.0        | Internal            |       |
|                   |        | sma3     | adenylate monophosphate kinase 5                                                                                          | 5.146e-10  |                     |       |
| rp11_olive_026977 | 966    | FL-Next  | sp=Calmodulin-binding transcription activator 5 Arabidopsis thaliana (Mouse-ear cress)                                    | 0.0        | Putative C-terminus | REF   |
| rp11_olive_026978 | 339    | FL-Next  | sp=50S ribosomal protein L36, chloroplastic (ECO:0000255 HAMAP-Rule:MF_00251) Morus indica (Mulberry) Plastid Chloroplast | 1.0e-07    | Complete            | REF   |
| rp11_olive_026979 | 966    | FL-Next  | sp=Uncharacterized protein (ECO:0000313 EMBL:EYU27020.1) Erythranthe guttata (Yellow monkey flower) (Mimulus guttatus)    | 0.0        | Internal            | REF   |
| rp11_olive_026981 | 966    | FL-Next  | sp=uncharacterized protein (eco:0000313 EMBL:EYU44200.1) Erythranthe guttata (Yellow monkey flower) (Mimulus guttatus)    | 0.0        | Putative C-terminus | REF   |
|                   |        | sma3     | uncharacterized protein (eco:0000313 EnsemblPlants:PGSC0003DMT400032213)                                                  | 3.358e-07  |                     |       |
| rp11_olive_026982 | 339    | FL-Next  | sp=Protein IQ-DOMAIN 14 Arabidopsis thaliana (Mouse-ear cress)                                                            | 3.0e-05    | Internal            |       |
| rp11_olive_026984 | 339    | FL-Next  | sp=Putative uncharacterized protein (ECO:0000313 EMBL:CAN72408.1) Vitis vinifera (Grape)                                  | 4.2039e-45 | Internal            | REF   |
| rp11_olive_026985 | 966    | FL-Next  | sp=Protein argonaute 4 Arabidopsis thaliana (Mouse-ear cress)                                                             | 0.0        | Internal            |       |
|                   |        | sma3     | Protein argonaute (ECO:0000256 RuleBase:RU361178)                                                                         | 1.042e-25  |                     |       |

**Figura 4.2: Capturas de pantalla de una vista general de la base de datos ReprOlive.** (A) El panel “All assemblies” muestra los cuatro transcriptomas incluidos en la base de datos. (B) Primera página del panel “Tentative transcripts (TT)” del transcriptoma reproductivo. Más información en el texto. Tomada de [57].

Full-LengtherNext, junto con la posición de los codones de inicio y parada de la transcripción, si el ORF está completo y el alineamiento que permite tales predicciones, información que resulta muy útil en el laboratorio, por ejemplo para el diseño de cebadores para clonar estos TT; (4) si el TT forma parte del transcriptoma de referencia, se añaden datos de expresión (véase el apartado 3.4.4) que incluyen tanto los recuentos brutos, que se pueden descargar y analizar con otro *software*, como valores de RPKM con fines comparativos para identificar claramente TT específicos o no de polen y/o pistilo.

## SSR, Descriptions, GO, EC e InterPro

La información sobre los microsatélites (SSR) detectados se puede consultar en la pestaña “SSR” para cada transcriptoma (figura 4.2), donde se despliega una tabla con todos los TT que tienen al menos

All assemblies Assembly info Tentative Transcripts SSR Descriptions GO EC KEGG Pathways InterPro Expression

Assembly Name: reproductive\_tissues\_transcriptome\_v1.1  
Transcript Name: rp11\_olive\_022501  
Length: 1060 nt  
Where:

Transcript Fasta [Download Fasta](#)

Annotations

| Source  | Description                                                                 | Term | Type | E-value  | Identity % |
|---------|-----------------------------------------------------------------------------|------|------|----------|------------|
| sma3    | ATP-AMP phosphotransferase                                                  | -    | -    | 5.73e-11 | -          |
| FL-Next | sp=Adenylate kinase 2, chloroplastic Arabidopsis thaliana (Mouse-ear cress) | -    | -    | 0.0      | 75         |

| Source | EC               | Term    | Type | E-value | Identity % |
|--------|------------------|---------|------|---------|------------|
| sma3   | Adenylate kinase | 2.7.4.3 | -    | 0.0     | -          |

| Source | KEGG Pathways     | Term  | Type | E-value | Identity % |
|--------|-------------------|-------|------|---------|------------|
| sma3   | Purine metabolism | 00230 | -    | 0.0     | -          |

| Source | Gene name |
|--------|-----------|
| sma3   |           |

| Source | GO                              | Term       | Type               | E-value   | Identity % |
|--------|---------------------------------|------------|--------------------|-----------|------------|
| sma3   | chloroplast stroma              | GO:0009570 | Cellular Component | 1.629e-19 | -          |
| sma3   | chloroplast envelope            | GO:0009941 | Cellular Component | 6.647e-18 | -          |
| sma3   | adenylate kinase activity       | GO:0004017 | Molecular Function | 0.0       | -          |
| sma3   | ATP binding                     | GO:0005524 | Molecular Function | 0.0       | -          |
| sma3   |                                 | GO:0019205 | Molecular Function | 5.229e-12 | -          |
| sma3   |                                 | GO:0006139 | Biological Process | 2.83e-10  | -          |
| sma3   |                                 | GO:0006354 | Biological Process | 0.0       | -          |
| sma3   | regulation of proton transport  | GO:0010155 | Biological Process | 4.775e-32 | -          |
| sma3   | carotenoid biosynthetic process | GO:0016117 | Biological Process | 1.618e-26 | -          |
| sma3   | protein autophosphorylation     | GO:0046777 | Biological Process | 6.695e-31 | -          |
| sma3   | nucleotide phosphorylation      | GO:0046939 | Biological Process | 1.384e-07 | -          |

| Source | InterPro                                            | Term      | Type | E-value | Identity % |
|--------|-----------------------------------------------------|-----------|------|---------|------------|
| sma3   | Adenylate kinase/UMP-CMP kinase                     | IPR000850 | -    | 0.0     | -          |
| sma3   | Adenylate kinase subfamily                          | IPR006259 | -    | 0.0     | -          |
| sma3   | Adenylate kinase, active site lid domain            | IPR007862 | -    | 0.0     | -          |
| sma3   | Domain of unknown function DUF1995                  | IPR018962 | -    | 0.0     | -          |
| sma3   | P-loop containing nucleoside triphosphate hydrolase | IPR027417 | -    | 0.0     | -          |

| Source  | DB - Species              | ID | Descriptions                                                                                  | E-value | Identity % |
|---------|---------------------------|----|-----------------------------------------------------------------------------------------------|---------|------------|
| FL-Next | Tair_Athaliana_database   |    | sp=symbols: AMK2   adenosine monophosphate kinase   chr5:19375488-19378058 FORWARD LENGTH=283 | 0.0     | 75         |
| FL-Next | RefSeq_Athaliana_database |    | sp=adenosine monophosphate kinase [Arabidopsis thaliana]                                      | 0.0     | 75         |

ORF Prediction

Status: Complete  
Subject seq: Q9F1J7  
Warning info: Selected protein is the only available option.  
Coding index: 0  
Coding index was used to find complete genes when there was not found a reliable orthologue. A Coding index value >= 0 means the ORF is probably coding. A Coding index value < 0 means the ORF is probably non-coding. Test Code values in between 0.74 and 0.95 mean it is uncertain whether the ORF is coding or not.

Predicted protein: MACSGCYSVNFMVSSNPXKPLTSLTSLKLSQSYSLQIFSSLKFNKSLSPQLRISKSPNPFVVGKAEPLKVMISGAPASGKGTQCELTXYDLVHIAAGDLLRAEVAAGTENGRAKE YMEKQLVDEIVTMVKRLSQPSQKEGWLMDGYPRSSQATALKGRGFHFILFVPEELVDRVGRRLDPVYGTIVHLKSPPEETEEIARLTQRFDTEKKVRLRLTHNKNVAVLSMYEDTVRVVDGSLK EDVFTQDIALRLVGRQAPIGSVAM\*

Length: 295

Coding sequence: CGCTAGAGCGAAA...ATGGCTCTAGCGGTTGCTTCACTGCAATTCATGGCAGTATCTCCATCCAAACAGCCCTTAATCTCACTCTCTCCATCTCAAAATATCTCAATGCTA CTCCCAATCCCCCTTTCTTCTTGAATTCGAATTCGAAATATCCCAATCCAGTTGCGGTATATCAAAAGTCTGCTCCAAATTCATTCGTTGGTATCTGGGATCAAGGCGAGCGCTTGAAGGTAATGATATCGG GAGCTCTGCTCTGTTAAGGAACCAATGCGAGCTCATCACTAAGAAGTACGATTTGGTGACATTGCGAGCTGGAGATCTCTCAGGGCTCAAGTTGCTGCTGGGACAGAAATGGGACAGAGCGAAGGAA TACATGGAGAGGGGCGAGTTGGTCCAGATGAATATGTTGTCAGCATGTGCAAGGAGCGTTTATCAACACAGATCTTAAGAAAGGAGCTGCTTATGACCGATATCAAGGAGCTCTGCTCAAGGAGCTGCTG TCTTAAGGATTTGGCTCGAGCCACAGATTTTATCTCTGGAAGTCCCTGAAGAGATCTGTTGACAGAGATTGTTGGCCGACAGATAGATCTGTTACGGGGAATATATATCATCTAAATATCTCCGCTG AAACGAGAGAGATGCTGCTAGGCTTACTCAAGCTTTGATGACACTGAGAAAGGTGAACTGCGTTTGTCTACACATACAAATATGAGAGGCGGCTTCTTCAATGATAGAGGATGAAAGTCAAGGCTGG ATGGAAGCTTTTCAAGGAGAGCGTATTTACCAATACGACAGCGTTGAAATGTAATCTTGACAAAGGCGAGCTCCCATAGGATCAGTGGCAATGTGA...AAACATCAAGAAATCAATGAAGAAAGGCA TATGACGCTCTAGACGAGCTGACCACTCTATTGCTGCTGATCCACACCTGAGTTGATCCGCAATGGGCTATCGGTGTTTCAAGGCCGACGAGCACTGCTGCGTTTGAAGCTGCTCTTCCAC

Alignment:

rp11\_olive\_022501.R188-----APNSFVYVGGIKAEPLKVMISGAPASGKGTQCELTXYDLVHIAAGDLLRAEVAAGTENGRAKE  
Q9F1J7 RYSSSPSTAPN-PQIVAAERSEPLKVMISGAPASGKGTQCELTXYDLVHIAAGDLLRAEVAAGTENGRAKE

Expression

| Class value | Pist  | Pollen | Vegetative |
|-------------|-------|--------|------------|
| raw         | 11    | 3      | -          |
| rpkm        | 0.033 | 0.009  | -          |

**Figura 4.3: Captura de pantalla de la información en ReprOlive para un único TT, por ejemplo, 'rp11\_olive\_022501', que corresponde a una adenilato cinasa del transcriptoma reproductivo. La información de las pestañas "Annotation", "ORF prediction" y "Expression" aparece desplegada. Más detalles en el texto. Tomada de [57].**

un SSR. Para cada motivo SSR se muestra su secuencia agrupada según el periodo (tríos, tétradas...), el TT que lo contiene y las posiciones de inicio y fin dentro del mismo. Es posible filtrar los SSR por el número de nucleótidos del motivo y por su longitud. De este modo se ha visto que el transcriptoma

reproductivo tiene repeticiones de hexanucleótidos en 501 de sus TT, de tetranucleótidos en 575 TT, y que 493 SSR tienen más de 20 nt de longitud. Estos SSR pueden ser empleados como marcadores moleculares, ya que son fácilmente convertibles en cebadores [274] que presentan codominancia y proporcionan resultados estables [275], y que además superan las limitaciones de otros tipos de marcadores [276].

La recopilación de las descripciones asignadas a los TT se puede consultar en la pestaña “Descriptions” (figura 4.2). Debido a que las descripciones han sido escritas por humanos, es frecuente encontrar diferentes descripciones para una misma secuencia, como puede verse en la figura 4.4A, lo cual se deduce de que haya 82 334 descripciones para 63 965 TT con anotación funcional en el transcriptoma reproductivo. Si se clic en una de esas descripciones se despliega toda la lista de TT que comparten dicha descripción. Asimismo existen pestañas para navegar por los 45 781 términos GO, 10 003 códigos EC y los 187 899 InterPro del transcriptoma reproductivo (figura 4.2) y siguen el mismo comportamiento que la pestaña “Descriptions” (figura 4.4).

## Rutas KEGG

Las rutas KEGG de los distintos transcriptomas pueden consultarse en la pestaña “KEGG pathways” (figuras 4.2 y 4.4B). El transcriptoma reproductivo presenta 146 rutas (143 en el polen y 146 en el pistilo) y en el tejido vegetativo 145. Las rutas aparecen ordenadas por el número de EC que están presentes en el transcriptoma (columna “Present ECs” en la figura 4.4B). También aparecen representados los códigos de las rutas, el total de enzimas de la ruta (columna “Total ECs” en la figura 4.4B) y la cobertura total de la ruta (columna “Coverage (%)” en la figura 4.4B). Algunas de las rutas mejor cubiertas del transcriptoma reproductivo son la biosíntesis de glucosinolatos (mapa 00966) con un 91 % de cobertura, la biosíntesis de betalainas (mapa 00965) con un 80 % y la biosíntesis de brasinoesteroides (mapa 00905) y degradación de DDT (mapa 00351), ambas con un 75 % de cobertura.

Al clicar sobre el nombre de la ruta se despliega la lista con las EC presentes en el transcriptoma para esa ruta. A su vez, si se clic sobre los EC se muestran todos los TT del transcriptoma que tienen esa función asignada. Esto permite averiguar si hay algún TT que codifique la proteína completa (figura 4.4C).

ReprOlive proporciona también una localización gráfica de las enzimas, ya que muestra una imagen de la ruta donde las enzimas encontradas aparecen destacadas en verde (figura 4.4C). Dicha imagen es interactiva y permite conectar con la base de datos KEGG para obtener más información. Los mapas KEGG son muy útiles cuando se llevan a cabo análisis funcionales, ya que permiten relacionar entre sí los genes que participan en una misma ruta.

## Datos de expresión

En la pestaña de expresión aparece una lista de los TT del transcriptoma de referencia con los recuentos brutos de lecturas mapeadas sobre cada TT y los valores de RPKM (figuras 4.2 y 4.4D). La primera línea de la tabla hace referencia al número total de lecturas usadas en el mapeo (si se elige la opción “raw”) y a la longitud total del transcriptoma (si se elige la opción “RPKM”) con los que se calcula un valor medio de RPKM para cada transcriptoma (cociente entre el número total de lecturas y la longitud total del transcriptoma de referencia) que se utiliza para colorear los valores de RPKM de cada TT en función de si están por encima (color verde) o por debajo (color rojo) del RPKM medio del transcriptoma. Estos valores solo deben ser considerados a modo orientativo, ya que el cálculo de RPKM no es un método de normalización adecuado [277] y los valores de RNA-seq corresponden a mapeo de lecturas procedentes de pirosecuenciación y no de lecturas cortas. No obstante, estos valores pueden ayudar

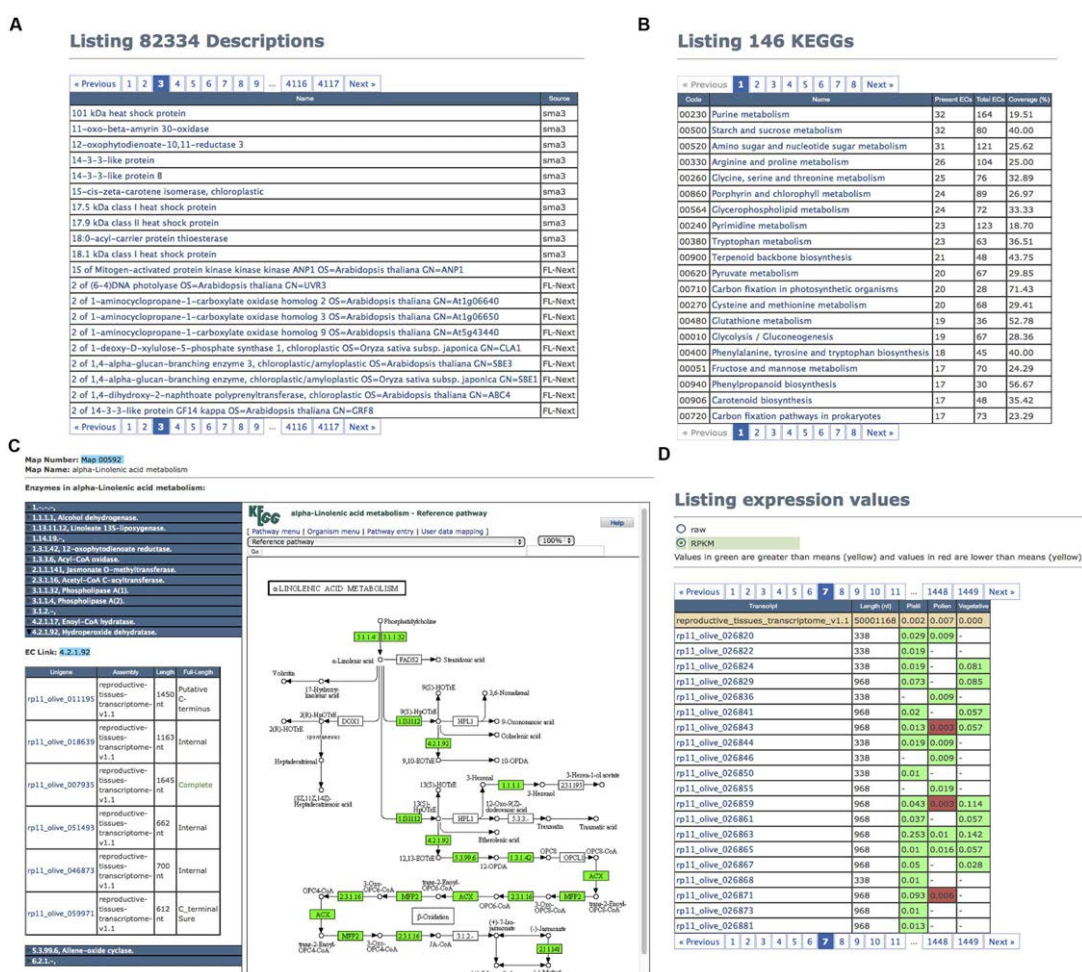


**Figura 4.4: Capturas de pantalla de las pestañas “Descriptions”, “KEGG Pathways” y “Expression” para el transcriptoma reproductivo. (A)** Lista paginada de todas las descripciones de TT. **(B)** Primera página de las rutas KEGG, donde se muestran las más completas de ReprOlive. **(C)** Ejemplo de la ruta metabólica del ácido  $\alpha$ -linoleico; las enzimas presentes en ReprOlive aparecen destacadas en verde; los TT correspondientes al código EC 4.2.1.92 (hidroperóxido deshidratasa aparecen desplegados para mostrar que al menos uno (rp11\_olive\_007935) codifica la proteína completa. **(D)** La séptima página de la pestaña de “Expression” ilustra los distintos tipos de TT basados en RPKM. Más detalles en el texto. Tomada de [57].

a determinar en qué tejido se expresa preferentemente el TT. Por ejemplo, el TT rp11\_olive\_005693 (correspondiente a una proteína que contiene un dominio POZ/BTB) se expresa en los tres tejidos, principalmente en el pistilo; el polen parece expresar específicamente una proporción significativa de proteínas no caracterizadas, como rp11\_olive\_052211, y una probable hidrolasa (rp11\_olive\_026947) parece ser expresada solo en tejidos reproductivos, principalmente en el polen.

La utilidad de la combinación de las anotaciones de ReprOlive junto con los valores de expresión expresados como RPKM queda reflejada en la figura 4.5, donde se ha llevado a cabo un enriquecimiento biológico de términos GO de los 1655 TT expresados exclusivamente en el polen y que tienen un ortólogo

RefSeq en *Arabidopsis*. Los TT se han ordenado en función de su RPKM y se han analizado con la herramienta GOrilla (véase el apartado 3.2.4). En cuanto a los procesos biológicos, los términos GO enriquecidos de los TT específicos de polen más representativos corresponden al crecimiento del tubo polínico, la organización de filamentos de actina, la organización y modificación de la pared celular vegetal, el catabolismo de los polisacáridos y el transporte de carbohidratos (parte superior de la figura 4.5). En relación al componente celular, estos TT se expresan principalmente en la región extracelular, seguida del tubo polínico y la pared celular vegetal (parte inferior de la figura 4.5). Estos resultados respaldan el procedimiento experimental de las librerías secuenciadas, su ensamblaje, la anotación con ortólogos, e incluso los valores de RPKM, a pesar de que se han calculado a partir de lecturas 454/Roche y no de lecturas cortas.

### Búsquedas por secuencia, anotación y semántica

Los TT se pueden recuperar por similitud de secuencia mediante BLAST [104] a través del botón “BLAST” de la barra superior de navegación (figuras 4.2A y 4.6A). La búsqueda basada en BLAST puede hacerse tanto para nucleótidos (BLASTn) como para aminoácidos (BLASTx), con un *E* personalizable. El tipo de secuencia (aminoácidos o nucleótidos) se detecta automáticamente y es posible elegir contra qué transcriptoma/s se quiere lanzar la búsqueda. Las ejecuciones se encolan y la URL con los resultados se muestra por pantalla una vez que la tarea finaliza y se almacena durante un mes. El usuario puede descargar los resultados en formato HTML o un archivo FASTA con los alineamientos.

También es posible realizar búsquedas por cadenas de texto basadas en anotación accediendo al botón “Annotation Search” de la barra superior de navegación (figuras 4.2A y 4.6B). ReprOlive permite realizar búsquedas por anotación combinando GO, EC, InterPro, ruta KEGG, descripciones, ortólogos y nombres de genes. Se pueden concatenar cadenas de texto usando las conjunciones “AND” o “OR”. Se puede seleccionar en qué transcriptomas se quiere realizar la búsqueda. Como resultado de la búsqueda se despliega una lista paginada de los TT que cumplen los requisitos y sus secuencias se pueden descargar en formato FASTA.

Una novedad de ReprOlive con respecto a otras bases de datos de plantas es que su información se ha publicado también estructurada en formato RDF, lo cual permite su interconexión con otras bases de datos semánticas, como por ejemplo UniProtKB. Como la mayoría de los TT en ReprOlive se han anotado también con un ortólogo en *Arabidopsis* en TAIR10 o RefSeq, es posible realizar una búsqueda de datos vinculados que permite recuperar información sobre estructuras 3D (de la base de datos PDB<sup>1</sup>), alérgenos (de la base de datos Allergome<sup>2</sup>), interacciones (de las bases de datos STRING<sup>3</sup> e IncAct<sup>4</sup>), enzimas (de las bases de datos KEGG, BRENDA<sup>5</sup>, BioCyc<sup>6</sup> y Reactome<sup>7</sup>) y especies reactivas de oxígeno (ROS, de la base de datos PeroxiBase<sup>8</sup>). La principal ventaja de esta búsqueda semántica consiste en tener acceso a información actualizada procedente de bases de datos externas, de forma que es posible complementar y extender la información almacenada en ReprOlive.

Se puede acceder a la búsqueda semántica mediante el botón “Linked-data search” de la barra superior de navegación (figuras 4.2A y 4.6C). En primer lugar se elige el TT (semilla) a partir del cual

<sup>1</sup><https://www.rcsb.org>

<sup>2</sup><http://www.allergome.org>

<sup>3</sup><https://string-db.org>

<sup>4</sup><https://www.ebi.ac.uk/intact>

<sup>5</sup><https://www.brenda-enzymes.org>

<sup>6</sup><https://biocyc.org>

<sup>7</sup><https://reactome.org>

<sup>8</sup><http://peroxibase.toulouse.inra.fr>

se quiere iniciar la búsqueda semántica. La primera línea del panel de búsqueda (figura 4.6C) permite recopilar TT que comparten identificador de secuencia, términos GO, códigos EC o InterPro. Si se clica en el botón “Get local data”, se proporciona información relacionada que hay en ReprOlive, mientras que si se clica en “Search” se extrae el ortólogo en *Arabidopsis* de cada TT y se usa para recuperar información externa descargable relativa a estructuras, interacciones, alérgenos..., en función de la base de datos que haya sido seleccionada.

Como ejemplo se llevó a cabo una búsqueda semántica a partir del TT rp11\_olive\_029403 que se expresa solo en ambos tejidos reproductivos y no en tejidos vegetativos. Se anota como un factor de transcripción similar a UNE12 de *Arabidopsis* y con el término GO:0080147, que corresponde a desarrollo celular de los pelos radiculares. Usando este término GO como semilla, se recuperaron de la base de datos STRING (versión 9.1) siete redes de interacción proteica en *Arabidopsis*, una de ellas centrada en el gen UNE12 (figura 4.7). Se sabe que UNE12 es una proteína del saco embrionario no fertilizado, involucrada en la doble fertilización que genera el cigoto y el endospermo. Además, este ortólogo de *Arabidopsis* interactúa con: otros factores de transcripción de función conocida (AT1G03040) o desconocida (AT5G03500, AT5G03495), una proteína involucrada en la escasez de fosfato (SPX4), una proteína responsable de respuesta de los fitocromos (PAR1) y otras proteínas de función desconocida (AT3G27420, AT5G11980, AT4G24840). En función de la homología entre UNE12 y rp11\_olive\_029403, este TT probablemente sea un buen candidato como factor de transcripción en tejidos reproductivos.

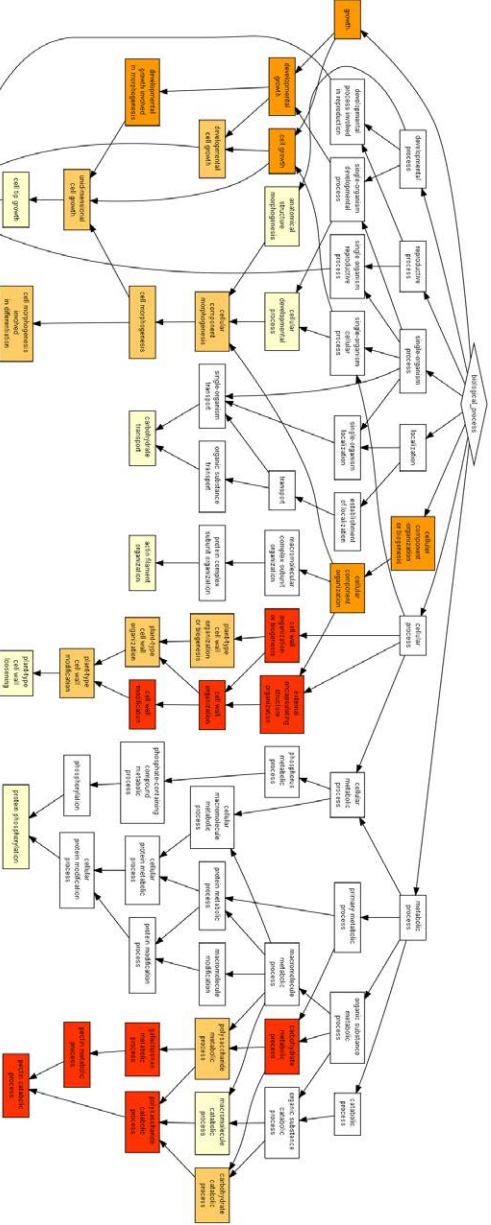


Figura 4.5: Enriquecimiento de términos GO obtenido con GOrilla de los 1 655 TT que son específicos de polen y tienen un ortólogo en *Arabidopsis* en RefSeq, ordenados por su RPKM. La parte superior **Suplementario** File 2.5. Los procesos biológicos significativos y la parte inferior muestra los componentes celulares donde tienen lugar dichos procesos. Más detalles en el texto. Tomada de [57].

**A**

**B**

**C**

**Figura 4.6: Distintos tipos de búsqueda que admite ReprOlive: (A) Búsqueda por similitud de secuencia mediante BLAST. (B) Búsqueda por cadenas de texto basadas en anotación. (C) Búsqueda semántica mediante datos vinculados (Linked-data). Más detalles en el texto.**

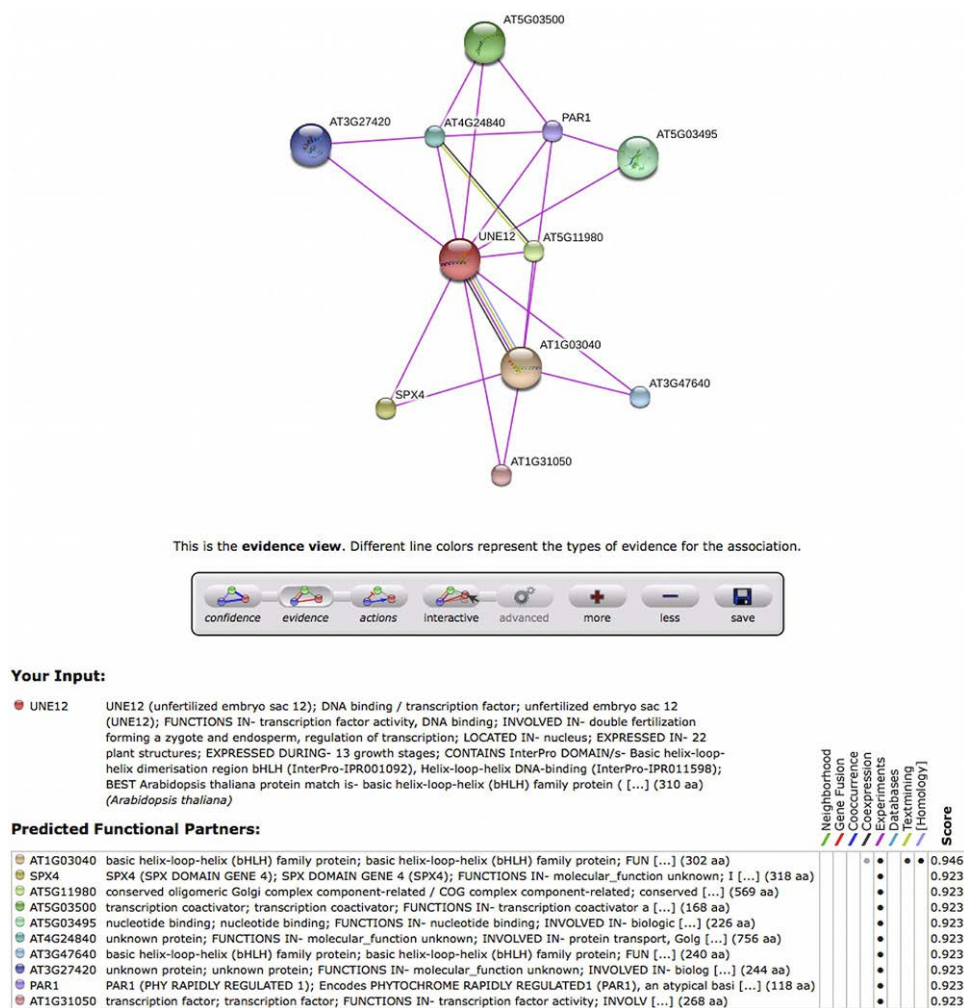


Figura 4.7: Captura de pantalla de las interacciones de UNE12, el ortólogo en *Arabidopsis* para el *TT rp11\_olive\_029403*. Esta información se obtuvo de la base de datos STRING (versión 9.1). El color magenta de las conexiones indica que estas interacciones han sido determinadas experimentalmente. Más detalles en el texto. Tomada de [57].

## 4.2. Transcriptoma mejorado de polen maduro

### 4.2.1. Ensamblaje mejorado

Con el objetivo de conseguir un transcriptoma de polen de la variedad Picual, más completo que el obtenido en ReprOlive y centrado exclusivamente en el polen maduro, se utilizó la herramienta TransFlow [126], que combinó las lecturas 454/Roche de polen maduro empleadas en ReprOlive con lecturas pareadas de Illumina. TransFlow es capaz de combinar distintos programas de ensamblaje, y con ello dar lugar a ensamblajes diferentes, que son evaluados y puntuados mediante la comparación con transcriptomas de referencia bien caracterizados (*Arabidopsis* y *Populus* en este caso), para escoger finalmente el mejor de ellos. El archivo HTML que contiene todas las tablas, figuras y gráficas resultantes está disponible en el fichero complementario 2 de [126]. Varias de esas figuras se han empleado en la presente tesis.

El ensamblaje *de novo* partió de 40 488 002 lecturas pareadas de Illumina ( $2 \times 75$  nt) y 216 497 lecturas 454/Roche, de las cuales resultaron útiles tras el preprocesamiento 32 529 229 (80.3 %) lecturas pareadas de Illumina y 111 242 (51.4 %) de Sanger (tabla 3.3). Los módulos 1, 2 y 3 de TransFlow (figura 3.1) produjeron 181 ensamblajes distintos, que se agruparon en tres *clusters* diferentes (figura 4.8). El primer *cluster* contiene los transcriptomas de referencia, los ensamblajes con *scaffolding* de lecturas de Illumina fusionados con lecturas 454/Roche y solo los ensamblajes reconciliados de lecturas 454/Roche, lo cual sugiere que las lecturas 454/Roche tienen un papel complementario no determinante en el ensamblaje. Los ensamblajes con RAY aparecen en el *cluster* más lejano (*cluster* 3), lo cual indica que este ensamblador no produce transcriptomas apropiados.

La tabla 4.2 muestra los cinco mejores ensamblajes del *ranking* establecido en la evaluación del módulo 5, basado en distancia media (MD) a los transcriptomas de referencia. Estos cinco ensamblajes presentan diferencias mínimas entre ellos (MD varía entre 0,297 y 0,367). Cuatro de ellos incluyen ensamblajes que usan OASES y *scaffolding*, y el mejor de ellos (con menor MD, scOases\_cat\_cd) involucra solo lecturas de Illumina, mientras que los otros contienen una combinación de lecturas de Illumina y 454/Roche con diferentes aproximaciones. Cabe destacar que el primer ensamblaje que contiene solo lecturas 454/Roche, aunque se encuentra también en el primer *cluster* del PCA (figura 4.8), queda relegado a la posición 64 del *ranking* con MD = 0,708 (véase el fichero complementario 2 de [126]), con la etiqueta ctMIRA\_ctEulK29\_rcCAP3, que corresponde a los ensamblajes primarios con MIRA4 y Euler-SR (*k*-mero 29) reconciliados con CAP3 y que es el equivalente a la estrategia seguida en la obtención de los transcriptomas de ReprOlive (apartado 3.4). Por tanto, las lecturas Illumina bastaron para proporcionar el mejor ensamblaje, mientras que las lecturas 454/Roche parecen ser complementarias. Por otra parte, las combinaciones de todas las lecturas con RAY no parecen proporcionar resultados satisfactorios, ya que estos ensamblajes se localizan en el *cluster* más alejado de los transcriptomas de referencia (*cluster* número 3).

Para ver cómo contribuyen los parámetros de evaluación (tabla 3.1) a la hora de decidir cuál es el mejor ensamblaje, se estableció la correlación con las dos primeras dimensiones del PCA. La tabla 4.3 muestra los tres parámetros más significativos ( $P < 0,05$ ) en las primeras dos dimensiones del PCA. Los parámetros más significativos en la primera dimensión fueron ComplOrth, Contigs500 y DiffComplProts, los cuales son indicativos de un transcriptoma bastante completo y sin redundancia.

### 4.2.2. Comparación del transcriptoma mejorado con el de ReprOlive

La comparación entre ambos transcriptomas no puede hacerse directamente entre la versión obtenida en ReprOlive (apartado 4.1.1) y el transcriptoma mejorado de TransFlow por varios motivos: (1)

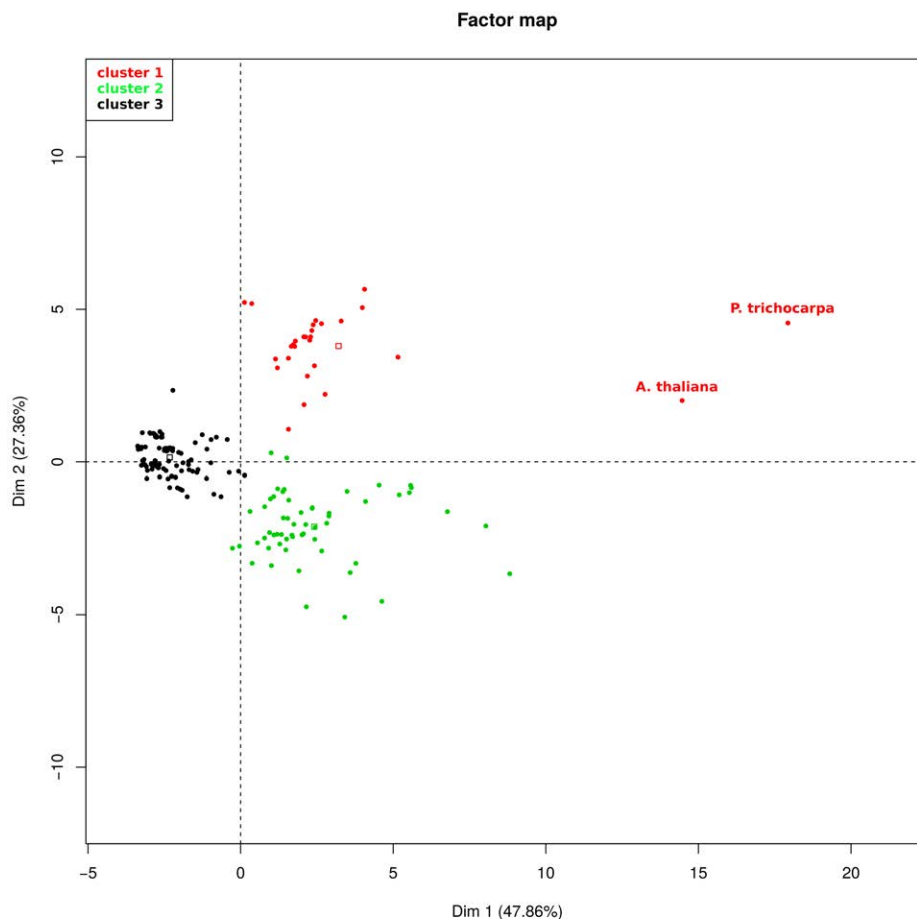


**Figura 4.8: Evaluación de los ensamblajes del transcriptoma mejorado de polen maduro de olivo obtenido con TransFlow.** Gráfica del PCA de las dos primeras dimensiones con los tres clusters significativos encontrados con el análisis HCPC. **Cluster 1** (puntos rojos): comprende los dos transcriptomas de referencia, todos los ensamblajes de Illumina combinados con lecturas 454/Roche ensamblados con MIRA4 y dos ensamblajes de 454/Roche (uno usando solo MIRA4 y el otro es el ensamblaje reconciliado con CAP3). **Cluster 2** (puntos verdes): formado por todos los ensamblajes de Illumina (primarios y de scaffolding), excepto el que usa RAY con k-mero 35, los dos ensamblajes que combinan las lecturas de Illumina con las de 454/Roche (lecturas-lecturas) obtenidos con RAY con k-meros 25 y 35, todas las combinaciones ensamblaje-ensamblaje usando Minimus2 y las dos reconciliaciones de todo con todo usando Minimus2. **Cluster 3** (puntos negros): formado por el ensamblaje primario con EULER-SR, el ensamblaje de Illumina con k-mero 35 hecho con RAY, todos los ensamblajes de Illumina fusionados con las lecturas 454/Roche usando RAY, y todas las combinaciones ensamblaje-ensamblaje llevadas a cabo con RAY. Adaptada de [57].

el transcriptoma de polen de ReprOlive comprende lecturas de polen en varios estadios de desarrollo, mientras que los transcriptomas obtenidos con TransFlow se basan solo en lecturas procedentes de polen maduro; (2) aunque en ambos casos se utiliza Full-LengtherNext para anotar y obtener estadísticas que

**Tabla 4.2:** *Ranking de los cinco mejores ensamblajes obtenidos con TransFlow, con sus respectivas distancias medias a los transcriptomas de referencia (MD). Adaptada de [126].*

| Identificador del ensamblaje         | Módulo 1                                                                                                                                                                   | Módulo 2                           | Módulo 3                                         | MD    |
|--------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------|--------------------------------------------------|-------|
| scOases_cat_cd                       | Concatenación de los dos ensamblajes con <i>scaffolding</i> de OASES con diferentes <i>k</i> -meros y eliminación de redundancia de secuencia                              | -                                  | -                                                | 0,297 |
| aaMin2/scALL/454Cap3                 | Todos los ensamblajes con <i>scaffolding</i>                                                                                                                               | Ensamblaje reconciliación con CAP3 | Combinación de los ensamblajes con Minimus2      | 0,323 |
| scOases_cat                          | Concatenación de los dos ensamblajes con <i>scaffolding</i> de OASES con diferentes <i>k</i> -meros                                                                        | -                                  | -                                                | 0,324 |
| arMIRA/scOases_cat_cd                | Concatenación de los dos ensamblajes con <i>scaffolding</i> de OASES con diferentes <i>k</i> -meros y eliminación de redundancia de secuencia                              | Preprocesamiento de las lecturas   | Combinación de ensamblaje con lecturas con MIRA4 | 0,352 |
| aaMin2/scOases_cat_cd_rcMin2/454Cap3 | Concatenación de los dos ensamblajes con <i>scaffolding</i> de OASES con diferentes <i>k</i> -meros, eliminación de redundancia de secuencia y reconciliación con Minimus2 | Ensamblaje reconciliación con CAP3 | Combinación de los ensamblajes con Minimus2      | 0,367 |

**Tabla 4.3:** *Ranking de los tres parámetros de evaluación más significativos de las primeras dos dimensiones del PCA para los ensamblajes del transcriptoma de olivo. El coeficiente de correlación R (Coef R) mide la correlación de la variable con cada dimensión del PCA. Adaptada de [126].*

| Dimensión 1    |        | Dimensión 2   |        |
|----------------|--------|---------------|--------|
| Nombre         | Coef R | Nombre        | Coef R |
| ComplOrth      | 0,959  | N90           | 0,923  |
| Contigs500     | 0,940  | MeanContigLen | 0,869  |
| DiffComplProts | 0,940  | MeanContigCov | 0,811  |

permitan la comparación de los ensamblajes, en el caso de ReprOlive Full-LengtherNext utiliza Swiss-Prot y TrEMBL como bases de datos para anotar, mientras que en TransFlow, Full-LengtherNext solo emplea Swiss-Prot; y, además, (3) aún cuando en ambos transcriptomas Full-LengtherNext se hubiese basado exclusivamente en Swiss-Prot, para la obtención del transcriptoma mejorado se empleó una versión más actualizada de Swiss-Prot, lo cual daría lugar presumiblemente a un número mayor de proteínas anotadas. En todos los casos, los números obtenidos no serían comparables. Para evitar estos problemas se ha optado por comparar los mejores transcriptomas según TransFlow con el transcriptoma equivalente al de ReprOlive generado también con TransFlow (ensamblaje con la etiqueta ctMIRA\_ctEulK29\_rcCAP3).

En el ensamblaje de ReprOlive (apartado 4.1.1) se emplearon únicamente lecturas de 454/Roche. El hecho de que el mejor ensamblaje obtenido con TransFlow involucre exclusivamente lecturas Illumina, mientras que el ensamblaje resultante de la estrategia equivalente a la seguida en ReprOlive queda relegado al puesto 64 del *ranking*, parece indicar que el nuevo transcriptoma de polen ha mejorado con respecto a ReprOlive. Para profundizar más en esta afirmación, el *heatmap* de la figura 4.10 representa gráficamente los valores numéricos de los distintos parámetros de evaluación (tabla 4.4) obtenidos para: los transcriptomas de referencia, los cinco mejores ensamblajes de TransFlow, y el ensamblaje equivalente a ReprOlive. Se observa que para los mejores ensamblajes de TransFlow, estos valores se suelen acercar mucho más a los observados en los transcriptomas de referencia que en el caso del equivalente a ReprOlive. Los histogramas de la figura 4.9 permiten hacer una comparación de los valores de los parámetros en los

distintos ensamblajes más en detalle.

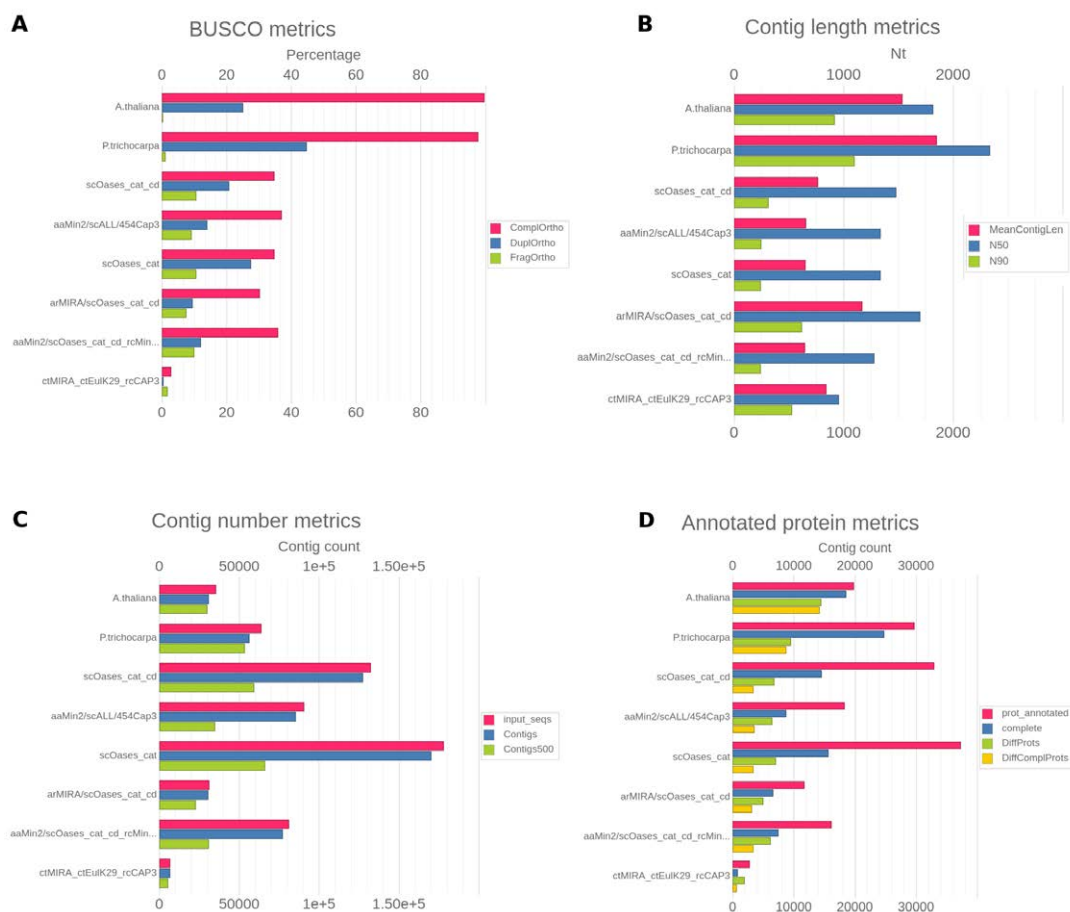


**Figura 4.9: Histogramas representativos de los valores de los parámetros de evaluación empleados en TransFlow.**

El transcriptoma equivalente a ReprOlive (ctMIRA\_ctEulK29\_rcCAP3) obtenido con TransFlow está formado por un total de 6 421 *contigs* con una longitud media de 840 nucleótidos, de los cuales se han anotado 2 722 proteínas (42 %). De ellas, 1 899 son diferentes y 573 contienen un ORF completo (véase tabla 4.4). En comparación, el mejor transcriptoma según TransFlow (scOases\_cat\_cd) comprende un total de 127 252 *contigs/scaffolds*, de longitud media 760 nucleótidos. Se han conseguido anotar 32 897 proteínas (26 %), que corresponden a 6 717 proteínas diferentes y 3 307 contienen un ORF completo. En comparación, estos números evidencian un transcriptoma de polen maduro mucho más completo y por tanto mejor que el que obtenido con la estrategia de ReprOlive.

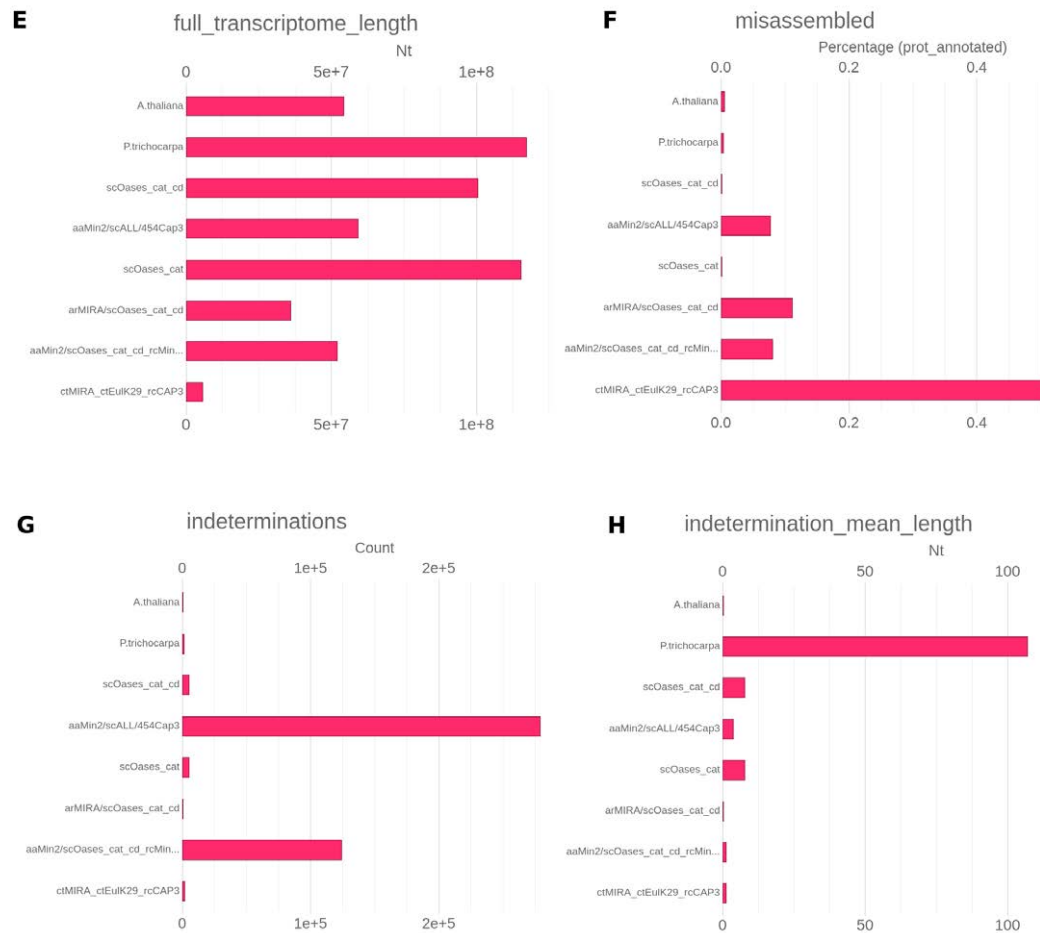


**Figura 4.9: Histogramas representativos de los valores de los parámetros de evaluación empleados en TransFlow (continuación).** Se muestran los valores obtenidos para los transcriptomas de referencia (*A. thaliana* y *P. trichocarpa*), los cinco mejores ensamblajes obtenidos con TransFlow (*scOases\_cat\_cd*, *aaMin2/scALL/454Cap3*, *scOases\_cat*, *arMIRA/scOases\_cat\_cd* y *aaMin2/scOases\_cat\_cd\_rcMin2/454Cap3*) y el ensamblaje equivalente a la estrategia seguida en ReprOlive (*ctMIRA\_ctEulK29\_rcCAP3*). Se muestran parámetros que devuelve TransFlow para evaluar la completitud del transcriptoma (A), la longitud (B) y cantidad (C) de los *contigs* obtenidos; la cantidad, completitud y redundancia de las proteínas anotadas (D); la longitud del transcriptoma completo (E); el porcentaje de proteínas mal deducidas (F) y la cantidad y longitud de las indeterminaciones encontradas (G y H, respectivamente) en los *contigs* o *scaffolds*. Para una descripción más detallada de los parámetros de evaluación representados en cada caso, consúltese la tabla 3.1. Adaptada de [126] (Additional file 2).

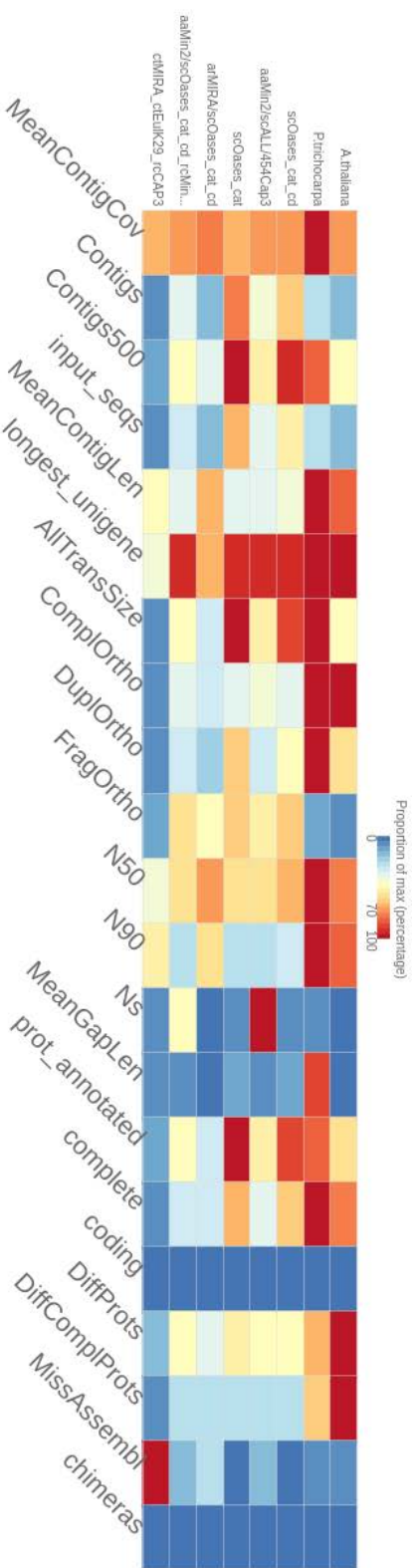


Figura 4.10: *Heatmap* de los parámetros de evaluación representados como un porcentaje del máximo valor obtenido para todos los ensamblajes del transcriptoma de polen maduro de olivo, incluidos los *transcriptomas de referencia*. Se muestran únicamente los transcriptomas de referencia (*A.thaliana* y *P.trichocarpa*), los cinco mejores ensamblajes obtenidos con TransFlow (*scOases\_cat\_cd*, *aaMin2/scALL/454Cap3*, *scOases\_cat*, *arMIRA/scOases\_cat\_cd* y *aaMin2/scOases\_cat\_cd*) y el ensamblaje equivalente a la estrategia seguida en ReprOlive (*ctMIRA\_ctEuk29\_rcCAP3*). Adaptada del fichero complementario 2 de [126].

**Tabla 4.4:** Valores de los parámetros de evaluación para los siguientes ensamblajes obtenidos con TransFlow: los transcriptomas de referencia (*A.thaliana* y *P.trichocarpa*), los cinco mejores ensamblajes obtenidos con TransFlow (*scOases\_cat\_cd*, *aaMin2/scALL/454Cap3*, *scOases\_cat*, *arMIRA/scOases\_cat\_cd* y *aaMin2/scOases\_cat\_cd*) y el ensamblaje equivalente a la estrategia seguida en ReprOlive (*ctMIRA\_ctEuk29\_rcCAP3*). 'Ind': indeterminaciones. Adaptada del fichero complementario 2 de [126].

|                              | Basic Metrics (%) |           |           | Contigs Length Metrics |      |      | Contigs Number Metrics |         |            | Annotated Protein Metrics |          |           | Full Transcriptome |              |             |        |
|------------------------------|-------------------|-----------|-----------|------------------------|------|------|------------------------|---------|------------|---------------------------|----------|-----------|--------------------|--------------|-------------|--------|
|                              | CompOrtho         | DuplOrtho | FragOrtho | MeanContigLen          | N50  | N90  | input_seqs             | Contigs | Contigs500 | prot_annotated            | complete | DiffProts | DiffCompProts      | AllTransSize | MissAssembl | Ind.   |
| <i>A.thaliana</i>            | 99.5              | 21.9      | 0.2       | 1535                   | 1814 | 914  | 35386                  | 30336   | 29609      | 19790                     | 18523    | 14366     | 14170              | 54305826     | 0.005       | 1284   |
| <i>P.trichocarpa</i>         | 97.7              | 44.7      | 1         | 1846                   | 2335 | 1093 | 63498                  | 56240   | 53044      | 29622                     | 24718    | 9466      | 8657               | 117241571    | 0.003       | 1284   |
| <i>scOases_cat_cd</i>        | 34.7              | 20.7      | 10.4      | 760                    | 1475 | 310  | 132140                 | 127293  | 59055      | 32897                     | 14522    | 6717      | 3307               | 100469457    | 0.076       | 27961  |
| <i>aaMin2/scALL/454Cap3</i>  | 36.9              | 13.9      | 9         | 655                    | 1331 | 243  | 90538                  | 85239   | 34703      | 18502                     | 8700     | 6367      | 3487               | 59148670     | 0.113       | 0      |
| <i>scOases_cat</i>           | 34.7              | 27.4      | 10.4      | 647                    | 1335 | 236  | 177923                 | 169387  | 65977      | 37239                     | 16615    | 6365      | 3077               | 115189988    | 0.084       | 124254 |
| <i>arMIRA/scOases_cat_cd</i> | 30.1              | 9.4       | 7.4       | 1167                   | 1697 | 616  | 30849                  | 30334   | 22663      | 11663                     | 6365     | 4938      | 3077               | 36007908     | 0.084       | 124254 |
| <i>aaMin2/scOases_cat_cd</i> | 35.7              | 11.9      | 9.8       | 642                    | 1277 | 240  | 80838                  | 77140   | 30784      | 16139                     | 7453     | 6176      | 3321               | 51929654     | 0.516       | 1674   |
| <i>ctMIRA_ctEuk29_rcCAP3</i> | 2.7               | 0.3       | 1.6       | 840                    | 954  | 526  | 6627                   | 6421    | 5231       | 2722                      | 748      | 1899      | 573                | 5565564      | 0.516       | 1674   |

### 4.3. Marcadores para la expresión génica por RT-PCR de los tejidos reproductivos de olivo

Los análisis de expresión génica requieren genes de referencia (RG) para llevar a cabo la normalización necesaria para obtener resultados fiables. Los niveles de expresión de estos RG deberían permanecer constantes en todas las células, tejidos o condiciones experimentales en estudio. Tradicionalmente se han usado con este propósito los genes constitutivos o *housekeeping*, por cumplir este requisito de forma general. Sin embargo, se ha visto que la expresión de estos genes no es tan invariable como se pensaba [278, 279, 280], por lo que habrá que analizar y validar estos RG en cada escenario experimental concreto. Los datos de *microarrays* se han usado para proponer nuevos RG, pero solo están disponibles para especies modelo y condiciones concretas. Por el contrario, los experimentos de RNA-seq son cada vez más frecuentes, por lo que constituyen una nueva fuente de candidatos a RG.

Con este objetivo en mente hemos aprovechado la disponibilidad de lecturas 454/Roche correspondientes a estadios distintos de desarrollo de maduración tanto para polen como para pistilo, para diseñar un flujo automatizado de trabajo con la herramienta AutoFlow [125], en el cual las lecturas se mapean sobre los transcritos obtenidos en ReprOlive y se calcula el coeficiente de variación de un valor de expresión normalizado por millón de lecturas mapeadas (RPMM), que permite filtrar y seleccionar aquellos transcritos que se expresan de forma homogénea y consistente en todas las condiciones experimentales de interés, y que pueden ser utilizados como posibles marcadores para RT-PCR específicos para estos tejidos y en estas condiciones experimentales [266]. La idoneidad de algunos de los marcadores propuestos ha sido validada por PCR semicuantitativa. Además, para probar la validez de la herramienta en otras especies y para lecturas de otro tipo de tecnología, el flujo de trabajo se ha utilizado también sobre lecturas pareadas de Illumina correspondientes a dos variedades diferentes flores de *Arabidopsis thaliana* correspondientes a estudio sobre el efecto de la temperatura en el momento de la floración [265], y también en este caso se han obtenido candidatos a RG.

#### 4.3.1. Tiempos de ejecución del flujo de trabajo

Para evaluar el tiempo que emplea el flujo de detección de RG en función de la especie o tejido, la tecnología de secuenciación de los datos, el tamaño del transcriptoma y de la librería, y la longitud de las lecturas, se han medido los tiempos de ejecución de las diferentes tareas del flujo completo (preprocesamiento, mapeo y análisis con *findRGs*; figura 3.4) usando dos conjuntos de datos muy diferentes: las lecturas 454/Roche de los tres estadios de desarrollo de pistilo de olivo (767 963 lecturas en total, tabla 3.3.1) y las lecturas de Illumina de las variedades Columbia y Killeen de una de las réplicas de *Arabidopsis* (23 821 198  $\times$  2 lecturas en total, correspondientes a la réplica 2 de la tabla 3.5). Como se observa en la tabla 4.5, usando el mismo número de CPUs la tarea de preprocesamiento es con mucho la etapa más larga tanto en el olivo como en *Arabidopsis* (98 % del tiempo total del flujo en el olivo y casi el 70 % en *Arabidopsis*), siendo mucho más duradera en el olivo que en *Arabidopsis*, debido principalmente a la mayor longitud de las lecturas de 454/Roche. Por el contrario, la etapa más rápida corresponde al análisis con *findRGs* (menos de un minuto por cada 100 000 lecturas en ambos casos), independientemente del tipo o longitud de lectura, la especie y el número de condiciones experimentales.

#### 4.3.2. Candidatos a genes de referencia en el olivo

Puede ocurrir que un transcrito sea un RG muy apropiado dentro de un determinado tejido, y sin embargo no sea adecuado para análisis de expresión génica que involucren varios tejidos [281]. Por

**Tabla 4.5: Tiempos de ejecución estimados para el flujo de detección de marcadores de RT-PCR en dos conjuntos de datos. Adaptada de [266].**

| Especie/<br>Tejido | N.º lecturas<br>brutas | Longitud<br>(nt) | N.º<br>transcritos | Preprocesamiento                 | Mapeo                    | <i>findRGs</i>          | Total       |
|--------------------|------------------------|------------------|--------------------|----------------------------------|--------------------------|-------------------------|-------------|
| Olivo (pistilo)    | 767 963                | 525              | 9 157              | 24 min 45 s<br>8 nodos, 192 cpus | 26 s<br>3 nodos, 72 cpus | 5 s<br>1 nodo, 9 cpus   | 25 min 16 s |
| <i>Arabidopsis</i> | 23 821 198 (×2)        | 100 (×2)         | 35 386             | 43 s<br>8 nodos, 192 cpus        | 19 s<br>2 nodos, 48 cpus | 0,2 s<br>1 nodo, 9 cpus | 1 min 2s    |

Todos los tiempos se refieren a 100 000 lecturas y las ejecuciones se han llevado a cabo en servidores SUSE® Linux v12 con procesadores Opteron con 4 GB/core de RAM.

esta razón se llevaron a cabo tres ejecuciones del flujo en el olivo: una en el polen, otra en el pistilo y una última en ambos tejidos juntos. Se retuvieron los transcritos con expresión menos variable usando dos valores de corte de CV: 10 % (defecto) y 20 % (no estricto). Teniendo en cuenta que las lecturas utilizadas vienen de la plataforma 454/Roche, y que la cantidad de lecturas obtenidas es relativamente baja en comparación con Illumina, el número mínimo de lecturas mapeadas por transcrito se estableció en 10, 50 y 100. La figura 4.11 muestra un resumen comparativo de los resultados de las combinaciones de estos valores de los parámetros de filtrado. Mientras que se obtiene un significativo número de RG para polen y pistilo por separado, incluso con las condiciones más estrictas ( $>100$  lecturas y  $CV < 10\%$ ), no se obtuvo ningún candidato para polen y pistilo juntos, y tan solo uno con parámetros menos rigurosos ( $>50$  lecturas y  $CV < 20\%$ ), lo cual es un indicativo de que los análisis de expresión en tejidos reproductivos requieren una cuidadosa selección de RG.

La tabla 4.6 contiene los candidatos a RG obtenidos con nuestro flujo para tejido reproductivo (polen y pistilo juntos) con  $CV < 20\%$  y como mínimo 10 lecturas mapeadas, mientras que los RG resultantes de la ejecución en el pistilo (con  $CV < 10\%$  y como mínimo 100 lecturas mapeadas) y los de la ejecución en el polen (con  $CV < 10\%$  y como mínimo 100 lecturas mapeadas) corresponden a los ficheros complementarios 2<sup>9</sup> y 3<sup>10</sup> de [266], respectivamente.

Muchos de los RG obtenidos como candidatos en nuestro flujo de trabajo han sido ya sugeridos o usados como RG en otras especies. En el caso de pistilo (figura 4.11 A; fichero complementario 2 de [266]) uno de los mejores RG según nuestros resultados es la ubicuitina (con tres secuencias diferentes: rp11\_olive\_006479, rp11\_olive\_031243 y rp11\_olive\_045557; fichero complementario 2 de [266]). Se trata de un gen muy usado tradicionalmente como RG en las plantas, como por ejemplo el plátano [282], el melocotón [283] y el arroz [284], entre otras. La poliubicuitina 10 (rp11\_olive\_006473; fichero complementario 2 de [266]) también parece un buen candidato según nuestros resultados, de hecho ha demostrado tener una expresión muy estable en *Arabidopsis* [285]. Se ha validado como RG en el arándano [286], el algodón [287] y el chopo [288], y se ha usado para normalización en un trabajo sobre maduración y desarrollo del fruto en el olivo [289]. El factor de elongación 1- $\alpha$  (dos secuencias diferentes: rp11\_olive\_008243 and rp11\_olive\_009319; fichero complementario 2 de [266]) también emerge como RG candidato en el pistilo. Este gen ya se evaluó como RG en la patata (*Solanum tuberosum*), donde resultó ser el más estable en condiciones de estrés biótico y abiótico [290] y ha sido también validado como buen RG en otras muchas especies. Otro candidato obtenido es la gliceraldehído-3-fosfato deshidrogenasa (rp11\_olive\_003751; fichero complementario 2 de [266]), que fue identificado como uno de los mejores RG en desarrollo y maduración del fruto en el olivo [291] y usado como normalizador para el análisis de cDNA asociado con la vecería en el olivo [223]. Por otra parte, los resultados de nuestro análisis

<sup>9</sup>[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5568602/bin/12938\\_2017\\_356\\_MOESM2\\_ESM.docx](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5568602/bin/12938_2017_356_MOESM2_ESM.docx)

<sup>10</sup>[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5568602/bin/12938\\_2017\\_356\\_MOESM3\\_ESM.docx](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5568602/bin/12938_2017_356_MOESM3_ESM.docx)

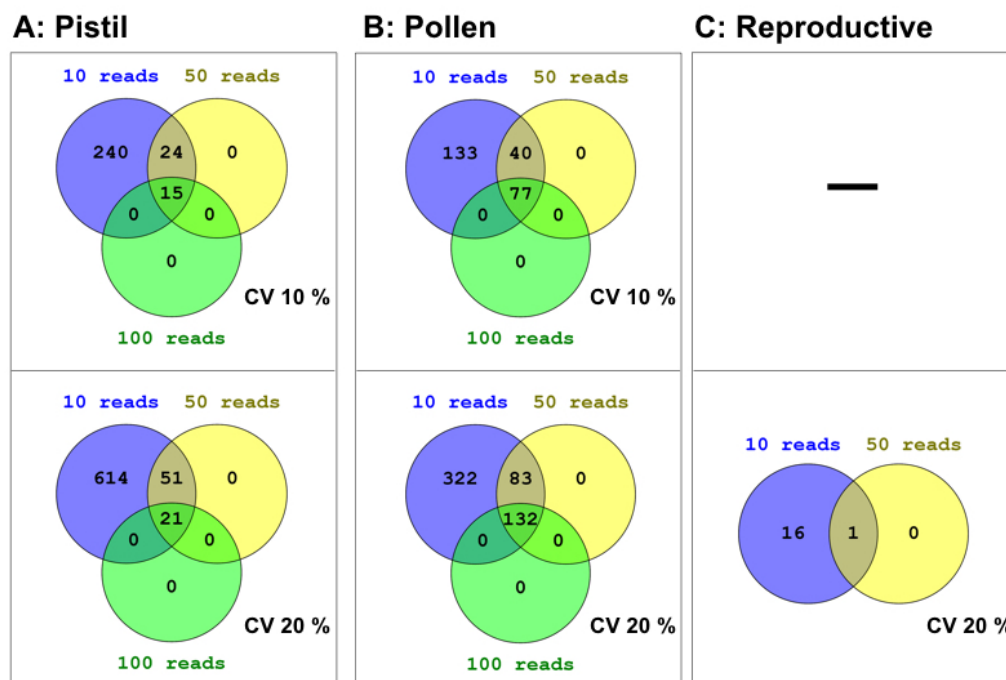


**Figura 4.11:** Diagramas de Venn que resumen el número de RG obtenidos para tejidos reproductivos de olivo. Se usaron dos valores de corte para CV (10 % y 20 %), y tres valores de mínimo de lecturas mapeadas (10, 50 y 100). Los RG de reproductivo se obtuvieron combinando las lecturas de polen y pistilo. Tomada de [266].

sugieren también otros candidatos a RG que no se han usado nunca antes como RG. Sin embargo, dados sus destacados valores de RPM y bajos CV, resultan prometedores y creemos que deberían ser considerados y evaluados cuidadosamente en un futuro. Tal es el caso de la proteína 2 de unión a ácido salicílico y la metilesterasa 1 (rp11\_olive\_003751 y rp11\_olive\_015883, respectivamente; fichero complementario 2 de [266]), entre otros.

En el caso de polen, hemos obtenido un número más alto de candidatos a RG (figura 4.11 B; fichero complementario 3 de [266]). Para la mayoría de ellos, no encontramos referencias bibliográficas sobre su uso como RG, a pesar de la aparente expresión homogénea en este tejido. Tal es el caso del activador cinasa MOB (rp11\_olive\_000239; fichero complementario 3 de [266]) o el citocromo P450 (rp11\_olive\_006957; fichero complementario 3 de [266]). Según nuestra aproximación, serían los mejores RG en este tejido y en estas condiciones, por lo que creemos que merecen ser evaluados. El único gen propuesto por nuestro flujo como un RG apropiado para polen de olivo que ha sido usado previamente es la cisteína proteasa (rp11\_olive\_005653; fichero complementario 3 de [266]), validado como RG estable en *Coffea arabica* [292]. En el olivo también se evaluó como RG en el fruto, pero no estaba entre los genes analizados más estables [293], al menos en este tejido.

En la tabla 4.6 se muestran los candidatos a RG propuestos por nuestro flujo en el análisis de ambos tejidos juntos, polen y pistilo (“Reproductive” en la figura 4.11 C). Dos de ellos corresponden

a RG muy reconocidos y de uso muy extendido. Uno de ellos, la S-adenosilmetionina descarboxilasa (rp11\_olive\_005197\_split\_1; tabla 4.6), ha sido señalado como una de las secuencias más abundantes en librerías de EST de patata (*Solanum tuberosum*) [294]. De hecho, este gen también se sugiere como candidato según nuestro flujo tanto en el polen (fichero complementario 3 de [266]) como en el pistilo (fichero complementario 2 de [266]). Por otra parte, la actina 7 (rp11\_olive\_005099; tabla 4.6) se ha usado extensamente como RG en muchas especies, como por ejemplo la achicoria [295], la vid [296] y el guisante [297]. Al igual que antes, la actina 7 también se sugiere como candidato en nuestro análisis en el polen (fichero complementario 3 de [266]) y en el pistilo, aunque en condiciones algo menos restrictivas en este último (>50 lecturas mapeadas, CV <10 %; resultados no mostrados). Según nuestro análisis, la proteína con una expresión más homogénea a lo largo de todos los estadios de desarrollo de polen y pistilo examinados es la cinasa *Shaggy-related η* (rp11\_olive\_006695; tabla 4.6), por lo que sería interesante un estudio más profundo. Cabe destacar que, al igual que hemos observado en el polen y en el pistilo por separado, se obtienen como candidatos RG correspondientes a proteínas de función desconocida. Su consideración más adelante, cuando se amplíe el conocimiento acerca de estas proteínas, sería también necesaria.

Como hemos visto, solo obtenemos candidatos a RG válidos para ambos tejidos reproductivos cuando usamos valores de los parámetros menos estrictos que los usados en el polen o el pistilo por separado (figura 4.11), por lo que estos candidatos deben ser evaluados con cuidado. Este hecho además puede ser un indicativo de la existencia de diferencias sustanciales en la diferenciación de ambos tejidos, hasta el punto de que casi ningún gen presenta el mismo nivel de expresión en ambos tejidos.

Podemos concluir que la aproximación seguida aquí para los tejidos reproductivos, ya sean solos o en combinación, produce un conjunto de candidatos a RG, muchos de los cuales están respaldados por resultados previamente descritos en la bibliografía, lo cual aporta fiabilidad a nuestro método.

### Validación experimental de los candidatos a RG en el olivo

Es posible pensar que el número de lecturas que se obtiene en un experimento de secuenciación con la tecnología 454/Roche puede no ser suficiente para hacer una predicción fiable de RG. Por eso decidimos hacer una validación experimental que pudiera respaldar los RG obtenidos con *findRGs*. Esta validación preliminar corrió a cargo de María José Jiménez-Quesada y consistió en el análisis de algunos de los candidatos a RG mediante RT-PCR semicuantitativa en tiempo real, como se describe en el apartado 3.6.2. En el resultado de la amplificación (figura 4.12) vemos que la poliubicitina y la actina se validan en diferentes tejidos de olivo en comparación con 18S, un gen *housekeeping* muy usado como RG. Ambos genes presentan niveles de expresión similares en los tejidos reproductivos (polen maduro y pistilo). Sin embargo, mientras que la ubicuitina parece ser un buen RG también en la inflorescencia, la hoja y la semilla, los niveles de actina fueron buenos en todos los tejidos menos en la semilla. Dos de los transcritos que mostraron mayor expresión y menor variación en el polen según *findRGs*, el transportador de prolina (rp11\_olive\_009589) y el homólogo de la L-ascorbato oxidasa (rp11\_olive\_004773) mostraron un buen nivel de expresión tanto en el polen maduro como en la inflorescencia, mientras que presentaron menor o nula expresión en el pistilo, la hoja y la semilla. Respecto a los transcritos propuestos como RG en el pistilo por su baja variación: el ensayo de la gliceraldehído-3-fosfato deshidrogenasa (rp11\_olive\_003751) resultó en una alta expresión en el pistilo y baja en el resto de tejidos, incluido el vegetativo (hoja); la proteína de unión a ácido salicílico (rp11\_olive\_019507) generó una intensa banda de amplificación en el pistilo, el polen maduro y la inflorescencia, y no amplificó en el tejido vegetativo ni en la semilla. Un patrón muy similar se observó para la proenzima S-adenosilmetionina descarboxilasa, un RG propuesto por *findRGs* para tejido reproductivo (polen y pistilo juntos). En este caso se observan bandas de idéntica intensidad en el polen y el pistilo, y una de menor intensidad en la inflorescencia. Por el contrario, el

**Tabla 4.6: Los mejores RG en tejido reproductivo de olivo (combinación de polen y pistilo) según la figura 4.11C, ordenados por CV.** Se obtuvieron para diferente estadios de desarrollo de polen y pistilo, con CV <20 % y como mínimo 10 lecturas mapeadas. Transcript\_id: identificador del transcrito en el transcriptoma reproductivo de ReprOlive. Tomada de [266].

| transcript_id             | PM  | PG1 | RPMM |     |     |     | CV(%) | Mean RPMM | Best hit   | Description                                                                 |
|---------------------------|-----|-----|------|-----|-----|-----|-------|-----------|------------|-----------------------------------------------------------------------------|
|                           |     |     | PG5  | S2  | S3  | S4  |       |           |            |                                                                             |
| rp11_olive_006695         | 205 | 209 | 206  | 208 | 177 | 154 | 10.72 | 193.2     | Q39011     | Shaggy-related protein kinase eta<br><i>Arabidopsis thaliana</i>            |
| rp11_olive_000781         | 105 | 140 | 112  | 104 | 108 | 127 | 11.36 | 116       | Q94A41     | Alpha-amylase 3, chloroplastic<br><i>Arabidopsis thaliana</i>               |
| rp11_olive_006061         | 327 | 272 | 238  | 343 | 285 | 253 | 13.16 | 286.3     | Q8VZ80     | Polyol transporter 5<br><i>Arabidopsis thaliana</i>                         |
| rp11_olive_006091         | 283 | 221 | 211  | 208 | 177 | 190 | 15.65 | 215       | A0A022R151 | Uncharacterized protein<br><i>Erythranthe guttata</i>                       |
| rp11_olive_010107         | 228 | 213 | 184  | 250 | 295 | 199 | 15.98 | 228.2     | O23254     | Serine hydroxymethyltransferase 4<br><i>Arabidopsis thaliana</i>            |
| rp11_olive_005197_split_1 | 366 | 430 | 381  | 343 | 423 | 552 | 16.37 | 415.8     | Q42679     | S-adenosylmethionine decarboxylase proenzyme<br><i>Catharanthus roseus</i>  |
| rp11_olive_003279         | 122 | 179 | 153  | 114 | 118 | 145 | 16.64 | 138.5     | Q9LV37     | Mitogen-activated protein kinase 9<br><i>Arabidopsis thaliana</i>           |
| rp11_olive_000623         | 94  | 94  | 58   | 104 | 108 | 100 | 17.68 | 93        | A0A068V6W8 | <i>Coffea canephora</i><br>DH200—94 genomic scaffold, scaffold_132          |
| rp11_olive_007981         | 144 | 149 | 108  | 156 | 147 | 91  | 18.2  | 132.5     | Q93Y40     | Oxysterol-binding protein-related protein 3C<br><i>Arabidopsis thaliana</i> |
| rp11_olive_005099         | 888 | 728 | 678  | 530 | 550 | 579 | 18.84 | 658.8     | P53492     | Actin-7<br><i>Arabidopsis thaliana</i>                                      |
| rp11_olive_005815         | 311 | 272 | 256  | 322 | 364 | 444 | 19.03 | 328.2     | P17598     | Catalase isozyme 1<br><i>Gossypium hirsutum</i>                             |
| rp11_olive_000209_split_1 | 161 | 115 | 117  | 166 | 118 | 100 | 19.16 | 129.5     | Q67YI9-2   | 2 of Clathrin interactor EPSIN 2<br><i>Arabidopsis thaliana</i>             |
| rp11_olive_001245         | 161 | 175 | 108  | 166 | 187 | 118 | 19.16 | 152.5     | A5A7I7     | Calcium-dependent protein kinase 4<br><i>Solanum tuberosum</i>              |
| rp11_olive_008079         | 239 | 204 | 197  | 343 | 275 | 263 | 19.34 | 253.5     | M1AVD3     | Uncharacterized protein<br><i>Solanum tuberosum</i>                         |
| rp11_olive_008883         | 128 | 119 | 144  | 187 | 108 | 118 | 19.51 | 134       | Q9LZI2     | UDP-glucuronic acid decarboxylase 2<br><i>Arabidopsis thaliana</i>          |
| rp11_olive_035033         | 178 | 166 | 224  | 177 | 285 | 199 | 19.76 | 204.8     | P62201     | Calmodulin<br><i>Lilium longiflorum</i>                                     |
| rp11_olive_029725         | 211 | 128 | 184  | 177 | 157 | 118 | 19.82 | 162.5     | O04834     | GTP-binding protein SAR1A<br><i>Arabidopsis thaliana</i>                    |

ensayo con proteína cinasa activada por mitógeno (rp11\_olive\_002359), también propuesto como RG para tejidos reproductivos, solo produjo algo de amplificación en el polen maduro, y muy poca o nula amplificación en el resto de tejidos.

En general, las validaciones por RT-PCR mostraron un alto grado de concordancia con los resultados obtenidos por métodos bioinformáticos, a pesar del escaso número de lecturas que proporciona la plataforma 454/Roche. Además, esta validación preliminar de algunos de los candidatos a RG propuestos por *findRGs* proporciona fiabilidad a nuestra herramienta. En vista de los resultados creemos que es conveniente el uso de nuestra herramienta antes de llevar a cabo cualquier experimento de RT-PCR o qPCR.

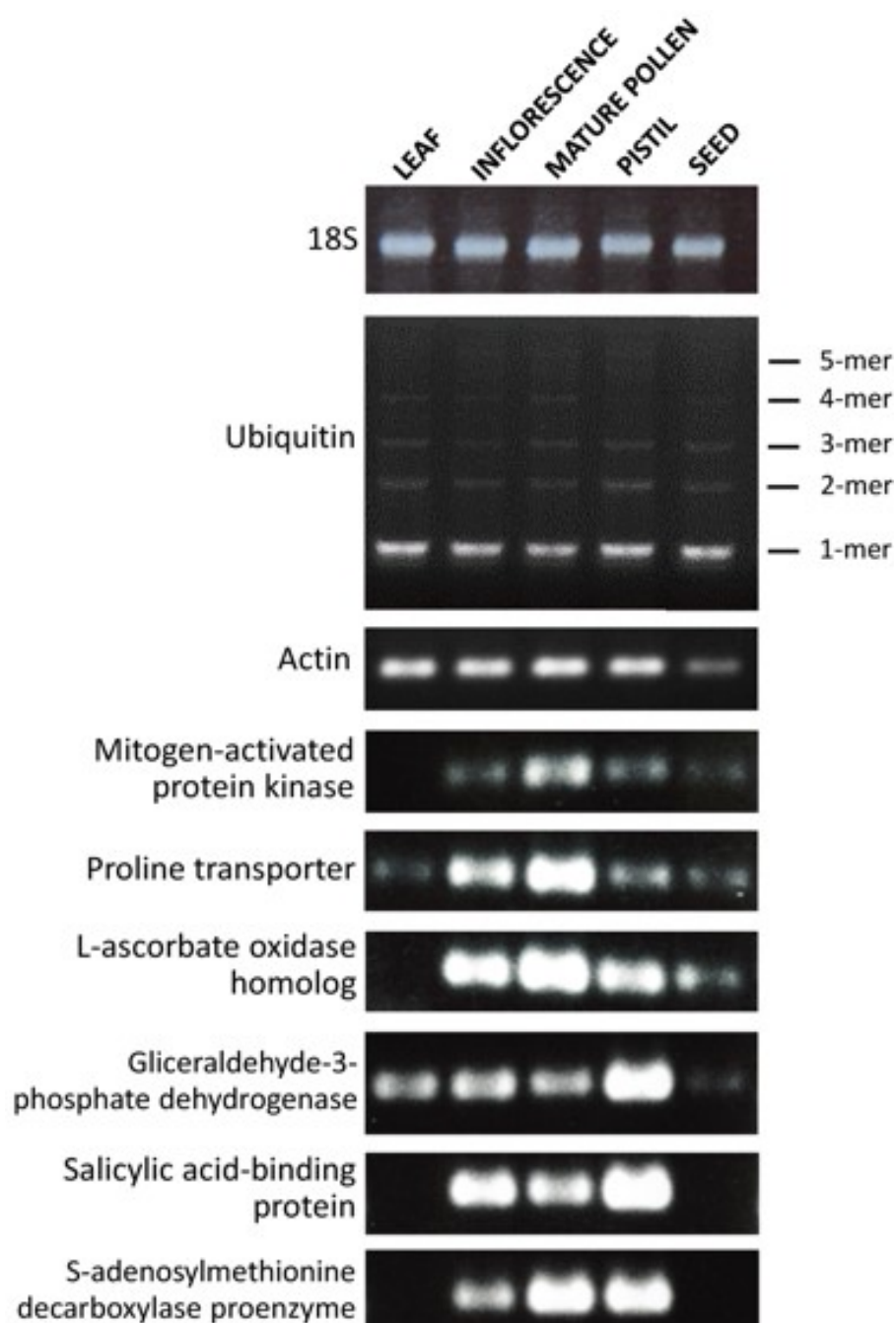


**Figura 4.12:** Validación preliminar mediante RT-PCR semicuantitativa en tiempo real en tejidos de olivo de algunos de los candidatos a RG predichos por *findRGs*. Este trabajo ha sido llevado a cabo por María José Jiménez-Quesada. Tomada de [266].

### 4.3.3. Candidatos a genes de referencia en *Arabidopsis*

Tras los buenos resultados obtenidos con el uso de *findRGs* en tejidos reproductivos de olivo, decidimos demostrar que nuestro algoritmo también funcionaba con lecturas de Illumina en experimentos más complejos que el mostrado antes con el olivo. Para ello localizamos en las bases de datos los de *Arabidopsis* descritos en el apartado 3.3.3 y ejecutamos tres veces el flujo, una por réplica (figura 4.13). Al igual que en el olivo, los transcritos con expresión menos variable se retuvieron usando dos valores de corte de CV: 10 % (defecto) y 20 % (no estricto). Dado que la plataforma Illumina genera una cantidad mucho mayor de lecturas, establecimos valores mínimos de lecturas mapeadas por transcrito muy superiores (10 000, 30 000 y 100 000 lecturas). El número más bajo de candidatos a RG obtenidos en *Arabidopsis* (figura 4.13) con respecto a olivo (figura 4.11) puede explicarse por los diferentes umbrales utilizados en función de la tecnología de secuenciación. El número de candidatos a RG obtenido bajo las condiciones más estrictas (>100 000 lecturas y CV <10 %) es muy homogéneo entre las tres réplicas (figura 4.13 A-C) y se refieren casi exclusivamente al mismo gen: ribulosa-1,5-bisfosfato carboxilasa/o-xigenasa (rubisco) (AT1G67090, AT5G38410, AT5G38420, AT5G38430; tabla 4.7), que es de lejos el mejor candidato en todos los casos, ya que su RPMM es mucho mayor que el de otros candidatos a RG obtenidos en condiciones menos estrictas (fichero complementario 4<sup>11</sup> de [266]). Esta proteína ya se ha usado previamente como RG en tejidos de la hoja en té [298]. No obstante, se ha visto que no es un buen candidato para tejidos no fotosintéticos como los de la antera, ni para polen [299].

Si comparamos los candidatos a RG obtenidos en condiciones menos estrictas (>10 000 lecturas y CV <10 %) observamos 77 candidatos comunes a las tres réplicas (fichero complementario 4 de [266]). Varios de estos genes ya se han nombrado previamente en este capítulo por su uso extendido como RG, como es el caso de la gliceraldehído-3-fosfato deshidrogenasa (AT3G26650, AT1G42970, AT3G04120 y AT1G13440; fichero complementario 4 de [266]) y la S-adenosilmetionina descarboxilasa (AT3G02470; fichero complementario 4 de [266]). Aparte nuestro flujo sugiere como buenos candidatos otros RG de uso común en otras especies, como es el caso de la fosfoglicerato cinasa 1 (AT3G12780; fichero complementario 4 de [266]), usada típicamente como control. Este gen muestra niveles de expresión constantes en hojas, fruto y flores de tomate [281] y también se ha descrito como uno de los mejores RG en especies de *Chrysanthemum* sometidas a distintos tipos de estrés [300]. Entre los candidatos encontramos también varias proteínas ribosómicas (AT1G43170, AT1G02780, AT3G25520, AT5G39740, AT1G56070 y AT5G20290; fichero complementario 4 de [266]) que han sido clasificadas como genes *housekeeping* y sugeridas como RG basándose en análisis de datos de *microarrays* [301]. No obstante, ya que se ha visto que estos genes presentan diferencias de expresión entre tejidos, su validez debería evaluarse en cada situación concreta. Otro gen propuesto como candidato es la  $\alpha$ -tubulina (AT1G50010 y AT1G04820; fichero complementario 4 de [266]). Este gen se ha usado bastante como RG, aunque existe controversia en cuanto a su fiabilidad, pues en algunas especies se considera el mejor, y en otras el peor [291]. Además de los mencionados, hay otros muchos candidatos a RG bien posicionados en todas las ejecuciones del flujo que merecerían ser evaluados experimentalmente, como es el caso de la fructosa-bisfosfato aldolasa (AT2G21330 y AT4G38970; fichero complementario 4 de [266]) o el factor de elongación 1- $\alpha$  (AT1G07920, AT1G07930, AT1G07940 y AT5G60390; fichero complementario 4 de [266]).

El hecho de que el mismo gen aparezca como RG candidato en las tres réplicas por separado indica que la posible variabilidad que existe entre réplicas no parece estar afectando a la estimación. Adicionalmente llevamos a cabo una ejecución del flujo combinando todas las réplicas (resultados no publicados) y obtuvimos exactamente los mismos candidatos a RG, lo cual sugiere que un número medio de lecturas por réplica de  $\sim 11\,000\,000$  (tabla 3.5) podría ser suficiente para el objetivo del flujo.

<sup>11</sup>[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5568602/bin/12938\\_2017\\_356\\_MOESM4\\_ESM.docx](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5568602/bin/12938_2017_356_MOESM4_ESM.docx)

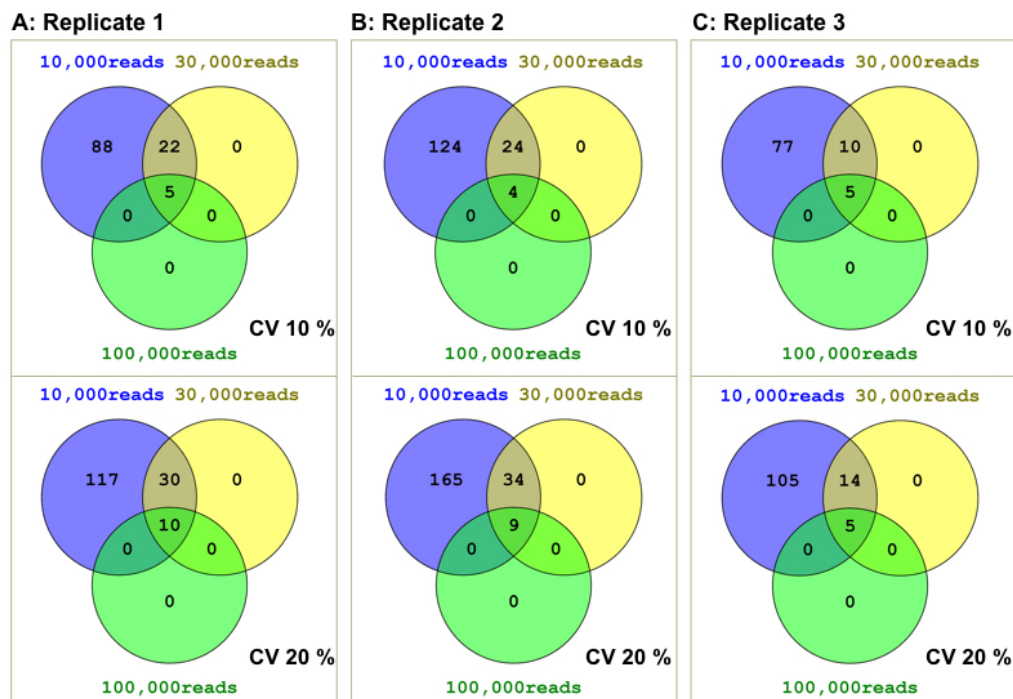


**Figura 4.13:** Diagramas de Venn que resumen el número de RG obtenidos en *Arabidopsis*. Se usaron dos valores de corte para CV (10 % y 20 %), y tres valores de mínimo de lecturas mapeadas (10 000, 30 000 y 100 000). Tomada de [266].

En vista de los resultados podemos concluir que la lista de candidatos a RG que produce *findRGs* en este análisis ofrece una primera y fiable estimación de los RG más apropiados para los posibles estudios de expresión que se lleven a cabo entre estas dos variedades de *Arabidopsis*, y para estas condiciones experimentales concretas.

**Tabla 4.7: Los mejores RG *Arabidopsis thaliana* según la figura 4.13, ordenados por CV.** Se obtuvieron para las tres réplicas, con CV <10 % y como mínimo 100 000 lecturas mapeadas. *Transcript\_id*: identificadores en la base de datos TAIR. Tomada de [266].

| <i>Arabidopsis</i> | Transcript_id | RPMM   |       | CV(%) | Mean<br>RPMM | Description                                                   |
|--------------------|---------------|--------|-------|-------|--------------|---------------------------------------------------------------|
|                    |               | Col_0  | Kil_0 |       |              |                                                               |
| Replicate 1        | AT1G67090.1   | 9,476  | 9,191 | 1.53  | 9,333.5      | Ribulose biphosphate carboxylase small chain 1A               |
|                    | AT5G38410.1   | 7,760  | 8,135 | 2.36  | 7,947.5      | Ribulose biphosphate carboxylase (small chain) family protein |
|                    | AT5G38430.1   | 7,054  | 7,548 | 3.38  | 7,301        | Ribulose biphosphate carboxylase (small chain) family protein |
|                    | AT2G39730.1   | 4,051  | 4,343 | 3.48  | 4,197        | Rubisco activase                                              |
|                    | AT5G38420.1   | 7,329  | 7,889 | 3.68  | 7,609        | Ribulose biphosphate carboxylase (small chain) family protein |
| Replicate 2        | AT2G39730.1   | 3,906  | 4,323 | 5.07  | 4,114.5      | Rubisco activase                                              |
|                    | AT1G67090.1   | 8,523  | 9,636 | 6.13  | 9,079.5      | Ribulose biphosphate carboxylase small chain 1A               |
|                    | AT1G21310.1   | 7,013  | 7,976 | 6.42  | 7,494.5      | Extensin 3                                                    |
|                    | AT5G38410.1   | 7,047  | 8,438 | 8.98  | 7,742.5      | Ribulose biphosphate carboxylase (small chain) family protein |
| Replicate 3        | AT5G38420.1   | 8,708  | 8,526 | 1.06  | 8,617        | Ribulose biphosphate carboxylase (small chain) family protein |
|                    | AT5G38430.1   | 8,424  | 8,169 | 1.54  | 8,296.5      | Ribulose biphosphate carboxylase (small chain) family protein |
|                    | AT2G39730.1   | 4,365  | 4,524 | 1.79  | 4,444.5      | Rubisco activase                                              |
|                    | AT5G38410.1   | 9,172  | 8,822 | 1.95  | 8,997        | Ribulose biphosphate carboxylase (small chain) family protein |
|                    | AT1G67090.1   | 11,051 | 9,694 | 6.54  | 10,372.5     | Ribulose biphosphate carboxylase small chain 1A               |

## 4.4. Detección bioinformática de nuevos alérgenos

A pesar de que ya se han descrito 15 alérgenos distintos en el olivo, 14 de ellos en el polen y uno en el fruto (tabla 1.2), se sabe que todavía quedan alérgenos de esta planta por identificar, puesto que los alergogramas (transferencias de tipo Western correspondientes a extractos de polen de olivo sometidas a incubación con el suero de pacientes alérgicos) revelan un número de bandas reactivas superior al número de alérgenos descrito. La disponibilidad de un transcriptoma reproductivo de olivo permite, entre otras muchas cosas, explorar la existencia de estos posibles alérgenos nuevos [302]. Con este objetivo, se diseñó un algoritmo con la herramienta AutoFlow (apartado 3.7), que se aplicó sobre los  $TT_{pollen}$  de ReprOlive y que ha permitido, además de encontrar posibles alérgenos nuevos, examinar posibles casos de reactividad cruzada con otras especies y detectar probablemente isoformas nuevas de los alérgenos de olivo que ya se conocen.

### 4.4.1. Alérgenos de polen de olivo ya conocidos

En la primera etapa del algoritmo (figura 3.5A), los 27 823  $TT_{pollen}$  que componen el transcriptoma de polen de ReprOlive (véase la tabla 4.1) se compararon mediante BLASTX con la base de datos *Olive allergens DB* (apartado 3.7.1) para detectar cualquiera de los 13 alérgenos de olivo que ya se conocían, que se hayan conseguido ensamblar en ReprOlive. Se obtuvieron 5 814 coincidencias o alineamientos, de los cuales se mantuvieron 923 tras el filtrado (véase tabla complementaria 1 en <https://figshare.com/s/7159a5d5b8c1cf93ba88>). Estos alineamientos involucraron a 99 *contigs* que, según su anotación con Full-LengthNext, corresponden a 26 transcritos diferentes y representan los alérgenos de olivo que contiene ReprOlive ( $TT_{pollen}^{known}$ ). Que haya tantos  $TT$  para tan solo 13 alérgenos conocidos se debe a que un mismo  $TT_{pollen}$  puede generar una o más coincidencias con un mismo o diferentes alérgenos de la base de datos. De estos 26  $TT_{pollen}^{known}$ , 21 contienen un ORF completo. En la tabla 4.8 se muestran las correspondencias de mayor porcentaje de identidad entre estos transcritos completos y su homólogo *Ole e* en la base de datos. Cabe destacar que los alineamientos entre estos  $TT_{pollen}^{known}$  y las secuencias de los alérgenos de la base de datos alcanzan el 100 % de la longitud del alérgeno en todos los casos, y los porcentajes de identidad son muy próximos o iguales a 100. Aquellos casos con porcentajes de identidad

inferiores 100 pueden ser debidos a polimorfismos de uno o unos cuantos nucleótidos, o bien a diferentes isoformas del alérgeno. Algunos de los alérgenos de olivo son altamente polimórficos. Tal es el caso de *Ole e 2* [173, 174, 175], con un alto número de secuencias distintas recogidas en UniProt, muchas de las cuales se identifican completas en ReprOlive (tabla 4.8).

La base de datos *Olive allergens DB* utilizada contenía los alérgenos de olivo identificados hasta la fecha de realización del análisis (*Ole e 1* a *Ole e 13*). En función de los resultados observados en la tabla 4.8, podemos concluir que, de los alérgenos mencionados, el transcriptoma de polen de ReprOlive contiene 19 transcritos completos correspondientes solo a 10 de ellos, mientras que los alérgenos *Ole e 4*, *Ole e 7* y *Ole e 13* no fueron identificados.

En el caso concreto de *Ole e 7*, la base de datos *Olive allergens DB* utilizada contenía solo un fragmento de 21 aminoácidos de este alérgeno (UniProt P81430) [183]. Posteriormente a la realización de nuestro análisis se publicó su secuencia completa (95 aminoácidos) [184], por lo que decidimos analizar más en detalle la presencia o ausencia de este alérgeno, no solo en el transcriptoma de polen de ReprOlive, sino también en el transcriptoma mejorado de polen (descrito en el apartado 4.2) y en los genomas de las variedades Farga, acebuche y Picual (tabla 3.6). La figura 4.14 muestra los resultados de la comparación BLAST del fragmento de 21 aminoácidos de *Ole e 7*, mientras que las figuras 4.15, 4.16, 4.17 y 4.18 corresponden a las comparaciones con el alérgeno completo.

En la comparación entre el fragmento de 21 aminoácidos de *Ole e 7* y el transcriptoma de polen de ReprOlive se obtienen algunas coincidencias, aunque con valores de identidad y cobertura del alérgeno bastante bajos ( $< 70\%$  y 13-15 aminoácidos respectivamente, figura 4.14A). La misma tendencia se observa para el transcriptoma de polen mejorado (véase figura 4.14B). En la comparación con la secuencia completa de *Ole e 7*, disminuye la media de identidad con ReprOlive o con el transcriptoma mejorado de polen (40/42 %, figuras 4.15A y 4.15B, respectivamente), aunque en este caso se cubre un área significativamente mayor de la secuencia (hasta 35/49 aminoácidos, figuras 4.15A y 4.15B, respectivamente). Sin embargo, en la búsqueda del alérgeno completo en los transcriptomas y proteomas derivados de genoma se obtiene una identidad cercana o igual a 100 % sobre toda la secuencia del alérgeno en las variedades Farga y Picual (véanse figuras 4.16 y 4.18). Sin embargo, la identidad sigue siendo baja en el caso del genoma del acebuche (véase figura 4.17).

#### 4.4.2. Reactividad cruzada con otras especies

Para estudiar la posible reactividad cruzada de los alérgenos de polen de olivo con otras especies, los 99  $TT_{pollen}^{known}$  de la etapa anterior (figura 3.5A) que coincidieron con algún alérgeno de olivo se compararon con la base de datos de alérgenos de otras especies (*Non-olive allergens DB*) en la segunda etapa del algoritmo. Se obtuvieron un total de 4 024 coincidencias, de las cuales 3 333 cumplieron los requisitos de filtrado (figura 3.5B) (véase tabla complementaria 2 en <https://figshare.com/s/7159a5d5b8c1cf93ba88>). Todos los  $TT_{pollen}^{known}$  analizados produjeron al menos un alineamiento con algún alérgeno de otra especie ( $AP_{non-olive}$ ). Se alinearon con algún  $TT_{pollen}^{known}$  94  $AP_{non-olive}$  y son, por tanto, candidatos a protagonizar casos de reactividad cruzada con polen de olivo. En la tabla 4.9 se muestran las posibles relaciones de reactividad cruzada. Dado que cada  $AP_{non-olive}$  puede haberse alineado con más de un  $TT_{pollen}^{known}$ , se indica en la tabla el rango de porcentajes de identidad encontrados para cada uno. Resulta destacable el hecho de que *Ole e 2* parece ser con diferencia el alérgeno con más reactividad cruzada, pues coincide con 53  $AP_{non-olive}$ , seguido de *Ole e 1* (14) y *Ole e 3* (12). Esto concuerda con el carácter ubicuo de la familia de proteínas similares a *Ole e 1* y a la cualidad de panalérgenos de *Ole e 2* y *Ole e 3*. Esta tabla recoge también las referencias bibliográficas de los casos en los que se ha estudiado dicha reactividad

**Tabla 4.8: Transcritos únicos del transcriptoma de polen de ReprOlive que corresponden a alérgenos Ole e.** *Transcrito*: identificador del TT en la base de datos ReprOlive; *Genoma id*: identificador del transcrito correspondiente en el transcriptoma deducido del borrador del genoma de olivo ‘Farga’ [135]; *Alérgeno*: identificador del alérgeno en UniProt; *Allergome id*: identificador del alérgeno en la base de datos Allergome; *Ole*: identificador del alérgeno según la base de datos de nomenclatura de alérgenos WHO/IUIS (<http://www.allergen.org>); *Familia proteica*: familia proteica del alérgeno; *% Id*: porcentaje de identidad. Para todos los casos el porcentaje de la longitud del alérgeno que alinea con el transcrito fue del 100 %.

| Transcrito                | Genoma id    | Alérgeno | Allergome id        | Ole             | Familia proteica                            | % Id  |
|---------------------------|--------------|----------|---------------------|-----------------|---------------------------------------------|-------|
| po11_olive_008927         | OE6A083939P1 | G0YM01   | 482                 | <i>Ole e 1</i>  | Proteínas similares a <i>Ole e 1</i>        | 100   |
| po11_olive_002425         | OE6A083939P1 | P19963   | 482;483;487;488;489 | <i>Ole e 1</i>  | Proteínas similares a <i>Ole e 1</i>        | 95,86 |
| po11_olive_015796         | OE6A006158P1 | A4KA52   | 490                 | <i>Ole e 2</i>  | Profilinas                                  | 99,24 |
| po11_olive_016426         | OE6A027931P1 | A4GE55   | 490                 | <i>Ole e 2</i>  | Profilinas                                  | 99,25 |
| po11_olive_013589         | OE6A018554P2 | B2BGV4   | 490                 | <i>Ole e 2</i>  | Profilinas                                  | 100   |
| po11_olive_015514         | OE6A027931P1 | A4KA54   | 490                 | <i>Ole e 2</i>  | Profilinas                                  | 100   |
| po11_olive_015754         | OE6A027931P1 | A4GFC3   | 490                 | <i>Ole e 2</i>  | Profilinas                                  | 100   |
| po11_olive_015758         | OE6A006158P1 | A4GFC0   | 490                 | <i>Ole e 2</i>  | Profilinas                                  | 100   |
| po11_olive_016414         | OE6A027931P1 | A4GDR8   | 490                 | <i>Ole e 2</i>  | Profilinas                                  | 100   |
| po11_olive_016420         | OE6A027931P1 | A4GFB7   | 490                 | <i>Ole e 2</i>  | Profilinas                                  | 100   |
| po11_olive_010847_split_0 | OE6A027931P1 | O24170   | 3383;490            | <i>Ole e 2</i>  | Profilinas                                  | 100   |
| po11_olive_021619         | OE6A026280P1 | O81092   | 3384;491            | <i>Ole e 3</i>  | Polcalcinas                                 | 98,81 |
| po11_olive_011703         | OE6A103554P1 | Q8L5E0   | 493                 | <i>Ole e 5</i>  | Superóxido dismutasas                       | 100   |
| po11_olive_025638         | OE6A022228P1 | O24172   | 3387;494            | <i>Ole e 6</i>  | Desconocida                                 | 100   |
| po11_olive_006357         | OE6A049638P1 | Q9M7R0   | 3389;496            | <i>Ole e 8</i>  | Proteínas de unión a calcio                 | 97,66 |
| po11_olive_002121         | OE6A043576P1 | Q94G86   | 3391;497            | <i>Ole e 9</i>  | 1,3-β-glucanasas                            | 99,13 |
| po11_olive_002069         | OE6A069575P2 | Q84V39   | 1319;3382           | <i>Ole e 10</i> | Proteínas de unión a carbohidratos (CBM-43) | 100   |
| po11_olive_001877         | OE6A111519P1 | B2VPR8   | 5966;8780           | <i>Ole e 11</i> | Pectina metilesterasas                      | 98,9  |
| po11_olive_004891         | OE6A055073P1 | E1U332   | 9073                | <i>Ole e 12</i> | Proteínas similares a isoflavona reductasas | 99,35 |

cruzada, para lo cual se ha consultado la base de datos Allerbase<sup>12</sup> [303], entre otras fuentes bibliográficas complementarias.

<sup>12</sup><http://bioinfo.unipune.ac.in/AllerBase>

**Tabla 4.9: Posibles casos de reactividad cruzada entre alérgenos de polen de olivo y otras especies.** Alérgeno: identificador del alérgeno según la base de datos de nomenclatura de alérgenos WHO/IUIS (<http://www.allergen.org>) Organismo: especie fuente del alérgeno; Identidad (%): rango de porcentaje de identidad entre el alérgeno y los distintos contigs con los que alinea. Entre paréntesis se indica el total de contigs; Referencias RC: referencias bibliográficas que analizan la reactividad cruzada del alérgeno con otras especies. Las referencias marcadas con un asterisco (\*) indican que existe reactividad cruzada con el alérgeno de olivo en cuestión, mientras que las marcadas con dos asteriscos (\*\*) describen una reactividad cruzada baja o nula con dicho alérgeno de olivo.

| Alérgeno           | Organismo                       | Identidad (%) | Referencias RC             |
|--------------------|---------------------------------|---------------|----------------------------|
| <b>Ole e 1</b>     |                                 |               |                            |
| Ama r 1            | <i>Amaranthus retroflexus</i>   | 36-44 % (70)  | [304]                      |
| Cha o 2            | <i>Chamaecyparis obtusa</i>     | 46 % (1)      | [305]                      |
| Che a 1            | <i>Chenopodium album</i>        | 35-41 % (70)  | [306]** [307]**            |
| Cro s 1            | <i>Crocus sativus</i>           | 35-41 % (68)  | [308]                      |
| Fra e 1            | <i>Fraxinus excelsior</i>       | 71-92 % (72)  | [309]* [310]*              |
| Koc s 1            | <i>Bassia scoparia</i>          | 35-42 % (66)  | [311]                      |
| Lig v 1            | <i>Ligustrum vulgare</i>        | 70-92 % (72)  | [309]*                     |
| Lol p 11           | <i>Lolium perenne</i>           | 35-40 % (9)   | [312]* [313]** [307]**     |
| Ory s Glyoxalase I | <i>Oryza sativa</i>             | 87 % (1)      | [314]                      |
| Phl p 11           | <i>Phleum pratense</i>          | 35-39 % (7)   | [313]** [307]**            |
| Pla l 1            | <i>Plantago lanceolata</i>      | 35-43 % (68)  | [315]*                     |
| Pro j 1            | <i>Prosopis juliflora</i>       | 35-47 % (71)  | [316]                      |
| Sal k 5            | <i>Kali turgidum</i>            | 36-43 % (70)  | [317]**                    |
| Vac fa 1           | <i>Vachellia farnesiana</i>     | 37-50 % (72)  | [318]                      |
| <b>Ole e 2</b>     |                                 |               |                            |
| Ama r 2            | <i>Amaranthus retroflexus</i>   | 75-85 % (12)  | [319]                      |
| Amb a 8            | <i>Ambrosia artemisiifolia</i>  | 64-85 % (12)  | [320]                      |
| Ana c 1            | <i>Ananas comosus</i>           | 73-82 % (12)  | [321]                      |
| Api g 4            | <i>Apium graveolens</i>         | 75-82 % (12)  | [322] [323]                |
| Ara h 5            | <i>Arachis hypogaea</i>         | 71-85 % (12)  | [324]                      |
| Art v 4            | <i>Artemisia vulgaris</i>       | 66-78 % (12)  | [325] [326]                |
| Beta v 2           | <i>Beta vulgaris</i>            | 100 % (5)     | [327]                      |
| Bet v 2            | <i>Betula pendula</i>           | 71-86 % (12)  | [322] [326] [172]* [328]*  |
| Bra n 8            | <i>Brassica napus</i>           | 66-81 % (12)  | [329]                      |
| Cap a 2            | <i>Capsicum annuum</i>          | 76-85 % (12)  | [330] [331]                |
| Car mi 2           | <i>Caryota mitis</i>            | 72-81 % (12)  | [332]                      |
| Che a 2            | <i>Chenopodium album</i>        | 68-82 % (12)  | [333] [334]* [335]* [328]* |
| Citr l 2           | <i>Citrullus lanatus</i>        | 73-82 % (12)  | [336]                      |
| Cit s 2            | <i>Citrus sinensis</i>          | 71-83 % (12)  | [337] [338]                |
| Cor a 2            | <i>Corylus avellana</i>         | 73-95 % (12)  | [339]                      |
| Cro s 2            | <i>Crocus sativus</i>           | 70-75 % (12)  | [340]                      |
| Cuc m 2            | <i>Cucumis melo</i>             | 72-83 % (12)  | [341] [342]* [328]*        |
| Cyn d 12           | <i>Cynodon dactylon</i>         | 75-88 % (12)  | [343] [172]* [344]*        |
| Dau c 4            | <i>Daucus carota</i>            | 73-82 % (12)  | [345]                      |
| Der f 36           | <i>Dermatophagoides farinae</i> | 35-37 % (6)   | [346]                      |

Table 4.9 (continuación)

| Alérgeno       | Organismo                             | Identidad (%) | Referencias RC             |
|----------------|---------------------------------------|---------------|----------------------------|
| Fra a 4        | <i>Fragaria ananassa</i>              | 72-83 % (12)  | [347]                      |
| Fra e 2        | <i>Fraxinus excelsior</i>             | 74-96 % (12)  | [172]*                     |
| Gly m 3        | <i>Glycine max</i>                    | 71-86 % (12)  | [348]                      |
| Hel a 2        | <i>Helianthus annuus</i>              | 65-82 % (12)  | [344]*                     |
| Hev b 8        | <i>Hevea brasiliensis</i>             | 74-92 % (12)  | [349]* [175]* [328]*       |
| Hor v 12       | <i>Hordeum vulgare</i>                | 72-92 % (12)  | [350]                      |
| Hum j 2        | <i>Humulus scandens</i>               | 74-88 % (12)  | [351]                      |
| Koc s 2        | <i>Bassia scoparia</i>                | 73-85 % (12)  | [352]                      |
| Lit c 1        | <i>Litchi chinensis</i>               | 77-88 % (12)  | [353]                      |
| Mal d 4        | <i>Malus domestica</i>                | 71-88 % (12)  | [354] [355]                |
| Man i 3        | <i>Mangifera indica</i>               | 72-83 % (12)  | [356] [357]                |
| Mer a 1        | <i>Mercurialis annua</i>              | 73-84 % (12)  | [358] [359]* [344]* [360]* |
| Mus a 1        | <i>Musa acuminata</i>                 | 71-84 % (12)  | [321]                      |
| Nic t 8        | <i>Nicotiana tabacum</i>              | 75-83 % (12)  | [361]                      |
| Ory s 12       | <i>Oryza sativa</i>                   | 74-94 % (12)  | [362] [314] [363]          |
| Par j 3        | <i>Parietaria judaica</i>             | 72-83 % (12)  | [364]                      |
| Pet c 2        | <i>Petroselinum crispum</i>           | 72-82 % (12)  | -                          |
| Phl p 12       | <i>Phleum pratense</i>                | 73-98 % (12)  | [365]* [366]*              |
| Pho d 2        | <i>Phoenix dactylifera</i>            | 76-83 % (12)  | [360]                      |
| Pro j 2        | <i>Prosopis juliflora</i>             | 72-82 % (12)  | [367]                      |
| Pru av 4       | <i>Prunus avium</i>                   | 73-84 % (12)  | [368]                      |
| Pru du 4       | <i>Prunus dulcis x Prunus persica</i> | 71-85 % (12)  | [369]                      |
| Pru p 4        | <i>Prunus persica</i>                 | 71-87 % (12)  | [328]*                     |
| Pyr c 4        | <i>Pyrus communis</i>                 | 76-87 % (12)  | [328]*                     |
| Ric c 8        | <i>Ricinus communis</i>               | 74-86 % (12)  | [351]                      |
| Sal k 4        | <i>Kali turgidum</i>                  | 71-82 % (12)  | [370]                      |
| Sin a 4        | <i>Sinapis alba</i>                   | 74-81 % (12)  | [371]                      |
| Sola l 1       | <i>Solanum lycopersicum</i>           | 75-86 % (12)  | [372]                      |
| Sola t 8       | <i>Solanum tuberosum</i>              | 77-86 % (12)  | [351]                      |
| Tri a 12       | <i>Triticum aestivum</i>              | 74-83 % (12)  | [373]                      |
| Tyr p 36       | <i>Tyrophagus putrescentiae</i>       | 35-40 % (8)   | [374] [375]                |
| Vac fa 2       | <i>Vachellia farnesiana</i>           | 72-81 % (12)  | [351]                      |
| Vit v 4        | <i>Vitis vinifera</i>                 | 73-85 % (12)  | [351]                      |
| <b>Ole e 3</b> |                                       |               |                            |
| Aln g 4        | <i>Alnus glutinosa</i>                | 82-84 % (3)   | [376]*                     |
| Amb a 9        | <i>Ambrosia artemisiifolia</i>        | 70 % (2)      | [320]                      |
| Art v 5        | <i>Artemisia vulgaris</i>             | 64 % (3)      | [320] [325]                |
| Bet v 4        | <i>Betula pendula</i>                 | 79-80 % (3)   | [376]* [377]*              |
| Bra n 7        | <i>Brassica napus</i>                 | 71-75 % (3)   | [378] [379]                |
| Bra r 7        | <i>Brassica campestris</i>            | 71-75 % (3)   | [380]                      |
| Che a 3        | <i>Chenopodium album</i>              | 82 % (3)      | [381]* [376]* [335]*       |
| Cup a 4        | <i>Hesperocyparis arizonica</i>       | 50-51 % (3)   | [382]                      |
| Fra e 3        | <i>Fraxinus excelsior</i>             | 83-90 % (3)   | [383]                      |
| Jun o 4        | <i>Juniperus oxycedrus</i>            | 52 % (3)      | [382] [384]                |

Table 4.9 (continuación)

| Alérgeno         | Organismo                       | Identidad (%) | Referencias RC       |
|------------------|---------------------------------|---------------|----------------------|
| Par j 4          | <i>Parietaria judaica</i>       | 74-77 % (3)   | [385]                |
| Phl p 7          | <i>Phleum pratense</i>          | 67-68 % (2)   | [381]* [376]* [377]* |
| <b>Ole e 5</b>   |                                 |               |                      |
| Neo fi Cu/Zn-SOD | <i>Neosartorya fischeri</i>     | 53-57 % (5)   | -                    |
| <b>Ole e 8</b>   |                                 |               |                      |
| Amb a 10         | <i>Ambrosia artemisiifolia</i>  | 55 % (1)      | [320]                |
| Cup a 4          | <i>Hesperocyparis arizonica</i> | 42 % (1)      | [382] [386]          |
| <b>Ole e 9</b>   |                                 |               |                      |
| Cap a Glucanase  | <i>Capsicum annuum</i>          | 36 % (1)      | [387]* [388] [389]   |
| Fra e 10         | <i>Fraxinus excelsior</i>       | 57 % (1)      | [390]                |
| Fra e 9          | <i>Fraxinus excelsior</i>       | 43-44 % (1)   | [387]* [391]         |
| Hev b 2          | <i>Hevea brasiliensis</i>       | 35-36 % (1)   | [387]* [392]*        |
| Mus a 5          | <i>Musa acuminata</i>           | 37-40 % (1)   | [387]* [393]         |
| Sola l Glucanase | <i>Solanum lycopersicum</i>     | 37 % (1)      | [387]* [394]         |
| Sola t Glucanase | <i>Solanum tuberosum</i>        | 36-37 % (1)   | [387]* <sup>13</sup> |
| <b>Ole e 10</b>  |                                 |               |                      |
| Cap a Glucanase  | <i>Capsicum annuum</i>          | 36 % (1)      | [388] [389]          |
| Fra e 10         | <i>Fraxinus excelsior</i>       | 94 % (1)      | [188]* [390]         |
| Fra e 9          | <i>Fraxinus excelsior</i>       | 90-94 % (1)   | [188]* [391]         |
| Hev b 2          | <i>Hevea brasiliensis</i>       | 35 % (1)      | [188]*               |
| Mus a 5          | <i>Musa acuminata</i>           | 36-38 % (1)   | [393]                |
| Sola l Glucanase | <i>Solanum lycopersicum</i>     | 37-38 % (1)   | [188]* [394]         |
| Sola t Glucanase | <i>Solanum tuberosum</i>        | 36-39 % (1)   | [188]* <sup>13</sup> |
| <b>Ole e 11</b>  |                                 |               |                      |
| Fra e 11         | <i>Fraxinus excelsior</i>       | 79 % (2)      | [390]                |
| <b>Ole e 12</b>  |                                 |               |                      |
| Bet v 6          | <i>Betula pendula</i>           | 81 % (1)      | [395] [192]          |
| Cor a 6          | <i>Corylus avellana</i>         | 79 % (1)      | [396]                |
| Cry j IFR        | <i>Cryptomeria japonica</i>     | 61 % (1)      | [397] [192]          |
| Dau c 5          | <i>Daucus carota</i>            | 72 % (1)      | [398] [192]          |
| Fra e 12         | <i>Fraxinus excelsior</i>       | 96 % (1)      | [390] [192]          |

<sup>13</sup><https://www.astmalitomysl.cz/ALERGIE%20publikace%202014.pdf#page=33>

### 4.4.3. Posibles alérgenos nuevos y nuevas isoformas de los alérgenos ya conocidos

Los 27 724  $TT_{pollen}^{source}$  que en la primera etapa del algoritmo (figura 3.5A) no coincidieron con ningún alérgeno de olivo constituyen una fuente de posibles alérgenos nuevos. Para evaluar esto, se compararon con *Non-olive allergens DB* y se obtuvieron 36 160 coincidencias, de las cuales solo se conservaron 5 616 tras el filtrado (véase tabla complementaria 3 en <https://figshare.com/s/7159a5d5b8c1cf93ba88>), que corresponden a 663  $TT_{pollen}^{newcand}$ , de los cuales 309 son transcritos únicos (figura 3.5C), que pueden ser buenos candidatos a alérgenos nuevos de olivo.

Para descartar posibles variantes de los alérgenos de polen de olivo ya conocidos, pero más distantes, los 663  $TT_{pollen}^{newcand}$  se comparan de nuevo con *Olive allergens DB*, pero con unos valores de los parámetros de filtrado más relajados (figura 3.5D). El resultado son 1 595 alineamientos, de los cuales solo 233 permanecen tras el filtrado (véase tabla complementaria 4 en <https://figshare.com/s/7159a5d5b8c1cf93ba88>) e involucran a 35  $TT_{pollen}$ , 15 de ellos únicos y pueden representar variantes nuevas ( $TT_{pollen}^{iso}$ ), más distantes de algún/os *Ole e*. Estos  $TT_{pollen}^{iso}$  quedan descartados como candidatos a alérgenos nuevos y se deberían añadir a los  $TT_{pollen}^{known}$  identificados en la primera etapa (figura 3.5A). Así, por ejemplo, en nuestro laboratorio Zafra et al. [150] ha conseguido identificar nuevas variantes del alérgeno *Ole e 5*, que han sido experimentalmente validadas mediante amplificación por PCR. Una de ellas presenta una delección de 24 nt, probablemente debida a un ajuste alternativo, que implica diferencias en la alergia (así como en la actividad enzimática y la localización celular) con respecto a las isoformas que ya se conocían.

Los 628  $TT_{pollen}^{newcand}$  restantes que no coincidieron con ningún alérgeno de olivo representan 294 transcritos únicos y se confirman como posibles alérgenos nuevos de olivo ( $TT_{pollen}^{new}$ , figura 3.5D). Presentaron similitud significativa con 302  $AP_{non-olive}$ , por lo que se consideraron para validación experimental como probables alérgenos nuevos de polen de olivo. De ellos, 60 se encontraron también en el polen, 126 en alimentos, y los restantes están presentes en hongos, ácaros y otras fuentes (tabla 4.10 y para más detalle véase tabla complementaria 5 en <https://figshare.com/s/7159a5d5b8c1cf93ba88>).

**Tabla 4.10: Alérgenos de especies distintas al olivo con similitud significativa con algún  $TT_{pollen}$ .** Descripción: categoría funcional del alérgeno; Alérgenos: identificadores de los alérgenos según la base de datos Allergome; Fuente: categoría del alérgeno en función de la fuente o ruta de exposición, siendo (P) polen, (F) fruta o alimento, (H) hongos, (A) ácaros y (O) otros.

| Descripción                                | Alérgenos                                                                                                   | Fuente        |
|--------------------------------------------|-------------------------------------------------------------------------------------------------------------|---------------|
| Aldehído deshidrogenasas                   | Alt a 10, Bea b Ald, Cla h 10, Tyr p 35                                                                     | H, A          |
| Aldolasas                                  | Sal s 3, Thu a 3                                                                                            | F             |
| Amilasas                                   | Hor v 16, Hor v 17, Tri a bA                                                                                | F             |
| Amina oxidasas                             | Zea m PAO                                                                                                   | O             |
| Aminotransferasas                          | Asp f AT V, Bos d AATr                                                                                      | F, H          |
| Antígeno 5                                 | Ves f 5, Ves g 5, Ves m 5, Vesp m 5, Vesp ma 5, Ves v 5                                                     | O             |
| Aspartil-proteasas                         | Aed a 11                                                                                                    | O             |
| Bromelaína                                 | Ana c 2                                                                                                     | F             |
| Calcitonina                                | Sal s Calcitonin                                                                                            | O             |
| Calreticulina                              | Pen ch 31                                                                                                   | H             |
| Catalasas                                  | Asp f Catalase, Asp n 30, Met a 30, Pen c 30                                                                | H             |
| Ciclofilinas                               | Asp f 11, Asp f 27, Bet v 7, Cand a CyP, Cat r 1, Dau c CyP, Der f 29, Hev b CyP, Malas 6, Rhi o 2, Sola 15 | F, H, A, O, P |
| Cisteín-proteasas y proteínas relacionadas | Act c 4, Act d 1, Act d 4, Amb a 11, Amb a CPI, Blot 1, Car p Endoproteínase, Der f 1, Der p 1, Gly m Bd30K | F, A, P       |
| Citocromo C y proteínas relacionadas       | Ani s CCOS3, Cur l 3                                                                                        | F, H          |
| Cofilina                                   | Der f 31                                                                                                    | A             |

Table 4.10 (continuación)

| Descripción                               | Alérgenos                                                                                                                                                                                                                                                                                                                        | Fuente        |
|-------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| Defensinas                                | Amb a 4, Ara h 13, Art v 1                                                                                                                                                                                                                                                                                                       | F, P          |
| Endopeptidasas                            | Rhi o 1                                                                                                                                                                                                                                                                                                                          | H             |
| Enolasas                                  | Alt a 6, Asp f 22, Bea b Enolase, Bla g Enolase, Bos d Enolase, Cand a Enolase, Cla h 6, Cur l 2, Hev b 9, Ore m 2, Pen c 22, Rho m 1, Sac c Enolase, Sals 2, Thu a 2, Zea m 22                                                                                                                                                  | F, H, O, P    |
| Enzimas puente de berberina               | Api g 5, Cyn d 4, Dac g 4, Hor v 4, Lol p 4, Phl p 4, Sec c 4, Tri a 4, Zea m 4                                                                                                                                                                                                                                                  | F, P          |
| Factores de elongación                    | Der f EF, Hev b 1, Pen c 24, Tri a 45                                                                                                                                                                                                                                                                                            | F, H, A, O    |
| Factores de inicio de la traducción       | Homs eIF6                                                                                                                                                                                                                                                                                                                        | O             |
| Ferritina                                 | Der f 30, Pro cl 21kD                                                                                                                                                                                                                                                                                                            | F, A          |
| Flavodoxinas                              | Alt a 7, Cla h 7                                                                                                                                                                                                                                                                                                                 | H             |
| Fosfatasas                                | Der f 32, Tri a PAP                                                                                                                                                                                                                                                                                                              | A, O          |
| Fosfoglicerato deshidrogenasa             | Cyn d 22                                                                                                                                                                                                                                                                                                                         | P             |
| Fructofuranosidasas                       | Sola l 2                                                                                                                                                                                                                                                                                                                         | F             |
| Fructosiltransferasas                     | Lol p FT                                                                                                                                                                                                                                                                                                                         | P             |
| Galactosidasas                            | Homs alpha-Galactosidase                                                                                                                                                                                                                                                                                                         | O             |
| Gliceraldehído-3-fosfato deshidrogenasa   | Asp v GAPDH, Tri a 34                                                                                                                                                                                                                                                                                                            | H, O          |
| Glucosa-6-fostato isomerasa               | Asp f GPI, Gal d GPI                                                                                                                                                                                                                                                                                                             | F, H          |
| Glutación transferasa                     | Tri a GST                                                                                                                                                                                                                                                                                                                        | O             |
| Hidrofobinas                              | Cla h HCh1                                                                                                                                                                                                                                                                                                                       | H             |
| Inhibidor de xilanasas                    | Tri a XI                                                                                                                                                                                                                                                                                                                         | F             |
| Malato deshidrogenasas                    | Asp f MDH, Citr l MDH, Mala f 4                                                                                                                                                                                                                                                                                                  | F, H          |
| Metalotioneninas                          | Cof a 2                                                                                                                                                                                                                                                                                                                          | O             |
| Metionina sintasas                        | Salk 3                                                                                                                                                                                                                                                                                                                           | P             |
| Miosinas                                  | Der f 26, Gal d 7                                                                                                                                                                                                                                                                                                                | F, A          |
| N-acetilglucosaminidasas                  | Pen ch 20                                                                                                                                                                                                                                                                                                                        | H             |
| Oleosinas                                 | Ara h 10, Ara h 11, Ara h 14, Ara h 15, Cor a 12, Cor a 13, Ses i 4, Ses i 5                                                                                                                                                                                                                                                     | F             |
| Papaína                                   | Carp 1                                                                                                                                                                                                                                                                                                                           | F             |
| Parvalbúminas                             | Ang ja 1, Bos d PRVB, Clu h 1, Cyp c 1, Equ c PRVB, Evy j 1, Gad c 1, Gad m 1, Gal d 8, Hyp no 1, Kat p 1, Lat c 1, Lep w 1, Lit ca 1, Mel g PRVB, Mer ca 1, Mer hu 1, Mer mr 1, Onc m 1, Ran e 1, Ran e 2, Ras k 1, Salf 1, Sals 1, Sarm 1, Sars a 1, Sco j 1, Sco s 1, Seb m 1, Suss PRVB, The ch 1, Thu a 1, Tra j 1, Xip g 1 | F             |
| Pectato liasas                            | Amb a 1, Art v 6, Cha o 1, Cry j 1, Cup a 1, Cups 1, Jun a 1, Jun o 1, Jun v 1                                                                                                                                                                                                                                                   | P             |
| Pepsina                                   | Sus s Pepsin                                                                                                                                                                                                                                                                                                                     | O             |
| Peroxidasas                               | Cap a 30kD, Mus a 6, Sola l Peroxidase, Tri a Bd36K, Tri a Peroxidase 1, Tri a Peroxidase                                                                                                                                                                                                                                        | F, O          |
| Peroxirredoxinas                          | Hor v 32, Ory s 32, Tri a 32, Zea m 32                                                                                                                                                                                                                                                                                           | F, O          |
| Poligalacturonasas                        | Bra n PG, Cry j 2, Jun a 2, Lill PG, Pas n 13, Phl p 13, Pla a 2, Pla or 2, Sola l PG, Sor h 13, Zea m 13                                                                                                                                                                                                                        | F, P          |
| Proteínas de almacenamiento de tubérculos | Dio p TSP                                                                                                                                                                                                                                                                                                                        | F             |
| Proteínas de choque térmico               | Aed a 8, Alt a 3, Cass 9, Cla h HSP70, Cor a 10, Der f 28, Der f HSP70, Mala g 10, Mala s 10, Pen c 19, Tyr p 28                                                                                                                                                                                                                 | F, H, A, O, P |
| Proteínas del proteasoma                  | Zea m 20S                                                                                                                                                                                                                                                                                                                        | F             |
| Proteínas de membrana del peroxisoma      | Cand b 2, Mala f 3                                                                                                                                                                                                                                                                                                               | H             |
| Proteínas de tipo taumatina               | Cap a 1, Nic t Osmotin                                                                                                                                                                                                                                                                                                           | F, O          |
| Proteínas de unión a clorofila            | Api g 3                                                                                                                                                                                                                                                                                                                          | F             |

Table 4.10 (continuación)

| Descripción                                          | Alérgenos                                                                                                                                                                                                                                                                                | Fuente     |
|------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| Proteínas relacionadas con patogénesis               | Act c 8, Act d 8, Aln g 1, Art v 2, Bet ch 1, Bet co 1, Bet le 1, Bet n 1, Bet p 1, Bet v 1, Cap a 4, Car b 1, Cass 1, Cass 1, Cor a 1, Cor a 1, Cor he 1, Cuc m 3, Cyn d 24, Fags 1, Fra a 1, Jugr 5, Mal d 1, Ost c 1, Pru ar 1, Pru av 1, Pru p 1, Pru p 7, Pyr c 1, Que a 1, Sola 14 | F, P       |
| Proteínas ribosómicas                                | Alt a 12, Alt a 5, Asp f 23, Asp f 8, Asp f RPS3, Cla h 12, Cla h 5, Cul n 1, Fus c 1, Fus c 1, Hom s P2, Pen b 26, Pen cr 26, Pru du 5, Sac c P2                                                                                                                                        | F, H, O    |
| Proteínas similares a germina                        | Ara t GLP                                                                                                                                                                                                                                                                                | P          |
| Proteína tumoral controlada traduccionalmente (TCTP) | Alt a TCTP, Cla h TCTP, Hom s TCTP                                                                                                                                                                                                                                                       | H, O       |
| Quimopapaína                                         | Car p Chymopapain                                                                                                                                                                                                                                                                        | F          |
| Quimosina                                            | Bos d Chymosin                                                                                                                                                                                                                                                                           | O          |
| Cinasas y proteínas relacionadas                     | Blag RACK1, Cand a PGK, Gal d Phosvitin, Sal k 2                                                                                                                                                                                                                                         | F, H, O, P |
| Quitinasas                                           | Cas s 5, Cof a 1, Hev b 11, Hev b 14, Mus a 2, Mus xp 2, Pun g 14                                                                                                                                                                                                                        | F, O       |
| Ribulosa bifosfato carboxilasa                       | Spi o RuBisCO                                                                                                                                                                                                                                                                            | F          |
| Serín-proteasas                                      | Cuc m 1                                                                                                                                                                                                                                                                                  | F          |
| Tiol reductasa                                       | Tri a 27                                                                                                                                                                                                                                                                                 | O          |
| Tiorredoxinas                                        | Asp f 28, Asp f 29, Cop c 2, Fus c 2, Hev b Trx, Hom s Trx, Mala s 13, Plo i 2, Tri a 25, Tri a Trx, Zea m 25                                                                                                                                                                            | F, H, O    |
| Triosafosfato isomerasas                             | Arc s 8, Blag TPI, Bos d TPI, Cra c 8, Der f 25, Pen m 8, Tri a 31                                                                                                                                                                                                                       | F, A, O    |
| Tropomiosinas                                        | Art fr 1, But ju 10, Har sa 7, Mac h 7, Tur c 1                                                                                                                                                                                                                                          | F, O       |
| Troponinas                                           | Blag 6, Cra c 6, Hom a 6, Pen m 6, Per a 6, Tyr p 24                                                                                                                                                                                                                                     | F, A, O    |
| Tubulinas                                            | Lep d 33, Tyr p 33                                                                                                                                                                                                                                                                       | A          |

A)

| Transcrito        | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del transcrito (nt) |                      |
|-------------------|---------------|----------------------------|--------------------------------|------------------------------|----------------------|
| po11_olive_021429 | 69,23         | 61,90                      | 13                             | 583                          | sp P81430 ALL7_OLEEU |
| po11_olive_014249 | 61,54         | 61,90                      | 13                             | 687                          | po11_olive_021429    |
| po11_olive_006615 | 66,67         | 71,43                      | 15                             | 1 001                        | po11_olive_014249    |
| po11_olive_004011 | 66,67         | 71,43                      | 15                             | 1 260                        | po11_olive_006615    |
| po11_olive_017495 | 61,54         | 61,90                      | 13                             | 630                          | po11_olive_004011    |
| po11_olive_027273 | 61,54         | 61,90                      | 13                             | 532                          | po11_olive_017495    |
|                   |               |                            |                                |                              | po11_olive_027273    |

APSQSTVTALLTSCVSYIDQ

APSMILTATIALTS-----

-----VISLLTSCVPYFE--

SPSQ-MTTALSTSCL-----

SPSQ-ATTALSTSCL-----

APTPSRITIALTS-----

-----TWLLTSPISKIDE-

B)

| Transcrito  | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del transcrito (nt) |                      |
|-------------|---------------|----------------------------|--------------------------------|------------------------------|----------------------|
| 00000110045 | 66,67         | 71,43                      | 15                             | 2 010                        | sp P81430 ALL7_OLEEU |
| 00000110046 | 66,67         | 71,43                      | 15                             | 2 332                        | 0000110045           |
| 00000110048 | 66,67         | 71,43                      | 15                             | 2 371                        | 0000110046           |
|             |               |                            |                                |                              | 0000110048           |

APSQSTVTALLTSCVSYIDQ

SPSQ-ATTALSTSCL-----

SPSQ-ATTALSTSCL-----

SPSQ-ATTALSTSCL-----

SPSQ-ATTALSTSCL-----

Figura 4.14: Búsqueda del fragmento de Ole e 7 de 21 aminoácidos [183] (secuencia sp|P81430|ALL7\_OLEEU) en los transcriptomas de polen de ReprOlive (A) y de polen mejorado (B). La tabla de la izquierda recoge las coincidencias con los valores más altos de porcentaje de identidad y porcentaje de cobertura de alineamiento de la longitud del alérgeno (aa: aminoácidos; nt: nucleótidos). A la derecha se muestra el alineamiento con Clustal Ω de estos transcritos con el alérgeno, realizado con la plataforma SeaView.

A)

| Transcrito        | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del transcrito (nt) |
|-------------------|---------------|----------------------------|--------------------------------|------------------------------|
| po11_olive_024232 | 40            | 36,84                      | 35                             | 490                          |
| po11_olive_026548 | 40            | 36,84                      | 35                             | 511                          |
| po11_olive_000909 | 40            | 36,84                      | 35                             | 2 074                        |
| po11_olive_014079 | 40            | 36,84                      | 35                             | 691                          |

|                            |                                                                                                                                                                                                                                                                                                                              |
|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>ole7_OeoSantos_2018</b> | <b>RES</b> CSAVTA <b>K</b> ITSCVSYLDD <b>S</b> AK <b>FP</b> S <b>D</b> CC <b>AG</b> VK <b>TV</b> IAQA <b>TS</b> K <b>PD</b> K <b>L</b> AT <b>C</b> N <b>CL</b> K <b>S</b> ALAI <b>V</b> G <b>N</b> K <b>VD</b> T <b>G</b> RV <b>SS</b> I <b>PK</b> CG <b>MS</b> VD <b>IP</b> PL <b>D</b> EN <b>Y</b> D <b>C</b> SK <b>VP</b> |
| po11_olive_024232          | -----                                                                                                                                                                                                                                                                                                                        |
| po11_olive_026548          | -----                                                                                                                                                                                                                                                                                                                        |
| po11_olive_000909          | -----                                                                                                                                                                                                                                                                                                                        |
| po11_olive_014079          | -----                                                                                                                                                                                                                                                                                                                        |

B)

| Transcrito | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del transcrito (nt) |
|------------|---------------|----------------------------|--------------------------------|------------------------------|
| 0000055391 | 40,82         | 51,58                      | 49                             | 244                          |
| 0000042648 | 42,5          | 42,11                      | 40                             | 1 914                        |
| 0000042649 | 42,5          | 42,11                      | 40                             | 1 051                        |
| 0000035421 | 40            | 47,37                      | 45                             | 1 122                        |
| 0000118270 | 40            | 47,37                      | 45                             | 1 046                        |

|                            |                                                                                                                                                                                                                                                                                                                                    |
|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>ole7_OeoSantos_2018</b> | <b>AP</b> SC <b>SV</b> TA <b>K</b> ITSCVSYLDD <b>S</b> AK <b>FP</b> S <b>D</b> CC <b>AG</b> VK <b>TV</b> IAQA <b>TS</b> K <b>PD</b> K <b>L</b> AT <b>C</b> N <b>CL</b> K <b>S</b> AL <b>V</b> G <b>N</b> K <b>VD</b> T <b>G</b> RV <b>SS</b> I <b>PK</b> CG <b>MS</b> VD <b>IP</b> PL <b>D</b> EN <b>Y</b> D <b>C</b> SK <b>VP</b> |
| 0000055391                 | -----                                                                                                                                                                                                                                                                                                                              |
| 0000042648                 | -----                                                                                                                                                                                                                                                                                                                              |
| 0000042649                 | -----                                                                                                                                                                                                                                                                                                                              |
| 0000035421                 | -----                                                                                                                                                                                                                                                                                                                              |
| 0000118270                 | -----                                                                                                                                                                                                                                                                                                                              |

Figura 4.15: Búsqueda de la secuencia completa de Ole e 7 [184] (secuencia ole7\_OeoSantos\_2018) en los transcritos de polen de ReprOlive (A) y de polen mejorado (B). Las tablas recogen las coincidencias con los valores más altos de porcentaje de identidad y porcentaje de cobertura de alineamiento de la longitud del alérgeno (aa: aminoácidos; nt: nucleótidos). Bajo las tablas se visualiza el alineamiento con Clustal  $\Omega$  de estos transcritos con el alérgeno, realizados con la plataforma SeaView.

A)

| Transcrito   | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del transcrito (nt) |
|--------------|---------------|----------------------------|--------------------------------|------------------------------|
| OE6A002042T1 | 75,27         | 97,89                      | 93                             | 960                          |
| OE6A005344T1 | 87,23         | 98,95                      | 94                             | 366                          |
| OE6A014487T1 | 76,09         | 96,84                      | 92                             | 5 778                        |
| OE6A031005T1 | 87,23         | 98,95                      | 94                             | 366                          |

**ole7\_OeoSantos\_2018**

|               |                 |                       |                       |                   |            |               |
|---------------|-----------------|-----------------------|-----------------------|-------------------|------------|---------------|
| APSCSTIVAKLTS | CVSYLDDKSAKPTSD | CCAGKTVLAQATSKPD      | KLATCNCLKSALALVGNKVD  | TGRVSSL           | PKKCGMSVDL | PPIDKNYDCSKVP |
| -PSCPTVI      | AKLASCL         | SYIDDDQSTQ            | PASSCCAGLNGILAQAKSKQD | KIATVNCNCIKSALANL | GNKVD      | TGRVSGL       |
| APSCSTIVAKL   | TSCVSYIDDDQSAQ  | PTSDCCAGKSTILAQATSKPD | KIATCNCIKSALATVGNKVD  | TGRVSSL           | PKKCGMSVNL | PPIDKNYDCSKII |
| -PSCPTVI      | AKLASCL         | SYIDDDQSTQ            | PASSCCAGLNGILAQAKSKQD | KIATVNCNCIKSALANL | GNKVD      | TGRVSGL       |
| APSCSTIVAKL   | TSCVSYIDDDQSAQ  | PTSDCCAGKSTILAQATSKPD | KIATCNCIKSALATVGNKVD  | TGRVSSL           | PKKCGMSVNL | PPIDKNYDCSKII |
| APSCSTIVAKL   | TSCVSYIDDDQSAQ  | PTSDCCAGKSTILAQATSKPD | KIATCNCIKSALATVGNKVD  | TGRVSSL           | PKKCGMSVNL | PPIDKNYDCSKII |

B)

| Péptido      | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del péptido (aa) |
|--------------|---------------|----------------------------|--------------------------------|---------------------------|
| OE6A002042P1 | 75,27         | 97,89                      | 93                             | 122                       |
| OE6A005344P1 | 87,23         | 98,95                      | 94                             | 122                       |
| OE6A031005P1 | 87,23         | 98,95                      | 94                             | 122                       |

**ole7\_OeoSantos\_2018**

|               |                 |                       |                       |                   |            |               |
|---------------|-----------------|-----------------------|-----------------------|-------------------|------------|---------------|
| APSCSTIVAKLTS | CVSYLDDKSAKPTSD | CCAGKTVLAQATSKPD      | KLATCNCLKSALALVGNKVD  | TGRVSSL           | PKKCGMSVDL | PPIDKNYDCSKVP |
| -PSCPTVI      | AKLASCL         | SYIDDDQSTQ            | PASSCCAGLNGILAQAKSKQD | KIATVNCNCIKSALANL | GNKVD      | TGRVSGL       |
| APSCSTIVAKL   | TSCVSYIDDDQSAQ  | PTSDCCAGKSTILAQATSKPD | KIATCNCIKSALATVGNKVD  | TGRVSSL           | PKKCGMSVNL | PPIDKNYDCSKII |
| APSCSTIVAKL   | TSCVSYIDDDQSAQ  | PTSDCCAGKSTILAQATSKPD | KIATCNCIKSALATVGNKVD  | TGRVSSL           | PKKCGMSVNL | PPIDKNYDCSKII |

Figura 4.16: Búsqueda de la secuencia completa de Ole e 7 [184] (secuencia ole7\_OeoSantos\_2018) en el trans-criptoma (A) y el proteoma (B) deducidos del genoma de la variedad Farga. Las tablas recogen las coincidencias con los valores más altos de porcentaje de identidad y porcentaje de cobertura de alineamiento de la longitud del alérgeno (aa: aminoácidos; nt: nucleótidos). Bajo las tablas se visualiza el alineamiento con Clustal Ω de estos transcritos/péptidos con el alérgeno, realizados con la plataforma SeaView.

A)

| Transcrito  | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del transcrito (nt) |
|-------------|---------------|----------------------------|--------------------------------|------------------------------|
| Oeu000556.1 | 40            | 47,37                      | 45                             | 1 713                        |
| Oeu024795.1 | 43,59         | 41,05                      | 39                             | 357                          |
| Oeu024945.1 | 43,01         | 97,89                      | 93                             | 351                          |
| Oeu030974.2 | 40,91         | 46,32                      | 44                             | 2 850                        |

|                            |                    |                    |                       |                           |                     |                                      |
|----------------------------|--------------------|--------------------|-----------------------|---------------------------|---------------------|--------------------------------------|
| <b>ole7_OeoSantos_2018</b> | <b>APSCSTVTAKL</b> | <b>ITSCVSYLDDK</b> | <b>SAKPTSDCCAGVKT</b> | <b>VLQAATSKPKDKLATCNC</b> | <b>LSALALVGNKVD</b> | <b>TRVSSLPKKCGMSVDLPPLDKNYDCSKVP</b> |
| Oeu000556.1                | SSSCSFVARVSVPL     | ILCISALDDP         | ICTSEAIPEDEC          | BANKNTVLA                 |                     |                                      |
| Oeu024795.1                |                    |                    |                       | SDCLLVKTM                 | QALPHILET           | IPCLFK*FGLTLVENR                     |
| Oeu024945.1                | ATICSDVIKDL        | KPCLNYLKSG         | SGSPSPDCCAGASLS       | VSATSTADKQAA              | CACLKTAAKSISFK      | PELAKSLPGKCGTSLPYP-ITSPSVDCSK        |
| Oeu030974.2                |                    | AKLTSCA            | RSAMPCCAGL            | RLRTASAWM                 |                     | *N**RGWR                             |

B)

| Péptido     | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del péptido (aa) |
|-------------|---------------|----------------------------|--------------------------------|---------------------------|
| Oeu002216.1 | 40            | 42,11                      | 40                             | 441                       |
| Oeu005510.1 | 42,11         | 40                         | 38                             | 819                       |
| Oeu012188.1 | 40,48         | 44,21                      | 42                             | 3 138                     |
| Oeu016368.1 | 41,03         | 41,05                      | 39                             | 867                       |
| Oeu031557.2 | 42,11         | 40                         | 38                             | 990                       |
| Oeu055791.1 | 40            | 42,11                      | 40                             | 2 094                     |

|                            |                    |                    |                       |                           |                       |                                      |
|----------------------------|--------------------|--------------------|-----------------------|---------------------------|-----------------------|--------------------------------------|
| <b>ole7_OeoSantos_2018</b> | <b>APSCSTVTAKL</b> | <b>ITSCVSYLDDK</b> | <b>SAKPTSDCCAGVKT</b> | <b>VLQAATSKPKDKLATCNC</b> | <b>LSALALVGNKVD</b>   | <b>TRVSSLPKKCGMSVDLPPLDKNYDCSKVP</b> |
| Oeu002216.1                |                    | ATCCTTATAGT        | ITTCAGGTA             | CTTAAITCCAGG              | TATATA                |                                      |
| Oeu005510.1                |                    | AAATGACCAAGGAT     | CACATCA               |                           | G--GCCATCACTGGACA     |                                      |
| Oeu012188.1                |                    | ATCACCAAGTCTT      | CAAAATTT              |                           | TCAGATCCCCGAAATGGAAC  | TT                                   |
| Oeu016368.1                |                    |                    | CAGTGATAAA            | TGT                       | TTAAATCCCCCAAACTGTAAT | TGA                                  |
| Oeu031557.2                |                    | CCAGTCTTCAAA       | TJA                   |                           | TCAGATCCCCGAAATGGAAC  | TT                                   |
| Oeu031557.2                | TCAAAAGTTAAAT      | TCAGATTCGAGAAT     |                       |                           | GATGTATCTTGAACTGGACA  | TGAT                                 |
| Oeu055791.1                |                    | AGTCTTCTA          | CCTGCTT               |                           | AAAGTACCAAGTCTTACA    |                                      |

Figura 4.17: Búsqueda de la secuencia completa de Ole e 7 [184] (secuencia ole7\_OeoSantos\_2018) en el transcriptoma (A) y el proteoma (B) deducidos del genoma del acebuche. Las tablas recogen las coincidencias con los valores más altos de porcentaje de identidad y porcentaje de cobertura de alineamiento de la longitud del alérgeno (aa: aminoácidos; nt: nucleótidos). Bajo las tablas se visualiza el alineamiento con Clustal  $\Omega$  de estos transcritos/péptidos con el alérgeno, realizados con la plataforma SeaView.

A)

| Transcrito      | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del transcrito (nt) |
|-----------------|---------------|----------------------------|--------------------------------|------------------------------|
| Scf1152g02039.1 | 74.19         | 97.89                      | 93                             | 366                          |
| Scf3225g03032.1 | 64.89         | 98.95                      | 94                             | 306                          |
| Scf3787g01003.1 | 87.23         | 98.95                      | 94                             | 366                          |
| Scf7393g03037.1 | 75.27         | 97.89                      | 93                             | 366                          |

**ole7\_OeoSantos\_2018**  
Scf1152g02039.1  
Scf3225g03032.1  
Scf3787g01003.1  
Scf7393g03037.1

APSCSTVTAKLTSCVSYLDDKSAKPTSDCCAGVKTYLAQAATSKPDKIATCNCIKSALALVGNKVD TGRVSSLPKCGMSVDLPIDKNIYDCSKIP  
-PSCPTVTAKLASCLSYIDDDSTQPASSCCAGLNGILAQAQSKSDKIATVAVCNCIRSAALANLGNKVD TGRVSGLPKCGMSVDLPIDKNIYDCSKI-  
APSCSTVTAKLTSE-----EYSCATATSKPDKIATCNCIKSALAIVGNKVD TGRVSSLPK0CGMSVNLPPIDKNIYDCSKI-  
APSCSTVTAKLTSCVSYIDDDSAQPTSDCCAGVKSTILAQAATSKPDKIATCNCIKSALAIVGNKVD TGRVSSLPK0CGMSVNLPPIDKNIYDCSKI-  
-PSCPTVTAKLASCLSYIDDDSTQPASSCCAGLNGILAQAQSKSDKIATVAVCNCIKSALANLGNKVD TGRVSGLPKCGMSVDLPIDKNIYDCSKI-

B)

| Péptido         | Identidad (%) | Cobertura del alérgeno (%) | Longitud del alineamiento (aa) | Longitud del péptido (aa) |
|-----------------|---------------|----------------------------|--------------------------------|---------------------------|
| Scf1152g02039.1 | 74.19         | 97.89                      | 93                             | 122                       |
| Scf3225g03032.1 | 65.96         | 98.95                      | 94                             | 102                       |
| Scf3787g01003.1 | 87.23         | 98.95                      | 94                             | 122                       |
| Scf7393g03037.1 | 75.27         | 97.89                      | 93                             | 122                       |

**ole7\_OeoSantos\_2018**  
Scf1152g02039.1  
Scf3225g03032.1  
Scf3787g01003.1  
Scf7393g03037.1

APSCSTVTAKLTSCVSYLDDKSAKPTSDCCAGVKTYLAQAATSKPDKIATCNCIKSALALVGNKVD TGRVSSLPKCGMSVDLPIDKNIYDCSKIP  
-PSCPTVTAKLASCLSYIDDDSTQPASSCCAGLNGILAQAQSKSDKIATVAVCNCIRSAALANLGNKVD TGRVSGLPKCGMSVDLPIDKNIYDCSKI-  
APSCSTVTAKLTSE-----EYSCATATSKPDKIATCNCIKSALAIVGNKVD TGRVSSLPK0CGMSVNLPPIDKNIYDCSKI-  
APSCSTVTAKLTSCVSYIDDDSAQPTSDCCAGVKSTILAQAATSKPDKIATCNCIKSALAIVGNKVD TGRVSSLPK0CGMSVNLPPIDKNIYDCSKI-  
-PSCPTVTAKLASCLSYIDDDSTQPASSCCAGLNGILAQAQSKSDKIATVAVCNCIKSALANLGNKVD TGRVSGLPKCGMSVDLPIDKNIYDCSKI-

Figura 4.18: Búsqueda de la secuencia completa de Ole e 7 [184] (secuencia ole7\_OeoSantos\_2018) en el transcrito (A) y el proteoma (B) deducidos del genoma de la variedad Picual. Las tablas recogen las coincidencias con los valores más altos de porcentaje de identidad y porcentaje de cobertura de alineamiento de la longitud del alérgeno (aa: aminoácidos; nt: nucleótidos). Bajo las tablas se visualiza el alineamiento con Clustal Omega de estos transcritos/péptidos con el alérgeno, realizados con la plataforma SeaView.

## 4.5. Predicción *in silico* de posibles modificaciones postraduccionales (*S*-nitrosilación y Tyr-nitración) en el polen de olivo

El NO es una molécula de señalización muy importante en numerosos procesos fisiológicos. Se sabe que en las plantas desempeña una función muy destacada en el crecimiento, el desarrollo, la inmunidad, la respuesta a estrés y la reproducción sexual [157, 158, 143]. Además, el NO es responsable de modificaciones postraduccionales (PTM) como la *S*-nitrosilación y la Tyr-nitración. Con el fin de conocer mejor el papel del NO a través de estas PTM en el desarrollo y la fisiología del polen del olivo, el transcriptoma de polen obtenido en ReprOlive (apartado 4.1.1) ha sido analizado de forma masiva *in silico* con diversos programas bioinformáticos, para evaluar la posible capacidad de modificación de sus transcritos.

### 4.5.1. Predicción *in silico* del *S*-nitrosoproteoma

El proteoma deducido del transcriptoma de polen de ReprOlive se analizó de forma independiente con tres herramientas de predicción de *S*-nitrosilación: iSNO-AAPair, GPS-SNO y SNOsite, con las condiciones más restrictivas en los casos en los que es posible elegir.

De las 21 607 proteínas que constituyen el proteoma de polen (tabla 4.1), se identificaron las siguientes proteínas propensas a *S*-nitrosilación: 9 266 (42,9 %) según el programa iSNO-AAPair, correspondientes a 5 899 (27,3 %) proteínas diferentes (ortólogos únicos) (véase la tabla 1 en [399]<sup>14</sup>); 7 788 (36 %) según el programa GPS-SNO, correspondientes a 5 258 (24,3 %) proteínas diferentes (tabla 2 en [399]<sup>15</sup>), y 15 285 (70,7 %) según el programa SNOsite, que corresponden a 8 829 (40,9 %) proteínas diferentes (tabla 3 en [399]<sup>16</sup>).

### 4.5.2. Predicción *in silico* del Tyr-nitroproteoma

El proteoma de polen también se analizó con la herramienta GPS-YNO2 para identificar las proteínas propensas a Tyr-nitración. De los tres umbrales que permite elegir el programa, se usaron dos: el recomendado por omisión (medio) y el más restrictivo (alto), con el fin de comparar los resultados y evaluar la robustez de las predicciones.

De las 21 607 proteínas que constituyen el proteoma se identificaron 12 977 (60,1 %) proteínas propensas a Tyr-nitración con el umbral por omisión (medio) de la herramienta, que corresponden a 7 883 (36,5 %) proteínas diferentes (ortólogos únicos) (tabla 4 en [399]<sup>17</sup>). En el caso del umbral más restrictivo, el número de proteínas propensas a Tyr-nitración estimado por el programa bajó hasta 10 836 (50,2 %) y corresponden a 6 860 (31,7 %) proteínas diferentes (tabla 5 en [399]<sup>18</sup>).

### 4.5.3. Proteínas propensas a ambos tipos de modificación

Cada una de las tablas obtenidas en los apartados anteriores (tablas 1-5 de [399]), contiene además del identificador de la proteína en ReprOlive y la posición/es de los aminoácidos propensos a modifi-

<sup>14</sup><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647517/bin/mmc2.xlsx>

<sup>15</sup><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647517/bin/mmc3.xlsx>

<sup>16</sup><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647517/bin/mmc4.xlsx>

<sup>17</sup><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647517/bin/mmc5.xlsx>

<sup>18</sup><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647517/bin/mmc6.xlsx>

cación según la herramienta en cuestión, la predicción del resto de herramientas para dicha proteína. Esto permite evaluar si las distintas herramientas o el uso de umbrales diferentes coinciden o no en la predicción, y también si una proteína es propensa simultáneamente a ambos tipos de modificación. De acuerdo con esto se llevó a cabo una comparación entre las proteínas propensas a *S*-nitrosilación según el programa iSNO-AAPair (tabla 1 de [399]) y las propensas a Tyr-nitración según el programa GPS-YNO2, con umbral medio recomendado por defecto (tabla 4 de [399]), para determinar qué proteínas podrían ser sujeto de ambos tipos de modificación (figura 4.19). De las 21 607 proteínas que constituyen el proteoma de polen se identificaron 6 056 (28 %) proteínas potencialmente susceptibles a ambos tipos de PTM, que corresponden a 4 180 (19,3 %) proteínas diferentes (ortólogos únicos).

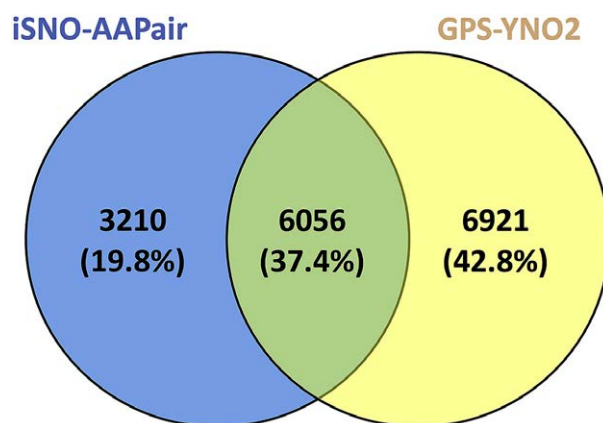


**Figura 4.19:** Comparación entre las proteínas propensas a *S*-nitrosilación según el programa iSNO-AAPair y las propensas a Tyr-nitración según el programa GPS-YNO2 (umbral medio) en el polen.

#### 4.5.4. Funciones celulares afectadas por las PTM

Las tablas 1-5 de [399] que contienen las posiciones potencialmente modificables según todas las herramientas de predicción, se han complementado además con la siguiente información obtenida con la herramienta de anotación Full-LengtherNext: el ortólogo más cercano, el estado del ORF (si es una proteína completa o incompleta), una descripción y el ortólogo en *Arabidopsis thaliana* de la base de datos TAIR10.

A partir de la lista de identificadores TAIR de las tablas 1 y 4 de [399], y con el propósito de dar significado biológico a los resultados obtenidos, se llevó a cabo un análisis de enriquecimiento de términos GO con la herramienta GeneCodis, determinando aquellos procesos biológicos, funciones moleculares, componentes celulares y rutas KEGG más significativos en los que las proteínas potencialmente modificables están involucradas.

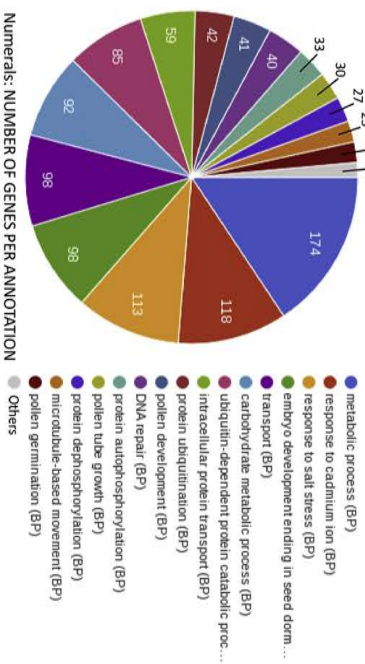
La figura 4.20 muestra el resultado del enriquecimiento llevado a cabo a partir de las proteínas potencialmente *S*-nitrosilables según el programa iSNO-AAPair. El análisis de los términos GO asociados a procesos biológicos incluye un amplio espectro de categorías, con una gran proporción de proteínas involucradas en varios procesos metabólicos, transporte, respuesta a distintos tipos de estrés (salino, por cadmio), etc. Cabe notar que numerosas de las proteínas destacadas como significativas están involucradas en procesos relacionados con la biología del polen (desarrollo, germinación y crecimiento del tubo polínico) (figura 4.20). En cuanto a las funciones moleculares destacan la unión a proteínas, ATP,

cinc, ácidos nucleicos y nucleótidos, así como diversas actividades enzimáticas. Las proteínas predichas como susceptibles a *S*-nitrosilación presentan además localizaciones celulares muy variadas (membrana plasmática, núcleo, cloroplastos, plasmodesmos, citoplasma...), mientras que las rutas KEGG más importantes incluyen la biosíntesis de metabolitos secundarios, con la mayor proporción de proteínas involucradas, seguida de otras rutas relevantes como el procesamiento de proteínas en el retículo endoplasmático, la fosforilación oxidativa y el metabolismo de moléculas como el almidón, la sacarosa, aminoácidos, nucleótidos, piruvato y glicerofosfolípidos, entre otros (figura 4.20).

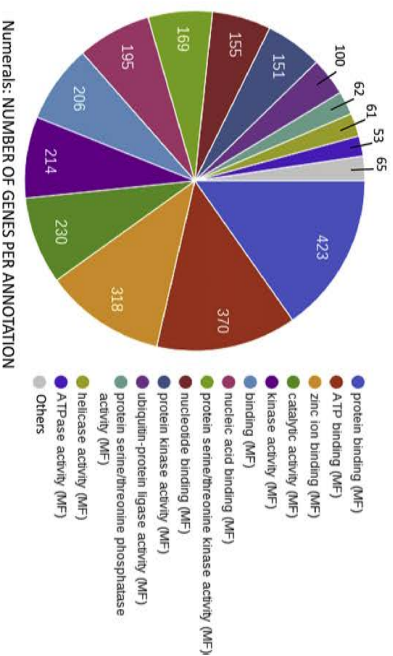
En cuanto al enriquecimiento llevado a cabo sobre la lista de proteínas potencialmente Tyr-nitrables obtenida con GPS-YNO2 con el umbral por omisión (medio), los procesos biológicos más destacados son también muy variados, con una gran proporción de proteínas involucradas en fosforilación de proteínas, diferentes procesos metabólicos, respuesta a diversos tipos de estrés (por cadmio, salino), transporte, etc. Además, al igual que ocurre con las proteínas *S*-nitrosiladas, un número relevante de las proteínas potencialmente Tyr-nitrables están involucradas en el desarrollo del grano de polen y el crecimiento del tubo polínico (figura 4.21). Con respecto al análisis de términos GO de funciones moleculares, destacaron numerosas categorías que incluyen la unión a proteínas, ATP, cinc, ácidos nucleicos (DNA y RNA) y nucleótidos, así como diversas actividades enzimáticas. Las proteínas predichas como propensas a esta PTM están asociadas a numerosas localizaciones celulares, con una gran proporción de ellas descritas como componentes de la membrana plasmática, núcleo, cloroplastos, membranas, plasmodesmos, etc. Las rutas KEGG más relevantes en las que intervienen estas proteínas incluyen la biosíntesis de metabolitos secundarios (con la mayor proporción de proteínas involucradas), junto con otras rutas relevantes como el procesamiento de proteínas en el retículo endoplasmático, el ayustosoma y el metabolismo de moléculas como al almidón, la sacarosa, aminoácidos y nucleótidos, entre otros (figura 4.21).

# ENRICHMENT ANALYSIS OF S-NITROSYLATION SITES IN OLIVE POLLEN PROTEINS

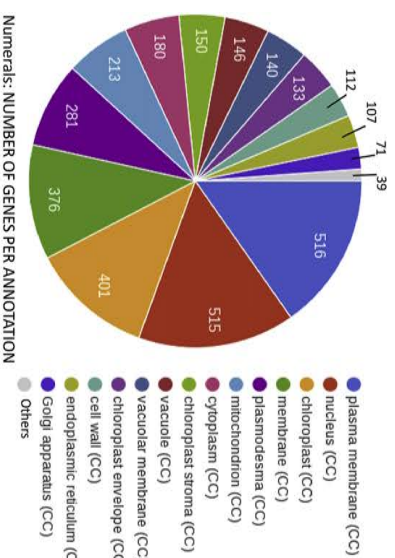
### Singular Enrichment Analysis of GO Biological Process



### Singular Enrichment Analysis of GO Molecular Function



### Singular Enrichment Analysis of GO Cellular Component



### Singular Enrichment Analysis of GO KEGG Pathways

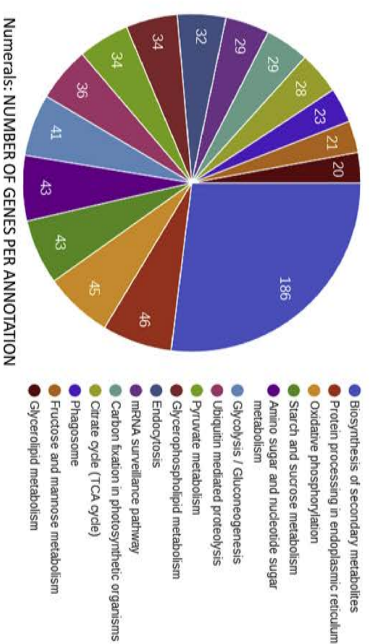


**Figura 4.20: Distribución de términos GO de las proteínas potencialmente S-nitrosilables del transcriptoma de polen obtenidas con iSNO-AAPair.** Se representa la distribución de procesos biológicos, funciones moleculares, componentes celulares y rutas KEGG. Se indica el número de genes correspondiente a cada categoría. Los gráficos se han generado utilizando la herramienta web GeneCodis. Más detalles en el texto. Tomada de [160].

## ENRICHMENT ANALYSIS OF TYR-NITRATION SITES IN OLIVE POLLEN PROTEINS

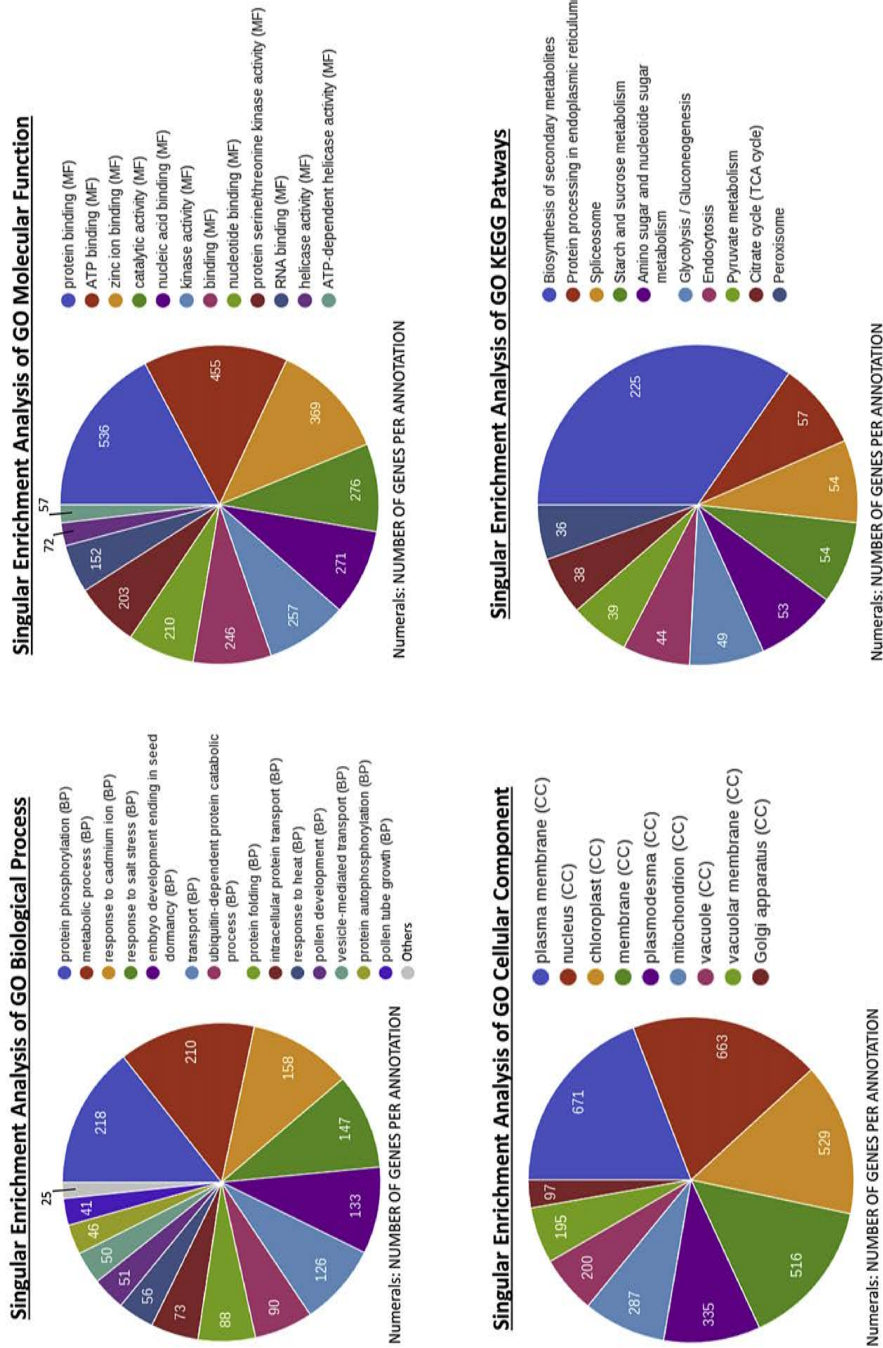


**Figura 4.21: Distribución de términos GO de las proteínas potencialmente Tyr-nitrables del transcriptoma de polen obtenidas con GPS-YNO2 (umbral medio).** Se representa la distribución de procesos biológicos, funciones moleculares, componentes celulares y rutas KEGG. Se indica el número de genes correspondiente a cada categoría. Los gráficos se han generado utilizando la herramienta web GeneCodis. Más detalles en el texto. Tomada de [160].

#### 4.5.5. Comparación entre resultados predictivos y validación experimental

Los resultados obtenidos en este análisis *in silico* han sido posteriormente complementados con una validación experimental inicial por parte de María José Jiménez-Quesada [160]. Para ello, cuantificó las proteínas de extractos de polen maduro y polen germinado *in vitro* mediante nano-LC-MS (que combina cromatografía líquida con espectrometría de masas), tras inmunoprecipitación con anticuerpos anti-*S*-nitrosoCys y anti-nitrotirosina. Esto le permitió identificar numerosas proteínas sometidas a uno o ambos tipos de modificación. Así, tanto el *S*-nitrosoproteoma como el Tyr-nitrosoproteoma de polen de olivo determinados experimentalmente pueden consultarse respectivamente en las tablas 6<sup>19</sup> y 7<sup>20</sup> de [399]. En ellas aparece información sobre los péptidos identificados, detalles de anotación sobre la proteína a la que corresponde, su identificador en ReprOlive y si la modificación de la proteína ha sido predicha o no *in silico* por alguna de las herramientas, tanto de *S*-nitrosilación, como de Tyr-nitración. En el caso del *S*-nitrosoproteoma determinado experimentalmente se identificaron 812 péptidos correspondientes a 491 proteínas, que representan 417 proteínas diferentes (ortólogos únicos), mientras que en la obtención del Tyr-nitrosoproteoma se identificaron 633 péptidos, pertenecientes a 345 proteínas, que representan 292 proteínas diferentes. Además al comparar los resultados experimentales se ha visto que 187 proteínas (165 únicas) presentan ambos tipos de PTM.

Para cada tipo de PTM hemos cruzado los listados de proteínas potencialmente modificables según la predicción *in silico* con las listas de proteínas cuya modificación se ha determinado experimentalmente. En el caso de la *S*-nitrosilación (figura 4.22), de las 491 proteínas determinadas experimentalmente, 423 (86,2 %) fueron predichas por al menos una de las tres herramientas (iSNO-AAPair, GPS-SNO y SNOSite) y 154 (31,4 %) por todas las herramientas simultáneamente.

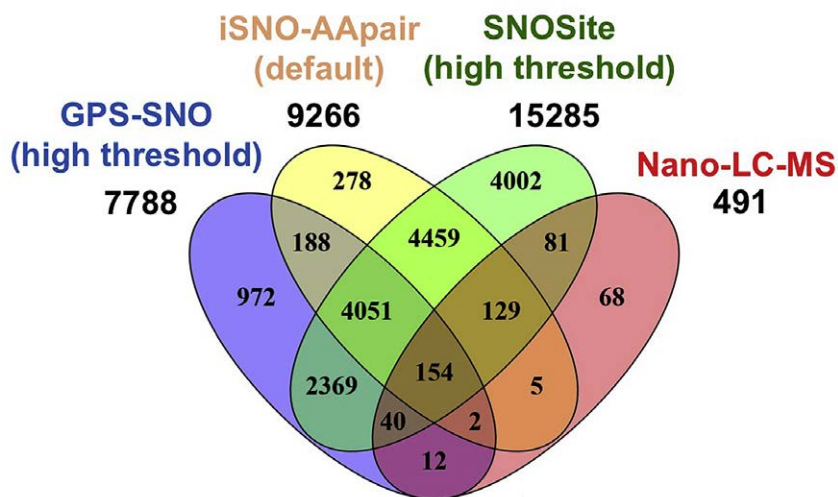


**Figura 4.22:** Distribución de los identificadores de los transcritos o productos génicos propensos a *S*-nitrosilación en función de la identificación *in silico* por las herramientas iSNO-AAPair, GPS-SNO (umbral alto) y SNOSite (umbral alto), y de la validación experimental (Nano-LC-MS).

Por otro lado, en el caso de la Tyr-nitración (figura 4.23), de las 345 proteínas determinadas experimentalmente, 272 (78,8 %) fueron predichas por el programa GPS-YN02 en alguno de los dos umbrales

<sup>19</sup><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647517/bin/mmc7.xlsx>

<sup>20</sup><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647517/bin/mmc8.xlsx>

utilizados (alto y medio), y 232 (67,2 %) por ambos umbrales simultáneamente.

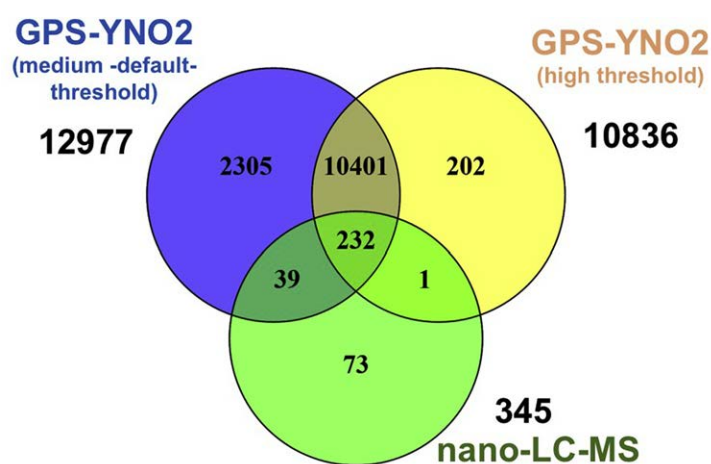


**Figura 4.23:** Distribución de los identificadores de los transcritos o productos génicos propensos a Tyr-nitración en función de la identificación *in silico* por la herramienta GPS-YN02 (umbrales medio y alto), o de la validación experimental (Nano-LC-MS).



## Capítulo 5

# DISCUSIÓN

### 5.1. Transcriptoma reproductivo del olivo

Los tejidos reproductivos presentan en general características especiales referentes a su accesibilidad (e.g. meiocitos masculinos y femeninos englobados en otros tejidos), presencia de componentes y metabolitos secundarios (e.g. exina en el caso del polen) y otros factores que dificultan el aislamiento y extracción de RNA. Varios estudios pioneros sobre la arquitectura de los transcriptomas reproductivos [272, 273], muestran que los transcriptomas reproductivos son sustancialmente diferentes de los correspondientes a las partes vegetativas, ya que presentan una alta proporción de transcritos específicos: entre un 4 % y un 11 % del número total de genes expresados [273]. Además, existen grandes diferencias en términos de expresión temporal, tanto entre estadios de desarrollo distintos, como a nivel de localización espacio-temporal, y de hecho se ha visto que los cambios más notables ocurren en el polen maduro, tras la hidratación y la germinación [400, 401].

Por otra parte, en las especies de plantas no modelo ha sido muy habitual llevar a cabo aproximaciones transcriptómicas [53] para obtener información biológica. En el momento en que se inicia esta investigación, el genoma del olivo aún no estaba publicado, y los estudios transcriptómicos en el olivo realizados hasta entonces se habían centrado principalmente en el fruto y las partes vegetativas (véase la tabla 1.3), dejando de lado las partes reproductivas de la planta.

Todos estos motivos unidos al notable desarrollo y abaratamiento de las tecnologías y plataformas de secuenciación, llevaron a nuestro grupo a abordar la construcción del transcriptoma reproductivo del olivo, ReprOlive, a partir de polen y pistilo en distintas fases de desarrollo, de forma que los distintos genes que se expresan en los diferentes órganos y en cada estadio estuviesen representados. Los distintos transcriptomas se han construido utilizando un flujo de trabajo automático fácilmente reproducible y reutilizable.

Como cabía esperar, polen y pistilo presentan TT comunes, lo cual se deduce de: (1) que el número de TT en el transcriptoma reproductivo es inferior a la suma de los TT de polen y pistilo (tabla 4.1); (2) los TT de los transcriptomas reproductivo y de pistilo están implicados en un total de 146 rutas, mientras que los de polen participan solo en 143; y (3) los datos de expresión indican que hay TT que se expresan tanto en el polen como en el pistilo (apartado 4.1.2 y figura 4.4). Un 87,2 % del total de TT de ReprOlive se anotaron funcionalmente (fila “Number of TTs with annotation” de la tabla 4.1), aunque no todos ellos tienen ortólogo. Además, el número de ortólogos únicos (fila “Unique IDs” de la tabla 4.1)

sugiere que el transcriptoma de pistilo parece ser bastante más complejo que el de polen.

Del total de TT del transcriptoma reproductivo, 17 445 no se consiguieron anotar (fila “Number of TTs without ortholog” de la tabla 4.1). De estos TT de función desconocida el análisis con TransDecoder contenido en Full-LengtherNext reveló que 7 114 TT podrían codificar alguna proteína, lo cual sugiere que los 10 331 restantes se deberían descartar del transcriptoma reproductivo. De los TT probablemente codificantes, 628 codifican una proteína completa y los restantes 6 486 una probablemente incompleta. Tras la anotación funcional con Sma3s, 80 (12,7 %) de los TT probablemente completos y 1750 (27 %) de los probablemente incompletos siguieron sin anotación, mientras que otros 134 (21,3 %) y 1 612 (24,8 %) respectivamente se anotaron como proteínas no caracterizadas. Es probable que estos subconjuntos de TT codificantes y sin ortólogo incluyan TT específicos de olivo, por lo que la posibilidad de descifrar su función abre nuevas oportunidades de investigación.

Tras la publicación de los primeros borradores del genoma del olivo ya se conoce el número total aproximado de genes de esta planta (56 349 en la variedad cultivada Farga [135] y 50 684 en el olivo silvestre [201]), por lo que nuestro transcriptoma reproductivo, si atendemos al número de TT con ortólogo único (fila “With unique ortholog IDs” de la tabla 4.1), representa un subconjunto de aproximadamente un 42 % a un 47 % del transcriptoma completo, es decir, en los tejidos reproductivos se expresa casi la mitad de los genes del olivo.

El tamaño de los transcriptomas obtenidos está sobrestimando el número total de genes del genoma de olivo. Esto puede deberse a la presencia de alelos, parálogos, secuencias fragmentadas, isoformas, *splicing* alternativo, mRNA inmaduros o una combinación de ellos. Para simplificar esta redundancia se construyeron transcriptomas de referencia, muy útiles para llevar a cabo estudios de expresión (véase apartado 3.4.4) como los que se incluyeron en la base de datos. De hecho, se ha visto que la tasa de mapeo de lecturas en análisis de expresión por RNA-seq es mayor cuando se usa el transcriptoma representativo que cuando se usa el transcriptoma completo [402]. El número de identificadores de RefSeq y TAIR10 únicos en el transcriptoma reproductivo completo (17 612 y 15 503, respectivamente) y en el transcriptoma reproductivo de referencia (14 706 y 13 584, respectivamente) están bastante próximos, lo que sugiere que, aunque falten algunos genes, el transcriptoma de referencia es representativo de los genes expresados en los tejidos reproductivos. Además, el número de identificadores únicos TAIR10/RefSeq es aproximadamente un 65 % del total de identificadores TAIR10/RefSeq, lo que parece indicar que en los transcriptomas de referencia hay presentes dos alelos por cada locus (filas debajo de “Reference transcriptome” en la tabla 4.1).

La secuenciación y el ensamblaje fueron bastante buenos y fiables, ya que el número de quimeras es muy bajo y el número de ORF completos obtenidos asciende a un 12,6 % del transcriptoma total, superando a otros transcriptomas contruidos con la misma estrategia [56, 54]. La disponibilidad de un gran número de secuencias proteicas completas resulta crucial para: una anotación precisa, realizar análisis comparativos entre transcriptomas y obtener perfiles de expresión génica precisos relacionados por ejemplo con el crecimiento, desarrollo y cambios ambientales. De hecho, una importante colección de estos ORF completos (como numerosas formas de los alérgenos *Ole e 1* y *Ole e 2*, Cu-Zn-superóxido dismutasas [150], catalasas, peroxidasas, NADPH oxidasas [403], enzimas del ciclo glutatión-ascorbato [404] y tioredoxinas) han sido clonados por nuestro grupo basándose en las secuencias contenidas en la base de datos. Adicionalmente, a partir del transcriptoma de polen contenido en ReprOlive se ha llevado a cabo un análisis sobre el metabolismo de los lípidos durante el desarrollo del grano de polen [405]. También hemos llevado a cabo en semilla de olivo una aproximación muy similar a la de ReprOlive, que ya nos está permitiendo caracterizar mejor algunas de sus proteínas [406, 153]. Todo esto confirma la fiabilidad y utilidad de ReprOlive.

Por eso, la publicación de ReprOlive supone una fuente de información complementaria a todo lo

que ya se ha publicado sobre el olivo, especialmente útil para investigadores dedicados a la reproducción en plantas. Ofrece información transcriptómica relacionada con los tejidos reproductivos del olivo (con hoja y raíz como tejidos vegetativos control) con acceso público sin restricciones. Contiene abundante información sobre TT que puede ser usada para genómica, estudios moleculares, mapas genéticos, análisis de expresión, detección de nuevos alérgenos, estudios para la mejora de la producción, entre otras muchas aplicaciones. Así mismo, los TT descritos en ReprOlive pueden ser muy útiles para corroborar y complementar las anotaciones de los borradores del genoma de olivo publicados recientemente, como ya se ha hecho con algunos de ellos en [213]. También como ejemplo de herramienta específica y complementaria, Alagna et al. [140] recomiendan el uso de ReprOlive en combinación con la secuencia del genoma como una valiosa herramienta para la identificación de genes implicados en el complejo sistema de autoincompatibilidad de esta planta. Una mejora interesante a tener en cuenta de nuestra base de datos sería incorporar la correspondencia entre los TT identificados y su gen en el genoma.

En ReprOlive hemos optado por un diseño que permite al usuario acceder a la descarga de la información para facilitar su análisis con herramientas externas, en vez de embeber herramientas bioinformáticas voluminosas en la propia base de datos. Además, ReprOlive ha sido publicada en un formato de conceptualización semántica estándar en RDF que posibilita su integración con otras bases de datos también basadas en RDF. De esta forma, la información estática que contiene ReprOlive puede explotarse de manera más dinámica, puesto que muchas de las bases de datos semánticas con las que está vinculada están siendo constantemente actualizadas. De hecho, una mejora sustancial de ReprOlive conllevaría la conexión con nuevas bases de datos que compartan esta estructura.

ReprOlive ha sido construida de forma que permite la importación automática de nuevos conjuntos de datos, así, en un futuro próximo se prevé añadir secuencias de otros tipos de células reproductivas, tejidos y órganos de interés (como meiocitos aislados, tapetum, endospermo, mesocarpo, ovario y saco embrionario), de estadios de desarrollo adicionales e incluso datos de RNA-seq. Dado que la tecnología 454 hace ya tiempo que quedó obsoleta, estas nuevas secuencias se han obtenido con lecturas pareadas de Illumina, lo cual, además de permitir análisis de expresión por RNA-seq ayudará entre otras cosas a identificar mejor las distintas isoformas de los transcritos y las variantes estructurales. Todo ello completará la información y el conocimiento que se tiene actualmente sobre el transcriptoma reproductivo del olivo. Por lo tanto creemos que ReprOlive (tanto en su versión actual como en desarrollos futuros), seguirá siendo una herramienta muy útil que permitirá a los investigadores seguir ampliando el conocimiento acerca de la transcriptómica de esta planta.

## 5.2. Transcriptoma mejorado de polen maduro

Dependiendo de la estrategia de secuenciación o del programa de ensamblaje elegido se pueden obtener diferentes transcriptomas. Cabe esperar que mediante la combinación de distintos tipos de lectura y ensambladores sea posible obtener un mejor transcriptoma que usando un solo tipo de ellos, ya que permite aprovechar los puntos fuertes de cada uno, si bien esta estrategia requiere algún tipo de validación. Cuando el genoma de referencia está disponible, dicha validación es fácilmente abordable, aunque en las especies no modelo, o bien no existe genoma con el cual comparar, o bien la variedad a la que corresponde dista mucho de la nuestra. En estos casos se suele optar por una validación empírica, o se calculan parámetros como la precisión, la completitud o la contigüidad para evaluar la calidad de un transcriptoma *de novo* [55, 56]. La herramienta TransFlow (apartado 3.2.7) surge con el fin de abordar esta forma de proceder y es capaz de combinar distintos tipos de lecturas y ensambladores, generando múltiples ensamblajes que se evaluarán y puntuarán según una serie de parámetros de calidad que se compararán con transcriptomas públicos de gran fiabilidad, para establecer un *ranking* y seleccionar cuál

de ellos es el mejor ensamblaje final.

El desarrollo de la investigación que ha dado lugar a la presente tesis, coincidió con la publicación de los dos primeros borradores del genoma de ‘Farga’ [135] y el acebuche [201], y por supuesto es anterior al genoma de ‘Picual’ [129]. A pesar de ello, esta planta se sigue considerando una especie no modelo, puesto que estos genomas corresponden a un ejemplar de la variedad Farga, de mas de 1000 años de antigüedad de escaso interés agronómico, el olivo silvestre o acebuche, que ya presentan bastantes diferencias entre ellas en el porcentaje de secuencias repetitivas y número de genes (J.D. Alché y M.G. Claros, comunicación personal), y seguramente con ‘Picual’, la variedad más importante en España y Andalucía. Por este motivo, y aprovechando una secuenciación RNA-seq de lecturas pareadas de Illumina ( $2 \times 75$  nt) de polen maduro de olivo llevada a cabo en nuestro laboratorio, se decidió utilizar TransFlow para intentar mejorar el transcriptoma de polen que se obtuvo en ReprOlive.

El mejor ensamblaje obtenido con TransFlow (scOases\_cat\_cd) involucró exclusivamente lecturas de Illumina, mientras que el ensamblaje resultante de la estrategia equivalente a la seguida en ReprOlive (ctMIRA\_ctEulK29\_rcCAP3) quedó relegado al puesto 64 del *ranking*. Esto parece indicar que el nuevo transcriptoma de polen ha mejorado con respecto a ReprOlive. De hecho, el número de proteínas diferentes anotadas en scOases\_cat\_cd (6717) es significativamente más alto que para ctMIRA\_ctEulK29\_rcCAP3 (1899), lo cual refleja un transcriptoma mucho más completo (tabla 4.4 y figura 4.9 D). Según estos datos, con las lecturas de 454/Roche solo conseguimos un 28 % del transcriptoma obtenido con Illumina. Con respecto al número total de genes estimado para esta planta a partir del genoma (56349 en la variedad cultivada Farga [135] y 50684 en el olivo silvestre [201]), el mejor transcriptoma de polen maduro según TransFlow representa un subconjunto de aproximadamente un 12 % a un 13 % del transcriptoma completo, frente al transcriptoma equivalente a ReprOlive, que solo contiene entre un 3 % y un 4 %.

Los valores de los parámetros más significativos según la primera dimensión del PCA: ComplOrtho, Contigs500 y DiffComplProts (tabla 4.3) fueron notablemente más altos para el ensamblaje scOases\_cat\_cd que en el ensamblaje ctMIRA\_ctEulK29\_rcCAP3, lo cual es también indicativo de que el primero es un transcriptoma mucho más completo (tabla 4.4 y figuras 4.9 A, C y D).

A pesar de que el ensamblaje scOases\_cat\_cd conlleva una etapa de *scaffolding*, la longitud media de *contig/scaffold* (760 nucleótidos) no consigue superar la obtenida en el ensamblaje ctMIRA\_ctEulK29\_rcCAP3 (840 nucleótidos) (tabla 4.4 y figura 4.9 B). Además, el valor de N90 es algo mayor en el ensamblaje equivalente a ReprOlive, lo que indica que los *contigs* generados son en general más largos que los del mejor ensamblaje según TransFlow (figura 4.9 B). No obstante, el uso de lecturas pareadas es muy útil a la hora de detectar isoformas de un transcrito [36], variantes estructurales [37] o fusiones génicas quiméricas [38]. De hecho, el porcentaje de proteínas anotadas mal ensambladas (*misassembled*), aunque es bajo en general, es destacadamente más alto en ctMIRA\_ctEulK29\_rcCAP3 que en scOases\_cat\_cd, lo cual indica que la estrategia seguida en ReprOlive produjo mayor número de artefactos (figura 4.9 F).

A pesar de que TransFlow partió de lecturas 454/Roche e Illumina, el mejor ensamblaje de los 181 generados solo involucró lecturas de Illumina, mientras que los que utilizaban lecturas 454/Roche recibieron peor puntuación. De aquí se deduce que en este caso las lecturas pareadas de Illumina son suficientes para proporcionar el mejor ensamblaje, mientras que las lecturas 454/Roche parecen ser complementarias. Por el contrario, se ha visto que la contribución de las lecturas más largas es significativa cuando se utilizan lecturas cortas y simples (*single*) de Illumina en vez de pareadas, como se ha podido comprobar en el ensamblaje también con TransFlow del transcriptoma del castaño (*Castanea sativa*) [126]. Esto demuestra que datos de partida diferentes requerirán distintas estrategias de ensamblaje.

El ensamblaje equivalente a ReprOlive quedó bastante alejado del mejor transcriptoma obtenido con

TransFlow en el proceso de clusterización basado en distancias medias a los transcriptomas de referencia de *Arabidopsis* y *Populus*. No obstante, es el primer ensamblaje en el *ranking* que involucra solo lecturas 454/Roche, lo cual indica que la estrategia que se siguió en ReprOlive es la mejor cuando únicamente se dispone de este tipo de lecturas.

En cuanto a los programas de ensamblaje utilizados por TransFlow en la estrategia seguida para polen de olivo se ha visto que OASES proporciona los mejores transcriptomas en comparación con el resto de ensambladores, mientras que RAY ofrece los peores ensamblajes (véase el apartado 4.2.1). Esta observación coincide con lo que ocurre en otros casos de estudio analizados con TransFlow (por ejemplo para vid y castaño [126]).

Está claro que TransFlow nos ha permitido obtener una versión sustancialmente mejorada del transcriptoma de polen maduro, más completa y con menos artefactos, que será incorporada en breve a la base de datos ReprOlive y que resulta mucho más apropiada para ser usada como transcriptoma de referencia para análisis de expresión de RNA-seq, por ejemplo. De hecho, en nuestro laboratorio se llevarán a cabo en breve análisis de expresión por RNA-seq en el polen que usarán este transcriptoma mejorado como transcriptoma de referencia, salvo que demostremos que el deducido del genoma de ‘Picual’ [129] da mejores resultados.

No obstante, los valores de los parámetros de evaluación para el conjunto transcriptomas obtenido quedan todavía bastante lejos de los obtenidos para los transcriptomas de referencia. Esto sugiere que el transcriptoma generado, a pesar de mejorar bastante a ReprOlive, puede ser todavía optimizado. De este modo, queda abierta la posible incorporación de nuevos sets de lecturas, nuevas estrategias y programas de ensamblaje que puedan mejorar el transcriptoma final.

En vista de los resultados obtenidos en nuestro caso y como se ha visto para otras especies, podemos concluir que el uso de la herramienta TransFlow para el ensamblaje *de novo* de transcriptomas de especies no modelo es altamente recomendable. Cada caso particular, en función de la naturaleza de los datos brutos de partida, tiene su propio comportamiento y pondera de forma diferente los parámetros de evaluación, por lo que puede requerir aproximaciones diferentes de ensamblaje [126]. TransFlow aporta una estrategia que se adapta a cada escenario y supone una forma objetiva y automatizada (por tanto fácilmente reproducible) para detectar los mejores ensamblajes transcriptómicos dentro de un conjunto de ellos.

TransFlow sirve también para evaluar cómo se comportan los distintos programas de ensamblaje de transcriptomas en función del tipo, cantidad y calidad de las lecturas de partida. Además tiene la ventaja de ser una herramienta flexible y adaptable, ya que permite incorporar nuevos programas de ensamblaje al flujo de trabajo, así como adaptarlo para lecturas procedentes de nuevas plataformas de secuenciación (como Nanopore o las procedentes de secuenciación de molécula única) y deja con ello el camino abierto a la obtención de un transcriptoma cada vez más óptimo.

### 5.3. Idoneidad del algoritmo de detección de genes de referencia

Tradicionalmente, los estudios de expresión génica se han llevado a cabo mediante análisis de expresión no cuantitativa o semicuantitativa en geles de RNA, y posteriormente mediante análisis de reacción en cadena de la polimerasa con transcripción inversa (RT-PCR). El posterior desarrollo de la PCR cuantitativa (qPCR) en tiempo real [407] ha ido sustituyendo estas técnicas debido a su mayor especificidad, sensibilidad e intervalo dinámico de identificación. Este aumento de la sensibilidad de la qPCR con respecto a los geles tradicionales y la RT-PCR supuso una mayor preocupación sobre la necesidad de genes de referencia apropiados para evitar resultados erróneos y proporcionar una interpretación adecuada de

los datos de expresión génica. Los primeros RG se obtuvieron a partir de ensayos de hibridación *Northern* y correspondían normalmente a proteínas con funciones estructurales y del metabolismo básico de la célula, debido a la teórica invariabilidad de su expresión en la mayoría de los tejidos (son los denominados genes *housekeeping*). Sin embargo, en muchos casos se ha visto que esta primera elección no es la más adecuada [278, 279, 280] y se han dedicado numerosos estudios a la búsqueda de RG más apropiados [408, 409, 410, 411].

Las conclusiones de cualquier experimento de qPCR dependen de la elección tanto de unos RG apropiados como de un método de normalización adecuado. La selección de ambos es crítica para obtener medidas de expresión génica cuantitativa fiables, que permitan corregir la variación inespecífica debida por ejemplo a diferencias en la cantidad o calidad de RNA. El método de normalización más extendido es la cuantificación relativa, donde el nivel de expresión de un gen se normaliza en función de un RG interno, cuya expresión debería permanecer constante en todas las condiciones experimentales que se estén evaluando. Los métodos más populares para la normalización y confirmación de RG son BestKeeper [412], geNorm [413] y NormFinder [414]. El método geNorm calcula el número mínimo de RG para cada experimento en función de cantidades brutas relativas. NormFinder proporciona además un valor de estabilidad para cada gen. BestKeeper realiza un análisis de correlación por pares basado en una media geométrica para determinar RG óptimos. Pero todos estos métodos están basados en datos de qPCR, lo cual supone un problema recursivo, ya que se necesita llevar a cabo una qPCR para decidir si un RG es apropiado para qPCR. Además se ha visto que la mayoría de los genes *housekeeping* más comúnmente usados como RG realmente presentan variabilidad en los datos de *microarrays* o bajo condiciones experimentales diferentes [415, 416]. Por lo tanto, cuando comparemos diferentes estadios de desarrollo, tejidos, tipos celulares o condiciones ambientales, la elección de los mejores RG debería ser minuciosamente validada [288] y basarse en pruebas experimentales anteriores.

En la bibliografía encontramos estudios que han utilizado repositorios públicos de datos de *microarrays* para la búsqueda masiva de RG estables, con resultados satisfactorios [410, 285]. Pero por lo general estos datos solo están disponibles para determinadas especies, normalmente organismos modelo, y para condiciones experimentales concretas. En el caso de especies no modelo, como el olivo o para condiciones experimentales inusuales hay que recurrir a otras estrategias para la búsqueda de RG. Así, en el caso del olivo se han publicado varios estudios con el objetivo de identificar los RG más apropiados [207, 57], que evalúan genes de olivo ortólogos de RG bien caracterizados en otras plantas y seleccionan los que son estables en los tejidos del olivo. Estos estudios se han llevado a cabo para diferentes condiciones experimentales: estadios de desarrollo diferentes del mesocarpio en distintas variedades de olivo [417], y varios estadios de desarrollo y maduración del fruto y daños foliares [291, 293]. En nuestro caso además, las peculiaridades de los tejidos reproductivos de las plantas complican la búsqueda de RG, pues se ha visto que muchos de los genes *housekeeping* tradicionales presentan diferencias de expresión en el polen, el pistilo y otros órganos florales de la planta [285]. Sin embargo, otros análisis indican que una gran proporción de los genes constitutivos son compartidos por la mayoría de tejidos reproductivos, somáticos y haploides [418]. Ante esta problemática, creemos que los RG para la validación por qPCR deberían inferirse de los propios datos experimentales que se van a analizar.

Con este objetivo hemos diseñado un flujo de trabajo automático que permite obtener RG altamente e invariablemente expresados en unas condiciones experimentales concretas [419, 266]. El método se basa en el cálculo del coeficiente de variación (CV) de valores de expresión normalizados por RPMM y permite filtrar los resultados en función de estos dos parámetros.

Los sets de datos elegidos para el diseño y confirmación de este flujo de trabajo fueron intencionadamente muy heterogéneos, tanto en tecnología de secuenciación, como en tamaño de librería y lectura, con el fin de poner a prueba la herramienta en distintos escenarios. Como puede observarse en la tabla 3.7, el porcentaje de lecturas útiles fue bastante homogéneo tanto en el caso de lecturas 454/Roche (>51 %,

con una media de 54 %) como en el caso de lecturas de Illumina (>96 %, con una media de 97 %), lo cual refleja un preprocesamiento adecuado. El porcentaje de lecturas mapeadas frente a lecturas útiles es también homogéneo dentro de cada tecnología (>45 %, con una media de 57 % en el caso de 454/Roche, y >69 %, con una media de 85 % para Illumina). La heterogeneidad en el tamaño de las réplicas o librerías también es regular dentro de cada experimento: en el olivo, las lecturas brutas oscilan entre 217 163 y 262 749 (1,2 veces) y las lecturas útiles van desde 111 760 hasta 150 185 (1,3 veces), mientras que en *Arabidopsis* las lecturas brutas oscilan entre 9 107 610 y 13 076 233 (1,4 veces) y las lecturas útiles van desde 8 859 088 hasta 12 678 437 (1,4 veces). Sin embargo, puede darse el caso de situaciones experimentales con muestras muy heterogéneas, y por ello nuestra herramienta incluye una normalización de la abundancia de RNA dentro de las muestras, para eliminar el sesgo debido a la profundidad de secuenciación de la muestra.

En cuanto a los métodos de normalización de los recuentos de lecturas, existen dos de uso muy extendido: RPKM (lecturas por kilobase de transcrito, por millón de lecturas mapeadas, véase la ecuación 3.1) para lecturas simples, y el FPKM (fragmentos por kilobase de transcrito, por millón de lecturas mapeadas) para lecturas pareadas [420], a pesar de que se ha visto que son inestables para las comparaciones dentro de la misma muestra [277]. Ambos métodos normalizan por longitud del transcrito, lo cual es necesario cuando se llevan a cabo comparaciones entre transcritos distintos. Sin embargo, en el caso que aquí nos ocupa, se analiza la expresión de cada gen de forma individual, mediante la comparación de la expresión de ese gen entre las distintas muestras. Como en ningún momento comparamos genes o transcritos entre sí, no es necesaria esta normalización por longitud. Además, en los organismos no modelo como el olivo, es posible que la longitud de sus transcritos no esté aún bien determinada, y solo se hayan conseguido reconstruir fragmentos del mismo, o bien puede que el transcrito esté dividido en varios *contigs* independientes. Por estos motivos decidimos utilizar como método de normalización en nuestro algoritmo el RPMM (lecturas por millón de lecturas mapeadas; fórmula 3.2), una simplificación del método RPKM, que es independiente de la longitud del transcrito.

El flujo de trabajo que presentamos aquí aprovecha los datos de NGS públicamente disponibles, cada vez menos costosos y más abundantes, para predecir RG apropiados para cada situación experimental. Estos candidatos a RG pueden ser útiles en validaciones mediante qPCR en análisis de expresión. Nuestro análisis resulta muy útil en el caso de especies no modelo, en las que se han identificado muy pocos o ningún RG, o en condiciones experimentales nuevas, para las que no existen datos disponibles.

El flujo de trabajo parece ser independiente de la especie, la tecnología de secuenciación que genera las lecturas, del número de lecturas y de la longitud de las mismas, ya que parece ser igual de eficiente cuando trabaja con muchas lecturas cortas (como hemos visto con Illumina en *Arabidopsis*), que con pocas lecturas largas (454/Roche en el olivo) y del número de condiciones experimentales. Además, los tiempos de ejecución para las diferentes tareas del flujo son razonablemente cortos, teniendo en cuenta que las partes que consumen más tiempo (preprocesamiento y mapeo) se necesitan en cualquier análisis de NGS y se llevan a cabo una sola vez. Por ello opinamos que nuestra herramienta puede ser utilizada prácticamente en cualquier organismo y diseño experimental (si bien los tiempos de ejecución podrían alargarse en el caso de lecturas más largas o transcriptomas más grandes) y podría ofrecerse como una herramienta web.

Nuestro flujo es personalizable y adaptable a los requerimientos de cada caso experimental particular. Además, al estar basado en AutoFlow permite utilizar y acoplar al flujo otras estrategias y programas de preprocesamiento o mapeo. El algoritmo de *findRGs* se ha utilizado con éxito en diferentes tejidos de tres especies: olivo y *Arabidopsis*, como se ha descrito en este trabajo, y también en humano (resultados no mostrados en esta tesis, pero incluidos en el artículo en el que se basa este capítulo), en este último caso con lecturas pareadas de Illumina correspondientes a tres tipos diferentes de cánceres, de próstata y de pulmón, un escenario con un volumen mucho mayor de lecturas, un número de muestras superior

y un transcriptoma bastante más grande [266]. Se trata de estudios comparativos enfocados a aspectos biológicos muy diversos, que incluyen diferentes estadios de desarrollo, fisiológicos y patológicos. En todos los casos se han generado listas de candidatos a RG, algunos de los cuales ya se habían descrito en la bibliografía, y otros se han validado preliminarmente a raíz de este trabajo, y ambas circunstancias dan peso a nuestra aproximación. Pensamos que es más destacable aún la posibilidad de descubrir RG nuevos y más apropiados. A pesar de que los niveles de expresión y la estabilidad de estos RG nuevos podría requerir alguna validación experimental antes de su uso para normalización, recomendamos la utilización de *findRGs* donde sea posible, ya que puede ser bastante útil como una aproximación preliminar sobre los mejores candidatos a RG, antes de cada análisis de expresión.

#### 5.4. El análisis bioinformático ha profundizado en la heterogeneidad alergénica del polen de olivo

El uso del genoma, el transcriptoma o estrategias basadas en el proteoma para examinar el repertorio alergénico de especies de plantas es una práctica habitual [362, 421, 422, 423, 424]. En nuestro área, el polen del olivo es uno de los principales responsables de los casos de sensibilización por polen, por lo que se han llevado a cabo numerosos esfuerzos para caracterizar el alergograma de esta planta [425, 307, 238, 426, 166]. A pesar de ello, se sabe que todavía quedan alérgenos por identificar y con este objetivo hemos puesto a punto en nuestro laboratorio una estrategia basada en transcriptómica que nos ha permitido explorar masivamente el transcriptoma de polen de ReprOlive, en la búsqueda de posibles alérgenos nuevos, además de indagar sobre la reactividad cruzada de esta planta con otras especies.

##### Alérgenos de polen de olivo ya conocidos

En el momento en el que se realizó el análisis, se habían identificado 13 alérgenos de olivo que se utilizaron para analizar el transcriptoma de polen de ReprOlive (apartado 4.1.1). Se consiguieron identificar transcritos completos correspondientes a todos los alérgenos descritos hasta la fecha, a excepción de *Ole e 4*, *Ole e 7* y *Ole e 13* (tabla 4.8). La ausencia de *Ole e 13* queda justificada, dado que es un alérgeno presente en el fruto, y no en el polen. Por su parte, *Ole e 4* [178] se describió inicialmente como una secuencia consecutiva de 24 aminoácidos. Posteriormente se vio que una sección de 10 y otra de 14 aminoácidos de esta secuencia presentan un 100 % de identidad con dos fragmentos no adyacentes de *Ole e 9*, por lo que *Ole e 4* podría ser un producto de degradación de *Ole e 9* [166]. De hecho, el transcrito po11\_olive\_002121 identificado como *Ole e 9* en ReprOlive (tabla 4.8) también produjo un alineamiento con la sección de 14 aminoácidos de *Ole e 4*, aunque fue descartado por no cumplir el filtro de cobertura de alineamiento con el alérgeno. Esto explicaría la ausencia de *Ole e 4* en ReprOlive.

En cuanto a *Ole e 7*, fue descrito inicialmente como un fragmento de 21 aminoácidos de una proteína de transferencia de lípidos no específica (nsLTP) [183] que corresponde a los transcritos OE6A005344T1 y OE6A031005T1 deducidos del primer borrador del genoma de esta planta [135]. Esta secuencia parcial de *Ole e 7* no fue identificada en nuestro transcriptoma (salvo identidades ligeramente inferiores a un 70 %, sobre fragmentos de 13-15 aminoácidos (véase figura 4.14A), lo que supone un reto a la fiabilidad del sistema de predicción e identificación utilizado. Se trata de una proteína altamente polimórfica, con un mRNA bastante inestable [183], lo cual dificulta su detección. Para descartar una falta de sensibilidad de la secuenciación con 454/Roche o de completitud del transcriptoma de ReprOlive, evaluamos también, aunque sin éxito, la presencia de *Ole e 7* en el transcriptoma mejorado de polen maduro descrito en el apartado 4.2.1 (véase figura 4.14B). Posteriormente a la realización de estos ensayos, una aproximación proteómica basada en el uso de fragmentos de péptidos obtenidos por secuenciación múltiple *de novo*

de péptidos mediante espectrometría de masas, ha permitido el ensamblaje de la secuencia completa de este alérgeno [184]. En este caso, y utilizando esta nueva secuencia (95 aminoácidos), las identidades con ReprOlive o con el transcriptoma mejorado de polen maduro disminuyen en su valor medio (40/42 %, figuras 4.15A y 4.15B, respectivamente), aunque en este caso cubren un área significativamente mayor de la secuencia (hasta 35/49 aminoácidos, figuras 4.15A y 4.15B, respectivamente). La reciente accesibilidad a diferentes genomas de olivo [135, 201, 129] nos ha permitido realizar nuevos análisis sobre la representatividad de las secuencias de *Ole e 7* en los transcriptomas y proteomas deducidos de dichos genomas. En este caso sí que obtenemos una identidad prácticamente absoluta sobre la totalidad de la secuencia, al menos en los genomas de las variedades Farga [135] y Picual [129] (véanse figuras 4.16 y 4.18), mientras que se mantiene la identidad reducida detectada anteriormente cuando se comparó la secuencia de *Ole e 7* con el genoma del acebuche [201] (véase figura 4.17).

Estos relativamente bajos niveles de identidad detectados pueden atribuirse a distintas causas, que podrían ser incluso coincidentes, y que deben ser investigadas en detalle. Por una parte, la elevada variabilidad intervarietal mostrada por numerosos alérgenos podría encontrarse exacerbada en el caso de *Ole e 7*, de forma que resulte imposible localizar fácilmente sus secuencias. Nuestros estudios iniciales fueron realizados con la variedad Picual, mientras que la determinación de la secuencia completa de *Ole e 7* [184] se realizó en una variedad no determinada. Análisis posteriores realizados en transcriptomas de otras variedades como Arbequina, Manzanilla o Meski (resultados no mostrados) han fallado igualmente en la identificación de las secuencias de *Ole e 7*. En los genomas analizados, la identidad total en las variedades Farga y Picual contrasta con la baja identidad en oleaster. Todo ello son evidencias inequívocas de la existencia de elevada variabilidad intervarietal en este alérgeno. La mayor tasa de identificación de *Ole e 7* en transcriptomas y/o proteomas deducidos de genomas completos frente a los transcriptomas de polen analizados aquí, podría deberse igualmente al menos a dos causas diferentes: los transcriptomas podrían incluir dominios en sus secuencias muy parecidas a otros parálogos de olivo, capaces de generar errores en el ensamblaje *de novo* (e.g. quimeras) que dificultarían la identificación de *Ole e 7*. Por otra parte, aunque el alérgeno *Ole e 7* está presente en el polen maduro, sus transcritos podrían no estar representados en los transcriptomas de polen maduro de las variedades analizadas, si su expresión no tuviera lugar de forma significativa en dicho material. Es frecuente la expresión de productos génicos relevantes para la fisiología del polen (incluidos alérgenos) en tejidos adyacentes al polen como puede ser el tapetum [427, 428, 429], como ocurre con *Ole e 1* [170]. En el caso de las nsLTP, Liu et al. [430] revisan la localización ultraestructural de estas proteínas, que suele asociarse frecuentemente al espacio extracelular, apoplasto y pared celular, tratándose frecuentemente de proteínas secretadas. Desconocemos en la actualidad la localización subcelular de *Ole e 7* en el polen de olivo, y si su síntesis tiene lugar en el polen propiamente dicho o en otros tejidos de la antera. La localización celular de los transcritos de *Ole e 7* (por ejemplo, mediante hibridación *in situ*, o tras el análisis de transcritos en tejidos accesorios de la antera mediante extracción de RNA con microdissección láser) ayudaría a resolver estas cuestiones. En cualquier caso, los resultados obtenidos muestran que la combinación de la información transcriptómica mostrada en este trabajo, junto con la información proteómica y genómica disponible en algunos casos permite refinar el conocimiento específico sobre la fisiología de los productos génicos analizados. Además, en el caso del olivo, sería altamente deseable la generación de un pangenoma de la especie, así como ampliar la obtención de información varietal individualizada, con fines comparativos. A este respecto, la publicación del genoma de la variedad Picual [129] incluye la resecuenciación de 40 variedades cultivadas de olivo y 10 variedades silvestres, información que puede ser aprovechada para estos menesteres. Todo esto podría ayudar a diseñar estrategias agronómicas, adaptativas, alérgicas, diagnósticas, etc.

## Reactividad cruzada entre el polen de olivo y otras especies

Desde que comenzó a elucidarse el complejo alergograma del olivo, se han descrito numerosos casos de reactividad cruzada de esta planta con otras especies, algunas de ellas taxonómicamente relacionadas, y otras bastante alejadas, pero con las que comparten proteínas con funciones vitales básicas y evolutivamente muy conservadas (panalérgenos).

En nuestro trabajo hemos evaluado la posible reactividad cruzada de los alérgenos de polen de olivo encontrados en ReprOlive con alérgenos de especies diferentes (apartados 3.7.2 y 4.4.2). Muchos de los posibles casos de reactividad cruzada obtenidos en nuestra aproximación han sido ya previamente descritos (tabla 4.9), como por ejemplo aquellos con especies muy relacionadas como el fresno (*Fraxinus excelsior*) o el aligustre (*Ligustrum vulgare*), u otros con especies muy distantes, como el abedul (*Betula pendula*) o la herbácea *Phleum pratense*, protagonizados por los panalérgenos *Ole e 2* y *Ole e 3*.

También hemos detectado candidatos a casos de reactividad cruzada entre los alérgenos de polen de olivo y alérgenos presentes en algunos alimentos que ya han sido demostrados en el laboratorio. Tal es el caso de la reacción cruzada entre *Ole e 2* y los alérgenos *Cuc m 2*, *Pru p 4* y *Pyr c 4* de melón, melocotón y pera, respectivamente [328]. Todas estas reacciones cruzadas forman parte del denominado síndrome polen-fruta o síndrome de alergia oral. Muchas de estas reacciones están enmascaradas y son difíciles de detectar, debido a que los alérgenos de los alimentos son sensibles al calor durante el cocinado y a la acción de las proteasas durante el proceso digestivo [328]. Nuestros resultados sugieren otros alérgenos presentes en alimentos cuya posible reactividad cruzada con alérgenos de polen podría haber pasado desapercibida y cuyo estudio resultaría interesante.

Igualmente, existe un síndrome de alergia a polen-látex-fruta debido al gran parecido entre determinados panalérgenos que están presentes a la vez en el polen, alimentos vegetales y el látex. El olivo participa de este síndrome, como evidencia la reacción cruzada descrita entre las profilinas *Ole e 2* y *Hev b 8*, y entre las proteínas glucosiladas *Ole e 9*, *Ole e 10* y *Hev b 2*, alérgenos de *Hevea brasiliensis* (planta de la que se extrae el látex), detectadas en nuestro flujo y ya comprobadas en el laboratorio [349, 175, 328, 387, 392].

Cabe reseñar que la mayoría de alérgenos que presentan reactividad cruzada con *Ole e 9* la presentan también para *Ole e 10* [376, 188], lo cual se debe a que éstos comparten la parte de su secuencia correspondiente a los epítomos responsables. Lo mismo ocurre con las proteínas de unión a calcio *Ole e 3* y *Ole e 8*, que según nuestro algoritmo podrían compartir reactividad cruzada con *Cup a 4* (*Hesperocyparis arizonica*), aunque todavía no se ha demostrado. Estos tres alérgenos guardan similitud en las regiones de unión a calcio [382]. En este caso, también se ha demostrado que hay reactividad cruzada entre los propios *Ole e 3* y *Ole e 8* [376].

Hemos obtenido también posibles casos de reactividad cruzada con alérgenos procedentes de fuentes bastantes más alejadas, como son los ácaros. Así, la profilina *Ole e 2* podría reaccionar con la proteína de unión a actina *Tyr p 36* del ácaro del moho o del queso *Tyrophagus putrescentiae*, y con el alérgeno *Der f 36* del ácaro del polvo *Dermatophagoides farinae*, de función hasta el momento desconocida. De hecho, son frecuentes los casos de pacientes alérgicos al polen del olivo que lo son además a los ácaros del polvo y del moho. El análisis más exhaustivo de estos casos podría servir para caracterizar mejor estos casos de alergia con organismos tan distantes.

Para los alérgenos *Che a 1*, *Lol p 11*, *Phl p 11* y *Sal k 5*, aunque cumplen alguno de los requisitos fijados por la FAO/WHO para ser candidato a producir reactividad cruzada con *Ole e 1*, se ha demostrado que estos niveles de reactividad son bastante bajos o nulos (tabla 4.9) [306, 307, 313, 317], lo cual indica que estos criterios de homología son siempre orientativos y una vez que se obtiene un candidato es necesario centrar el estudio de esta homología en las regiones de los epítomos si se conocen. Los candidatos

deben ser finalmente cuestionados y valorados en el laboratorio para validar esta putativa alergia.

### Nuevos alérgenos de polen de olivo

Como resultado de la búsqueda de posibles alérgenos nuevos de polen de olivo llevada a cabo con nuestro algoritmo (figura 3.5) se obtuvieron numerosos candidatos pertenecientes a muy diversas fuentes (tabla 4.10). Entre los candidatos encontrados resultan de especial interés los alérgenos correspondientes a las familias alérgicas de las poligalacturonasas y ciclofilinas, puesto que recientemente han sido descritos dos nuevos alérgenos de olivo correspondientes a estas dos categorías funcionales, aún desconocidos cuando se llevó a cabo nuestro análisis: la poligalacturonasa *Ole e 14* [167] y la ciclofilina *Ole e 15* [197].

La obtención de transcritos de polen de olivo en ReprOlive con homología suficiente con alérgenos correspondientes a poligalacturonasas de polen de otras especies (*Cry j 2*, *Jun a 2*, *Lil l PG*, *Pas n 13*, *Phl p 13*, *Pla a 2*, *Pla or 2*, *Zea m 13*, *Bra n PG* y *Sor h 13*) supone la presencia de *Ole e 14* en nuestro transcriptoma y además sugiere una posible reactividad cruzada de este alérgeno con estas especies. De hecho, ya se ha demostrado que se trata de un alérgeno con amplia reactividad cruzada [167]. Del mismo modo encontramos transcritos de polen en ReprOlive que guardan similitud con ciclofilinas de alérgenos de otras especies de muy diversas fuentes: *Bet v 7* y *Cat r 1* en el polen, *Sola l 5* en el fruto y *Asp f 27*, *Mala s 6* y *Rhi o 2* en hongos y que indican la presencia de *Ole e 15* en nuestro transcriptoma. La evidente ubicuidad de este alérgeno lo señala como un posible panalérgeno, capaz de contribuir al denominado síndrome polen-fruta. Los numerosos casos de reactividad cruzada de este alérgeno con especies muy alejadas ya demostrados apoyan esta afirmación [197].

En su estudio San Segundo-Acosta et al. [197] siguen una estrategia similar a la nuestra, pero en lugar de partir de datos transcriptómicos, llevan a cabo una aproximación proteómica inversa: comparan el proteoma de polen de olivo y el proteoma deducido del genoma de acebuche [201] con la base de datos de Allergome para encontrar posibles nuevos alérgenos de polen. Obtuvieron candidatos correspondientes a 47 familias de alérgenos, de las cuales 10 representan alérgenos de olivo ya identificados. De las 37 familias restantes que proponen, nuestro algoritmo consigue detectar también 30 familias. Las otras 7 familias que nosotros no identificamos podrían corresponder a alérgenos que no se estén expresando en nuestro escenario particular en ese momento, o bien lo hacen a muy bajo nivel, o puede que alguna de esas familias alérgicas no estuviese recogida el Allergome en el momento del análisis. Entre los candidatos que obtuvieron San Segundo-Acosta et al. [197] centraron su atención en las ciclofilinas, enolasas y malato deshidrogenasas mitocondrial y citosólica para una posible validación. A excepción de la malato deshidrogenasa citosólica, nuestro algoritmo también sugiere estos alérgenos como candidatos. Los autores consiguieron validar la alergia de la ciclofilina de olivo detectada, lo cual dio lugar a la identificación del alérgeno *Ole e 15*. Sin embargo, los pacientes alérgicos testados en su análisis no reaccionaron frente a los otros tres alérgenos ensayados, aunque afirman que no deben ser descartados como posibles alérgenos y deberían ser analizados en otras poblaciones de pacientes alérgicos.

Estos resultados coincidentes, resultantes de aproximaciones alternativas a las descritas en nuestro trabajo, nos hacen incidir de nuevo en la existencia de ventajas añadidas del uso de aproximaciones ómicas como ya hemos indicado en el caso de *Ole e 7*. En nuestro caso, seleccionamos varios alérgenos candidatos de la lista sugerida por nuestro algoritmo para una validación preliminar, sobre la base de la existencia de herramientas moleculares (anticuerpos, proteínas recombinantes), bien disponibles comercialmente, o previamente generadas en el grupo de investigación. Se han llevado a cabo en nuestro laboratorio ensayos experimentales con los siguientes alérgenos: tropomiosina, bromelaína y ferritina, usando como control la profilina *Ole e 2* recombinante. Los resultados indican que existe reactividad a estos tres alérgenos en el suero de pacientes alérgicos a polen de olivo, lo que nos permite proponerlos como posibles nuevos alérgenos de polen de olivo, desconocidos hasta la fecha (resultados todavía no

publicados). De estos tres alérgenos, solo la ferritina se incluye también en la lista de candidatos de San Segundo-Acosta et al. [197].

La bromelaína (*Ana c 2*) es una proteasa presente en el fruto o el tallo de la piña (*Ananas comosus*). Ya existían resultados que sugerían que esta proteína es reconocida por la IgE del suero de pacientes alérgicos al polen de olivo [431], debido a un caso de reactividad cruzada con *Ole e 1*, con quien comparte un epítipo. Puesto que en la etapa de búsqueda de nuevos alérgenos de nuestro algoritmo los *Ole e* ya han sido retirados del análisis, el hallazgo de transcritos en ReprOlive con homología significativa con este alérgeno sugieren la presencia en el polen de olivo de alguna proteína distinta a *Ole e 1*, con características funcionales y estructurales presumiblemente similares a la bromelaína. Este posible nuevo alérgeno, podría contribuir al síndrome polen-fruta, dada la reactividad cruzada con esta proteína alergénica del fruto de la piña.

Las tropomiosinas son proteínas relacionadas con la contracción muscular presentes en crustáceos, moluscos y otros invertebrados y muy conservadas filogenéticamente. Están involucradas en reacciones alérgicas ante el consumo de marisco y provocan reactividad cruzada con ácaros del polvo, cucarachas y nematodos como el *Anisakis*, por lo que se consideran panalérgenos. Por su parte, la ferritina es una proteína ampliamente distribuida y especializada en el almacenamiento de hierro. Libera hierro durante la germinación de la semilla. Hasta la fecha no se ha descrito ningún caso de reactividad cruzada entre polen y ferritina o entre polen y tropomiosina, pero en vista de los resultados, cabe sospechar la presencia en el polen de olivo de alérgenos correspondientes a estas familias proteicas, o bien de proteínas similares, capaces de reaccionar con estos alérgenos.

## Fiabilidad y reusabilidad del flujo de trabajo

Como resumen, podemos concluir que la identificación con nuestro algoritmo de transcritos de polen de olivo homólogos a poligalacturonasas y ciclofilinas, que posteriormente fueron identificados y validados como los alérgenos *Ole e 14* [167] y *Ole e 15* [197] de olivo deja clara la validez de nuestro método en la búsqueda de posibles nuevos alérgenos. También hemos aportado nueva información acerca del repertorio alergénico de este polen, con hallazgos relacionados con los alérgenos bromelaína, tropomiosina y ferritina. Además, nuestro algoritmo propone posibles casos de reactividad cruzada con otras especies todavía no descritos, cuyo análisis creemos que resulta interesante.

La disponibilidad de la secuencia del genoma de olivo ha facilitado mucho el estudio del alergograma de esta planta, como ha quedado demostrado en [197]. De hecho, la recientísima publicación del genoma de la variedad Picual (Jiménez-Ruiz et al., 2020), en la que se basa ReprOlive, supone el mejor de los sustratos para la correcta caracterización de los alérgenos de esta variedad, tanto de los ya conocidos, como de los que están por descubrir, y seguramente dará sus frutos en breve. No obstante, cabe recalcar que nuestra aproximación es una muy buena opción para estudiar el alergograma de especies cuyo genoma no esté aún disponible.

Esta investigación se ha llevado a cabo sobre el transcriptoma de polen obtenido en ReprOlive. No obstante, hemos visto que la nueva versión del transcriptoma obtenida con TransFlow mejora bastante (véase apartado 5.2), siendo mucho más completo y correcto que ReprOlive, por lo que creemos que al análisis de este transcriptoma mejorado con nuestro algoritmo de búsqueda de alérgenos daría mejores resultados que los obtenidos hasta ahora.

La estrategia que hemos seguido para la búsqueda de alérgenos por homología es eficaz sobre todo para detectar alérgenos con epítopos lineales continuos, aquellos en los que la secuencia primaria es suficiente para que se una la IgE [432]. Sin embargo, en los epítopos conformacionales las secuencias

de aminoácidos correspondientes suelen ser discontinuas, están alejadas en la estructura primaria, y se acercan tras el plegamiento, y es en esa conformación cuando son reconocidos por la IgE. La detección por homología de este último tipo de epítomos es menos fructífera y requiere la combinación con datos estructurales y conformacionales de la proteína. Una mejora sustancial de nuestra estrategia conllevaría introducir este tipo de aproximación para poder explorar de forma masiva la presencia de alérgenos con epítomos conformacionales, además de los lineales.

Nuestro algoritmo ha sido diseñado con la herramienta AutoFlow (apartado 3.2.7), por lo que se trata de un flujo de trabajo automático y fácilmente reproducible. La posibilidad de poder repetir el análisis a lo largo del tiempo supone una ventaja, ya que cuanto más se sepa del alergograma de la planta en cuestión, y aumente el conocimiento acumulado en bases de datos como UniProt y Allergome, mejorará también la capacidad de este algoritmo para establecer posibles casos de reactividad cruzada entre especies distantes y encontrar posibles nuevos alérgenos.

Este flujo de trabajo es aplicable a distintos sets de datos transcriptómicos, de cualquier tejido. Además puede utilizarse para explorar el alergograma de cualquier otra planta u organismo, simplemente basta con adaptar a la especie en cuestión las bases de datos de secuencias de alérgenos de entrada que necesita el algoritmo.

Finalmente, dejamos abierta la puerta a investigar los posibles casos de nuevos alérgenos de polen de olivo o eventos de reactividad cruzada con otras especies sugeridos por nuestra aproximación. Su consideración y evaluación en el laboratorio servirá, además de para aumentar el conocimiento del papel del olivo en la alergia, para mejorar tanto el diagnóstico como el tratamiento de esta extendida enfermedad.

## 5.5. Predicción *in silico* de posibles modificaciones postraduccionales (*S*-nitrosilación y Tyr-nitración) sobre el transcriptoma de polen

Las modificaciones postraduccionales de las proteínas (PTM) son un mecanismo clave en numerosos procesos biológicos. La determinación experimental de los residuos que sufren tales modificaciones resulta laboriosa y cara, y además consume mucho tiempo. Hoy en día, con el crecimiento de la bioinformática han emergido numerosas herramientas para la predicción *in silico* de estas PTM, por lo que su uso representa una primera aproximación bastante conveniente, que permite generar resultados preliminares y valorar su posterior verificación experimental.

Los algoritmos de predicción de PTM han ido evolucionando con el tiempo. Inicialmente, trataban todos los aminoácidos de la proteína como entidades independientes, centrándose en la propensión a modificación de cada aminoácido por separado. Sin embargo, sabemos que existen correlaciones entre ellos que se deben tener en cuenta. Por ello, los algoritmos empezaron a asumir en su estrategia que pequeños péptidos similares, podrían presentar estructuras 3D y características bioquímicas comunes [259], e incluyen factores estructurales como la composición aminoacídica flanqueante al residuo modificable, el área de superficie accesible y propiedades fisicoquímicas como la carga de los aminoácidos colindantes como factor en su toma de decisiones [260]. La incorporación de estos efectos de correlación a los algoritmos mejora notablemente la calidad de la predicción [258], y mejorarán presumiblemente aún más cuando sean capaces de integrar todas las propiedades estereológicas que emergen de la estructura 3D completa de la proteína [433]. No obstante, este sigue siendo un factor limitante, ya que la estructura 3D de muchas proteínas se desconoce por el momento.

Aunque los algoritmos de predicción *in silico* de PTM empleados en esta tesis utilizan ya algunas

técnicas de aprendizaje automático (*machine learning*), en los últimos años se ha observado un creciente auge de este tipo de herramientas basadas en aprendizaje de máquina, que permiten hacer predicciones cada vez más complejas [434, 435, 436].

En nuestro caso, con el objetivo de conocer mejor el papel de dos PTM concretas mediadas por NO (la *S*-nitrosilación y la Tyr-nitración) en el desarrollo y la fisiología del polen del olivo, el transcriptoma de polen obtenido en ReprOlive se ha analizado *in silico* con distintos programas y umbrales para evaluar la capacidad de sus transcritos de ser potencialmente modificados por *S*-nitrosilación/Tyr-nitración, de acuerdo con criterios bien establecidos. El resultado obtenido para cada programa ha sido en todos los casos una lista bastante extensa de productos génicos susceptibles de dichas modificaciones. Existen diferencias notables en cuanto al número total de proteínas potencialmente modificables predicho por cada herramienta. Esto es lógico, puesto que, aunque todos los programas seleccionados se basan en algoritmos de predicción estadística que están entrenados con conjuntos de datos de secuencias proteicas con sitios de modificación validados experimentalmente, se trata de algoritmos diferentes y los resultados obtenidos pueden variar de unos a otros, como se ha podido comprobar. No obstante, los distintos métodos se complementan entre sí y la integración de todos ellos permite aprovechar los puntos fuertes de cada uno y produce mejores resultados en consonancia con lo recientemente publicado [437]. Por esta razón decidimos interrogar todas las herramientas de forma independiente y tener en cuenta preferiblemente aquellas modificaciones predichas simultáneamente por todas las herramientas, como los resultados más fiables.

Una vez identificado *in silico* un conjunto de proteínas como potencialmente modificables, hay que validar experimentalmente los resultados mediante métodos aceptados como el de traspaso de biotinas, transferencia de tipo Western, inmunoprecipitación o identificación por espectrometría de masas. Con este objetivo, María José Jiménez-Quesada llevó a cabo una validación experimental de estas modificaciones sobre extractos de polen maduro y polen germinado *in vitro*, como parte de un trabajo que evalúa la producción de NO durante la germinación de polen de olivo *in vitro*, e indaga en las implicaciones de esta molécula en parámetros reproductivos tales como la viabilidad del polen y la elongación del tubo polínico [160]. Tras comparar los resultados obtenidos *in silico* con la validación experimental observamos que un alto porcentaje de los productos génicos modificables determinados experimentalmente se han predicho como candidatos por al menos una de las herramientas o umbrales utilizados (86,2 % y 78,8 % para *S*-nitrosilación y Tyr-nitración respectivamente). Además, las distintas herramientas o umbrales utilizados simultáneamente predicen un notable porcentaje de ellos (31,4 % y 67,2 % para *S*-nitrosilación y Tyr-nitración respectivamente). En ausencia de validación experimental, o como paso previo, estas proteínas que concurren como candidatas en todas las predicciones son especialmente interesantes a la hora de tomar decisiones sobre la evolución de una investigación. Por otra parte, un porcentaje nada despreciable de las proteínas *S*-nitrosiladas y Tyr-nitradas identificadas experimentalmente (el 13,8 % y el 21,2 % respectivamente) no aparecieron en las predicciones de ninguno de los algoritmos. Esto puede deberse a la todavía baja sensibilidad de los programas utilizados, o simplemente a efectos de contaminación proteica o interacciones no específicas de los anticuerpos usados en la inmunoprecipitación, combinado con la extrema sensibilidad de los métodos de nano-LC-MS.

Aunque se usaran los umbrales más restrictivos de las herramientas, la mayoría de las proteínas que se ha predicho *in silico* que pueden sufrir una PTM se consiguen validar experimentalmente, como puede corroborarse en las figuras 4.22 y 4.23. No obstante, estos métodos de predicción presentan todavía una elevada tasa de falsos positivos. ¿A qué puede deberse? En primer lugar, cabe recordar que, tanto la baja abundancia o la inestabilidad de ciertas proteínas, como los problemas asociados con las técnicas de identificación comentados previamente, pueden dificultar mucho su detección experimental y quizás algunos de los supuestos falsos positivos identificados *in silico*, en realidad no lo sean. Por otra parte hay que tener en cuenta que el número de proteínas potencialmente modificables que se obtiene en nuestro

análisis está sobreestimado: a pesar de que los autores recomiendan usar como entrada la secuencia completa de la proteína problema, en vez de fragmentos, para una predicción más adecuada [258], dado que el número de proteínas completas identificadas en el transcriptoma de polen fue relativamente bajo (2 809 proteínas, al haberse realizado este análisis antes de la secuenciación del genoma de olivo), el análisis se hizo sobre todas las proteínas presentes (21 607 proteínas), incluidas las proteínas incompletas. En muchos casos, estas proteínas o péptidos incompletos pueden corresponder a un mismo ortólogo. No obstante se decidió no eliminar esta redundancia por dos motivos principalmente: puede ocurrir que una proteína esté fragmentada en nuestro proteoma, y a no ser que los fragmentos sean solapantes, nos interesa buscar posibles puntos de modificación a lo largo de toda la secuencia de la que dispongamos y, en segundo lugar, podrían existir en nuestro transcriptoma variantes o isoformas distintas de una misma proteína, que difieran en cuanto a sus dianas de modificación postraduccional, con lo cual, es interesante conservarlas y analizarlas todas. En resumen, el uso exclusivamente de proteínas completas y no redundantes reduciría notablemente la tasa de falsos positivos, pero perderíamos información con respecto a isoformas distintas y proteínas incompletas en nuestro proteoma.

Cabe mencionar también que en un estudio de predicción computacional masiva de proteínas propensas a *S*-nitrosilación [433], anterior al nuestro, llevado a cabo en *Arabidopsis thaliana* con la herramienta GPS-SNO se obtuvieron resultados muy similares a los nuestros: un alto número de proteínas candidatas a *S*-nitrosilación y muchas de ellas coincidentes con las nuestras.

Estas altas cifras de proteínas supuestamente modificables también nos llevan a pensar si realmente la *S*-nitrosilación o la Tyr-nitración son procesos tan corrientes *in vivo*, y si efectivamente afectan a tantas proteínas. Con respecto a la *S*-nitrosilación, antes de que decidiéramos abordar esta predicción se venía observando un aumento considerable en el número de proteínas *S*-nitrosiladas identificadas en las plantas [438, 439, 440, 441, 442, 443, 444], hasta tal punto que esta PTM podría ser considerada tan importante como otras modificaciones bien consolidadas y caracterizadas como puede ser la fosforilación [157, 445]. Por otra parte, el estudio de la Tyr-nitración en las proteínas vegetales es todavía escaso, pero prometedor [164], y el número de proteínas de plantas Tyr-nitradas en condiciones de estrés identificadas va en aumento, así como las pruebas de la función reguladora de esta modificación en los procesos fisiológicos [446]. Además, tanto la predicción *in silico* como la determinación experimental han revelado la existencia de proteínas propensas a ambos tipos de modificación, hecho ampliamente documentado en la bibliografía [447].

Ambas modificaciones postraduccionales están involucradas en multitud de procesos fisiológicos, como se describe en la bibliografía [157, 158, 143, 159, 160, 155, 164] y como ha quedado demostrado en el gran abanico de categorías obtenidas en los análisis de enriquecimiento biológico llevados a cabo en nuestro trabajo (apartado 4.5.4), muchas de ellas coincidentes para ambos tipos de PTM. La clasificación funcional de los listados de proteínas obtenidos abre además una amplia discusión respecto a su posible intervención en la reproducción de las plantas. De hecho, todavía quedan muchos aspectos por conocer sobre la función del NO en la reproducción sexual.

En cuanto a los resultados obtenidos en la validación experimental preliminar mediante nano-LC-MS los siguientes pasos a dar consisten en determinar varias cosas: si estos productos génicos son realmente modificables por NO *in vivo*, en qué condiciones pueden ser modificados, qué efectos tienen tales modificaciones, la reversibilidad de la modificación (en el caso de *S*-nitrosilación), y muchos otros aspectos. En nuestro caso, la validación experimental de los resultados obtenidos en este trabajo ha permitido demostrar la existencia de diferencias importantes entre los perfiles de proteínas *S*-nitrosiladas y Tyr-nitradas de polen maduro y polen germinado de olivo [160], e identificar un gran número de proteínas específicas sujetas a estas modificaciones, miembros de numerosas familias proteicas y con múltiples funciones. A modo de ejemplo, se han identificado numerosos péptidos de actina entre las proteínas reconocidas por los anticuerpos anti-*S*-nitrosoCys. Ya se ha demostrado que la actina se modifica por

*S*-nitrosilación y esta PTM interfiere con la polimerización de la actina, lo que ocasiona alteraciones en el funcionamiento y la distribución del citoesqueleto [163]. De hecho, se ha propuesto un mecanismo de regulación del citoesqueleto del polen madiado por NO a través de *S*-nitrosilación de la actina [448], y la integridad del citoesqueleto de actina se ha descrito como un factor clave en el control de los sistemas de autoincompatibilidad en *Nicotiana glauca* [449].

Conforme avance el conocimiento, aumentará el número de proteínas modificables validadas experimentalmente. Esto enriquecerá las bases de datos con las que se entrenan los algoritmos de predicción, permitirá reajustar los valores de sus umbrales, optimizará su toma de decisiones y con ello mejorarán presumiblemente sus resultados. Las propias listas de proteínas generadas en este trabajo pueden contribuir a ello. De hecho, las proteínas modificables identificadas por nano-LC-MS, pero no predichas *in silico* por las herramientas, serían un buen punto de partida a tener en cuenta por los desarrolladores para tal mejora.

Este análisis de PTM se ha llevado a cabo sobre el transcriptoma de polen de ReprOlive. Como futuros pasos a dar, sería interesante repetir el mismo estudio sobre el transcriptoma mejorado de polen (apartado 5.2), mucho más completo, con menos artefactos y mejor resolución de las isoformas, lo cual posiblemente se traduzca en mejores resultados en la predicción. Así mismo, sería muy útil llevar a cabo una aproximación similar en el pistilo, para evaluar el papel de estas PTM en este tejido. Durante la escritura de esta tesis apareció el genoma del olivo ‘Picual’, por lo que sería aún más interesante usar el algoritmo con las proteínas completas obtenidas de dicho genoma [129], e incluso comparar si hay diferencias de PTM entre las proteínas de ‘Picual’ y de acebuche (el olivo silvestre, cuya secuencia genómica ya está también disponible [201]).

En general, los resultados que producen las herramientas bioinformáticas de predicción de PTM son fácilmente manejables cuando se trabaja con pequeños sets de secuencias proteicas. Sin embargo, para predicciones computacionales masivas, en las que se parte de un conjunto ingente de secuencias (por ejemplo un proteoma deducido de un transcriptoma, como es nuestro caso), la información se vuelve inmanejable. Por este motivo y a raíz de este trabajo, hemos publicado un pequeño manual [450] para la predicción bioinformática de sitios de *S*-nitrosilación en grandes sets de proteínas, que propone varias estrategias junto con un sencillo flujo de trabajo que permite a un usuario no experto en bioinformática (aunque sí con conocimientos básicos de bash) utilizar los distintos programas descritos en esta tesis y extraer la información de interés de los archivos de salida generados. Además, la estructura en bloques de este flujo de trabajo permite incorporar fácilmente nuevos programas o algoritmos al flujo ya diseñado e integrar nuevos resultados.

En nuestra aproximación hemos corroborado que el uso simultáneo de varias herramientas con algoritmos diferentes permite ampliar el espectro de detección y además proporciona fiabilidad sobre los resultados en los que coinciden todas las herramientas [437]. Los datos resultantes obtenidos aquí pueden ser comparados con datos procedentes de otras fuentes o predicciones. Así por ejemplo se pueden comparar tejidos vegetativos y reproductivos, material haploide y diploide o distintos estadios de desarrollo.

Podemos concluir que la identificación preliminar *in silico* de proteínas propensas a alguna de estas PTM que se describe en esta tesis para determinar el *S*-nitrosoproteoma y el Tyr-nitroproteoma de polen de olivo supone una valiosa herramienta complementaria al uso de aproximaciones proteómicas, necesarias éstas últimas para su verificación experimental, y es además aplicable a cualquier otra especie. El uso combinado de ambas aproximaciones supondrá un aumento del conocimiento de los mecanismos y la dinámica de la *S*-nitrosilación y la Tyr-nitración en los procesos biológicos.

## Capítulo 6

# CONCLUSIONES

De la presente tesis se pueden extraer las siguientes conclusiones finales:

- El diseño de un flujo de trabajo automático nos ha permitido ensamblar y anotar un transcriptoma reproductivo de olivo a partir de polen y pistilo, y otro transcriptoma vegetativo como control, a partir de raíz y hoja, a la vez que un transcriptoma exclusivo de polen y otro de pistilo. La automatización del proceso ha permitido reejecutarlo y validarlo tantas veces como ha sido necesario hasta obtener la versión que consideramos definitiva.
- ReprOlive constituye una poderosa herramienta de acceso libre para el manejo de la información transcriptómica en el olivo, que está permitiendo la identificación y el análisis en detalle de numerosas enzimas y productos génicos, así como la identificación de nuevas isoformas, entre otras aplicaciones.
- Hemos obtenido y validado un transcriptoma mejorado de polen maduro de olivo mediante la combinación de distintos tipos de lecturas y ensambladores. El resultado es un transcriptoma mucho más completo y con menos artefactos.
- El algoritmo *findRGs* desarrollado es adecuado para la detección a partir de datos de NGS de los genes de referencia para análisis de expresión más apropiados en cada escenario. Se ha demostrado su potencia en olivo, *Arabidopsis* y humano, y nos consta que el grupo de investigación de la PAB lo ha usado ya para otras especies.
- El análisis bioinformático llevado a cabo ha permitido profundizar en la heterogeneidad alérgica del polen de olivo, con la propuesta de posibles nuevos alérgenos, así como posibles casos de reactividad cruzada con otras especies. Algunos de los nuevos alérgenos propuestos han sido validados experimentalmente.
- El flujo de trabajo puesto a punto para la predicción de modificaciones postraduccionales nos ha permitido determinar el *S*-nitrosoproteoma y el Tyr-nitroproteoma de polen de olivo mediante la combinación de varios programas con algoritmos diferentes. Los datos obtenidos permitirán ampliar el conocimiento sobre la función del NO en la biología reproductiva de esta planta.



# APÉNDICES



## Apéndice A

# Flujo de trabajo basado en AutoFlow para la obtención del transcriptoma reproductivo de olivo



Listado de Código A.1: *autoflow\_template\_reprolive.txt*

```

1 #####
2 # Declaring variables
3 #####
4 # static variables
5 $original_454_reads=../../454_reads.fastq
6 $original_sanger_reads=../../Sanger_reads.fastq
7 $assembly_name=reproductive
8 $transcript_name=rp11_olive_
9 $db_orthologues_path=../../
10 $db_orthologues=[Tair_Athaliana_database;RefSeq_Athaliana_database]
11 # dynamic variable where the best assembly path will be stored
12 @best_assembly=NULL
13 #####
14 # Pre-processing 454/Roche reads using SeqTrimNext
15 #####
16 SeqTrimNext_454){
17     module load seqtrimnext/last
18     ?
19     seqtrimnext -t transcriptomics_454_plants.txt -Q $original_454_reads -w [lcpu] -s
    ↪ 10.243 > stn_454_reads.txt
20     cat output_files/RL*/sequences*.fastq > input_mira_in.454.fastq
21 }
22 #####
23 # Pre-processing Sanger reads using SeqTrimNext
24 #####
25 SeqTrimNext_Sanger){
26     module load seqtrimnext/last
27     ?
28     seqtrimnext -t sanger.txt -Q $original_sanger_reads -w [lcpu] -s 10.243 >
    ↪ stn_sanger_reads.txt
29     mv output_files/sequences_.fastq input_mira_in.sanger.fastq
30 }
31 #####
32 # Preparing reads for assembling
33 #####
34 Preparing_Input_Reads){
35     module load seqtrimnext/last
36     ln -s SeqTrimNext_454)/input_mira_in.454.fastq
37     ln -s SeqTrimNext_Sanger)/input_mira_in.sanger.fastq
38     ?
39     fastq2fasta.rb SeqTrimNext_454)/input_mira_in.454.fastq input_euler_454
40     fastq2fasta.rb SeqTrimNext_Sanger)/input_mira_in.sanger.fastq input_euler_sanger
41     cat input_euler_454.fasta input_euler_sanger.fasta > input_euler.fasta
42 }
43 #####
44 # Assembling with MIRA3 (OLC assembler)
45 #####
46 Mira_assembly){

```

```

47 module load mira/3.2.0
48 ln -s Preparing_Input_Reads)/input_mira_in.454.fastq
49 ln -s Preparing_Input_Reads)/input_mira_in.sanger.fastq
50 ln -s Preparing_Input_Reads)/input_euler.fasta
51 mkdir -p $SCRATCH/$assembly_name
52 ?
53 mira -fastq -project=input_mira --job=denovo,est,normal,454,sanger -CL:asc dc
  ↳ SANGER_SETTINGS -CO:fnicpst=yes
54 454_SETTINGS -CO:fnicpst=yes -notraceinfo COMMON_SETTINGS -GE:not=4 -DI:lrt=$SCRATCH/
  ↳ $assembly
55 rm -rf $SCRATCH/$assembly_name
56 }
57 # Remove artifacts from MIRA3
58 Mira_remove_artifacts){
59     source ~soft_cvi_114/initializes/init_fln
60     gem list full_lengther_next
61     sort_fasta_list.rb Mira_assembly)/input_mira_assembly/input_mira_d_results/
  ↳ input_mira_out.unpadded.fasta
62     lista_to_fasta.rb Mira_assembly)/input_mira_assembly/input_mira_d_results/
  ↳ input_mira_out.unpadded.fasta list >
63 mira_contigs_dispersed.fasta
64 ?
65 full_lengther_next -f mira_contigs_dispersed.fasta -g plants -c 500 -z -w [lcpu]
66 }
67 # Recover sequences from debris of MIRA3
68 Recover_debris){
69     source ~soft_cvi_114/initializes/init_fln
70     lista_to_fasta.rb Mira_assembly)/input_euler.fasta
71 Mira_assembly)/input_mira_assembly/input_mira_d_info/input_mira_info_debrislist.txt >
  ↳ mira_debris.fasta
72     sort_fasta_list.rb mira_debris.fasta
73     lista_to_fasta.rb mira_debris.fasta list > mira_debris_dispersed.fasta
74     ?
75     full_lengther_next -f mira_debris_dispersed.fasta -g plants -c 500 -z -w [lcpu] -
  ↳ q d
76     table_header.rb -t fln_results/pt_seqs n > coding_debris_mira.list
77     table_header.rb -t fln_results/new_coding.txt >> coding_debris_mira.list
78     lista_to_fasta.rb mira_debris_dispersed.fasta coding_debris_mira.list >
  ↳ coding_debris_mira.fasta
79 }
80 #####
81 # Assembling with EULER (de Bruijn assembling using 25 & 29 k-mers)
82 #####
83 Euler_assembly_k_[25;29]){
84     module load euler/120408
85     ln -s Preparing_Input_Reads)/input_euler.fasta
86     ?
87     Assemble.pl input_euler.fasta (*)
88 }

```

```

89 # Remove artifacts from EULER
90 Euler_remove_artifacts_k_[25;29]){
91     source ~soft_cvi_114/initializes/init_fln
92     sort_fasta_list.rb !Euler_assembly_k_*/input_euler.fasta.contig
93     lista_to_fasta.rb !Euler_assembly_k_*/input_euler.fasta.contig list >
    ↪ dispersed_input_euler.fasta
94     ?
95     full_lengther_next -f dispersed_input_euler.fasta -g plants -c 500 -z -w [lcpu]
96 }
97 Validate_contigs_with_mapping_k_[25;29]){
98     module load bowtie/v2.2.0.0-beta7
99     bowtie2-build -f !Euler_remove_artifacts_k_*/fln_results/unigenes.fasta ref
100    ?
101    bowtie2 ref -f -U Preparing_Input_Reads)/input_euler.fasta -p [cpu] --very-fast -
    ↪ S mapeo.sam
102 }
103 Rescue_contigs_not_mapped_k_[25;29]){
104     module load samtools/0.1.16
105     ?
106     mapping_tool.rb -i !Validate_contigs_with_mapping_k_*/mapeo.sam -f -o
    ↪ euler_mapped_contigs.list
107     mapping_tool.rb -i !Validate_contigs_with_mapping_k_*/mapeo.sam -f -r -o
    ↪ euler_unmapped_contigs.list
108     lista_to_fasta.rb !Euler_remove_artifacts_k_*/fln_results/unigenes.fasta
    ↪ euler_mapped_contigs.list >
109 euler_mapped_contigs.fasta
110     lista_to_fasta.rb !Euler_remove_artifacts_k_*/fln_results/unigenes.fasta
    ↪ euler_unmapped_contigs.list > temp.fasta
111     table_header.rb -t !Euler_remove_artifacts_k_*/fln_results/pt_seqs >
    ↪ annot_coding_euler_unmapped.list
112     table_header.rb -t !Euler_remove_artifacts_k_*/fln_results/new_coding.txt >>
    ↪ annot_coding_euler_unmapped.list
113     lista_to_fasta.rb temp.fasta annot_coding_euler_unmapped.list >
    ↪ euler_coding_unmapped_contigs.fasta
114 }
115 #####
116 # CAP3 (final reconciliation of MIRA3 contigs with each Euler-25 and Euler-29 contigs)
117 #####
118 CAP3_reassembly_k_[25;29]){
119     module load cap3/101507
120     ?
121     cat Recover_debris)/coding_debris_mira.fasta Mira_remove_artifacts)/fln_results/
    ↪ unigenes.fasta
122 !Rescue_contigs_not_mapped_k_*/euler_mapped_contigs.fasta
123 !Rescue_contigs_not_mapped_k_*/euler_coding_unmapped_contigs.fasta > reassembly.fasta
124     cap3 reassembly.fasta -p 95 -o 40
125 }
126 FLN_analysis_of_CAP3_contigs_k_[25;29]){
127     module load ruby

```

```

128     source ~soft_cvi_114/initializes/init_flm
129     cat !CAP3_reassembly_k_*/reassembly.fasta.cap.contigs !CAP3_reassembly_k_*/
    ↪ reassembly.fasta.cap.singlets >
130 unigenes.fasta
131     sort_fasta_list.rb unigenes.fasta
132     lista_to_fasta.rb unigenes.fasta list > dispersed_unigenes.fasta
133     fasta_standard_renamer.rb dispersed_unigenes.fasta $transcript_name 6
134     ?
135     full_lengther_next -f dispersed_unigenes.fasta_new -g plants -c 300 -z -r -w [lcpu]
136 }
137 #####
138 # Deciding the best assembly
139 #####
140 Decide_best_assembly){
141     module load ruby
142     assembly=`flm_assembly_evaluator !FLM_analysis_of_CAP3_contigs_k_! `
143     ?
144     echo $assembly
145     env_manager "best_assembly=$assembly;"
146 }
147 #####
148 # Annotating the best assembly with sma3s
149 #####
150 Sma3s_annotate_best_assembly){
151     . ~soft_cvi_114/initializes/init_distributed_sma3s
152     ?
153     distributed_sma3s -c 'sma3s_v2.pl -a 123 -p F -v 2' -i $best_assembly/flm_results
    ↪ /unigenes.fasta -o olive.annot -d
154 /mnt/home/soft/blast_db/current/fmt/sma3/uniprot_plants.dat -g 250 -w [lcpu]
155 }
156 #####
157 # Obtaining ORTHOLOGUES in others Arabidopsis thaliana databases
158 #####
159 Orthologues_$db_orthologues){
160     source ~soft_cvi_114/initializes/init_flm
161     echo Decide_best_assembly)
162     mkdir db
163     makeblastdb -in $db_orthologues_path/(*) -dbtype prot -parse_seqids -out db/(*)
164     ?
165     full_lengther_next -f $best_assembly/flm_results/unigenes.fasta -a ' ' -u db/(*) -
    ↪ c 500 -z -w [lcpu] -q d -g plants
166 }
167 #####
168 # Obtaining SSRS (microsatellites)
169 #####
170 Obtaining_SSRS){
171     module load ruby
172     module load mreps/2.5
173     echo Decide_best_assembly)

```

```
174 ln -s $best_assembly/fln_results/unigenes.fasta
175 ?
176 replace_n.rb unigenes.fasta
177 rm_codigo_degenerado.rb replaced_n_unigenes.fasta
178 mreps -minsize 12 -minperiod 2 -exp 3.0 -fasta new_replaced_n_unigenes.fasta >
    ↪ $assembly_name_ssr.txt
179 }
180 #####
```



## Apéndice B

### Script de lanzamiento y plantillas de los distintos módulos de TransFlow



## Listado de Código B.1: Script de lanzamiento TransFlow\_olive.sh

```

1  #!/usr/bin/env bash
2
3  current_dir=`pwd`
4  OUTPUT_WORKFLOW=$current_dir"/workflow_olive"
5
6  mkdir $current_dir
7
8  TEMPLATES=" ../Templates/transcriptome_assembly_Illumina.txt, ../Templates/
   ↪ transcriptome_assembly_454_single.txt, ../Templates/
   ↪ transcriptome_mix_assembly_454_single_Illumina.txt, ../Templates/
   ↪ transcriptome_reference_Illumina_plants.txt, ../Templates/
   ↪ transcriptome_validation_454_single_Illumina.txt"
9  vars=`echo "
10  \\$ill_type=paired,
11  \\$kmers=[25;35],
12  \\$key_organisms=P.trichocarpa_A.thaliana,
13  \\$FLN_DATABASE=plants,
14  \\$single_illumina=$current_dir/ ../assembly_reads/FILE,
15  \\$read_illumina_pair_1=$current_dir/ ../assembly_reads/pollen_picual_1.fastq.gz,
16  \\$read_illumina_pair_2=$current_dir/ ../assembly_reads/pollen_picual_2.fastq.gz,
17  \\$read_454=$current_dir/ ../assembly_reads/RL1_polen_maduro_raw.fastq,
18  \\$reference=$current_dir/ ../transcriptome_references/fasta_files,
19  \\$reads=$current_dir/ ../transcriptome_references/read_files,
20  \\$BUSCO_db_path=$current_dir/ ../BUSCO_DB/embryophyta_odb9,
21  \\$report_template=$current_dir/ ../Templates/assembly_report.erb
22  " | tr -d [:space:]`
23
24  if [ $1 == '1' ]; then
25    AutoFlow -w $TEMPLATES -o $OUTPUT_WORKFLOW -c 1 -t 160:00:00 -V $vars
26  fi
27
28  if [ $1 == '2' ]; then
29    flow_logger -e $OUTPUT_WORKFLOW -r all $2
30  fi

```

## Listado de Código B.2: (MÓDULO 1) transcriptome\_assembly\_Illumina.txt

```

1  #= Autoflow transcriptome de novo of mature pollen from Olea europaea
2  #####
3
4  #= De novo transcriptome assembly from Illumina single-end reads.
5  #= This template includes Illumina assembly only.
6  #= This workflow has been designed to be used in plants.
7  #= Author: Rosario Carmona
8  #####
9  #Variables
10 #####
11
12
13 #####
14 #Pre-processing reads using SeqTrimNext
15 #####
16
17 Cleanup_ill){
18     #module load seqtrimnext/last
19     if [ "$ill_type" == "paired" ]; then
20         input_files="$paired_illumina_1,$paired_illumina_2"
21     elif [ "$ill_type" == "single" ]; then
22         input_files="$single_illumina"
23     else
24         echo 'Read input files not set.'
25         exit 1
26     fi
27     ?
28     seqtrimnext -t transcriptomics_short_reads.txt -Q $input_files -K -w [lcpu] -s
29     ↪ 10.243
30
31     if ( ([ ! -s output_files/paired_1.fastq ] || [ ! -s output_files/paired_2.fastq
32     ↪ ]) ) && [ "$ill_type" == "paired" ] ) || ([ ! -s output_files/sequences.fastq ]
33     ↪ && [ "$ill_type" == "single" ] ); then
34         echo "ERROR: fastq files not exists or are empty"
35         exit 1 # Fail
36     fi
37 }
38
39 #####
40 #Assembling Illumina reads
41 #####
42
43 ##### Oases assembly #####
44
45 Oases_primary_assembling_${kmers){
46
47     #resources: -s -n bigmem -c 1 -t '7-00:00:00' -m '300gb'
48     #-----

```

```

46 #module load oases/0.2.08
47 echo -e "ctOasesK(*)\toases\tprimary\t(*)\tI11" > tracker #create a file to track
    ↳ names in order to do the metrics
48
49     if [ "$ill_type" == "paired" ]; then
50         input_files='-shortPaired -separate Cleanup_ill)/output_files/
    ↳ paired_1.fastq Cleanup_ill)/output_files/paired_2.fastq'
51     elif [ "$ill_type" == "single" ]; then
52         input_files='-short Cleanup_ill)/output_files/sequences.fastq'
53     else
54         echo 'Read input files not set.'
55         exit 1
56     fi
57
58     ?
59     velveth autoflow_assembly (*) -fastq $input_files > outFile_1
60     velvetg autoflow_assembly -read_trkg yes -cov_cutoff 2 -min_contig_lgth 100 >
    ↳ outFile_2
61     oases autoflow_assembly -cov_cutoff 2 > outFile_3
62     ln -s autoflow_assembly/contigs.fa all_sequences.fasta
63
64     if [ ! -s autoflow_assembly/contigs.fa ] || [ ! -s autoflow_assembly/transcripts
    ↳ .fa ]; then
65     echo "ERROR: Oases primary assembly or scaffolding have failed"
66     exit 1 # Fail
67 else
68     rm autoflow_assembly/Graph2 autoflow_assembly/LastGraph autoflow_assembly/PreGraph
    ↳ autoflow_assembly/Roadmaps autoflow_assembly/Sequences
69 fi
70 }
71
72 scaffolding_oases_$kmers){
73     echo -e "scOasesK(*)\toases\tcaffolding\t(*)\tI11" > tracker
74     ?
75     ln -s !Oases_primary_assembling_*/autoflow_assembly/transcripts.fa all_sequences.
    ↳ fasta
76 }
77
78 #####join kmers#####
79
80 join_Oases){
81     #module load ruby
82     echo -e "ctOases_cat\tcat\tjoin\t$kmers\tI11" | sed 's/[] []//g' > tracker
83     ?
84     cat !Oases_primary_assembling_!/all_sequences.fasta > all.fasta
85     fasta_standard_renamer.rb all.fasta
86     ln -s all.fasta_new all_sequences.fasta
87     if [ ! -s all.fasta_new ]; then
88         echo "ERROR: Oases primary assembly join has failed"

```

```

89     exit 1 # Fail
90 fi
91 }
92
93 join_SC_Oases){
94     module load ruby
95     echo -e "scOases_cat\tcat\ttjoin\tt$kmers\ttI11" | sed 's/[] []//g' > tracker
96     ?
97     cat !scaffolding_oases_!/all_sequences.fasta > all.fasta
98     fasta_standard_renamer.rb all.fasta
99     ln -s all.fasta_new all_sequences.fasta
100    if [ ! -s all.fasta_new ]; then
101        echo "ERROR: Oases scaffolding join has failed"
102        exit 1 # Fail
103    fi
104 }
105
106 ##### SOAP assembly #####
107 SOAP_primary_assembling_$kmers){
108     # resources: -n cal -c 16 -t '7-00:00:00'
109     #-----
110     #module load soapdenovo_trans/1.03
111     echo -e "SOAPdenovo-Trans\t`SOAPdenovo-Trans-127mer |grep -oP '([0-9]*\.[0-9]*)'`"
112     ↪ >> ../versions
113     echo -e "ctSoapK(*)\tSOAPdenovo\tprimary\t(*)\tI11" > tracker
114     echo -e "
115     max_rd_len=151
116     [LIB]
117     reverse_seq=0
118     asm_flags=3
119     rank=1" > config.txt
120     if [ "$ill_type" == "paired" ]; then
121         echo -e 'q1=Cleanup_ill)/output_files/paired_1.fastq
122         q2=Cleanup_ill)/output_files/paired_2.fastq' >> config.txt
123     elif [ "$ill_type" == "single" ]; then
124         echo 'q=Cleanup_ill)/output_files/sequences.fastq' >> config.txt
125     else
126         echo 'Read input files not set.'
127         exit 1
128     fi
129     ?
130     SOAPdenovo-Trans-127mer all -s config.txt -o assembly -K (*) -p [cpu]
131     ln -s assembly.contig all_sequences.fasta
132
133     if [ ! -s assembly.contig ] || [ ! -s assembly.scafSeq ]; then
134         echo "ERROR: SOAP primary assembly or scaffolding have failed"
135         exit 1 # Fail
136     fi

```

```

137 }
138
139 scaffolding_SOAP_$kmers){
140     echo -e "scSoapK(*)\tSOAPdenovo\tsc scaffolding\t(*)\tI11" > tracker
141     ?
142     ln -s !SOAP_primary_assembling_*/assembly.scafSeq all_sequences.fasta
143 }
144
145 ###join kmers#####
146
147 join_SOAP){
148     # module load ruby
149     echo -e "ctSoap_cat\tcat\ttjoin\t$kmers\tI11" | sed 's/[] []//g' > tracker
150     ?
151     cat !SOAP_primary_assembling_*/all_sequences.fasta > all.fasta
152     fasta_standard_renamer.rb all.fasta
153     ln -s all.fasta_new all_sequences.fasta
154     if [ ! -s all.fasta_new ]; then
155         echo "ERROR: SOAP primary assembly join has failed"
156         exit 1 # Fail
157     fi
158 }
159
160 join_SC_SOAP){
161     # module load ruby
162     echo -e "scSoap_cat\tcat\ttjoin\t$kmers\tI11" | sed 's/[] []//g' > tracker
163     ?
164     cat !scaffolding_SOAP_*/all_sequences.fasta > all.fasta
165     fasta_standard_renamer.rb all.fasta
166     ln -s all.fasta_new all_sequences.fasta
167     if [ ! -s all.fasta_new ]; then
168         echo "ERROR: SOAP scaffolding join has failed"
169         exit 1 # Fail
170     fi
171 }
172
173 ##### RAY assembly #####
174
175 RAY_primary_assembling_$kmers){
176     # resources: -s -n bigmem -c 32 -t '7-00:00:00' -m '100gb' -u 1
177     module load ray/2.3.1
178     echo -e "ctRayK(*)\tRay\tprimary\t(*)\tI11" > tracker
179     if [ "$ill_type" == "paired" ]; then
180         input_files='-p Cleanup_ill)/output_files/paired_1_.fastq Cleanup_ill)
181         ↪ /output_files/paired_2_.fastq'
182     elif [ "$ill_type" == "single" ]; then
183         input_files='-s Cleanup_ill)/output_files/sequences_.fastq'
184     else
185         echo 'Read input files not set.'
```

```

185         exit 1
186     fi
187 ?
188 mpiexec -np [cpu] Ray -k (*) $input_files # -route-messages -connection-type
    ↳ debruijn -routing-graph-degree 4 (Message routing for large number of cores)
189 if [ ! -s RayOutput/Contigs.fasta ] || [ ! -s RayOutput/Scaffolds.fasta ]; then
190     echo "ERROR: Ray primary assembly or scaffolding have failed"
191     exit 1 # Fail
192 fi
193
194 ln -s RayOutput/Contigs.fasta all_sequences.fasta
195 }
196
197 scaffolding_RAY_${kmers}){
198     echo -e "scRayK(*)\tRay\tscaffolding\t(*)\tIll" > tracker
199     ?
200     ln -s !RAY_primary_assembling_*/RayOutput/Scaffolds.fasta all_sequences.fasta
201 }
202
203
204 ###join kmers#####
205
206 join_RAY){
207     # module load ruby
208     echo -e "ctRay_cat\tcat\tjoin\t${kmers}\tIll" | sed 's/[] []//g' > tracker
209     ?
210     cat !RAY_primary_assembling_!/all_sequences.fasta > all.fasta
211     fasta_standard_renamer.rb all.fasta
212     ln -s all.fasta_new all_sequences.fasta
213     if [ ! -s all.fasta_new ]; then
214         echo "ERROR: Ray primary assembly join has failed"
215         exit 1 # Fail
216     fi
217 }
218
219 join_SC_RAY){
220     #module load ruby
221     echo -e "scRay_cat\tcat\tjoin\t${kmers}\tIll" | sed 's/[] []//g' > tracker
222     ?
223     cat !scaffolding_RAY_!/all_sequences.fasta > all.fasta
224     fasta_standard_renamer.rb all.fasta
225     ln -s all.fasta_new all_sequences.fasta
226     if [ ! -s all.fasta_new ]; then
227         echo "ERROR: Ray scaffolding join has failed"
228         exit 1 # Fail
229     fi
230 }
231
232 ##remove redundancy#####

```

```

233 reduce_illumina_kmer_redundance_[JobRegExp:join:-]){
234 # module load cdhit
235 echo -e "`cut -f 1 (*)/tracker`_cd\tcd-hit\ttjoin\tt$kmers\ttI11" | sed 's/[] []//g' >
    ↪ tracker
236 ?
237 cd-hit-est -i (*)/all_sequences.fasta -o all_sequences.fasta -c 1
238 if [ ! -s all_sequences.fasta ]; then
239     echo "ERROR: CD-hit "`cat tracker| cut -f 1`" has failed"
240     exit 1 # Fail
241 fi
242 }
243
244 ##### kmers assembly#####
245
246 merge_illumina_[JobRegExp:kmer_redundance:-]){
247 # module load amos/3.1.0
248 echo -e "`cut -f 1 (*)/tracker`_rcMin2\tminimus\tmerge\tt$kmers\ttI11" | sed 's/[] []//g' > tracker
    ↪ g' > tracker
249 toAmos -s (*)/all_sequences.fasta -o input_minimus.afg
250 ?
251 minimus2 input_minimus
252 cat input_minimus.fasta input_minimus.singletons.seq > all_sequences.fasta
253 if [ ! -s input_minimus.fasta ] || [ ! -s all_sequences.fasta ]; then
254     echo "ERROR: MINIMUS "`cat tracker| cut -f 1`" has failed"
255     exit 1 # Fail
256 fi
257 }
258 }

```

Listado de Código B.3: (MÓDULO 2) *transcriptome\_assembly\_454\_single.txt*

```

1  #= Autoflow transcriptome de novo of mature pollen from Olea europaea
2  #####
3
4  #= De novo transcriptome assembly from 454 single-end reads.
5  #= This template includes 454 assembly only.
6  #= This workflow has been designed to be used in plants.
7  #= Author: Rosario Carmona
8  #####
9  #Variables
10 #####
11
12
13 #####
14 #Pre-processing reads using SeqTrimNext
15 #####
16
17 Cleanup_454){
18
19     #module load seqtrimnext/last
20     ?
21     seqtrimnext -t transcriptomics_454.txt -Q $read_454 -K -w [lcpu] -s 10.243
22     cat output_files/*/sequences_.fastq > sequences_.fastq
23
24     if [ ! -s sequences_.fastq ] ; then
25         echo "ERROR: fastq file not exists or is empty"
26         exit 1 # Fail
27     fi
28 }
29
30 #####
31 #Assembling 454 reads
32 #####
33
34 ##Input data prep#####
35
36 Input_data_prep){
37     #module load seqtrimnext/last
38     ln -s Cleanup_454)/sequences_.fastq
39     ?
40     fastq2fasta.rb Cleanup_454)/sequences_.fastq sequences_prep
41     if [ ! -s sequences_prep.fasta ] ; then
42         echo "ERROR: fasta file not exists or is empty"
43         exit 1 # Fail
44     fi
45 }
46
47 ##### MIRA_454_ assembly #####
48

```

```

49 MIRA_assembling_454){
50     #module load mira/4.0.2
51     #create a file to track names in order to do the metrics
52     echo -e "ctMIRA\tmira\tprimary\tnot\t454" > tracker
53     echo "
54     project = transcriptome_test
55     job = est,denovo,accurate
56     parameters = -GE:not=4
57     readgroup = SomeUnpaired454ReadsIGotFromaNCBI
58     data = Cleanup_454)/sequences_.fastq
59     technology = 454
60     " > mira_config_template
61     ?
62     mira mira_config_template
63     if [ ! -s transcriptome_test_assembly/transcriptome_test_d_results/
64     ↪ transcriptome_test_out.unpadded.fasta ] ; then
65     echo "ERROR: MIRA assembly has failed"
66     exit 1 # Fail
67     fi
68     ln -s transcriptome_test_assembly/transcriptome_test_d_results/
69     ↪ transcriptome_test_out.unpadded.fasta all_sequences.fasta
70 }
71 ##### EULER_454_ assembly (de Bruijn assembling using 25 & 29 k-mers) ###
72 Euler_assembly_454){
73     #module load euler/120408
74     ln -s Input_data_prep)/sequences_prep.fasta
75     #create a file to track names in order to do the metrics
76     echo -e "ctEulK29\tEuler\tprimary\tauto\t454" > tracker
77     ?
78     Assemble.pl sequences_prep.fasta 29
79     ln -s transformed/sequences_prep.fasta.contig all_sequences.fasta
80     if [ ! -s transformed/sequences_prep.fasta.contig ] ; then
81     echo "ERROR: Euler assembly has failed"
82     exit 1 # Fail
83     fi
84 }
85 #####CAP3_recontiliation (MIRA+EULER_29)#####
86 CAP3_454_k29){
87     #module load cap3/101507
88     #create a file to track names in order to do the metrics
89     echo -e "ctMIRA_ctEulK29_rcCAP3\tcap3\trec\t29-auto\t454" | sed 's/[] []//g' >
90     ↪ tracker
91     cat MIRA_assembling_454)/all_sequences.fasta Euler_assembly_454)/transformed/
92     ↪ sequences_prep.fasta.contig > reassembly_454.fasta
93     ?

```

```
94         cap3 reassembly_454.fasta -p 95 -o 40
95         cat reassembly_454.fasta.cap.contigs reassembly_454.fasta.cap.singlets >
    ↪ reconciliation_454.fasta
96 if [ ! -s reassembly_454.fasta.cap.contigs ] || [ ! -s reconciliation_454.fasta ];
    ↪ then
97     echo "ERROR: CAP3 reconciliation has failed"
98     exit 1 # Fail
99 fi
100     ln -s reconciliation_454.fasta all_sequences.fasta
101 }
```

**Listado de Código B.4:** (MÓDULO 3) *transcriptome\_mix\_assembly\_454\_single\_Illumina.txt*

```

1  #= Autoflow transcriptome de novo of mature pollen from Olea europaea
2  #####
3
4  #= De novo transcriptome assembly from Illumina and 454 single-end reads.
5  #= This template includes different strategies: i)Illumina reads+454 reads assembly,iv
   ↪ )Illumina assembly+454 reads, v)Illumina assembly+454 assembly.
6  #= This workflow has been designed to be used in plants.
7  #= Author: Rosario Carmona
8  #####
9  #Variables
10 #####
11
12
13 #####
14 ### ILLUMINA ASSEMBLY AND CAP3 DATA PREPARATION FROM PREVIOUS TEMPLATES
15 #####
16
17 ### Illumina assemblies, remove sequence redundancy
18 #####
19
20 # To PRIMARY illumina assembly level
21 #-----
22 assemb_remove_illumina_redundancy_[JobRegExp:primary:-]){
23   #module load ruby
24   #module load cdhit
25   #create a file to track names in order to do the metrics
26   echo -e "`cut -f1 (*)/tracker\tcd-hit\tprimary\t`cut -f4 (*)/tracker\tI11" >
   ↪ tracker
27   ?
28   cd-hit-est -i (*)/all_sequences.fasta -o all_sequences -c 1
29   if [ ! -s all_sequences ] ; then
30     echo "ERROR: Remove redundancy with cdhit has failed"
31     exit 1 # Fail
32   fi
33   fasta_standard_renamer.rb all_sequences
34   mv all_sequences_new all_sequences.fasta
35 }
36
37 # To SCAFFOLDING illumina level
38 #-----
39 SC_remove_illumina_redundancy_[JobRegExp:scaffolding:-]){
40   # module load ruby
41   # module load cdhit
42   #create a file to track names in order to do the metrics
43   echo -e "`cut -f1 (*)/tracker\tcd-hit\ttscaffolding\t`cut -f4 (*)/tracker\tI11" >
   ↪ tracker
44   ?
45   cd-hit-est -i (*)/all_sequences.fasta -o all_sequences -c 1

```

```

46  if [ ! -s all_sequences ] ; then
47      echo "ERROR: Remove redundancy with cdhit has failed"
48      exit 1 # Fail
49  fi
50  fasta_standard_renamer.rb all_sequences
51  mv all_sequences_new all_sequences.fasta
52 }
53
54 ### Assemblies prep, convert fastas to fastq
55 #####
56
57 Qual_454_prep){
58     # module load seqtrimnext/last
59     ln -s CAP3_454_k29)/all_sequences.fasta
60     gen_qual.rb CAP3_454_k29)/all_sequences.fasta all_sequences.qual
61     ?
62     fasta2fastq.rb all_sequences.fasta all_sequences.qual all_sequences
63     if [ ! -s all_sequences.fastq ]; then
64         echo "ERROR: 454 (ens) fasta2fastq qual preparation has failed"
65         exit 1 # Fail
66     fi
67 }
68
69 Qual_Ill_prep_[JobRegExp:kmer_redundance|merge_illumina|scaffolding|remove:-]){
70     # module load seqtrimnext/last
71     ln -s (*)/all_sequences.fasta
72     #create a file to track names in order to do the metrics
73     echo -e "`cut -f1 (*)/tracker`\tfasta2fastq\ttprep_qual\t`cut -f4 (*)/tracker`\tIll"
74     ↪ > tracker
75     gen_qual.rb (*)/all_sequences.fasta all_sequences.qual
76     ?
77     fasta2fastq.rb all_sequences.fasta all_sequences.qual all_sequences
78     if [ ! -s all_sequences.fastq ]; then
79         echo "ERROR: Illumina (ens) fasta2fastq qual preparation has failed"
80         exit 1 # Fail
81     fi
82 }
83 #####
84 ### ILLUMINA - 454 ASSEMBLIES
85 #####
86
87 ### 454 assembly against Illumina assembly
88 #####
89
90 # MINIMUS reassembly (one illumina ens to 454 ens) CAP3/SOAP-25 ; CAP3/SOAP-35 ; CAP3/
91 ↪ Oases-25 ; CAP3/Oases-35 ; CAP3/SOAP(25-35) ; CAP3/Oases(25-35)
92 #-----
93 reassembly_454_Illumina_[JobRegExp:kmer_redundance|merge_illumina|primary|scaffolding

```

```

    ↪ :-~){
93 # module load amos/3.1.0
94 # module load ruby
95 #create a file to track names in order to do the metrics
96 echo -e "aaMin2/~cut -f1 (*)/tracker~/454Cap3~tminimus~trs~t29/~cut -f4 (*)/tracker
    ↪ ~t454-Ill" > tracker
97 cat CAP3_454_k29)/reconciliation_454.fasta (*)/all_sequences.fasta > pool_sequences
98 fasta_standard_renamer.rb pool_sequences
99 toAmos -s pool_sequences_new -o assembly_minimus.afg
100 ?
101 minimus2 assembly_minimus
102 cat assembly_minimus.fasta assembly_minimus.singletons.seq > all_sequences.fasta
103
104 if [ ! -s assembly_minimus.fasta ] || [ ! -s all_sequences.fasta ]; then
105     echo "ERROR: 454 (ens)- illumina (ens) Minimus reconciliation has failed"
106     exit 1 # Fail
107 fi
108 }
109
110 # MINIMUS reassembly (all illumina ens to 454 ens) all CAP3/SOAP25/SOAP35/Oases25/
    ↪ Oases35
111 #-----
112
113 # To PRIMARY illumina assembly level
114 reassembly_all){
115 # module load amos/3.1.0
116
117 # module load ruby
118 #create a file to track names in order to do the metrics
119 echo -e "aaMin2/ctALL/454Cap3~tminimus~trs_all~t29/$kmers~t454-Ill" | sed 's/[ ] []//g
    ↪ ' > tracker
120
121 cat CAP3_454_k29)/reconciliation_454.fasta !JobRegExp:
    ↪ assemb_remove_illumina_redundancy:-!/all_sequences.fasta > input_minimus
122 fasta_standard_renamer.rb input_minimus
123 toAmos -s input_minimus_new -o assembly_minimus.afg
124 ?
125 minimus2 assembly_minimus
126 cat assembly_minimus.fasta assembly_minimus.singletons.seq > all_sequences.fasta
127 if [ ! -s assembly_minimus.fasta ] || [ ! -s all_sequences.fasta ]; then
128     echo "ERROR: 454 (ens)- illumina (ens) Minimus reconciliation has failed"
129     exit 1 # Fail
130 fi
131 }
132
133 # To SCAFFOLDING illumina level
134 SC_reassembly_all){
135 # module load amos/3.1.0
136

```

```

137 # module load ruby
138 #create a file to track names in order to do the metrics
139 echo -e "aaMin2/scALL/454Cap3\tminimus\tSC_trs_all\t29/\$kmers\t454_Ill" | sed 's
    ↪ /[] []//g' > tracker
140
141 cat CAP3_454_k29)/reconciliation_454.fasta !JobRegExp:SC_remove_illumina_redundancy
    ↪ :~/all_sequences.fasta > input_minimus
142 fasta_standard_renamer.rb input_minimus
143 toAmos -s input_minimus_new -o assembly_minimus.afg
144 ?
145 minimus2 assembly_minimus
146 cat assembly_minimus.fasta assembly_minimus.singletons.seq > all_sequences.fasta
147 if [ ! -s assembly_minimus.fasta ] || [ ! -s all_sequences.fasta ]; then
148     echo "ERROR: 454 (ens)- illumina (ens) Minimus reconciliation has failed"
149     exit 1 # Fail
150 fi
151
152 }
153
154
155 # RAY reassembly 454, SOAP and Oases assemblies
156 #-----
157
158 Assembly_454_Illumina_25_[JobRegExp:Qual_Ill_prep:-]){
159     #module load ray/2.3.1
160     echo -e "aaRayK25/~cut -f1 (*)/tracker~/454Cap3\tRay\ttrs\t25\t454-Ill" > tracker
161     ?
162     mpiexec -np [cpu] Ray -k 25 -s Qual_454_prep)/all_sequences.fastq -s (*)/
    ↪ all_sequences.fastq # -route-messages -connection-type debruijn -routing-graph
    ↪ -degree 4 (Message routing for large number of cores)
163 if [ ! -s RayOutput/Contigs.fasta ] || [ ! -s RayOutput/Scaffolds.fasta ]; then
164     echo "ERROR: Ray Illumina ens - 454 ens assembly or scaffolding have failed"
165     exit 1 # Fail
166 fi
167 ln -s RayOutput/Scaffolds.fasta all_sequences.fasta
168 }
169
170 Assembly_454_Illumina_35_[JobRegExp:Qual_Ill_prep:-]){
171     #module load ray/2.3.1
172     echo -e "aaRayK35/~cut -f1 (*)/tracker~/454Cap3\tRay\ttrs\t35\t454-Ill" > tracker
173     ?
174     mpiexec -np [cpu] Ray -k 35 -s Qual_454_prep)/all_sequences.fastq -s (*)/
    ↪ all_sequences.fastq # -route-messages -connection-type debruijn -routing-graph
    ↪ -degree 4 (Message routing for large number of cores)
175 if [ ! -s RayOutput/Contigs.fasta ] || [ ! -s RayOutput/Scaffolds.fasta ]; then
176     echo "ERROR: Ray Illumina ens - 454 ens assembly or scaffolding have failed"
177     exit 1 # Fail
178 fi
179 ln -s RayOutput/Scaffolds.fasta all_sequences.fasta

```

```

180 }
181
182
183 ### 454 reads reassembly with Illumina assembly
184 #####
185
186 # MIRA Reassembly 454reads (one illumina ens to 454 reads), SOAP and Oases
187 #-----
188 reassembling_Illumina_454reads_[JobRegExp:Qual_Ill:-]){
189 # module load mira/4.0.2
190 # module load ruby
191 ln -s (*)/all_sequences.fasta orig.fasta
192 #create a file to track names in order to do the metrics
193 echo -e "arMIRA/`cut -f1 (*)/tracker`\tmira\trs_454read/Ill\t`cut -f4 (*)/tracker`\
↪ t454-Ill" > tracker
194 echo "
195 project = transcriptome_test
196 job = est,denovo,accurate
197 parameters = -GE:not=4 -NW:cmrnl=warn
198 # the 454 data
199 readgroup = Reads454
200 data = Cleanup_454)/sequences_.fastq
201 technology = 454
202 # the Illumina data
203 readgroup = ContigsIllumina
204 data = (*)/all_sequences.fastq
205 technology = sanger
206 " > mira_config_template
207 ?
208 mira mira_config_template
209 if [ ! -s transcriptome_test_assembly/transcriptome_test_d_results/
↪ transcriptome_test_out.unpadded.fasta ] ; then
210 echo "ERROR: 454 (reads) - Illumina (ens) MIRA assembly has failed"
211 exit 1 # Fail
212 fi
213 rm_codigo_degenerado.rb transcriptome_test_assembly/transcriptome_test_d_results/
↪ transcriptome_test_out.unpadded.fasta all_sequences.fasta
214 }
215 }
216
217 # RAY reassembly 454reads (one illumina ens to 454 reads), SOAP and Oases
218 #-----
219
220 Assembly_454reads_Illumina_25_[JobRegExp:Qual_Ill_prep:-]){
221 # module load ray/2.3.1
222 echo -e "arRayK25/`cut -f1 (*)/tracker`\tRay\trs_454read/Ill\t25\t454-Ill" > tracker
223 ?
224 mpiexec -np [cpu] Ray -k 25 -s Cleanup_454)/sequences_.fastq -s (*)/all_sequences.
↪ fastq # -route-messages -connection-type debruijn -routing-graph-degree 4 (

```

```

    ↪ Message routing for large number of cores)
225 if [ ! -s RayOutput/Contigs.fasta ] || [ ! -s RayOutput/Scaffolds.fasta ]; then
226     echo "ERROR: Ray Illumina ens - 454 reads primary assembly or scaffolding have
    ↪ failed"
227     exit 1 # Fail
228 fi
229 ln -s RayOutput/Scaffolds.fasta all_sequences.fasta
230 }
231
232
233 Assembly_454reads_Illumina_35_[JobRegExp:Qual_Ill_prep:-]){
234 # module load ray/2.3.1
235 echo -e "arRayK35/`cut -f1 (*)/tracker\tRay\trs_454read/Ill\t35\t454-Ill" > tracker
236 ?
237 mpiexec -np [cpu] Ray -k 35 -s Cleanup_454)/sequences_.fastq -s (*)/all_sequences.
    ↪ fastq # -route-messages -connection-type debruijn -routing-graph-degree 4 (
    ↪ Message routing for large number of cores)
238 if [ ! -s RayOutput/Contigs.fasta ] || [ ! -s RayOutput/Scaffolds.fasta ]; then
239     echo "ERROR: Ray Illumina ens - 454 reads primary assembly or scaffolding have
    ↪ failed"
240     exit 1 # Fail
241 fi
242 ln -s RayOutput/Scaffolds.fasta all_sequences.fasta
243 }
244
245
246 ### 454 reads with Illumina reads assembly
247 #####
248
249 Assembly_454_Illumina_reads_${kmers}){
250 # module load ray/2.3.1
251 echo -e "rrRayK(*)\tRay\tas_reads\t(*)\t454-Ill" > tracker
252     if [ "$ill_type" == "paired" ]; then
253         input_files='-p Cleanup_ill)/output_files/paired_1_.fastq Cleanup_ill)
    ↪ /output_files/paired_2_.fastq'
254     elif [ "$ill_type" == "single" ]; then
255         input_files='-s Cleanup_ill)/output_files/sequences_.fastq'
256     else
257         echo 'Read input files not set.'
258         exit 1
259     fi
260 ?
261 mpiexec -np [cpu] Ray -k (*) -s Cleanup_454)/sequences_.fastq $input_files # -route-
    ↪ messages -connection-type debruijn -routing-graph-degree 4 (Message routing for
    ↪ large number of cores)
262 if [ ! -s RayOutput/Contigs.fasta ] || [ ! -s RayOutput/Scaffolds.fasta ]; then
263     echo "ERROR: Ray Illumina reads - 454 reads primary assembly or scaffolding have
    ↪ failed"
264     exit 1 # Fail

```

```
265 fi  
266 ln -s RayOutput/Scaffolds.fasta all_sequences.fasta  
267 }
```

Listado de Código B.5: (MÓDULO 4) *transcriptome\_reference\_Illumina\_plants.txt*

```

1  #= References transcriptome de novo of mature pollen from Olea europaea
2  #####
3
4  #= References for de novo transcriptome assembly from Illumina and 454 single-end
   ↪ reads.
5  #= This workflow has been designed to be used with plants.
6  #= Author: Rosario Carmona
7  #####
8  #Variables
9  #####
10
11
12 #####
13 ##### REFERENCES
14 #####
15
16 reference_arab){
17     ln -s $reference/Athaliana_167_TAIR10.transcript all_sequences.fasta
18     ?
19     echo -e "A.thaliana\tref_p\tref_t\tref_k\tref_t" > tracker
20 }
21
22 reference_populus){
23     ln -s $reference/Ptrichocarpa_444_v3.1.transcript.fa all_sequences.fasta
24     ?
25     echo -e "P.trichocarpa\tref_p\tref_t\tref_k\tref_t" > tracker
26 }
27
28
29 #####
30 ##### METRICS BLOCK
31 #####
32
33 FLN_metric_Arab){
34     #resources: -s -u 3 -n cal -c 48 -t '7-00:00:00' -m '20gb'
35     ln -s reference_arab)/tracker
36     ?
37     full_lengther_next -s 10.243 -f reference_arab)/all_sequences.fasta -a 's' -z -g
   ↪ $FLN_DATABASE -c 500 -q 'd' -w [lcpu] -M '$reads/arab_ill_1.fastq,$reads/
   ↪ arab_ill_2.fastq'
38
39     if [ ! -s fln_results/summary_stats.txt ] ; then
40         echo "ERROR: full_lengther_next has failed"
41         exit 1 # Fail
42     fi
43
44     #making metric file
45     echo -e "`cat tracker`\tMeanContigCov\t\ttail +2 fln_results/

```

```

46 ↪ coverage_per_transcript.txt | awk '{ if ($2 > 0){ sum += $2; n++; } END { if (n
47 ↪ > 0) print sum / n; }' " >> ../metrics
48 input_seqs=`grep -w "input_seqs" fln_results/summary_stats.txt | cut -f 1`
49 unmapped=`grep -w "unmapped" fln_results/summary_stats.txt | cut -f 1`
50 mapped_contigs=$((input_seqs - $unmapped))
51 echo -e "`cat tracker`\tContigs\t$mapped_contigs" >> ../metrics
52 input_seqs500=`grep -w "PRE_FLN_sequences_>500" fln_results/summary_stats.txt |
53 ↪ cut -f 1`
54 unmapped500=`awk '{if ($2 >= 500) sum +=1} END {print sum}' fln_results/unmapped.
55 ↪ txt`
56 mapped_contigs500=$((input_seqs500 - $unmapped500))
57 echo -e "`cat tracker`\tContigs500\t$mapped_contigs500" >> ../metrics
58
59 fln_absolute_metric=( 'input_seqs' 'PRE_FLN_mean_length' 'longest_unigene' '
60 ↪ PRE_FLN_full_transcriptome_length' 'PRE_FLN_n50' 'PRE_FLN_n90' '
61 ↪ PRE_FLN_indeterminations' 'PRE_FLN_indetermination_mean_length' 'prot_annotated
62 ↪ ' 'complete' 'coding' 'different_orthologues' 'different_completes' )
63 metric_names=( 'input_seqs' 'MeanContigLen' 'longest_unigene' 'AllTransSize' 'N50'
64 ↪ 'N90' 'Ns' 'MeanGapLen' 'prot_annotated' 'complete' 'coding' 'DiffProts' '
65 ↪ DiffComplProts' )
66 position=0
67 for i in "${fln_absolute_metric[@]}"
68 do
69     command=`grep -w "$i" fln_results/summary_stats.txt | cut -f 1`
70     echo -e "`cat tracker`\t${metric_names[$position]}\t$command" >> ../metrics
71     position=$((position + 1))
72 done
73
74 fln_stats_metric=( 'misassembled' 'chimeras' )
75 metric_names=( 'MissAssembl' 'chimeras' )
76 position=0
77 total=`grep "prot_annotated" fln_results/summary_stats.txt | cut -f 1`
78 for i in "${fln_stats_metric[@]}"
79 do
80     command=`grep -w "$i" fln_results/summary_stats.txt | cut -f 1`
81     echo -e "`cat tracker`\t${metric_names[$position]}\t`echo "scale=4; $command
82 ↪ *100/($total+$command)" | bc`" >> ../metrics
83     position=$((position + 1))
84 done
85 }
86
87 FLN_metric_Pop){
88     #resources: -s -u 3 -n cal -c 48 -t '7-00:00:00' -m '20gb'
89     ln -s reference_populus)/tracker
90     ?
91     full_lengther_next -s 10.243 -f reference_populus)/all_sequences.fasta -a 's' -z
92     ↪ -g $FLN_DATABASE -c 500 -q 'd' -w [lcpu] -M '$reads/populus_ill_1.fastq,$reads/
93     ↪ populus_ill_2.fastq'

```

```

83  if [ ! -s fln_results/summary_stats.txt ] ; then
84      echo "ERROR: full_lengther_next has failed"
85      exit 1 # Fail
86  fi
87
88  #making metric file
89  echo -e "`cat tracker`\tMeanContigCov\t\tail +2 fln_results/
↪ coverage_per_transcript.txt | awk '{ if ($2 > 0){ sum += $2; n++} } END { if (n
↪ > 0) print sum / n; }'`" >> ../metrics
90  input_seqs=`grep -w "input_seqs" fln_results/summary_stats.txt | cut -f 1`
91  unmapped=`grep -w "unmapped" fln_results/summary_stats.txt | cut -f 1`
92  mapped_contigs=$((input_seqs - $unmapped))
93  echo -e "`cat tracker`\tContigs\t$mapped_contigs" >> ../metrics
94  input_seqs500=`grep -w "PRE_FLN_sequences_>500" fln_results/summary_stats.txt |
↪ cut -f 1`
95  unmapped500=`awk '{if ($2 >= 500) sum +=1} END {print sum}' fln_results/unmapped.
↪ txt`
96  mapped_contigs500=$((input_seqs500 - $unmapped500))
97  echo -e "`cat tracker`\tContigs500\t$mapped_contigs500" >> ../metrics
98
99  fln_absolute_metric=( 'input_seqs' 'PRE_FLN_mean_length' 'longest_unigene' '
↪ PRE_FLN_full_transcriptome_length' 'PRE_FLN_n50' 'PRE_FLN_n90' '
↪ PRE_FLN_indeterminations' 'PRE_FLN_indetermination_mean_length' 'prot_annotated
↪ ' 'complete' 'coding' 'different_orthologues' 'different_completes' )
100 metric_names=( 'input_seqs' 'MeanContigLen' 'longest_unigene' 'AllTransSize' 'N50'
↪ 'N90' 'Ns' 'MeanGapLen' 'prot_annotated' 'complete' 'coding' 'DiffProts' '
↪ DiffComplProts' )
101 position=0
102 for i in "${fln_absolute_metric[@]}"
103 do
104     command=`grep -w "$i" fln_results/summary_stats.txt | cut -f 1`
105     echo -e "`cat tracker`\t${metric_names[$position]}\t${command}" >> ../metrics
106     position=$((position + 1))
107 done
108
109 fln_stats_metric=( 'misassembled' 'chimeras' )
110 metric_names=( 'MissAssembl' 'chimeras' )
111 position=0
112 total=`grep "prot_annotated" fln_results/summary_stats.txt | cut -f 1`
113 for i in "${fln_stats_metric[@]}"
114 do
115     command=`grep -w "$i" fln_results/summary_stats.txt | cut -f 1`
116     echo -e "`cat tracker`\t${metric_names[$position]}\t`echo "scale=4; $command
↪ *100/($total+$command)" | bc`" >> ../metrics
117     position=$((position + 1))
118 done
119 }

```

**Listado de Código B.6:** (MÓDULO 5) *transcriptome\_validation\_454\_single\_Illumina.txt*

```

1  #= Validation Autoflow transcriptome de novo of mature pollen from Olea europaea
2  #####
3
4  #= Validation of de novo transcriptome assembly from Illumina and 454 single-end reads
5  ↪ .
6  #= This workflow has been designed to be used in plants.
7  #= Author: Rosario Carmona
8  #####
9
10 #####
11 ##### METRICS BLOCK
12 #####
13
14 calculating_BUSCO_metric_[JobRegExp:primary|scaffolding|join|reduce|merge|MIRA|Euler|
15 ↪ CAP3|reassembly_454|reassembly_all|reassembling_Illumina|reference|Assembly|
16 ↪ remove_illumina_redundancy:-)]{
17 #resources: -n cal -c 10 -t '15:00:00' -m '30gb'
18 #module load busco/2.0
19 ln -s (*)/all_sequences.fasta
20 fasta_standard_renamer.rb all_sequences.fasta
21 ln -s (*)/tracker
22 var_tracker=`cat tracker`
23 ?
24 BUSCO.py -i all_sequences.fasta_new -l $BUSCO_db_path -o BUSCO_analysis -m tran -f
25 ↪ -c [cpu]
26 if [ ! -s run_BUSCO_analysis/short_summary_BUSCO_analysis.txt ] ; then
27     echo "ERROR: BUSCO has failed"
28     exit 1 # Fail
29 else
30     rm -rf run_BUSCO_analysis/hmmer_output run_BUSCO_analysis/translated_proteins
31 fi
32 grep -r -o 'C:[0-9\.]*\%[S:[0-9\.]*\%,D:[0-9\.]*\%,F:[0-9\.]*\%'
33 ↪ run_BUSCO_analysis/short_summary_BUSCO_analysis.txt | sed -r "s|C:([0-9\.]*)
34 ↪ \%[S:[0-9\.]*\%,D:([0-9\.]*)\%,F:([0-9\.]*)\%|$var_tracker\tComp1Ortho\t1\
35 ↪ n$var_tracker\tDuplOrtho\t2\n$var_tracker\tFragOrtho\t3|g" >> ../metrics
36 }
37
38 FLN_metric_[JobRegExp:primary|scaffolding|join|reduce|merge|MIRA|Euler|CAP3|
39 ↪ reassembly_454|reassembly_all|reassembling_Illumina|Assembly|
40 ↪ remove_illumina_redundancy:-)]{
41 #resources: -u 1 -s -n cal -c 16 -t '7-00:00:00' -m '30gb'
42 ln -s (*)/tracker
43 if [ "$ill_type" == "paired" ]; then
44     files_illumina='Cleanup_ill)/output_files/paired_1_.fastq,Cleanup_ill)
45 ↪ /output_files/paired_2_.fastq'
46 elif [ "$ill_type" == "single" ]; then
47     files_illumina="Cleanup_ill)/output_files/sequences_.fastq"

```

```

39         else
40             echo 'Read input files not set.'
41             exit 1
42         fi
43     files_454="Cleanup_454)/sequences_.fastq"
44     ?
45     full_lengther_next -s 10.243 -f (*)/all_sequences.fasta -a 's' -z -g $FLN_DATABASE
46     ↪ -c 5000 -q 'd' -w [lcpu] -M "$files_illumina;$files_454"
47     if [ ! -s fln_results/summary_stats.txt ] ; then
48         echo "ERROR: full_lengther_next has failed"
49         exit 1 # Fail
50     fi
51
52     #making metric file
53     echo -e "`cat tracker`\tMeanContigCov\t`tail +2 fln_results/
54     ↪ coverage_per_transcript.txt | awk '{ if ($2 > 0){ sum += $2; n++; } } END { if (n
55     ↪ > 0) print sum / n; }`" >> ../metrics
56     input_seqs=`grep -w "input_seqs" fln_results/summary_stats.txt | cut -f 1`
57     unmapped=`grep -w "unmapped" fln_results/summary_stats.txt | cut -f 1`
58     mapped_contigs=$((input_seqs - $unmapped))
59     echo -e "`cat tracker`\tContigs\t$mapped_contigs" >> ../metrics
60     input_seqs500=`grep -w "PRE_FLN_sequences_>500" fln_results/summary_stats.txt |
61     ↪ cut -f 1`
62     unmapped500=`awk '{if ($2 >= 500) sum +=1} END {print sum}' fln_results/unmapped.
63     ↪ txt`
64     mapped_contigs500=$((input_seqs500 - $unmapped500))
65     echo -e "`cat tracker`\tContigs500\t$mapped_contigs500" >> ../metrics
66
67     fln_absolute_metric=( 'input_seqs' 'PRE_FLN_mean_length' 'longest_unigene' '
68     ↪ PRE_FLN_full_transcriptome_length' 'PRE_FLN_n50' 'PRE_FLN_n90' '
69     ↪ PRE_FLN_indeterminations' 'PRE_FLN_indetermination_mean_length' 'prot_annotated
70     ↪ ' 'complete' 'coding' 'different_orthologues' 'different_completes' )
71     metric_names=( 'input_seqs' 'MeanContigLen' 'longest_unigene' 'AllTransSize' 'N50'
72     ↪ 'N90' 'Ns' 'MeanGapLen' 'prot_annotated' 'complete' 'coding' 'DiffProts' '
73     ↪ DiffComplProts' )
74     position=0
75     for i in "${fln_absolute_metric[@]}"
76     do
77         command=`grep -w "$i" fln_results/summary_stats.txt | cut -f 1`
78         echo -e "`cat tracker`\t${metric_names[$position]}\t$command" >> ../metrics
79         position=$((position + 1))
80     done
81
82     fln_stats_metric=( 'misassembled' 'chimeras' )
83     metric_names=( 'MissAssembl' 'chimeras' )
84     position=0
85     total=`grep "prot_annotated" fln_results/summary_stats.txt | cut -f 1`
86     for i in "${fln_stats_metric[@]}"
87     do

```

```

78     command=`grep -w "$i" fln_results/summary_stats.txt | cut -f 1`
79     echo -e "`cat tracker`\t${metric_names[$position]}`\t`echo "scale=4; $command
↪ *100/($total+$command)" | bc`" >> ../metrics
80     position=$((position + 1))
81     done
82 }
83
84 table_report){
85     #module load R/3.3.1
86     echo " !JobRegExp:metric:!/algo "
87     REFS=`echo $key_organisms | tr '_' ','`
88     ?
89     create_metric_table.rb ../metrics "Name,Program,Task,Kmer,Technology"
90     sed 's/>//g' metric_table.txt -i
91     sed 's/PRE_FLN_//g' metric_table.txt -i
92     ## PCA
93     #####
94
95     SUPPLEMENTARY_VARS="`head -n 1 metric_table.txt | cut -f 2,3,4,5|tr "\t" "="|sed "
↪ s/=/' , '/g`"
96
97     # Using general vars
98     # -----
99     echo "general vars"
100    echo "-----"
101    PCA_variables="
102    'Contigs',
103    'Contigs500',
104    'MeanContigLen',
105    'N50',
106    'N90',
107    'Ns',
108    'MeanGapLen',
109    'ComplOrtho',
110    'FragOrtho',
111    'DuplOrtho',
112    'MissAssembl',
113    'chimeras',
114    'DiffProts',
115    'DiffComplProts',
116    'MeanContigCov',
117    'AllTransSize'
118    "
119    factor_analysis.R -d metric_table.txt -l "$SUPPLEMENTARY_VARS" -c "$PCA_variables"
↪ -o general_pca.pdf -r $REFS > PCA_general_results
120    parse_pca.rb PCA_general_results > PCA_general_results_parsed
121    report_html -t $report_template -d metric_table.txt,PCA_general_results_parsed -o
↪ assembly_report
122 }

```



## Apéndice C

Script de lanzamiento y plantilla de  
AutoFlow de *findRGs* para el  
desarrollo de marcadores de RT-PCR



**Listado de Código C.1:** *Script de lanzamiento launch\_findRGs.sh*

```
1 #!/usr/bin/env bash
2
3 . ~soft_cvi_114/initializes/init_autoflow
4
5
6 AutoFlow -w findRGs -V '$input_file=read_count_table,$min_reads=[10000;30000;100000],
  ↪ $cv_filter=[10,20],$annot=annotation_table' -o output_findRGs
```

Listado de Código C.2: *findRGS.txt*

```

1 #####
2 #= AUTOFLOW TEMPLATE findRGS
3 #####
4 #= Detecting candidate reference genes from NGS data
5 #= INPUT = a read count table and a tab-delimited annotation table (optional)
6 #= Version = 1.0
7 #= Author = Rosario Carmona
8
9 #####
10 # Declaring variables
11 #####
12 #$input_file=
13 #$min_reads=
14 #$cv_filter=
15 #$annot=
16 #####
17
18 filter_coef_var_$cv_filter){
19
20   filter_reads_$min_reads){
21     module load ruby
22     ?
23
24     #Normalize by RPMM (reads per mapped million)
25     rpmm_itera.rb $input_file > rpmm_table
26     #Filter by minimum counted reads
27     min_reads.rb -f $input_file -m (*) > genes_filtered
28     #Link filtered genes and rpmm
29     table_linker.rb rpmm_table genes_filtered final_rpmm
30     #Calculate coefficient of variation and filters
31     coef_vari.rb -t final_rpmm -c filter_coef_var_(+) > coef_var_table
32
33
34     if [ "$annot" ]
35
36       #If annotation file exists
37       then
38       #Prepare input for filtering
39       table_header.rb -t coef_var_table -c 0,2 > gen_rpmm_table
40       table_linker.rb $annot gen_rpmm_table gen_rpmm_annot
41       #Filter by best ortholog
42       extract_best.rb gen_rpmm_annot > uniq_ort_table
43       table_linker.rb coef_var_table uniq_ort_table final_table
44
45       #Extract list of orthologs for intersections
46       table_header.rb -t final_table -c 1 > genes_list
47
48       #If annotation file doesn't exist

```

```

49     else
50         table_header.rb -t coef_var_table -c 0 > genes_list
51
52     fi
53
54 }
55
56     #Node of intersections
57 venn){
58     module load R/3.0.2
59     #Prepare file paths
60     files_list=`echo !filter_reads_!/genes_list |sed 's/ /,/g'`
61
62     #Prepare set names
63     set_names=`echo "$min_reads" | sed 's/\([ \(\) \[ \] \) //g' | sed 's/;/\n/g' | sed 's/$/
↪ _reads/' | sed ':a;N;$!ba;s/\n/,/g'`
64     ?
65     script_venn.R -f $files_list -n $set_names
66
67 }
68
69 }

```



## Apéndice D

Script de lanzamiento y plantilla de  
AutoFlow para la detección  
bioinformática de posibles nuevos  
alérgenos de olivo



**Listado de Código D.1:** *Script de lanzamiento launch\_findAllergens.sh*

```
1 #!/usr/bin/env bash
2
3 . ~/soft_cvi_114/initializes/init_autoflow
4
5
6 AutoFlow -w findAllergens -V '$allergen_olea=../../alergen_olea_11_7_2017.fasta,
  ↳ $transc_pollen=../../Total_transcriptome_pollen_v11.fasta,$allergens_noOlea
  ↳ =../../all_allergens_noOlea_11_7_2017.fasta,$perc_identity=40,$perc_align=40,
  ↳ $olea_secure_identity=80,$olea_secure_alignment=80,$evalue=1e-1,
  ↳ $control_identity=60,$control_alignment=60' -o output_findAllergens $1
```

Listado de Código D.2: *findAllergens.txt*

```

1 #####
2 #= AUTOFLOW TEMPLATE findAllergens
3 #####
4 #= Detecting allergens from NGS data
5 #= INPUT = fasta file of the olive pollen transcriptome (nucleotides); fasta file with
   ↳ olive tree allergens (amino acids); fasta file with non-olive allergens (amino
   ↳ acids)
6 #= Version = 1.0
7 #= Author = Rosario Carmona
8
9 #####
10 # Declaring variables
11 #####
12 #$allergen_olea=
13 #$allergens_noOlea=
14 #$transc_pollen=
15 #$evalue=
16 #$olea_secure_identity=
17 #$olea_secure_alignment=
18 #$perc_identity=
19 #$perc_align=
20 #$control_identity=
21 #$control_alignment=
22 #####
23
24
25
26 #####
27 ## PREPARE DATABASES
28 #####
29
30 making_databases){
31     module load blast_plus/2.2.29+
32     module load ruby
33     mkdir allergen_olea_db
34     mkdir allergens_noOlea_db
35     ?
36     makeblastdb -in $allergen_olea -dbtype prot -parse_seqids -out allergen_olea_db/
   ↳ allergen_olea_db
37     makeblastdb -in $allergens_noOlea -dbtype prot -parse_seqids -out
   ↳ allergens_noOlea_db/allergens_noOlea_db
38 }
39
40 #####
41 ## BLAST WITH OLE ALLERGENS
42 #####
43
44 ole_blast){

```

```

45 module load blast_plus/2.2.29+
46 module load ruby
47 ?
48 blastx -query $transc_pollen -db making_databases)/allergen_olea_db/allergen_olea_db
    ↪ -evalue $evalue -outfmt '7 qseqid sacc pident length mismatch gapopen qstart
    ↪ qend sstart send evalue bitscore score qframe sframe qseq sseq qlen slen stitle
    ↪ ' > blast_result
49 blast_filter_manipulated.rb -b blast_result -p $olea_secure_identity -s
    ↪ $olea_secure_alignment > blast_result_filtered
50
51 #Separate transcripts with/without ole match
52 cut -f 1 blast_result_filtered | sort | uniq > list_unigenes_ole
53 grep '>' $transc_pollen | sed 's/>//g' > lista_unigenes_total
54 sort list_unigenes_ole lista_unigenes_total | uniq -u > list_unigenes_noOle
55 lista_to_fasta.rb $transc_pollen list_unigenes_noOle > transc_pollen_noOle.fasta
56 lista_to_fasta.rb $transc_pollen list_unigenes_ole > transc_pollen_ole.fasta
57 }
58
59 #####
60 ## BLAST BETWEEN OLE AND NO OLE ALLERGENS (CROSS-REACTIVITY)
61 #####
62
63 cross_reactivity){
64 module load blast_plus/2.2.29+
65 module load ruby
66 ?
67 blastx -query ole_blast)/transc_pollen_ole.fasta -db making_databases)/
    ↪ allergens_noOlea_db/allergens_noOlea_db -evalue $evalue -outfmt '7 qseqid sacc
    ↪ pident length mismatch gapopen qstart qend sstart send evalue bitscore score
    ↪ qframe sframe qseq sseq qlen slen stitle' > blast_result
68
69 #Filtering blast
70 blast_filter_manipulated.rb -b blast_result -p 50 -s 100 > blast_result_filtered_1
71 blast_filter_manipulated.rb -b blast_result -p 35 -a 80 > blast_result_filtered_2
72 blast_filter_manipulated.rb -b blast_result -p 100 -a 8 > blast_result_filtered_3
73
74 cat blast_result_filtered* | sort | uniq > blast_result_filtered
75 }
76
77 #####
78 ## BLAST WITH NO OLE ALLERGENS AND CONTROL BLAST WITH OLE ALLERGENS
79 #####
80
81 no_ole_blast){
82 module load blast_plus/2.2.29+
83 module load ruby
84 ?
85 blastx -query ole_blast)/transc_pollen_noOle.fasta -db making_databases)/
    ↪ allergens_noOlea_db/allergens_noOlea_db -evalue $evalue -outfmt '7 qseqid sacc

```

```

86     ↪ pident length mismatch gapopen qstart qend sstart send evalue bitscore score
87     ↪ qframe sframe qseq sseq qlen slen stitle' > blast_result
88 blast_filter_manipulated.rb -b blast_result -p $perc_identity -s $perc_align >
89     ↪ blast_result_filtered
90
91 #Prepare transcripts with allergen match
92 cut -f 1 blast_result_filtered | sort | uniq > list_unigenes_withAllergen
93 lista_to_fasta.rb $transc_pollen list_unigenes_withAllergen > unigenes_withAllergen.
94     ↪ fasta
95 }
96
97 control_blast){
98     module load blast_plus/2.2.29+
99     module load ruby
100     ?
101     blastx -query no_ole_blast)/unigenes_withAllergen.fasta -db making_databases)/
102     ↪ allergen_olea_db/allergen_olea_db -evalue $evalue -outfmt '7 qseqid sacc pident
103     ↪ length mismatch gapopen qstart qend sstart send evalue bitscore score qframe
104     ↪ sframe qseq sseq qlen slen stitle' > blast_result
105 blast_filter_manipulated.rb -b blast_result -p $control_identity -s
106     ↪ $control_alignment > blast_result_filtered
107 }

```

## Apéndice E

### Script de parseo e intersección de las salidas de los programas de predicción de *S*-nitrosilación



Listado de Código E.1: *snitro\_parse.sh*

```

1  #!/usr/bin/env bash
2
3  #####
4  #1.- Declaring variables
5  #####
6  #Define input files (e.g. SNOSite_output=SNOSite_output1.htm,SNOSite_output2.htm)
7  iSNOAAPair_output=
8  GPSSNO_output=
9  SNOSite_output=
10 annotation=
11
12
13 #####
14 #2.- PARSING iSNO-AAPair output
15 #####
16
17 echo $iSNOAAPair_output | sed 's/,/\n/g' > list_files
18
19 #Extract protein identifier and S-nitrosylation position
20
21 for i in `cat list_files`; do
22
23     dos2unix $i
24     sed 's/protein_seq_/\nprotein_seq_/g' $i > file1
25     grep -v 'No Predicted' file1 > file2
26     grep 'protein_seq_' file2 > file3
27     sed -i 's/<tr align="center"><td><font color="#0000FF" face="Courier">/\n/g' file3
28     sed -i 's/<\font>.*;/g' file3
29     sed 's/<\td>.*//g' file3 | sed 's/~>/|' | sed 's/>protein_seq_//g' > file4
30     sed ':a;N;$!ba;s/\n>/\t/g' file4 | sed 's/;\t;/g' | sed 's/.$//g' > parsed_$i.txt
31
32     rm file*
33
34 done
35
36 cat parsed*.txt > list_iSNOAAPair_protein_position.txt
37
38
39 #Extract list of putative S-nitrosylated proteins
40
41 cut -f 1 list_iSNOAAPair_protein_position.txt > list_iSNOAAPair_protein.txt
42
43
44 rm list_files parsed*
45
46 #####
47
48

```

```

49 #####
50 #3.- PARSING GPS-SNO output
51 #####
52
53 echo $GPSSNO_output | sed 's/,/\n/g' > list_files
54
55 #Extract protein identifier and S-nitrosylation position
56
57 for i in `cat list_files`; do
58
59     dos2unix $i
60     cut -f 1 $i | sed '1d' | sed ':a;N;$!ba;s/\n>/\t>/g' | sed 's/>.*\t>/>/g' > output
61     sed -i 's/\t>/\n>/g' output
62
63     if [[ `tail -n 1 output` == ">"* ]]
64     then
65         sed -i '$d' output
66     fi
67
68
69     while read line; do
70         if [[ $line == [0-9]* ]]
71         then
72             echo $line";"
73         else
74             echo $line
75         fi
76     done < output > file
77
78     sed -i ':a;N;$!ba;s/\n/\t/g' file
79     sed -i 's/>protein_seq_/\n/g' file
80     sed 's/;\t/;/g' file | sed '1d' | sed 's/.$//g' > parsed_$i.txt
81
82     rm output file
83
84 done
85
86 cat parsed*.txt > list_GPSSNO_protein_position.txt
87
88
89 #Extract list of putative S-nitrosylated proteins
90
91 cut -f 1 list_GPSSNO_protein_position.txt > list_GPSSNO_protein.txt
92
93
94 rm list_files parsed*
95
96 #####
97

```

```

98
99 #####
100 #4.- PARSING SNOsite output
101 #####
102
103 echo $SNOsite_output | sed 's/,/\n/g' > list_files
104
105 # Extract protein identifier and S-nitrosylation position
106
107 for i in `cat list_files`; do
108
109     dos2unix $i
110     sed ':a;N;$!ba;s/\n//g' $i > output
111     sed -i 's/>protein_seq/\n>protein_seq/g' output
112     sed -i '1d' output
113     sed -i 's/<\font><\div><\b><\td><td><b><div align="center"><font color
↪ ="#999999" face="Courier New, Courier, mono" size="2">/\t/g' output
114     cut -f 1,2 output | sed 's/>protein_seq_//g' > position
115     cut -f 1 position | sort | uniq > id
116
117     # Concentrate S-nitrosylation positions per protein
118     for j in `cat id`; do
119         grep "${j}" position >> file1
120         cut file1 -f 2 | sed ':a;N;$!ba;s/\n//g' > file2
121         echo -e "${j}\t`cat file2`" >> parsed_${i}.txt
122         rm file*
123     done
124
125     rm output position id
126
127 done
128
129 cat parsed*.txt > list_SNOsite_protein_position.txt
130
131
132 #Extract list of putative S-nitrosylated proteins
133
134 cut -f 1 list_SNOsite_protein_position.txt > list_SNOsite_protein.txt
135
136
137 rm list_files parsed*
138
139 #####
140
141 #####
142 #5.- INTERSECTING lists of proteins
143 #####
144
145

```

```

146 #Extract list of common proteins to all tools
147
148 cat list_iSNOAAPair_protein.txt list_GPSSNO_protein.txt| sort| uniq -d| cat -
    ↪ list_SNOSite_protein.txt| sort| uniq -d > final_common_protein.txt
149
150 #####
151
152
153 #####
154 #6.- ADDING ANNOTATION DATA
155 #####
156
157 #Add annotation data to final list of proteins
158
159 if [ "$annotation" ]
160 then
161     sort -k 1 $annotation > annotation_sort.txt
162     join final_common_protein.txt annotation_sort.txt -t $'\t' >
    ↪ final_common_protein_annot.txt
163     rm annotation_sort.txt
164 fi
165
166 #####

```

## Apéndice F

### Publicaciones derivadas de la tesis

- F.1. ReprOlive: a database with linked data for the olive tree (*Olea europaea* L.) reproductive transcriptome (publicación aval)

# ReprOlive: a database with linked data for the olive tree (*Olea europaea* L.) reproductive transcriptome

Rosario Carmona<sup>1,2†</sup>, Adoración Zafra<sup>1†</sup>, Pedro Seoane<sup>3</sup>, Antonio J. Castro<sup>1</sup>, Darío Guerrero-Fernández<sup>2</sup>, Trinidad Castillo-Castillo<sup>4</sup>, Ana Medina-García<sup>4</sup>, Francisco M. Cánovas<sup>3</sup>, José F. Aldana-Montes<sup>4</sup>, Ismael Navas-Delgado<sup>4</sup>, Juan de Dios Alché<sup>1</sup> and M. Gonzalo Claros<sup>2,3\*</sup>

<sup>1</sup> Department of Biochemistry, Cell and Molecular Biology of Plants, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Granada, Spain, <sup>2</sup> Plataforma Andaluza de Bioinformática, Edificio de Bioinnovación, Universidad de Málaga, Málaga, Spain, <sup>3</sup> Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Málaga, Spain, <sup>4</sup> Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Málaga, Spain

## OPEN ACCESS

### Edited by:

Ariel Orellana,  
Universidad Andres Bello, Chile

### Reviewed by:

Matthew R. Willmann,  
University of Pennsylvania, USA  
Pere Puigdomènech,  
Consejo Superior de Investigaciones  
Científicas, Spain

### \*Correspondence:

M. Gonzalo Claros,  
Departamento de Biología Molecular  
y Bioquímica, Facultad de Ciencias,  
Universidad de Málaga, Campus  
de Teatinos, 29071 Málaga, Spain  
claros@uma.es

<sup>†</sup> These authors have contributed  
equally to this work.

### Specialty section:

This article was submitted to  
Plant Genetics and Genomics,  
a section of the journal  
Frontiers in Plant Science

**Received:** 21 April 2015

**Accepted:** 28 July 2015

**Published:** 11 August 2015

### Citation:

Carmona R, Zafra A, Seoane P,  
Castro AJ, Guerrero-Fernández D,  
Castillo-Castillo T, Medina-García A,  
Cánovas FM, Aldana-Montes JF,  
Navas-Delgado I, Alché JD  
and Claros MG (2015) ReprOlive:  
a database with linked data for the  
olive tree (*Olea europaea* L.)  
reproductive transcriptome.  
Front. Plant Sci. 6:625.  
doi: 10.3389/fpls.2015.00625

Plant reproductive transcriptomes have been analyzed in different species due to the agronomical and biotechnological importance of plant reproduction. Here we presented an olive tree reproductive transcriptome database with samples from pollen and pistil at different developmental stages, and leaf and root as control vegetative tissues (<http://reprolive.eez.csic.es>). It was developed from 2,077,309 raw reads to 1,549 Sanger sequences. Using a pre-defined workflow based on open-source tools, sequences were pre-processed, assembled, mapped, and annotated with expression data, descriptions, GO terms, InterPro signatures, EC numbers, KEGG pathways, ORFs, and SSRs. Tentative transcripts (TTs) were also annotated with the corresponding orthologs in *Arabidopsis thaliana* from TAIR and RefSeq databases to enable Linked Data integration. It results in a reproductive transcriptome comprising 72,846 contigs with average length of 686 bp, of which 63,965 (87.8%) included at least one functional annotation, and 55,356 (75.9%) had an ortholog. A minimum of 23,568 different TTs was identified and 5,835 of them contain a complete ORF. The representative reproductive transcriptome can be reduced to 28,972 TTs for further gene expression studies. Partial transcriptomes from pollen, pistil, and vegetative tissues as control were also constructed. ReprOlive provides free access and download capability to these results. Retrieval mechanisms for sequences and transcript annotations are provided. Graphical localization of annotated enzymes into KEGG pathways is also possible. Finally, ReprOlive has included a semantic conceptualisation by means of a Resource Description Framework (RDF) allowing a Linked Data search for extracting the most updated information related to enzymes, interactions, allergens, structures, and reactive oxygen species.

**Keywords:** olive, transcriptome, reproduction, database, pollen, pistil, annotation

## Introduction

Research in plant reproduction is accelerating rapidly as a direct consequence of the technological progresses (Dickinson and Franklin-Tong, 2011). Differential screening was initially used to identify abundant or specific transcripts of very specialized cells (Engel et al., 2003) and have been progressively replaced by the use of commercial microarrays and RNA-sequencing platforms. The

## F.2. AutoFlow: an easy way to build workflows

## AutoFlow: an easy way to build workflows

Pedro Seoane, Rosario Carmona, Rocío Bautista, Darío Guerrero-Fernández y M. Gonzalo Claros

Plataforma Andaluza de Bioinformática & Dpto de Biología Molecular y Bioquímica, Universidad de Málaga, 29071 Málaga (Spain)

**Abstract.** Many bioinformatics tasks require the use of different software, making workflows a current need in this research field. There are workflow builders that usually try to simplify the interface disregarding a complex use. This may lead to a non-scalability limitation, or the dependence on the facilities available. Here it is presented AutoFlow, a workflow builder that can handle most computer systems. It has been developed in Ruby and accepts any kind of software that can even use very specific resources (such as GPU or FPGA). AutoFlow has been designed to automatically launch tasks to the queue system. It can then handle big workflows that can overflow the maximum execution time of the queue system provided that each individual task can be finished within the maximum execution time. AutoFlow has been implemented with iterative task capability. Other interesting capability is an environment variable system that allows the persistence of certain data for all tasks. This allows the data transfer from a task to the next task and so on, enabling the inclusion of decisions that can affect downstream tasks. AutoFlow includes tools to monitor task status, graphic representations of workflows, file searching and timing. Two case-of use are presented to illustrate AutoFlow capabilities: one workflow for assembling and annotation of several libraries of Roche 454 sequences and another workflow for RNA-seq analysis.

### 1 Introduction

The high-throughput sequencing technologies produce a large amount of data that require the development of large and complex workflows with lots of instructions. For example, genome sequencing generates large files of reads that must be pre-processed, assembled, verified, and then annotated. Typically, assembly and annotation can be performed using various programs and parameters to obtain different results that require further reconciliation or selection. Some way of automation of this repetitive task will be beneficial. Moreover, when selection is required, downstream tools are depending on this decision, providing different results. Therefore, a workflow or a pipeline is designed to encompass all the programs and parameters used to get the final result, and allows the user to save time and efforts by not having to launch the different tasks on his own.

Pipelining tasks allows to automatise the use different software tools, where the output data of a tool is used as input for other tools, or different input files must

**F.3. Identification and *in silico* analysis of NADPH oxidase homologues involved in allergy from an olive pollen transcriptome**

## Identification and *in silico* Analysis of NADPH Oxidase Homologues Involved in Allergy from an Olive Pollen Transcriptome

María José Jiménez-Quesada<sup>1</sup>, Jose Ángel Traverso<sup>1,3</sup>, Adoración Zafra<sup>1</sup>,  
Jose C. Jimenez-Lopez<sup>1</sup>, Rosario Carmona<sup>1,2</sup>, M. Gonzalo Claros<sup>2</sup>,  
and Juan de Dios Alché<sup>1</sup>

<sup>1</sup> Estación Experimental del Zaidín (CSIC), Granada, Spain  
{mariajose.jimenez,dori.zafra,josecarlos.jimenez,  
juandedios.alche}@eez.csic.es

<sup>2</sup> Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain  
{rosario.carmona,claros}@uma.es

<sup>3</sup> Departamento Biología Celular. Universidad de Granada, Granada, Spain  
traverso@ugr.es

**Abstract.** Reactive oxygen species generated by pollen NADPH oxidases are present in numerous allergenic pollen species. The superoxide generated by this enzyme has been suggested as a key actor in the induction of allergic inflammation. However, this enzyme has been characterized in *Arabidopsis thaliana* pollen only, where two pollen-specific genes (*RbohH* and *RbohJ*) have been described. The olive (*Olea europaea* L.) pollen is an important source of allergy in Mediterranean countries. We have assembled and annotated an olive pollen transcriptome, which allowed us to determine the presence of at least two pollen-specific NADPH oxidase homologues. Primers were designed to distinguish between the two homologues, and full-length sequences were obtained through a PCR strategy. Complete *in silico* analysis of such sequences, including phylogeny, 3-D modeling of the N-terminus, and prediction of cellular localization and post-translational modifications was carried out with the purpose of shed light into the involvement of olive pollen-intrinsic NADPH oxidases in triggering allergy symptoms.

**Keywords:** allergy, NADPH oxidase, pollen, Rboh, ROS, superoxide.

### 1 Introduction

One of the most important causes of seasonal respiratory allergy in the Mediterranean area is olive pollen [1]. Allergens are proteins significantly expressed in pollen, many of them showing essential roles in pollen physiology. In the case of olive, twelve allergens have been identified and characterized to date [2].

Together with the classic view of the allergic inflammatory response triggered by pollen allergens, several studies have shown that pollen is also able to induce IgE-independent mast cell degranulation mediated by reactive oxygen species (ROS) [3].

F. Ortuño and I. Rojas (Eds.): IWBBIO 2015, Part I, LNCS 9043, pp. 450–459, 2015.  
© Springer International Publishing Switzerland 2015

**F.4. Identification of distinctive variants of the olive pollen allergen *Ole e 5* (Cu,Zn Superoxide Dismutase) throughout the analysis of the olive pollen transcriptome**

## Identification of Distinctive Variants of the Olive Pollen Allergen Ole e 5 (Cu,Zn Superoxide Dismutase) throughout the Analysis of the Olive Pollen Transcriptome

Adoración Zafra<sup>1</sup>, Rosario Carmona<sup>1,2</sup>, José C. Jimenez-Lopez<sup>1</sup>, Amada Pulido<sup>3</sup>,  
M. Gonzalo Claros<sup>2</sup>, and Juan de Dios Alché<sup>1</sup>

<sup>1</sup> Estación Experimental del Zaidín (CSIC), Granada, Spain

{dori.zafra, juandedios.alche, josecarlos.jimenez}@eez.csic.es

<sup>2</sup> Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain

{rosariocarmona, claros}@uma.es

<sup>3</sup> Departamento de Fisiología Vegetal, Universidad de Granada, Granada, Spain

amadapulido@ugr.es

**Abstract.** Ole e 5 is an olive pollen allergen displaying high identity with Cu,Zn superoxide dismutases. Previous studies characterized biochemical variability in this allergen, which may be of relevance for allergy diagnosis and therapy. The generation of an olive pollen transcriptome allowed us to identify eight Ole e 5 sequences, one of them including a 24 nt deletion. Further *in silico* analysis permitted designing primers for PCR amplification and cloning from both cDNA and gDNA. A large number of sequences were retrieved, which experimentally validated the predictive NGS sequences, including the deleted enzyme. The PCR-obtained sequences were used for further scrutiny, including sequence alignment and phylogenetic analysis. Two model sequences (a complete sequence and a deleted one) were used to perform 3-D modeling and a prediction of the T- and B-cell epitopes. These predictions interestingly foresee relevant differences in the antigenicity/allergenicity of both molecules. Clinical relevance of differences is discussed.

**Keywords:** allergen, deletion, Ole e 5, pollen, Superoxide Dismutase.

### 1 Introduction

Olive pollen is one of the most important causes of respiratory allergy in the Mediterranean area [1]. To date, twelve allergens have been identified in olive pollen, and one in olive mesocarp, named as Ole e 1 to Ole e 13 [2,3]. Several studies have shown that olive pollen allergens possess a relatively high level of microheterogeneity in their sequence, which has been extensively characterized for allergens Ole e 1, Ole e 2 and

**F.5. In the search of non described olive tree allergens by transcriptome-mining**

inhibitory effects which highlights the therapeutic potential of this agent for the treatment of allergic diseases.

*University of Kent, Chatham Maritime, ME4 4TB, UK.  
E-mail: B.F.Gibbs@kent.ac.uk*

### IN THE SEARCH FOR NON DESCRIBED OLIVE TREE ALLERGENS BY TRANSCRIPTOME-MINING

*R. Carmona, A. Zafra, J.C. Jiménez-López, P. Seoane, J.D. Alché, M.G. Claros*

The olive tree (*Olea europaea* L.) is an important crop plant in Spain and the Mediterranean basin, its pollen representing an important source of allergens. Although some transcriptomes have been obtained from vegetative tissues, the peculiarity of reproductive tissues in terms of gene expression deserves a dedicated study not only for biological reasons but also in the search for allergens. For this purpose, Sanger sequences and Roche/454 reads were obtained from pollen at different developmental stages. After pre-processing, sequences were assembled, and annotated using a complex workflow, including the corresponding orthologues in *Arabidopsis thaliana* from TAIR and RefSeq databases. Using AutoFlow, an automated workflow has been constructed to import the annotated pollen transcriptome (27,823 transcripts) into a freely-accessible, web-browseable database called ReprOlive (<http://reprolive.eez.csic.es/olivodb>). Functional annotations of such transcriptomes are offering the possibility of exploring the presence of new variants of already described allergens in the olive pollen. In this respect, distinctive variants of the olive pollen allergen Ole e 5 have been identified and pollen NADPH oxidase homologs involved in allergy inflammation have been identified and analysed (unpublished work). Furthermore, this annotated transcriptome also allows the assessment of the putative occurrence of not yet described allergens, which might account for the complex allergogram of the pollen of this species. Hence, a massive and strict BLAST comparison was carried out between ReprOlive pollen transcripts and UniProt allergens (including those belonging to the specific database ALLERGOME, <http://www.allergome.org>). A list of putative new allergens has been obtained and the identification and analysis of possible antigenic regions in these proteins is currently underway.

*Estación Experimental del Zaidín. Spanish Council for Scientific Research (CSIC). 18008 Granada, Spain. E-mail: juandedios.alche@eez.csic.es*

### EFFECT OF H<sub>1</sub>-ANTIHISTAMINES ON HISTAMINE H<sub>1</sub> RECEPTOR GENE EXPRESSION IN THE NASAL MUCOSA OF PATIENTS WITH POLLINOSIS INDUCED BY ARTIFICIAL EXPOSURE TO CEDAR POLLEN

*Y. Kitamura, H. Nakagawa, T. Fujii, T. Sakoda, T. Enomoto, H. Mizuguchi, H. Fukui, N. Takeda*

Stimulation of histamine H<sub>1</sub> receptors results in their up-regulation with a proliferation of receptor signalling. Elevation of histamine H<sub>1</sub> receptor mRNA has been observed in rat models of allergic rhinitis and treatment with H<sub>1</sub>-antihistamines reduce gene expression. Thus, the H<sub>1</sub> receptor is thought to act as an allergic rhinitis sensitive gene.

The effect of antihistamines were examined on histamine H<sub>1</sub> receptor mRNA in the nasal mucosa of pollinosis patients induced by artificial cedar pollen exposure. Symptom provocation was induced by artificial cedar pollen exposure using an environmental exposure unit. Histamine H<sub>1</sub> receptor mRNA was determined by real-time quantitative RT-PCR.

Artificial pollen exposure provoked nasal symptoms and elevation of histamine H<sub>1</sub> receptor mRNA expressions in the nasal mucosa of some pollinosis patients. Prophylactic antihistamine treatment improved symptoms and also suppressed the elevation of H<sub>1</sub> receptor mRNA.

The histamine H<sub>1</sub> receptor gene may play a role as a disease sensitive gene. Improvement of symptoms in pollinosis as well as suppression of H<sub>1</sub> receptor gene expression can be accomplished by use of H<sub>1</sub> antihistamines.

*Departments of Otolaryngology, of Molecular Pharmacology and of Molecular Studies for Incurable Diseases, Institute of Health Biosciences, Tokushima University Graduate School, Tokushima, 770-8503 Japan. E-mail: hfukui@tokushima-u.ac.jp*

### HISTAMINE MEDIATED REGULATION OF HUMAN MYELOID DENDRITIC CELL CYTOSKELETON AND FUNCTION

*M.B. Passani, E. Bonechi, A. Aldinucci, D. Nosi, E. Masini, C. Ballerini*

Histamine receptors (H<sub>1-4</sub>R) are differentially expressed in different cell types of the immune system. Among them, myeloid dendritic cells (DCs) were shown to be

**F.6. Transcriptome-based identification of a seed olive legumin (11S Globulin). Characterization of subunits, 3D modelling and molecular assessment of allergenicity**

## Transcriptome-Based Identification of a Seed Olive Legumin (11S Globulin). Characterization of Subunits, 3D Modelling and Molecular Assessment of Allergenicity

Adoración Zafra<sup>1</sup>, José Carlos Jimenez-Lopez<sup>1</sup>, Rosario Carmona<sup>1</sup>,  
Gonzalo Claros<sup>2</sup>, and Juan de Dios Alché<sup>1</sup>✉

<sup>1</sup> Plant Reproductive Biology Laboratory, Department of Biochemistry,  
Cell and Molecular Biology of Plants Estación Experimental Del Zaidín, CSIC, Granada, Spain  
{dori.zafra,josecarlos.jimenez,juandedios.alche}@eez.csic.es

<sup>2</sup> Department of Molecular Biology and Biochemistry, University of Málaga, Málaga, Spain  
{rosariocarmona,claros}@uma.es

**Abstract.** Seed storage proteins (SSPs) are fundamental molecules for seed germination as an important source of carbon and nitrogen. Among the main four protein families that integrate SSPs, legumins (11S globulins) are widely distributed in dicots, and represent the major contribution to the pool of seed proteins in olive. In the present study, we have used an olive seed transcriptome generated de novo by 454/Roche Titanium+ sequencing to identify a broad panel of 11S protein sequences. Among these identified legumin sequences, five were selected using their presence within the output results from the BLASTP alongside the whole NCBI database, and their clustering with previously-characterized 11S sequences in the phylogenetic analysis as the criteria. The selected sequences were identified as corresponding to the isoform 2 of the 11S protein precursor, and one of the sequences was used for further analysis. Individual acidic and basic subunits within this sequence were recognised, 3D-modeled and assessed as regard to their potential molecular interaction by docking methods. Furthermore, T-cell epitopes were forecast by using predictive software in order to evaluate the putative implications of the olive 11S proteins in food allergy. The potential use of this protein highly present in the olive seed as a food source is discussed.

**Keywords:** Electrostatic potential · Globulin · Legumin · Molecular modelling · Olive · T-cell epitopes · Transcriptome · 11S

### 1 Introduction

The olive tree (*Olea europaea* L.) is agriculturally and economically very important, mainly in the Mediterranean area where it represents 95 % of the total world production of green and black table olives (16 million tons) and olive oil (2.7 million tons). Among Mediterranean countries, Spain is the major producer of olive oil (44 % of the world production) [1, 2]. Olive oil is unavoidably rooted to a genuine diet that has been associated to protective features against cardiovascular disease and cancer [3–7].

**F.7. Automatic workflow for the identification of constitutively-expressed genes based on mapped NGS reads**

## Automatic Workflow for the Identification of Constitutively-Expressed Genes Based on Mapped NGS Reads

Rosario Carmona<sup>1</sup>, Pedro Seoane<sup>2</sup>, Adoración Zafra<sup>1</sup>, María José Jiménez-Quesada<sup>1</sup>, Juan de Dios Alché<sup>1</sup>, and M. Gonzalo Claros<sup>2</sup>✉

<sup>1</sup> Plant Reproductive Biology Laboratory, Department of Biochemistry, Cell and Molecular Biology of Plants, Estación Experimental Del Zaidín, CSIC, Granada, Spain  
{rosario.carmona,dori.zafra,mariajose.jimenez,juandedios.alche}@eez.csic.es

<sup>2</sup> Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain  
{seoanezonjic,claros}@uma.es

**Abstract.** Expression analyses such as quantitative and/or real-time PCR require the use of reference genes for normalization in order to obtain reliable assessments. The expression levels of these reference genes must remain constant in all different experimental conditions and/or tissues under study. Traditionally, housekeeping genes have been used for this purpose, but most of them have been reported to vary their expression levels under some experimental conditions. Consequently, the election of the best reference genes should be tested and validated in every experimental scenario. Microarray data are not always available for the search of appropriate reference genes, but NGS experiments are increasingly common. For this reason, an automatic workflow based on mapped NGS reads is presented with the aim of obtaining putative reference genes for a given species in the experimental conditions of interest. The calculation of the coefficient of variation (CV) and a simple, normalized expression value such as RPKM per transcript allows for filtering and selecting those transcripts expressed homogeneously and consistently in all analyzed conditions. This workflow has been tested with Roche/454 reads obtained from olive (*Olea europaea* L.) pollen and pistil at different developmental stages, as well as with Illumina paired-end reads from two different accessions of *Arabidopsis thaliana*. Some of the putative candidate reference genes have been experimentally validated.

**Keywords:** Housekeeping genes · Normalization · Real-time PCR · Olive (*Olea europaea* L.)

### 1 Introduction

For a long time, gene expression studies have been carried out by non-quantitative or semi-quantitative RNA gel blots and reverse transcription-polymerase chain reaction (RT-PCR) analyses. Later, real-time PCR revolutionized the field, because of its higher sensitivity, specificity and broad quantification range, in comparison to classical RT-PCR.

- F.8.** AutoFlow, a versatile workflow engine illustrated by assembling an optimised de novo transcriptome for a non-model species, such as faba bean (*Vicia faba*)

Send Orders for Reprints to [reprints@benthamscience.ae](mailto:reprints@benthamscience.ae)

Current Bioinformatics, 2016, 11, 000-000

1

## RESEARCH ARTICLE

## AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised *de novo* Transcriptome for a Non-Model Species, such as Faba Bean (*Vicia faba*)

Pedro Seoane<sup>1</sup>, Sara Ocaña<sup>2</sup>, Rosario Carmona<sup>3</sup>, Rocío Bautista<sup>3</sup>, Eva Madrid<sup>4</sup>, Ana M. Torres<sup>2</sup> and Manuel Gonzalo Claros<sup>1,3,\*</sup>

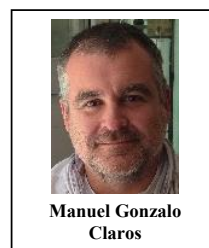
<sup>1</sup>Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, E-29071, Malaga, Spain; <sup>2</sup>Área de Mejora y Biotecnología, IFAPA Centro “Alameda del Obispo”, Apdo 3092, E-14080 Cordoba, Spain; <sup>3</sup>Plataforma Andaluza de Bioinformática, Universidad de Málaga, E-29071 Malaga, Spain and <sup>4</sup>Institute for Sustainable Agriculture, CSIC, Apdo 4084, E-14080 Cordoba, Spain

## ARTICLE HISTORY

Received: July 20, 2015  
Revised: September 30, 2015  
Accepted: October 2, 2015

DOI:  
10.2174/15748936116661602122351  
17

**Abstract:** The use of workflows to automate routine tasks is an absolute requirement in many bioinformatics fields. Current workflow manager systems usually compromise between providing a user-friendly interface and constructing complex, scalable pipelines. We present AutoFlow, a Ruby-based workflow engine devoid of graphic interface and tool repository, that is useful in most computer systems and most workflow requirements in any scientific field. It accepts any local or remote command-line software and converts one workflow into a series of independent tasks. It has been supplied with control patterns that allow for iterative task capability, supporting static and dynamic variables for decision-making or chaining workflows, as well as debugging utilities that include graphs, file searching, functional consistency and timing. Two proof-of-concept cases are presented to illustrate AutoFlow capabilities, and a case-of-use illustrates the automated construction of the best transcriptome for a non-model species (*Vicia faba*) after analysis of several combinations of Illumina reads and Sanger sequences with different assemblers and different parameters in a complex and repetitive workflow where branching and convergent tasks were used and internal, automated decisions were taken. The workflow finally produced an optimal transcriptome of 118,188 transcripts, of which 38,004 were annotated, 10,516 coded for a complete protein, 3,314 were putatively new faba-specific transcripts, and 23,727 were considered the representative transcriptome of *V. faba*.



Manuel Gonzalo  
Claros

**Keywords:** Workflow, pipeline, automation, Ruby, bash, transcriptome, assembly, non-model species.

## INTRODUCTION

The high-throughput technologies produce a large amount of data that require the combination of multiple bioinformatic tools to process data, validate results and provide scientific insights. For example, *de novo* transcriptome sequencing generates large files of reads that must be pre-processed, assembled, verified, and then annotated using multiple software tools and data sources where varying parameters provide different results that require further reconciliation or selection [1, 2]. Since different input data will require different software or parameters to provide the best result (i.e., assembly), this can become a long and cumbersome task. The development of sophisticated workflows, where the output of a tool is used as input for other tool(s) can save researchers time and effort

[3]. Workflows provide advantages related to effectiveness (by automation of repetitive tasks), reproducibility (analyses can be repeated over time), traceability (by storing intermediate results) and reusability (of complete workflows or of intermediate results). Many workflow manager systems can build and execute workflows (reviewed by [3-5]), with Taverna [6] and Galaxy [7] being the most popular and accepted.

Installation of workflow manager systems usually includes a large library or repository of tools and ready-made scripts for data processing. Ontology-guided exploitation of local tools, remote web services and computer grids by means of graphical, user-oriented interfaces aims to simplify the design of customised workflows. In return, interface usability limits flexibility and sometimes precludes the design of workflows with new software not included in repositories, or the use of data from non-model organisms, or even the use of large datasets, as occurs in high-throughput sequencing [8]. In spite of graphical front-ends, workflow managers still require knowledge of tools, data formats and programming skills, especially in complex computational

\*Address correspondence to this author at the Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, E-29071, Malaga, Spain; Tel: +34 95 213 72 84; Fax: +34 95 213 20 41; E-mail: [claros@uma.es](mailto:claros@uma.es)

**F.9. Identification and in silico Analysis of Glutathione Reductase Transcripts Expressed in Olive (*Olea europaea* L.) Pollen and Pistil**

## Identification and *in silico* Analysis of Glutathione Reductase Transcripts Expressed in Olive (*Olea europaea* L.) Pollen and Pistil

Estefanía García-Quirós<sup>1</sup>, Rosario Carmona<sup>1,2</sup>, Adoración Zafra<sup>1</sup>,  
M. Gonzalo Claros<sup>2</sup>, and Juan de Dios Alché<sup>1</sup>(✉)

<sup>1</sup> Estación Experimental del Zaidín (CSIC), Granada, Spain

{estefania.garcia, dori.zafra,  
juandedios.alche}@eez.csic.es

<sup>2</sup> Departamento de Biología Molecular y Bioquímica,

Universidad de Málaga, Málaga, Spain

{rosariocarmona, claros}@uma.es

**Abstract.** Glutathione (GSH) protects proteins against oxidation of their thiol-containing groups, by alternatively becoming the subject of oxidation, forming glutathione disulfide (GSSG). Appropriate GSH:GSSG levels are maintained by glutathione reductase (GR), a homodimeric flavoprotein which uses NADPH to reduce one GSSG molecule to two of GSH. This enzyme has been characterized in several species, and described as highly conserved, with two isoforms only. Heterogeneity and distinctiveness of plant reproductive tissues led us to investigate the presence of GR sequences. A *de novo* assembled and annotated olive reproductive transcriptome was subjected to screening, which allowed us to identify at least 11 GR homologues (1 pollen-specific and 10 from pistils). Primers were designed, and full-length sequences were obtained through PCR. *In silico* analysis, including phylogeny, 3-D modeling of N-terminus, and prediction of cellular localization and post-translational modifications was carried out to shed light into the involvement of olive pollen-intrinsic GR in reproductive development.

**Keywords:** Glutathione reductase · Pollen · Pistil · ROS · Glutathione

### 1 Introduction

Olive tree is one of the most important crops in the Mediterranean area, representing one 95% of globally cultivated olive trees, and the source of olive oil, accounting the sixth most abundant vegetable oil produced worldwide. It exhibits a high tolerance towards salinity and drought as compared to other salinity susceptible fruit trees as recently reviewed [1]. Resistance to different abiotic stresses is modulated by reactive oxygen species (ROS) metabolism in numerous plant species. Concomitantly, other metabolically related chemicals like total glutathione, reduced glutathione (GSH) and oxidized glutathione (GSSG) contents, GSH/GSSG ratio and glutathione reductase

**F.10. Automated identification of reference genes based on RNA-seq data**

Carmona et al. *BioMed Eng OnLine* 2017, **16**(Suppl 1):65  
DOI 10.1186/s12938-017-0356-5

BioMedical Engineering  
OnLine

## RESEARCH

## Open Access



# Automated identification of reference genes based on RNA-seq data

Rosario Carmona<sup>1</sup>, Macarena Arroyo<sup>2</sup>, María José Jiménez-Quesada<sup>1</sup>, Pedro Seoane<sup>3</sup>, Adoración Zafra<sup>1</sup>, Rafael Larrosa<sup>4</sup>, Juan de Dios Alché<sup>1</sup> and M. Gonzalo Claros<sup>3\*</sup>

From 4th International Work-Conference on Bioinformatics and Biomedical Engineering-IWBBIO 2016 Grenada, Spain. 20-22 April 2016

\*Correspondence:

claros@uma.es

<sup>3</sup> Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain

Full list of author information is available at the end of the article

## Abstract

**Background:** Gene expression analyses demand appropriate reference genes (RGs) for normalization, in order to obtain reliable assessments. Ideally, RG expression levels should remain constant in all cells, tissues or experimental conditions under study. Housekeeping genes traditionally fulfilled this requirement, but they have been reported to be less invariant than expected; therefore, RGs should be tested and validated for every particular situation. Microarray data have been used to propose new RGs, but only a limited set of model species and conditions are available; on the contrary, RNA-seq experiments are more and more frequent and constitute a new source of candidate RGs.

**Results:** An automated workflow based on mapped NGS reads has been constructed to obtain highly and invariantly expressed RGs based on a normalized expression in reads per mapped million and the coefficient of variation. This workflow has been tested with Roche/454 reads from reproductive tissues of olive tree (*Olea europaea* L.), as well as with Illumina paired-end reads from two different accessions of *Arabidopsis thaliana* and three different human cancers (prostate, small-cell cancer lung and lung adenocarcinoma). Candidate RGs have been proposed for each species and many of them have been previously reported as RGs in literature. Experimental validation of significant RGs in olive tree is provided to support the algorithm.

**Conclusion:** Regardless sequencing technology, number of replicates, and library sizes, when RNA-seq experiments are designed and performed, the same datasets can be analyzed with our workflow to extract suitable RGs for subsequent PCR validation. Moreover, different subset of experimental conditions can provide different suitable RGs.

**Keywords:** Reference genes, Normalization, Real-time PCR, Quantitative PCR, Olive (*Olea europaea* L.), Cancer

## Background

Traditionally, gene expression studies have been carried out by non-quantitative or semi-quantitative RNA gel blotting and later by reverse transcription-polymerase chain reaction (RT-PCR) analyses. Development of real-time, quantitative PCR (qPCR) [1] took

© The Author(s) 2017. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

- F.11. Identification and functional annotation of genes differentially expressed in the reproductive tissues of the olive tree (*Olea europaea* L.) through the generation of subtractive libraries**



## OPEN ACCESS

**Edited by:**

Mariela Torres,  
Instituto Nacional de Tecnología  
Agropecuaria (INTA), Argentina

**Reviewed by:**

Luciana Baldoni,  
Istituto di Bioscienze e Biorisorse  
(IBBR), Italy  
Rosario Muleo,  
Università degli Studi della Tuscia, Italy

**\*Correspondence:**

Juan D. Alche  
juandedios.alche@eez.csic.es

**†Present Address:**

José A. Traverso,  
Department of Cell Biology, University  
of Granada, Granada, Spain  
Simon J. Hiscock,  
Oxford Botanic Garden and Harcourt  
Arboretum, Department of Plant  
Sciences, University of Oxford,  
Oxford, United Kingdom

‡These authors have contributed  
equally to this work.

**Specialty section:**

This article was submitted to  
Crop Science and Horticulture,  
a section of the journal  
Frontiers in Plant Science

**Received:** 28 April 2017

**Accepted:** 28 August 2017

**Published:** 13 September 2017

**Citation:**

Zafra A, Carmona R, Traverso JA,  
Hancock JT, Goldman MHS,  
Claros MG, Hiscock SJ and Alche JD  
(2017) Identification and Functional  
Annotation of Genes Differentially  
Expressed in the Reproductive Tissues  
of the Olive Tree (*Olea europaea* L.)  
through the Generation of Subtractive  
Libraries. *Front. Plant Sci.* 8:1576.  
doi: 10.3389/fpls.2017.01576

# Identification and Functional Annotation of Genes Differentially Expressed in the Reproductive Tissues of the Olive Tree (*Olea europaea* L.) through the Generation of Subtractive Libraries

Adoración Zafra<sup>1‡</sup>, Rosario Carmona<sup>1‡</sup>, José A. Traverso<sup>1†</sup>, John T. Hancock<sup>2</sup>,  
Maria H. S. Goldman<sup>3</sup>, M. Gonzalo Claros<sup>4</sup>, Simon J. Hiscock<sup>5†</sup> and Juan D. Alche<sup>1\*</sup>

<sup>1</sup> Plant Reproductive Biology Laboratory, Department of Biochemistry, Cellular and Molecular Biology of Plants, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Granada, Spain, <sup>2</sup> Faculty of Health and Life Sciences, University of the West of England, Bristol, United Kingdom, <sup>3</sup> Departamento de Biología, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, São Paulo, Brazil, <sup>4</sup> Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain, <sup>5</sup> School of Biological Sciences, University of Bristol, Bristol, United Kingdom

The olive tree is a crop of high socio-economical importance in the Mediterranean area. Sexual reproduction in this plant is an essential process, which determines the yield. Successful fertilization is mainly favored and sometimes needed of the presence of pollen grains from a different cultivar as the olive seizes a self-incompatibility system allegedly determined of the sporophytic type. The purpose of the present study was to identify key gene products involved in the function of olive pollen and pistil, in order to help elucidate the events and signaling processes, which happen during the courtship, pollen grain germination, and fertilization in olive. The use of subtractive SSH libraries constructed using, on the one hand one specific stage of the pistil development with germinating pollen grains, and on the other hand mature pollen grains may help to reveal the specific transcripts involved in the cited events. Such libraries have also been created by subtracting vegetative mRNAs (from leaves), in order to identify reproductive sequences only. A variety of transcripts have been identified in the mature pollen grains and in the pistil at the receptive stage. Among them, those related to defense, transport and oxidative metabolism are highlighted mainly in the pistil libraries where transcripts related to stress, and response to biotic and abiotic stimulus have a prominent position. Extensive lists containing information as regard to the specific transcripts determined for each stage and tissue are provided, as well as functional classifications of these gene products. Such lists were faced up to two recent datasets obtained in olive after transcriptomic and genomic approaches. The sequences and the differential expression level of the SSH-transcripts identified here, highly matched the transcriptomic information. Moreover, the unique presence of a representative number of these transcripts has been validated by means of qPCR approaches. The construction

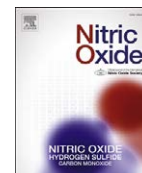
- F.12.** Generation of nitric oxide by olive (*Olea europaea* L.) pollen during in vitro germination and assessment of the *S*-nitroso and nitro-proteomes by computational predictive methods

Nitric Oxide 68 (2017) 23–37



Contents lists available at ScienceDirect

Nitric Oxide

journal homepage: [www.elsevier.com/locate/yniox](http://www.elsevier.com/locate/yniox)

## Generation of nitric oxide by olive (*Olea europaea* L.) pollen during *in vitro* germination and assessment of the S-nitroso- and nitro-proteomes by computational predictive methods



María José Jimenez-Quesada <sup>a</sup>, Rosario Carmona <sup>a</sup>, Elena Lima-Cabello <sup>a</sup>,  
José Ángel Traverso <sup>a,1</sup>, Antonio Jesús Castro <sup>a</sup>, M. Gonzalo Claros <sup>b</sup>, Juan de Dios Alché <sup>a,\*</sup>

<sup>a</sup> Plant Reproductive Biology Laboratory, Department of Biochemistry, Cellular and Molecular Biology of Plants, Estación Experimental del Zaidín (CSIC), Profesor Albareda 1, 18008 Granada, Spain

<sup>b</sup> Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain

### ARTICLE INFO

#### Article history:

Received 18 November 2016

Received in revised form

29 April 2017

Accepted 5 June 2017

Available online 21 June 2017

#### Keywords:

Germination

Olive

Pollen

S-nitrosylation

Transcriptome

Tyr-nitration

### ABSTRACT

Nitric oxide is recognized as a signaling molecule involved in a broad range of physiological processes in plants including sexual reproduction. NO has been detected in the pollen grain at high levels and regulates pollen tube growth. Previous studies demonstrated that NO as well as ROS are produced in the olive reproductive tissues in a stage- and tissue-specific manner. The aim of this study was to assess the production of NO throughout the germination of olive (*Olea europaea* L.) pollen *in vitro*. The NO fluorescent probe DAF-2DA was used to image NO production *in situ*, which was correlated to pollen viability. Moreover, by means of a fluorimetric assay we showed that growing pollen tubes release NO. GSNO -a mobile reservoir of NO, formed by the S-nitrosylation of NO with reduced glutathione (GSH) - was for the first time detected and quantified at different stages of pollen tube growth using a LC-ES/MS analysis. Exogenous NO donors inhibited both pollen germination and pollen tube growth and these effects were partially reverted by the specific NO-scavenger c-PTIO. However, little is known about how NO affects the germination process. With the aim of elucidating the putative relevance of protein S-nitrosylation and Tyr-nitration as important post-translational modifications in the development and physiology of the olive pollen, a *de novo* assembled and annotated reproductive transcriptome from olive was challenged *in silico* for the putative capability of transcripts to become potentially modified by S-nitrosylation/Tyr-nitration according to well-established criteria. Numerous gene products with these characteristics were identified, and a broad discussion as regards to their potential role in plant reproduction was built after their functional classification. Moreover, the importance of both S-nitrosylation/Tyr-nitrations was experimentally assessed and validated by using Western blotting, immunoprecipitation and proteomic approaches.

© 2017 Elsevier Inc. All rights reserved.

### 1. Introduction

Nitric oxide (NO) is an important signaling molecule in physiological and pathological processes in all living beings. It represents a gaseous, relatively high reactive free radical, soluble in aqueous solution while able to diffuse through biological membranes [1].

In mammals, the sources of NO are NO synthase (NOS) enzymes,

which catalyze the oxidation of arginine to NO and citrulline. However, plant NO origin is more complex, as NOS-like activity has been described although the gene encoding the enzyme did not. In fact, up to seven pathways for NO biosynthesis have been proposed in plants and this multiplicity difficult clarifying the route for NO production (reviewed in Refs. [2–4]).

In plants, NO has a central role in the wide range of mechanisms involved in growth and development as well as in plant immunity, the hypersensitive response and abiotic interactions (reviewed in Ref. [3]). Apparently contradictory functions are exerted by NO (i.e. in plant defence vs symbiotic interaction), and such functions seem to depend on NO concentration and location [4].

\* Corresponding author.

E-mail address: [juandedios.alche@eez.csic.es](mailto:juandedios.alche@eez.csic.es) (J.D. Alché).

<sup>1</sup> Present address: Department of Cell Biology, Faculty of Science, Granada University, Granada, Spain.

**F.13.** *S*-nitroso- and nitro- proteomes in the olive (*Olea europaea* L.) pollen. Predictive versus experimental data by nano-LC-MS

Data in Brief 15 (2017) 474–477



Contents lists available at ScienceDirect

## Data in Brief

journal homepage: [www.elsevier.com/locate/dib](http://www.elsevier.com/locate/dib)

## Data Article

S-nitroso- and nitro- proteomes in the olive (*Olea europaea* L.) pollen. Predictive versus experimental data by nano-LC-MS

Rosario Carmona<sup>a,1</sup>, María José Jimenez-Quesada<sup>a,1</sup>,  
 Elena Lima-Cabello<sup>a</sup>, José Ángel Traverso<sup>a,2</sup>,  
 Antonio Jesús Castro<sup>a</sup>, M. Gonzalo Claros<sup>b</sup>,  
 Juan de Dios Alché<sup>a,\*</sup>

<sup>a</sup> Plant Reproductive Biology Laboratory, Department of Biochemistry, Cellular and Molecular Biology of Plants, Estación Experimental del Zaidín (CSIC), Profesor Albareda 1, 18008 Granada, Spain

<sup>b</sup> Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain

## ARTICLE INFO

## Article history:

Received 8 June 2017

Received in revised form

13 July 2017

Accepted 26 September 2017

Available online 6 October 2017

## Keywords:

Immunoprecipitation

Proteome

S-nitrosylation

Transcriptome

Tyrosine nitration

## ABSTRACT

The data presented here are related to the research article entitled “Generation of nitric oxide by olive (*Olea europaea* L.) pollen during *in vitro* germination and assessment of the S-nitroso- and nitro-proteomes by computational predictive methods” doi:10.1016/j.niox.2017.06.005 (Jimenez-Quesada et al., 2017) [1]. Predicted cysteine S-nitrosylation and Tyr-nitration sites in proteins derived from a *de novo* assembled and annotated pollen transcriptome from olive tree (*Olea europaea* L.) were obtained after using well-established predictive tools *in silico*. Predictions were performed using both default and highly restrictive thresholds. Numerous gene products identified with these characteristics are listed here. An experimental validation of the data, consisting in nano-LC-MS (Liquid Chromatography-Mass Spectrometry) determination of olive pollen proteins after immunoprecipitation with antibodies to anti-S-nitrosoCys and anti-3-NT (NitroTyrosine) allowed identification of numerous proteins subjected to these

DOI of original article: <http://dx.doi.org/10.1016/j.niox.2017.06.005>

\* Corresponding author.

E-mail address: [juandedios.alche@eez.csic.es](mailto:juandedios.alche@eez.csic.es) (J. de Dios Alché).

<sup>1</sup> These authors contributed equally to the present work.

<sup>2</sup> Present address: Department of Cell Biology, Faculty of Science, Granada University, Granada, Spain

<http://dx.doi.org/10.1016/j.dib.2017.09.058>

2352-3409/© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

**F.14. Bioinformatic prediction of *S*-nitrosylation sites in large protein datasets**



## Chapter 19

### Bioinformatic Prediction of S-Nitrosylation Sites in Large Protein Datasets

Rosario Carmona, M. Claros, and Juan de Alché

#### Abstract

S-nitrosylation is an essential and reversible posttranslational modification of proteins involved in numerous biological processes. The experimental determination of S-nitrosylation sites is laborious and time-consuming. Therefore, the use of computational prediction tools of this modification represents a convenient first approach to generate useful information for subsequent experimental verification. Here we describe an in silico analysis pipeline to integrate the use of several bioinformatic tools while dealing with big query protein sets.

**Key words** Nitric oxide, Posttranslational modification, PTMs prediction, S-nitrosylation

#### 1 Introduction

Nitric oxide (NO) is an important gaseous signal molecule implicated in fundamental biological actions in living beings [1]. It is involved in several posttranslational modifications (PTMs), including cysteine S-nitrosylation, tyrosine nitration and metal nitrosylation [2]. In this chapter, we focus on S-nitrosylation modification (SNO), whereby the NO binding to cysteine residues may affect proteins by modulating fundamental features like their activity, location and/or stability. S-Nitrosylation is a reversible PTM, and consequently can act as an on/off switch control of cell signaling [2, 3], transcriptional regulation [4], chromatin remodeling [5], and apoptosis [6], among other processes. Hence, S-nitrosylated proteins so far identified correspond to quite diverse functional categories.

The determination of nitrosylation sites in proteins by means of experimental approaches is laborious, expensive, and time-consuming. With the growth of the bioinformatics, numerous tools have emerged for the in silico prediction of this PTM. The preliminary identification in silico of proteins prone to this modification is a convenient and valuable tool complementary to the use

**F.15. TransFlow: A modular framework for assembling and assessing accurate de novo transcriptomes in non-model organisms**

Seoane et al. *BMC Bioinformatics* 2018, **19**(Suppl 14):416  
<https://doi.org/10.1186/s12859-018-2384-y>

BMC Bioinformatics

RESEARCH

Open Access



# TransFlow: a modular framework for assembling and assessing accurate de novo transcriptomes in non-model organisms

Pedro Seoane<sup>1†</sup>, Marina Espigares<sup>1†</sup>, Rosario Carmona<sup>2</sup>, Álvaro Polonio<sup>3</sup>, Julia Quintana<sup>4</sup>, Enrico Cretazzo<sup>5</sup>, Josefina Bota<sup>6</sup>, Alejandro Pérez-García<sup>3</sup>, Juan de Dios Alché<sup>2</sup>, Luis Gómez<sup>7,8</sup> and M. Gonzalo Claros<sup>1\*</sup>

From 5th International Work-Conference on Bioinformatics and Biomedical Engineering  
 Granada, Spain. 26-28 April 2017

## Abstract

**Background:** The advances in high-throughput sequencing technologies are allowing more and more de novo assembling of transcriptomes from many new organisms. Some degree of automation and evaluation is required to warrant reproducibility, repetitivity and the selection of the best possible transcriptome. Workflows and pipelines are becoming an absolute requirement for such a purpose, but the issue of assembling evaluation for de novo transcriptomes in organisms lacking a sequenced genome remains unsolved. An automated, reproducible and flexible framework called TransFlow to accomplish this task is described.

**Results:** TransFlow with its five independent modules was designed to build different workflows depending on the nature of the original reads. This architecture enables different combinations of Illumina and Roche/454 sequencing data, and can be extended to other sequencing platforms. Its capabilities are illustrated with the selection of reliable plant reference transcriptomes and the assembling six transcriptomes (three case studies for grapevine leaves, olive tree pollen, and chestnut stem, and other three for haustorium, epiphytic structures and their combination for the phytopathogenic fungus *Podosphaera xanthii*). Arabidopsis and poplar transcriptomes revealed to be the best references. A common result regarding de novo assemblies is that Illumina paired-end reads of 100 nt in length assembled with OASES can provide reliable transcriptomes, while the contribution of longer reads is noticeable only when they complement a set of short, single-reads.

**Conclusions:** TransFlow can handle up to 181 different assembling strategies. Evaluation based on principal component analyses allows its self-adaptation to different sets of reads to provide a suitable transcriptome for each combination of reads and assemblers. As a result, each case study has its own behaviour, prioritises evaluation parameters, and gives an objective and automated way for detecting the best transcriptome within a pool of them. Sequencing data type and quantity (preferably several hundred millions of 2 × 100 nt or longer), assemblers (OASES for Illumina, MIRA4 and EULER-SR reconciled with CAP3 for Roche/454) and strategy (preferably scaffolding with OASES, and probably merging with Roche/454 when available) arise as the most impacting factors.

**Keywords:** Transcriptome, Assembling, Workflow, pipeline, PCA, Non-model organism

\*Correspondence: [claros@uma.es](mailto:claros@uma.es)

<sup>†</sup>Pedro Seoane and Marina Espigares contributed equally to this work.

<sup>1</sup>Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Campus de Teatinos s/n, 29071 Málaga, Spain

Full list of author information is available at the end of the article

**F.16. Structural and functional features of glutathione reductase transcripts from olive (*Olea europaea* L.) seeds**

### Structural and functional features of glutathione reductase transcripts from olive (*Olea europaea* L.) seeds.

Elena Lima-Cabello<sup>1</sup>[0000-0002-5602-0620], Isabel Martínez-Beas<sup>1</sup>, Estefanía García-Quirós<sup>1</sup>[0000-0002-8026-1547], Rosario Carmona<sup>1</sup>[0000-0002-4235-2886], M. Gonzalo Claros<sup>2</sup>[0000-0002-0112-3550], Jose C. Jimenez-Lopez<sup>1</sup>[0000-0002-1470-360X] and Juan de Dios Alché<sup>1</sup>[0000-0002-7547-6025]

<sup>1</sup> Estación Experimental del Zaidín. CSIC, 18008 Granada, Spain.

<sup>2</sup> Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain.  
juandedios.alche@eez.csic.es

**Abstract.** The olive seed is a promising byproduct generated in the olive oil related industries, with increasing interest because of its nutritional value and potential nutraceutical properties. Knowledge concerning the antioxidant capacity of this new alimentary material is scarce. Moreover, oxidative homeostasis and signaling involved physiological processes such as development, dormancy and germination in the olive seed are also unknown. Glutathione (one of the most abundant antioxidants in plant cells), is crucial for seeds physiology, and for defense and detoxification mechanisms. The availability of glutathione in its reduced (GSH) and oxidized (GSSG) forms, the ratio of both forms (GSH/GSSG), and their concurrence in other numerous metabolic pathways is tightly regulated by numerous enzymes. Prominent among these enzymes is glutathione reductase (GR), which has been considered essential for seedling growth and development. The present work aims to increase the knowledge about the functional insights of GR in olive seeds. Searching in the olive transcriptome, at least 19 GR homologues (10 from seed and 9 from vegetative tissue) were identified and retrieved. An *in silico* analysis was carried out, which included phylogeny, 3-D modelling of the N-terminus, and the prediction of cellular localization and post-translational modifications (PTM) for these gene products. The high variability of forms detected for this enzyme in olive seeds and their susceptibility to numerous PTMs suggest a relevant role for this enzyme in redox metabolism and signalling events.

**Keywords:** Antioxidant, Glutathione Reductase, Modelling, Olive, Post-Translational Modifications, Seed, Signalling

## 1 Introduction

Olive tree is the most economically important oil-producing crop in Mediterranean countries. Olive fruits are consumed only after processing, which ends up to either table olives or olive oil. Both types of products are well-known natural sources of

# Bibliografía

- [1] Rp Horgan and Lc Kenny. Ómic'technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, pages 189–195, 2011.
- [2] Sarah B. Ng, Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, sep 2009.
- [3] S. M. Brady. Unraveling the Dynamic Transcriptome. *THE PLANT CELL ONLINE*, 18(9):2101–2111, 2006.
- [4] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, feb 2001.
- [5] J. Craig Venter, Mark D. Adams, Eugene W. Myers, P. W. Li, Richard J. Mural, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, feb 2001.
- [6] Amalio Telenti, Levi C.T. Pierce, William H. Biggs, Julia Di Iulio, Emily H.M. Wong, et al. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(42):11901–11906, oct 2016.
- [7] Gary Saunders, Michael Baudis, Regina Becker, Sergi Beltran, Christophe Bérout, et al. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nature Reviews Genetics*, 20(11):693–701, nov 2019.
- [8] a M Maxam and W Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, 1977.
- [9] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain terminating inhibitors. *Proc Natl Acad Sci USA*, 74:5463–5467, 1977.
- [10] Mostafa Ronaghi, Samer Karamohamed, Bertil Pettersson, Mathias Uhlén, and Pål Nyrén. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242(1):84–89, nov 1996.
- [11] Karl V. Voelkerding, Shale Dames, and Jacob D. Durtschi. Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: A Paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology, 2010.
- [12] Michael Quail, Miriam E Smith, Paul Coupland, Thomas D Otto, Simon R Harris, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, 13(1):1, 2012.
- [13] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, et al. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 2012:251364, 2012.

- [14] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: Ten years of next-generation sequencing technologies, jun 2016.
- [15] Rafael Navajas-Pérez and Andrew H Paterson. Patterns of tandem repetition in plant whole genome assemblies. *Molecular Genetics and Genomics*, 281(6):579–590, 2009.
- [16] Manuel Gonzalo Claros, Rocío Bautista, Darío Guerrero-Fernández, Hicham Benzerki, Pedro Seoane, et al. Why Assembling Plant Genome Sequences Is So Challenging. *Biology*, 1(2):439–459, 2012.
- [17] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1):36–46, 2012.
- [18] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, et al. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature biotechnology*, 437(7057):376–380, 2006.
- [19] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.
- [20] Chengwei Luo, Despina Tsementzi, Nikos Kyrpides, Timothy Read, and Konstantinos T Konstantinidis. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS one*, 7(2):e30087, 2012.
- [21] Olena Morozova and Marco A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–64, nov 2008.
- [22] Yoon Seong Jeon, Sang Cheol Park, Jeongmin Lim, Jongsik Chun, and Bong Soo Kim. Improved pipeline for reducing erroneous identification by 16S rRNA sequences using the Illumina MiSeq platform. *Journal of Microbiology*, 53(1):60–69, jan 2015.
- [23] Dale J Hedges, Toumy Guettouche, Shan Yang, Guney Bademci, Ashley Diaz, et al. Comparison of Three Targeted Enrichment Strategies on the SOLiD Sequencing Platform. *PLoS ONE*, 6(4):e18595, 2011.
- [24] Yu Feng Huang, Sheng Chung Chen, Yih Shien Chiang, Tzu Han Chen, and Kuo Ping Chiu. Palindromic sequence impedes sequencing-by-ligation mechanism. *BMC Systems Biology*, 6(SUPPL.2):S10, dec 2012.
- [25] Barry Merriman and Jonathan M Rothberg. Progress in Ion Torrent semiconductor chip. *Electrophoresis*, 33(23):3397–3417, 2012.
- [26] John F. Thompson and Kathleen E. Steinmann. Single molecule sequencing with a HeliScope genetic analysis system, oct 2010.
- [27] Dmitry Pushkarev, Norma F Neff, and Stephen R Quake. Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, 27(9):847–850, sep 2009.
- [28] Adam C English, Stephen Richards, Yi Han, Min Wang, Vanesa Vee, et al. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE*, 7(11), 2012.
- [29] Benjamin A Flusberg, Dale Webster, Jessa Lee, Kevin Travers, Eric Olivares, et al. Direct detection of DNA methylation during single-molecule, real- time sequencing. *Nature methods*, 7(6):461–465, 2010.

- [30] Donald Sharon, Hagen Tilgner, Fabian Grubert, and Michael Snyder. A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology*, 31(11):1009–1014, nov 2013.
- [31] Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications, oct 2015.
- [32] Michael Eisenstein. Startups use short-read data to expand long-read sequencing market. *Nat Biotech*, 33(5):433–435, may 2015.
- [33] Winston Timp, Utkur M Mirsaidov, Deqiang Wang, Jeff Comer, Aleksei Aksimentiev, et al. Nanopore sequencing: Electrical measurements of the code of life. *IEEE Transactions on Nanotechnology*, 9(3):281–294, 2010.
- [34] Srilakshmi Srinivasan and Jyotsna Batra. Four Generations of Sequencing- Is it Ready for the Clinic Yet ? *Next Generation: Sequencing & Applications*, 1(1):1–7, 2014.
- [35] James Clarke, Hai Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4):265–270, apr 2009.
- [36] Cole Trapnell, Brian a Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.
- [37] Jo Korbelt, Ae Urban, Jp Affourtit, and B Godwin. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 13(5849):350–358, 2007.
- [38] Christopher a Maher, Nallasivam Palanisamy, John C Brenner, Xuhong Cao, Shanker Kalyana-Sundaram, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30):12353–12358, 2009.
- [39] Eric D. Green. Strategies for the systematic sequencing of complex genomes, aug 2001.
- [40] R Staden. A strategy of DNA sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610, 1979.
- [41] Juan E. Zalapa, Hugo Cuevas, Huayu Zhu, Shawn Steffan, Douglas Senalik, et al. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences, feb 2012.
- [42] Hicham Benzekri. *Aportación bioinformática a la biología de especies marinas*. PhD thesis, Universidad de Málaga, 2016.
- [43] Pedro Seoane. *Creación y uso de los flujos de trabajo bioinformáticos*. PhD thesis, Universidad de Málaga, 2016.
- [44] Xiangyang Xu and Guihua Bai. Whole-genome resequencing: changing the paradigms of SNP detection, molecular mapping and gene discovery, jan 2015.
- [45] Jie Xu, Yibing Yuan, Yunbi Xu, Gengyun Zhang, Xiaosen Guo, et al. Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC Plant Biology*, 14(1):83, apr 2014.
- [46] Gancho T. Slavov, Stephen P. Difazio, Joel Martin, Wendy Schackwitz, Wellington Muchero, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, 196(3):713–725, nov 2012.

- [47] Richard Cronn, Brian J. Knaus, Aaron Liston, Peter J. Maughan, Matthew Parks, et al. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*, 99(2):291–311, feb 2012.
- [48] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8):651–657, 2007.
- [49] Dustin R Masser, David R Stanford, and Willard M Freeman. Targeted DNA methylation analysis by next-generation sequencing. *Journal of visualized experiments : JoVE*, 96:1–11, 2015.
- [50] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):3, 2012.
- [51] H. P.J. Buermans and J. T. den Dunnen. Next generation sequencing technology: Advances and applications, oct 2014.
- [52] Diana L. Cox-Foster, Sean Conlan, Edward C. Holmes, Gustavo Palacios, Jay D. Evans, et al. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*, 318(5848):283–287, oct 2007.
- [53] Susan R Strickler, Aureliano Bombarely, and Lukas A Mueller. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American journal of botany*, 99(2):257–266, feb 2012.
- [54] Javier Canales, Rocio Bautista, Philippe Label, Josefa Gómez-Maldonado, Isabelle Lesur, et al. De novo assembly of maritime pine transcriptome: Implications for forest breeding and biotechnology. *Plant Biotechnology Journal*, 12(3):286–299, 2014.
- [55] Noé Fernández-Pozo, Javier Canales, Darío Guerrero-Fernández, David P Villalobos, Sara M Díaz-Moreno, et al. EuroPineDB: A high-coverage web database for maritime pine transcriptome. *BMC Genomics*, 12(1):366, dec 2011.
- [56] Hicham Benzekri, Paula Armesto, Xavier Cousin, Mireia Rovira, Diego Crespo, et al. De novo assembly, characterization and functional annotation of Senegalese sole (*Solea senegalensis*) and common sole (*Solea solea*) transcriptomes: integration in a database and design of a microarray. *BMC Genomics*, 15(1):952, 2014.
- [57] Rosario Carmona, Adoración Zafra, Pedro Seoane, Antonio J Castro, Darío Guerrero-Fernández, et al. ReprOlive: a database with linked data for the olive tree (*Olea europaea* L.) reproductive transcriptome. *Frontiers in plant science*, 6:625, jan 2015.
- [58] Salvador González-Gordo, Rocío Bautista, M. Gonzalo Claros, Alberto Paradela, Antonio Ramos, et al. Expression profiling of redox-metabolism-related genes and proteins during sweet pepper (*Capsicum annuum* L.) fruit ripening. *Free Radical Biology and Medicine*, 120:S106–S107, may 2018.
- [59] Tingting Lu, Lingling Cui, Yan Zhou, Chuanrang Zhu, Danlin Fan, et al. Transcriptome-wide investigation of circular RNAs in rice. *RNA*, 21(12):2076–2087, dec 2015.
- [60] Xiaoguang Zhou, Lufeng Ren, Qingshu Meng, Yuntao Li, Yude Yu, et al. The next-generation sequencing technology and application. *Protein and Cell*, 1(6):520–536, jun 2010.
- [61] John H Malone and Brian Oliver. Microarrays, deep sequencing and the true measure of the transcriptome, dec 2011.
- [62] Sunitha Kogenaru, Yan Qing, Yinping Guo, and Nian Wang. RNA-seq and microarray complement each other in transcriptome profiling. *BMC genomics*, 13(1):629, 2012.

- [63] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq : a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.
- [64] Olena Morozova, Martin Hirst, and Marco A Marra. Applications of New Sequencing Technologies for Transcriptome Analysis. *Annual Review of Genomics and Human Genetics*, 10(1):135–151, 2009.
- [65] R Ekblom and J Galindo. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15, 2010.
- [66] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell RNA sequencing technologies and bioinformatics pipelines, aug 2018.
- [67] Biao Huang and Rongxin Zhang. Regulatory non-coding RNAs: Revolutionizing the RNA world, 2014.
- [68] Sofie K.D. Nielsen, Thomas L. Koch, Frank Hauser, Anders Garm, and Cornelis J.P. Grimmelikhuijzen. De novo transcriptome assembly of the cubomedusa *Tripedalia cystophora*, including the analysis of a set of genes involved in peptidergic neurotransmission. *BMC Genomics*, 20(1):175, dec 2019.
- [69] Nam V. Hoang, Agnelo Furtado, Patrick J. Mason, Annelie Marquardt, Lakshmi Kasirajan, et al. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics*, 18(1):395, dec 2017.
- [70] Liangzhen Zhao, Hangxiao Zhang, Markus V. Kohnen, Kasavajhala V.S.K. Prasad, Lianfeng Gu, et al. Analysis of transcriptome and epitranscriptome in plants using pacbio iso-seq and nanopore-based direct RNA sequencing, mar 2019.
- [71] Marius Nicolae, Serghei Mangul, Ion I Măndoiu, and Alex Zelikovsky. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, 6(1):9, 2011.
- [72] Alyssa A. Gullledge, April D. Roberts, Hiral Vora, Ketan Patel, and Ann E. Loraine. Mining *Arabidopsis thaliana* RNA-seq data with integrated genome browser reveals stress-induced alternative splicing of the putative splicing regulator SR45A. *American Journal of Botany*, 99(2):219–231, feb 2012.
- [73] Douglas E. Soltis, Victor A. Albert, Jim Leebens-Mack, Charles D. Bell, Andrew H. Paterson, et al. Polyploidy and angiosperm diversification. *American Journal of Botany*, 96(1):336–348, jan 2009.
- [74] Juan Falgueras, Antonio J Lara, Noé Fernández-Pozo, Francisco R Cantón, Guillermo Pérez-Trabado, et al. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC bioinformatics*, 11(38):1–12, 2010.
- [75] Ram Krishna Shrestha, Baruch Lubinsky, Vijay B Bansode, Mónica B J Moinz, Grace P McCormack, et al. QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC bioinformatics*, 15:33, 2014.
- [76] Cristian Del Fabbro, Simone Scalabrin, Michele Morgante, and Federico M Giorgi. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS ONE*, 8(12):1–13, 2013.
- [77] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, aug 2014.

- [78] Sara El-Metwally, Taher Hamza, Magdi Zakaria, and Mohamed Helmy. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS computational biology*, 9(12):e1003345, 2013.
- [79] R Staden. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Research*, 8(16):3673–3694, aug 1980.
- [80] R Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, jun 1979.
- [81] Xiaoqi Huang and Anup Madan. CAP3: A DNA sequence assembly program. *Genome Research*, 9(9):868–877, 1999.
- [82] B Chevreux, T Pfisterer, B Drescher, A J Driesel, W E G Müller, et al. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research*, 14:1147–1159, 2004.
- [83] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.
- [84] P A Pevzner, H Tang, and M S Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9748–9753, 2001.
- [85] Daniel R Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 2008.
- [86] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, et al. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123, jun 2009.
- [87] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, 2010.
- [88] Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, jul 2011.
- [89] William R Jeck, Josephine A Reinhardt, David A Baltrus, Matthew T Hickenbotham, Vincent Magrini, et al. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 23(21):2942–2944, 2007.
- [90] René L Warren, Granger G Sutton, Steven J M Jones, and Robert A Holt. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23(4):500–501, 2007.
- [91] Chen Shan Chin, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569, apr 2013.
- [92] Chen Shan Chin, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, dec 2016.
- [93] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, et al. Canu: Scalable and accurate long-read assembly via adaptive  $\kappa$ -mer weighting and repeat separation. *Genome Research*, 27(5):722–736, may 2017.

- [94] Govinda M. Kamath, Ilan Shomorony, Fei Xia, Thomas A. Courtade, and David N. Tse. HINGE: Long-read assembly achieves optimal repeat resolution. *Genome Research*, 27(5):747–756, may 2017.
- [95] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, et al. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, apr 2013.
- [96] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, mar 2015.
- [97] Alexander Dobin and Thomas R. Gingeras. Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics*, 51(1):11.14.1–11.14.19, sep 2015.
- [98] Thomas D. Wu, Jens Reeder, Michael Lawrence, Gabe Becker, and Matthew J. Brauer. GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. In *Methods in Molecular Biology*, volume 1418, pages 283–334. Humana Press Inc., 2016.
- [99] Sophie Schbath, Véronique Martin, Matthias Zytnicki, Julien Fayolle, Valentin Loux, et al. Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. *Journal of Computational Biology*, 19(6):796–813, 2012.
- [100] Gerton Lunter and Martin Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939, 2011.
- [101] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 00(00):3, 2013.
- [102] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, 2012.
- [103] Kanako Kawaaura, Keiichi Mochida, Akiko Enju, Yasushi Totoki, Atsushi Toyoda, et al. Assessment of adaptive evolution between wheat and rice as deduced from full-length common wheat cDNA sequence data and expression patterns. *BMC genomics*, 10:271, 2009.
- [104] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, et al. BLAST+: architecture and applications. *BMC bioinformatics*, 10:421, 2009.
- [105] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(SUPPL. 1):D501–4, jan 2007.
- [106] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, et al. GenBank. *Nucleic Acids Research*, 45(D1):D37–D42, nov 2017.
- [107] Alex Bateman, Maria Jesus Martin, Claire O’Donovan, Michele Magrane, Emanuele Alpi, et al. UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.
- [108] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, et al. Blast2GO: A universal annotation and visualization tool in functional genomics research. Application note. *Bioinformatics*, 21:3674–3676, 2005.
- [109] Antonio Muñoz-Mérida, Enrique Viguera, M Gonzalo Claros, Oswaldo Trelles, and Antonio J Perez-Pulido. Sma3s: A three-step modular annotator for large sequence datasets. *DNA Research*, 21(4):341–353, 2014.

- [110] Carlos S. Casimiro-Soriguer, Antonio Muñoz-Mérida, and Antonio J. Pérez-Pulido. Sma3s: a universal tool for easy functional annotation of proteomes and transcriptomes. *PROTEOMICS*, 17(12):1700071, jun 2017.
- [111] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, et al. Gene ontology: Tool for the unification of biology, may 2000.
- [112] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, jan 2012.
- [113] A Bairoch. The ENZYME database in 2000. *Nucleic acids research*, 28(1):304–5, jan 2000.
- [114] Alex L Mitchell, Teresa K Attwood, Patricia C Babbitt, Matthias Blum, Peer Bork, et al. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360, jan 2019.
- [115] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, jan 2018.
- [116] Philippe Lamesch, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, et al. The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, 40(D1):1202–1210, 2012.
- [117] S T Sherry, M H Ward, M Kholodov, J Baker, L Phan, et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–11, jan 2001.
- [118] Alexander Hüttenhofer, Peter Schattner, and Norbert Polacek. Non-coding RNAs: Hope or hype?, may 2005.
- [119] Marco Cirilli, Tiziano Flati, Silvia Gioiosa, Ilario Tagliaferri, Angelo Ciacciulli, et al. PeachVar-DB: A Curated Collection of Genetic Variations for the Interactive Analysis of Peach Genome Data. *Plant & cell physiology*, 59(1):e2, jan 2018.
- [120] Ryoichi Yano, Satoko Nonaka, and Hiroshi Ezura. Melonet-DB, a Grand RNA-Seq Gene Expression Atlas in Melon (*Cucumis melo* L.). *Plant & cell physiology*, 59(1):e4, jan 2018.
- [121] Toru Kudo, Masaaki Kobayashi, Shin Terashima, Minami Katayama, Soichi Ozaki, et al. TOMA-TOMICS: A web database for integrated omics information in tomato. *Plant and Cell Physiology*, 58(1):e8, jan 2017.
- [122] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, et al. Scientific workflow management and the Kepler system. In *Concurrency Computation Practice and Experience*, volume 18, pages 1039–1065. John Wiley & Sons, Ltd, aug 2006.
- [123] K Wolstencroft, R Haines, D Fellows, A Williams, D Withers, et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(W1):W557–W561, 2013.
- [124] Jeremy Goecks, Anton Nekrutenko, and James Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, 2010.
- [125] Pedro Seoane, Sara Ocaña, Rosario Carmona, Rocio Bautista, Eva Madrid, et al. AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised de novo Transcriptome for a Non-Model Species, such as Faba Bean (*Vicia faba*). *Current Bioinformatics*, 11(September), 2016.

- [126] Pedro Seoane, Marina Espigares, Rosario Carmona, Álvaro Polonio, Julia Quintana, et al. TransFlow: A modular framework for assembling and assessing accurate de novo transcriptomes in non-model organisms. *BMC Bioinformatics*, 19(S14):416, nov 2018.
- [127] Concepcion M. Diez, Isabel Trujillo, Nieves Martinez-Urdiroz, Diego Barranco, Luis Rallo, et al. Olive domestication and diversification in the Mediterranean Basin. *New Phytologist*, 206(1):436–447, apr 2015.
- [128] Muriel Gros-Balthazard, Guillaume Besnard, Gautier Sarah, Yan Holtz, Julie Leclercq, et al. Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. *The Plant Journal*, page tpj.14435, jul 2019.
- [129] Jaime Jiménez-Ruiz, Jorge A. Ramírez-Tejero, Noé Fernández-Pozo, María de la O. Leyva-Pérez, Haidong Yan, et al. Transposon activation is a major driver in the genome evolution of cultivated olive trees (*Olea europaea* L.). *Plant Genome*, 13(1), mar 2020.
- [130] MINISTERIO DE AGRICULTURA PESCA Y ALIMENTACIÓN. Superficie y Producción Olivar. [https://www.mapa.gob.es/es/agricultura/temas/producciones-agricolas/2014%2002%2027%20datos%20WEB%20olivar-final\\_tcm30-135522.pdf](https://www.mapa.gob.es/es/agricultura/temas/producciones-agricolas/2014%2002%2027%20datos%20WEB%20olivar-final_tcm30-135522.pdf), 2013.
- [131] João Loureiro, Eleazar Rodriguez, Armando Costa, and Conceição Santos. Nuclear DNA content estimations in wild olive (*Olea europaea* L. ssp. *europaea* var. *sylvestris* Brot.) and Portuguese cultivars of *O. europaea* using flow cytometry. *Genetic Resources and Crop Evolution*, 54(1):21–25, jan 2007.
- [132] Rocío Bautista, Francisco M. Cánovas, and M. Gonzalo Claros. Genomic evidence for a repetitive nature of the RAPD polymorphisms in *Olea europaea* (olive-tree). *Euphytica*, 130(2):185–190, 2003.
- [133] Elena Barghini, Lucia Natali, Rosa Maria Cossu, Tommaso Giordani, Massimo Pindo, et al. The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome. *Genome Biology and Evolution*, 6(4):776–791, 2014.
- [134] Elena Barghini, Lucia Natali, Tommaso Giordani, Rosa Maria Cossu, Simone Scalabrin, et al. LTR retrotransposon dynamics in the evolution of the olive (*Olea europaea*) genome. *DNA Research*, 22(1):91–100, feb 2015.
- [135] Fernando Cruz, Irene Julca, Jèssica Gómez-Garrido, Damian Loska, Marina Marcet-Houben, et al. Genome sequence of the olive tree, *Olea europaea*. *GigaScience*, 5(1):29, dec 2016.
- [136] Victor Preedy and Ronald Watson. *Olives and Olive Oil in Health and Disease Prevention*. Elsevier Inc., 2010.
- [137] Genevieve Buckland and Carlos A. Gonzalez. The role of olive oil in disease prevention: A focus on the recent epidemiological evidence from cohort studies and dietary intervention trials, jul 2015.
- [138] Catherine M. Breton and André Bervillé. New hypothesis elucidates self-incompatibility in the olive tree regarding S-alleles dominance relationships as in the sporophytic model. *Comptes Rendus - Biologies*, 335(9):563–572, sep 2012.
- [139] Pierre Saumitou-laprade, Philippe Vernet, Xavier Vekemans, Sylvain Billiard, Sophie Gallina, et al. Elucidation of the genetic architecture of self-incompatibility in olive : Evolutionary consequences and perspectives for orchard management. *Evolutionary applications*, 10(October 2016):867–880, oct 2017.

- [140] F. Alagna, M. E. Caceres, S. Pandolfi, S. Collani, S. Mousavi, et al. The Paradox of Self-Fertile Varieties in the Context of Self-Incompatible Genotypes in Olive. *Frontiers in Plant Science*, 10:725, jun 2019.
- [141] Mine Turktas, Behcet Inal, Sezer Okay, Emine Gulden Erkilic, Ekrem Dundar, et al. Nutrition Metabolism Plays an Important Role in the Alternate Bearing of the Olive Tree (*Olea europaea* L.). *PLoS ONE*, 8(3), 2013.
- [142] Jose A. Traverso, Amada Pulido, María I. Rodríguez-García, and Juan D. Alché. Thiol-based redox regulation in sexual plant reproduction: New insights and perspectives, 2013.
- [143] Adoración Zafra, María Isabel Rodríguez-García, and Juan de Dios Alché. Cellular localization of ROS and NO in olive reproductive tissues during flower development. *BMC plant biology*, 10:36, jan 2010.
- [144] Adoración Zafra, Juan D. Rejón, Simon J. Hiscock, and Juan de Dios Alché. Patterns of ROS accumulation in the stigmas of angiosperms and visions into their multi-functionality in plant reproduction. *Frontiers in Plant Science*, 7(AUG2016), 2016.
- [145] María J. Jiménez-Quesada, José Traverso, and Juan de Dios Alché. NADPH oxidase-dependent superoxide production in plant reproductive tissues, mar 2016.
- [146] María José Jimenez-Quesada, José Angel Traverso, Martin Potocký, Viktor Žárský, and Juan de Dios Alché. Generation of Superoxide by OeRbohH, a NADPH Oxidase Activity During Olive (*Olea europaea* L.) Pollen Development and Germination. *Frontiers in Plant Science*, 10:1149, sep 2019.
- [147] Martin Potocký, Přemysl Pejchar, Małgorzata Gutkowska, María José Jiménez-Quesada, Andrea Potocká, et al. NADPH oxidase activity in pollen tubes is affected by calcium ions, signaling phospholipids and Rac/Rop GTPases. *Journal of Plant Physiology*, 169(16):1654–1663, nov 2012.
- [148] Juan D. Alché, Francisco J. Corpas, Maria I. Rodriguez-Garcia, and Luis A. Del Rio. Identification and immunolocalization of superoxide dismutase isoenzymes of olive pollen. In *Physiologia Plantarum*, volume 104, pages 772–776. John Wiley & Sons, Ltd (10.1111), dec 1998.
- [149] A. Zafra, M. J. Jiménez-Quesada, J. A. Traverso, F. J. Corpas, M. I. Rodríguez-García, et al. Peroxisomal localization of cuzn superoxide dismutase in the male reproductive tissues of the olive tree. In *Microscopy and Microanalysis*, volume 18, pages 33–34. Cambridge University Press, aug 2012.
- [150] Adoración Zafra, Rosario Carmona, José C Jimenez-Lopez, Amada Pulido, M Gonzalo Claros, et al. Identification of Distinctive Variants of the Olive Pollen Allergen Ole e 5 (Cu,Zn Superoxide Dismutase) throughout the Analysis of the Olive Pollen Transcriptome. In *Bioinformatics and Biomedical Engineering: Third International Conference, IWBBIO 2015*, volume 9043, pages 460–470. 2015.
- [151] Adoración Zafra, Antonio Jesús Castro, and Juan de Dios Alché. Identification of novel superoxide dismutase isoenzymes in the olive (*Olea europaea* L.) pollen. *BMC Plant Biology*, 18(1):114, dec 2018.
- [152] Francisco J Corpas, Juan D Alché, and Juan B Barroso. Current overview of S-nitrosoglutathione (GSNO) in higher plants, 2013.
- [153] Elena Lima-Cabello, Isabel Martínez-Beas, Estefanía García-Quirós, Rosario Carmona, M. Gonzalo Claros, et al. Structural and Functional Features of Glutathione Reductase Transcripts from Olive (*Olea europaea* L.) Seeds. In *Lecture Notes in Computer Science (including subseries Lecture*

- Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), volume 11465 LNBI, pages 178–191. Springer, Cham, may 2019.
- [154] Estefanía García-Quirós, Juan de Dios Alché, Barbara Karpinska, and Christine H Foyer. Glutathione redox state plays a key role in flower development and pollen vigour. *Journal of Experimental Botany*, sep 2019.
  - [155] Jeremy Astier and Christian Lindermayr. Nitric oxide-dependent posttranslational modification in plants: An update, nov 2012.
  - [156] Shamsul Hayat, Syed Aiman Hasan, Masaki Mori, Qazi Fariduddin, and Aqil Ahmad. Nitric Oxide: Chemistry, Biosynthesis, and Physiological Role. In *Nitric Oxide in Plant Physiology*, pages 1–16. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2009.
  - [157] Manda Yu, Lorenzo Lamattina, Steven H. Spoel, and Gary J. Loake. Nitric oxide function in plant biology: A redox cue in deconvolution, jun 2014.
  - [158] Ian D. Wilson, Simon J. Hiscock, Philip E. James, and John T. Hancock. Nitric oxide and nitrite are likely mediators of pollen interactions. *Plant Signaling and Behavior*, 4(5):416–418, may 2009.
  - [159] Patricia Domingos, Ana Margarida Prado, Aloysius Wong, Christoph Gehring, and Jose A. Feijo. Nitric oxide: A multitasked signaling gas in plants, apr 2015.
  - [160] María José Jimenez-Quesada, Rosario Carmona, Elena Lima-Cabello, José Ángel Traverso, Antonio Jesús Castro, et al. Generation of nitric oxide by olive (*Olea europaea* L.) pollen during in vitro germination and assessment of the S-nitroso- and nitro-proteomes by computational predictive methods. *Nitric Oxide*, 68:23–37, aug 2017.
  - [161] Jasmeet Kaur Abat and Renu Deswal. Posttranslational Modifications of Proteins by Nitric Oxide: A New Tool of Metabolome Regulation. In *Nitric Oxide in Plant Physiology*, pages 189–201. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2009.
  - [162] Thomas Keller, Howard G Damude, Dietrich Werner, Peter Doerner, Richard A Dixon, et al. A plant homolog of the neutrophil NADPH oxidase gp91(phox) subunit gene encodes a plasma membrane protein with Ca<sup>2+</sup> binding motifs. *Plant Cell*, 10(2):255–266, feb 1998.
  - [163] M. Rodríguez-Serrano, D. M. Pazmiño, I. Sparkes, A. Rochetti, C. Hawes, et al. 2,4-Dichlorophenoxyacetic acid promotes S-nitrosylation and oxidation of actin affecting cytoskeleton and peroxisomal dynamics. *Journal of Experimental Botany*, 65(17):4783–4793, sep 2014.
  - [164] Capilla Mata-Pérez, Juan C. Begara-Morales, Mounira Chaki, Beatriz Sánchez-Calvo, Raquel Valderrama, et al. Protein tyrosine nitration during development and abiotic stress response in plants, nov 2016.
  - [165] Rosario Carmona Muñoz. El olivo como fuente de alérgenos. *Encuentros en la Biología*, IX(158):111–114, 2016.
  - [166] Mayte Villalba, Rosalía Rodríguez, and Eva Batanero. The spectrum of olive pollen allergens. From structures to diagnosis and treatment. *Methods (San Diego, Calif.)*, 66(1):44–54, mar 2014.
  - [167] Carmen Oeo-Santos, Salvador Mas, Joaquín Quiralte, Carlos Colás, Miguel Blanca, et al. A hypoallergenic polygalacturonase isoform from olive pollen is implicated in pollen-pollen cross-reactivity. *International Archives of Allergy and Immunology*, 177(4):290–301, 2018.
  - [168] Mayte Villalba, Eva Batanero, Carlos López-Otín, Luis M. Sánchez, Rafael I. Monsalve, et al. The amino acid sequence of Ole e I, the major allergen from olive tree (*Olea europaea*) pollen. *European Journal of Biochemistry*, 216(3):863–869, sep 1993.

- [169] Mayte Villalba, Eva Batanero, Rafael I Monsalve, Manuel A. González De La Peña, Carlos Lahoz, et al. Cloning and Expression of Ole e I, the Major Allergen from Olive Tree Pollen: Polymorphism Analysis and Tissue Specificity. *Journal of Biological Chemistry*, 269(21):15217–15222, may 1994.
- [170] Juan de Dios Alché, a J Castro, A Olmedilla, M C Fernández, R Rodríguez, et al. The major olive pollen allergen (Ole e I) shows both gametophytic and sporophytic expression during anther development, and its synthesis and storage takes place in the RER. *Journal of cell science*, 112 (Pt 1):2501–9, 1999.
- [171] Juan de Dios Alché, Mohamed M'rani-Alaoui, Antonio Jesús Castro, and María Isabel Rodríguez-García. Ole e 1, the major allergen from olive (*Olea europaea* L.) pollen, increases its expression and is released to the culture medium during in vitro germination. *Plant & cell physiology*, 45(8):1149–1157, sep 2004.
- [172] A Ledesma, R Rodríguez, and M Villalba. Olive-pollen profilin. Molecular and immunologic properties. *Allergy: European Journal of Allergy and Clinical Immunology*, 53(5):520–526, may 1998.
- [173] Juan A Asturias, M Carmen Arilla, Nuria Gómez-Bayón, Jorge Martínez, Alberto Martínez, et al. Cloning and expression of the panallergen profilin and the major allergen (Ole e 1) from olive tree pollen. *Journal of Allergy and Clinical Immunology*, 100(3):365–372, sep 1997.
- [174] Jose C Jimenez-Lopez, Sonia Morales, Antonio J Castro, Dieter Volkmann, María I Rodríguez-García, et al. Characterization of profilin polymorphism in pollen with a focus on multifunctionality. *PloS one*, 7(2):e30878, jan 2012.
- [175] Jose C Jimenez-Lopez, María I Rodríguez-García, and Juan D Alché. Analysis of the effects of polymorphism on pollen profilin structural functionality and the generation of conformational, T- and B-cell epitopes. *PloS one*, 8(10):e76066, jan 2013.
- [176] E Batanero, M Villalba, A Ledesma, X S Puente, and R Rodríguez. Ole e 3, an olive-tree allergen, belongs to a widespread family of pollen proteins. *European journal of biochemistry*, 241(3):772–778, nov 1996.
- [177] Amalia Ledesma, Mayte Villalba, Eva Batanero, and Rosalía Rodríguez. Molecular cloning and expression of active Ole e 3, a major allergen from olive-tree pollen and member of a novel family of Ca<sup>2+</sup>-binding proteins (polcalcins) involved in allergy. *European Journal of Biochemistry*, 258(2):454–459, dec 1998.
- [178] L Boluda, C Alonso, and E Fernández-Caldas. Purification, characterization, and partial sequencing of two new allergens of *Olea europaea*. *The Journal of allergy and clinical immunology*, 101(2 Pt 1):210–6, feb 1998.
- [179] Cinzia Butteroni, Claudia Afferni, Bianca Barletta, Patrizia Iacovacci, Silvia Corinti, et al. Cloning and expression of the *Olea europaea* allergen Ole e 5, the pollen Cu/Zn superoxide dismutase. *International Archives of Allergy and Immunology*, 137(1):9–17, may 2005.
- [180] J D Alché, A J Castro, J C Jiménez-López, S Morales, A Zafra, et al. Differential characteristics of olive pollen from different cultivars: biological and clinical implications. *Journal of investigational allergology & clinical immunology : official organ of the International Association of Asthmology (INTERASMA) and Sociedad Latinoamericana de Alergia e Inmunología*, 17 Suppl 1:17–23, 2007.
- [181] Eva Batanero, Amalia Ledesma, Mayte Villalba, and Rosalía Rodríguez. Purification, amino acid sequence and immunological characterization of Ole e 6, a cysteine-enriched allergen from olive tree pollen. *FEBS Letters*, 410(2-3):293–296, jun 1997.

- [182] Miguel Ángel Treviño, María Flor García-Mayoral, Patricia Barral, Mayte Villalba, Jorge Santoro, et al. NMR solution structure of Ole e 6, a major allergen from olive tree pollen. *Journal of Biological Chemistry*, 279(37):39035–39041, sep 2004.
- [183] M.Luisa Tejera, Mayte Villalba, Eva Batanero, and Rosalía Rodríguez. Identification, isolation, and characterization of Ole e 7, a new allergen of olive tree pollen. *Journal of Allergy and Clinical Immunology*, 104(4):797–802, oct 1999.
- [184] Carmen Oeo-Santos, Salvador Mas, Sara Benedé, María López-Lucendo, Joaquín Quiralte, et al. A recombinant isoform of the Ole e 7 olive pollen allergen assembled by de novo mass spectrometry retains the allergenic ability of the natural allergen. *Journal of Proteomics*, 187(June):39–46, 2018.
- [185] Amalia Ledesma, Mayte Villalba, and Rosalía Rodríguez. Cloning, expression and characterization of a novel four EF-hand  $\text{Ca}^{2+}$ -binding protein from olive pollen with allergenic activity. *FEBS Letters*, 466(1):192–196, jan 2000.
- [186] A Ledesma, M Villalba, F Vivanco, and R Rodríguez. Olive pollen allergen Ole e 8: Identification in mature pollen and presence of Ole e 8-like proteins in different pollens. *Allergy: European Journal of Allergy and Clinical Immunology*, 57(1):40–43, jan 2002.
- [187] S Huecas, M Villalba, and R Rodríguez. Ole e 9, a major olive pollen allergen is a 1,3-beta-glucanase. Isolation, characterization, amino acid sequence, and tissue specificity. *The Journal of biological chemistry*, 276(30):27959–66, jul 2001.
- [188] Patricia Barral, Eva Batanero, Oscar Palomares, Joaquín Quiralte, Mayte Villalba, et al. A Major Allergen from Pollen Defines a Novel Family of Plant Proteins and Shows Intra- and Interspecie Cross-Reactivity. *The Journal of Immunology*, 172(6):3644–3651, mar 2004.
- [189] Patricia Barral, Cinthya Suárez, Eva Batanero, Carlos Alfonso, Juan de Dios Alché, et al. An olive pollen protein with allergenic activity, Ole e 10, defines a novel family of carbohydrate-binding modules and is potentially implicated in pollen germination. *The Biochemical journal*, 390(Pt 1):77–84, aug 2005.
- [190] Guillermo Salamanca, Rosalía Rodríguez, Joaquín Quiralte, Carmen Moreno, Cristina Y Pascual, et al. Pectin methylesterases of pollen tissue, a major allergen in olive tree. *The FEBS journal*, 277(13):2729–39, jul 2010.
- [191] Jose C Jimenez-Lopez, Simeon O Kotchoni, María I Rodríguez-García, and Juan D Alché. Structure and functional features of olive pollen pectin methylesterase using homology modeling and molecular docking methods. *Journal of molecular modeling*, 18(12):4965–84, dec 2012.
- [192] Jose C. Jimenez-Lopez, Simeon O. Kotchoni, Maria C. Hernandez-Soriano, Emma W. Gachomo, and Juan D. Alch?? Structural functionality, catalytic mechanism modeling and molecular allergenicity of phenylcoumaran benzylic ether reductase, an olive pollen (Ole e 12) allergen. *Journal of Computer-Aided Molecular Design*, 27(10):873–895, 2013.
- [193] Lourdes Castro, Jesús F. Crespo, Julia Rodríguez, Rosalía Rodríguez, and Mayte Villalba. Immunoproteomic tools are used to identify masked allergens: Ole e 12, an allergenic isoflavone reductase from olive (*Olea europaea*) pollen. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1854(12):1871–1880, dec 2015.
- [194] Oscar Palomares, Manuel Alcántara, Joaquín Quiralte, Mayte Villalba, Francisco Garzón, et al. Airway disease and thaumatin-like protein in an olive-oil mill worker. *The New England journal of medicine*, 358(12):1306–8, mar 2008.

- [195] Clara Esteve, Benito Cañas, Estefanía Moreno-Gordaliza, Carmen Del Río, María Concepción García, et al. Identification of olive (*Olea europaea*) pulp proteins by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and nano-liquid chromatography tandem mass spectrometry. *Journal of Agricultural and Food Chemistry*, 59(22):12093–12101, nov 2011.
- [196] J.C. Jimenez-Lopez, P. Robles-Bolivar, F.J. Lopez-Valverde, E. Lima-Cabello, S.O. Kotchoni, et al. Ole e 13 is the unique food allergen in olive: structure-functional, substrates docking, and molecular allergenicity comparative analysis. *Journal of Molecular Graphics and Modelling*, mar 2016.
- [197] Pablo San Segundo-Acosta, Carmen Oeo-Santos, Sara Benedé, Vivian de los Ríos, Ana Navas, et al. Delineation of the Olive Pollen Proteome and Its Allergenome Unmasks Cyclophilin as a Relevant Cross-Reactive Allergen. *Journal of Proteome Research*, 18(8):3052–3066, aug 2019.
- [198] Martin D. Chapman, Anna Pomés, Heimo Breiteneder, and Fatima Ferreira. Nomenclature and structural biology of allergens. *Journal of Allergy and Clinical Immunology*, 119(2):414–420, 2007.
- [199] Abdelmounim Hamman-Khalifa, Antonio Jesús Castro, José Carlos Jiménez-López, María Isabel Rodríguez-García, and Juan de Dios Alché. Olive cultivar origin is a major cause of polymorphism for Ole e 1 pollen allergen. *BMC plant biology*, 8(1):10, jan 2008.
- [200] Rosario Muleo, Michele Morgante, Riccardo Velasco, Andrea Cavallini, Gaetano Perrotta, et al. Olive Tree Genomic. In *Olive Germplasm - The Olive Cultivation, Table Olive and Olive Oil Industry in Italy*, number 2003. 2012.
- [201] Turgay Unver, Zhangyan Wu, Lieven Sterck, Mine Turktas, Rolf Lohaus, et al. Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences*, page 201708621, oct 2017.
- [202] Nehir Ozdemir Ozgenturk, Fatma Oru, Ugur Sezerman, Alper Kuçukural, Senay Vural Korkut, et al. Generation and analysis of expressed sequence tags from *Olea europaea* L. *Comparative and Functional Genomics*, 2010, 2010.
- [203] Mari Carmen García-López, Isabel Vidoy, Jaime Jiménez-Ruiz, Antonio Muñoz-Mérida, Ana Fernández-Ocaña, et al. Genetic changes involved in the juvenile-to-adult transition in the shoot apex of *Olea europaea* L. occur years before the first flowering. *Tree Genetics and Genomes*, 10(3):585–603, 2014.
- [204] Jaime Jiménez-Ruiz, Mari Carmen García-López, Isabel Vidoy, María de la O Leyva-Pérez, Ana Fernández-Ocaña, et al. Transcriptional analysis of adult cutting and juvenile seedling olive roots. *Tree Genetics and Genomes*, 11(4):1–12, aug 2015.
- [205] Juan J González-Plaza, Inmaculada Ortiz-Martín, Antonio Muñoz-Mérida, Carmen García-López, José F Sánchez-Sevilla, et al. Transcriptomic Analysis Using Olive Varieties and Breeding Progenies Identifies Candidate Genes Involved in Plant Architecture. *Frontiers in plant science*, 7:240, jan 2016.
- [206] Jaime Jiménez-ruiz, María De O Leyva-pérez, Isabel Vidoy-mercado, Araceli Barceló, and Francisco Luque. Transcriptomic time-series analysis of early development in olive from germinated embryos to juvenile tree. *BMC Genomics*, 19(1):1–14, dec 2018.
- [207] Giulio Galla, Gianni Barcaccia, Angelo Ramina, Silvio Collani, Fiammetta Alagna, et al. Computational annotation of genes differentially expressed along olive fruit development. *BMC plant biology*, 9:128, 2009.
- [208] Fiammetta Alagna, Nunzio D'Agostino, Laura Torchia, Maurizio Servili, Rosa Rao, et al. Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC genomics*, 10:399, 2009.

- [209] Antonio Muñoz-Mérida, Juan José González-Plaza, Andrés Cañada, Ana María Blanco, Maria del Carmen García-López, et al. De novo assembly and functional annotation of the olive (*Olea europaea*) transcriptome. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 20(1):93–108, feb 2013.
- [210] Fiammetta Alagna, Marco Cirilli, Giulio Galla, Fabrizio Carbone, Loretta Daddiego, et al. Transcript Analysis and Regulative Events during Flower Development in Olive (*Olea europaea* L.). *PLoS one*, 11(4):e0152943, jan 2016.
- [211] Domenico Iaria, Adriana Chiappetta, and Innocenzo Muzzalupo. De Novo Transcriptome Sequencing of *Olea europaea* L. to Identify Genes Involved in the Development of the Pollen Tube. *The Scientific World Journal*, 2016:1–7, 2016.
- [212] Domenico L. Iaria, Adriana Chiappetta, and Innocenzo Muzzalupo. A De novo Transcriptomic Approach to Identify Flavonoids and Anthocyanins “Switch-Off” in Olive (*Olea europaea* L.) Drupes at Different Stages of Maturation. *Frontiers in Plant Science*, 6, jan 2016.
- [213] Adoración Zafra, Rosario Carmona, José A. Traverso, John T. Hancock, Maria H. S. Goldman, et al. Identification and Functional Annotation of Genes Differentially Expressed in the Reproductive Tissues of the Olive Tree (*Olea europaea* L.) through the Generation of Subtractive Libraries. *Frontiers in Plant Science*, 8:1576, sep 2017.
- [214] Jose A Gil-Amado and Maria C Gomez-Jimenez. Transcriptome analysis of mature fruit abscission control in olive. *Plant & cell physiology*, 54(2):244–69, feb 2013.
- [215] Ruben Parra, Miguel A Paredes, Isabel M Sanchez-Calle, and Maria C Gomez-Jimenez. Comparative transcriptional profiling analysis of olive ripe-fruit pericarp and abscission zone tissues shows expression differences and distinct patterns of transcriptional regulation. *BMC genomics*, 14(1):866, dec 2013.
- [216] Jaime Jiménez-Ruiz, María de la O Leyva-Pérez, Elisabetta Schilirò, Juan Bautista Barroso, Aureliano Bombarely, et al. Transcriptomic Analysis of *Olea europaea* L. Roots during the *Verticillium dahliae* Early Infection Process. *The Plant Genome*, 10(1):0, nov 2017.
- [217] Christos Bazakos, Maria E Manioudaki, Ioannis Therios, Demetrios Voyiatzis, Dimitris Kafetzopoulos, et al. Comparative transcriptome analysis of two olive cultivars in response to NaCl-stress. *PLoS one*, 7(8):e42931, jan 2012.
- [218] Christos Bazakos, Maria E. Manioudaki, Elena Sarropoulou, Thodhoraq Spano, and Panagiotis Kalaitzis. 454 pyrosequencing of olive (*Olea europaea* L.) transcriptome in response to salinity. *PLoS ONE*, 10(11):e0143000, nov 2015.
- [219] Davide Guerra, Antonella Lamontanara, Paolo Bagnaresi, Luigi Orrù, Fulvia Rizza, et al. Transcriptome changes associated with cold acclimation in leaves of olive tree (*Olea europaea* L.). *Tree Genetics & Genomes*, 11(6):113, oct 2015.
- [220] María De La O Leyva-Pérez, Antonio Valverde-Corredor, Raquel Valderrama, Jaime Jiménez-Ruiz, Antonio Muñoz-Merida, et al. Early and delayed long-term transcriptional changes and short-term transient responses during cold acclimation in olive leaves. *DNA Research*, 22(1):1–11, feb 2015.
- [221] Ahmet İpek, Meryem İpek, Sezai Ercişli, and Nesrin Aktepe Tangu. Transcriptome-based SNP discovery by GBS and the construction of a genetic map for olive. *Functional & Integrative Genomics*, 2017.
- [222] Livia Donaire, Laia Pedrola, Raúl de la Rosa, and César Llave. High-throughput sequencing of rna silencing-associated small RNAs in olive (*Olea europaea* L.). *PLoS ONE*, 6(11):e27916, nov 2011.

- [223] Huriye Yanik, Mine Turktas, Ekrem Dundar, Pilar Hernandez, Gabriel Dorado, et al. Genome-wide identification of alternate bearing-associated microRNAs (miRNAs) in olive (*Olea europaea* L.). *BMC Plant Biology*, 13(1):10, 2013.
- [224] Fabrizio Carbone, Leonardo Bruno, Gaetano Perrotta, Maria B. Bitonti, Innocenzo Muzzalupo, et al. Identification of miRNAs involved in fruit ripening by deep sequencing of *Olea europaea* L. transcriptome. *PLOS ONE*, 14(8):e0221460, aug 2019.
- [225] Gautier Sarah, Felix Homa, Stéphanie Pointet, Sandy Contreras, François Sabot, et al. A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. *Molecular Ecology Resources*, 17(3):565–580, may 2017.
- [226] Linda Bianco, Fiammetta Alagna, Luciana Baldoni, Christine Finnie, Birte Svensson, et al. Proteome Regulation during *Olea europaea* Fruit Development. *PLoS ONE*, 8(1):e53563, jan 2013.
- [227] Anna Laura Capriotti, Chiara Cavaliere, Patrizia Foglia, Susy Piovesana, Roberto Samperi, et al. Proteomic platform for the identification of proteins in olive (*Olea europaea*) pulp. *Analytica Chimica Acta*, 800:36–42, oct 2013.
- [228] Juan David Rejón, François Delalande, Christine Schaeffer-Reiss, Christine Carapito, Krzysztof Zienkiewicz, et al. The plant stigma exudate. A biochemically active extracellular environment for pollen germination? *Plant signaling & behavior*, 9(4):e28274, 2014.
- [229] Noela Prado, Juan De Dios Alché, Juan Casado-Vela, Salvador Mas, Mayte Villalba, et al. Nanovesicles are secreted during pollen germination and pollen tube growth: A possible role in fertilization, 2014.
- [230] Juan Rejón, François Delalande, Christine Schaeffer-Reiss, Juan Alché, María Rodríguez-García, et al. The Pollen Coat Proteome: At the Cutting Edge of Plant Reproduction. In *Proteomes*, volume 4, page 5. Multidisciplinary Digital Publishing Institute, jan 2016.
- [231] Loukia Roka, Konstantinos Koudounas, Gerasimos Daras, Jerome Zoidakis, Antonia Vlahou, et al. Proteome of olive non-glandular trichomes reveals protective protein network against (a)biotic challenge. *Journal of Plant Physiology*, 231:210–218, dec 2018.
- [232] Giandomenico Corrado, Fiammetta Alagna, Mariapina Rocco, Giovanni Renzone, Paola Varrichio, et al. Molecular interactions between the olive and the fruit fly *Bactrocera oleae*. *BMC Plant Biology*, 12(1):86, jun 2012.
- [233] Fiammetta Alagna, Mario Kallenbach, Andrea Pompa, Francesca De Marchis, Rosa Rao, et al. Olive fruits infested with olive fly larvae respond with an ethylene burst and the emission of specific volatiles. *Journal of Integrative Plant Biology*, 58(4):413–425, apr 2016.
- [234] Rao Guodong, Liu Xiaoxia, Zha Weiwei, Wu Wenjun, and Zhang Jianguo. Metabolomics reveals variation and correlation among different tissues of olive (*Olea europaea* L.). *Biology Open*, 6(9):1317–1323, sep 2017.
- [235] Maurizio Servili, Beatrice Sordini, Sonia Esposto, Agnese Taticchi, Stefania Urbani, et al. Metabolomics of Olive Fruit: A Focus on the Secondary Metabolites. In *The Olive Tree Genome*, pages 123–139. Springer, Cham, 2016.
- [236] Lucía Olmo García. *Aplicaciones de interés en el campo de la metabolómica de alimentos: caracterización de productos y subproductos derivados del olivar*. PhD thesis, Universidad de Granada, oct 2018.

- [237] Damián Maestri, Damián Barrionuevo, Romina Bodoira, Adoración Zafra, José Jiménez-López, et al. Nutritional profile and nutraceutical components of olive (*Olea europaea* L.) seeds. *Journal of Food Science and Technology*, 56(9):4359–4370, sep 2019.
- [238] Eva Batanero, Rosalía Rodríguez, and Mayte Villalba. Olive Pollen Allergens: An Insight into Clinical, Diagnostic and Therapeutic Concepts of Allergy. In *Olives and Olive Oil in Health and Disease Prevention*, pages 1021–1031. Elsevier, 2010.
- [239] Enrico M. Bucci. *Xylella fastidiosa*, a new plant pathogen that threatens global farming: Ecology, molecular biology, search for remedies, jul 2018.
- [240] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009.
- [241] Ricardo H Ramirez-Gonzalez, Raoul Bonnal, Mario Caccamo, and Daniel MacLean. Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. *Source Code for Biology and Medicine*, 7(1):6, may 2012.
- [242] Naohisa Goto, Pjotr Prins, Mitsuteru Nakao, Raoul Bonnal, Jan Aerts, et al. BioRuby: Bioinformatics software for the Ruby programming language. *Bioinformatics*, 26(20):2617–2619, 2010.
- [243] Weizhong Li and Adam Godzik. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.
- [244] Manolo Gouy, Stéphane Guindon, and Olivier Gascuel. Sea view version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2):221–224, feb 2010.
- [245] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron, and Ewan Birney. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, apr 2012.
- [246] Yinlong Xie, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12):1660–1666, 2014.
- [247] Sébastien Boisvert, François Laviolette, and Jacques Corbeil. Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, 17(11):1401–1415, nov 2010.
- [248] Todd J. Treangen, Dan D. Sommer, Florent E. Angly, Sergey Koren, and Mihai Pop. Next generation sequence assembly with AMOS. *Current Protocols in Bioinformatics*, Chapter 11(SUPP.33):1–18, mar 2011.
- [249] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494–1512, 2013.
- [250] David Velasco, Pedro Seoane, and M. Gonzalo Claros. Bioinformatics Analyses to Separate Species Specific mRNAs from Unknown Sequences in de novo Assembled Transcriptomes. In *3rd International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 322–332, 2015.
- [251] Sara Ocaña, Pedro Seoane, Rocio Bautista, Carmen Palomino, Gonzalo M Claros, et al. Large-Scale Transcriptome Analysis in Faba Bean (*Vicia faba* L.) under *Ascochyta fabae* Infection. *PloS one*, 10(8):e0135143, 2015.
- [252] Roman Kolpakov, Ghizlane Bana, and Gregory Kucherov. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research*, 31(13):3672–3678, 2003.

- [253] Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, oct 2015.
- [254] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48, dec 2009.
- [255] Pedro Carmona-Saez, Monica Chagoyen, Francisco Tirado, Jose M Carazo, and Alberto Pascual-Montano. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome biology*, 8(1):R3, 2007.
- [256] Ruben Nogales-Cadenas, Pedro Carmona-Saez, Miguel Vazquez, Cesar Vicente, Xiaoyuan Yang, et al. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic acids research*, 37(Web Server issue):W317–22, jul 2009.
- [257] Daniel Tabas-Madrid, Ruben Nogales-Cadenas, and Alberto Pascual-Montano. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic acids research*, 40(Web Server issue):W478–83, jul 2012.
- [258] Yan Xu, Xiao-Jian Shao, Ling-Yun Wu, Nai-Yang Deng, and Kuo-Chen Chou. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine <i>S</i>-nitrosylation sites in proteins. *PeerJ*, 1:e171, oct 2013.
- [259] Yu Xue, Zexian Liu, Xinjiao Gao, Changjiang Jin, Longping Wen, et al. GPS-SNO: Computational Prediction of Protein S-Nitrosylation Sites with a Modified GPS Algorithm. *PLoS ONE*, 5(6):e11290, jun 2010.
- [260] Tzong Yi Lee, Yi Ju Chen, Tsung Cheng Lu, Hsien Da Huang, and Yu Ju Chen. Snosite: Exploiting maximal dependence decomposition to identify cysteine S-Nitrosylation with substrate site specificity. *PLoS ONE*, 6(7):e21849, jul 2011.
- [261] Zexian Liu, Jun Cao, Qian Ma, Xinjiao Gao, Jian Ren, et al. GPS-YN02: computational prediction of tyrosine nitration sites in proteins. *Molecular BioSystems*, 7(4):1197, 2011.
- [262] Pedro Seoane, Rosario Carmona, Rocío Bautista, Darío Guerrero-Fernández, and Gonzalo Claros. AutoFlow : an easy way to build workflows. In *Proceedings IWBBIO*, pages 342–349, 2014.
- [263] Evgeny M. Zdobnov, Fredrik Tegenfeldt, Dmitry Kuznetsov, Robert M. Waterhouse, Felipe A. Simao, et al. OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, 45(D1):D744–D749, jan 2017.
- [264] Sébastien Lê, Julie Josse, and François Husson. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, mar 2008.
- [265] Ulrich Lutz, David Posé, Matthias Pfeifer, Heidrun Gundlach, Jörg Hagmann, et al. Modulation of Ambient Temperature-Dependent Flowering in *Arabidopsis thaliana* by Natural Variation of FLOWERING LOCUS M. *PLOS Genetics*, 11(10):e1005588, 2015.
- [266] Rosario Carmona, Macarena Arroyo, María José Jiménez-Quesada, Pedro Seoane, Adoración Zafra, et al. Automated identification of reference genes based on RNA-seq data. *BioMedical Engineering OnLine*, 16(S1):65, aug 2017.
- [267] F Liang. An optimized protocol for analysis of EST sequences. *Nucleic Acids Research*, 28(18):3657–3665, sep 2000.
- [268] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.

- [269] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–212, oct 2014.
- [270] Adriano Mari, Chiara Rasi, Paola Palazzo, and Enrico Scala. Allergen databases: Current status and perspectives. *Current Allergy and Asthma Reports*, 9(5):376–383, aug 2009.
- [271] Adriano Mari, Enrico Scala, Paola Palazzo, Stefano Ridolfi, Danila Zennaro, et al. Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a model. *Cellular immunology*, 244(2):97–100, dec 2006.
- [272] Stefanie Dukowic-Schulze and Changbin Chen. The meiotic transcriptome architecture of plants. *Frontiers in plant science*, 5(June):220, 2014.
- [273] Nicholas Rutley and David Twell. A decade of pollen transcriptomics. *Plant reproduction*, 28(2):73–89, jun 2015.
- [274] Darío Guerrero, Rocío Bautista, David P Villalobos, Francisco R Cantón, and M Gonzalo Claros. AlignMiner: A Web-based tool for detection of divergent regions in multiple sequence alignments of conserved sequences. *Algorithms for Molecular Biology*, 5(1):24, jun 2010.
- [275] Kamal Foud Abdellatif and Yehia Abdallah Khidr. Genetic diversity of new maize hybrids based on SSR markers as compared with other molecular and biochemical markers. *Journal of Crop Science and Biotechnology*, 13(3):139–145, sep 2010.
- [276] Antonio A.F. Garcia, Luciana L. Benchimol, Antônia M.M. Barbosa, Isaias O. Geraldi, Cláudio L. Souza, et al. Comparison of RAPD, RFLP, AFLP and SSR markers for diversity studies in tropical maize inbred lines. *Genetics and Molecular Biology*, 27(4):579–588, 2004.
- [277] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften*, 131(4):281–5, dec 2012.
- [278] Maurício Rocha-Martins, Brian Njaine, and Mariana S Silveira. Avoiding pitfalls of internal controls: validation of reference genes for analysis by qRT-PCR and Western blot throughout rat retinal development. *PloS one*, 7(8):e43028, jan 2012.
- [279] Dave Lanoix, Andrée-Anne Lacasse, Joey St-Pierre, Sean C Taylor, Maude Ethier-Chiasson, et al. Quantitative PCR pitfalls: the case of the human placenta. *Molecular biotechnology*, 52(3):234–43, nov 2012.
- [280] Ying Wang, Yajuan Chen, Liping Ding, Jiewei Zhang, Jianhua Wei, et al. Validation of Reference Genes for Gene Expression by Quantitative Real-Time RT-PCR in Stem Segments Spanning Primary to Secondary Growth in *Populus tomentosa*. *PloS one*, 11(6):e0157370, jan 2016.
- [281] Jeffrey S. Coker and Eric Davies. Selection of candidate housekeeping controls in tomato plants using EST data. *BioTechniques*, 35(4):740–748, 2003.
- [282] Lei Chen, Hai-ying Zhong, Jian-fei Kuang, Jian-guo Li, Wang-jin Lu, et al. Validation of reference genes for RT-qPCR studies of gene expression in banana fruit under different experimental conditions. *Planta*, 234(2):377–90, 2011.
- [283] Zhaoguo Tong, Zhihong Gao, Fei Wang, Jun Zhou, and Zhen Zhang. Selection of reliable reference genes for gene expression studies in peach using real-time PCR. *BMC molecular biology*, 10:71, jan 2009.
- [284] Mukesh Jain, Aashima Nijhawan, Akhilesh K. Tyagi, and Jitendra P. Khurana. Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. *Biochemical and Biophysical Research Communications*, 345(2):646–651, 2006.

- [285] Tomasz Czechowski, Mark Stitt, Thomas Altmann, Michael K. Udvardi, and Wolf Rüdiger Scheible. Genome-wide identification and testing of superior reference genes for transcript normalization in arabidopsis, 2005.
- [286] Tripti Vashisth, Lisa Klima Johnson, and Anish Malladi. An efficient RNA isolation procedure and identification of reference genes for normalization of gene expression in blueberry. *Plant cell reports*, 30(12):2167–76, 2011.
- [287] Sinara Artico, Sarah M Nardeli, Osmundo Brilhante, Maria Fátima Grossi-de Sa, and Marcio Alves-Ferreira. Identification and evaluation of new reference genes in *Gossypium hirsutum* for accurate normalization of real-time quantitative RT-PCR data. *BMC plant biology*, 10(1):49, jan 2010.
- [288] Amy M Brunner, Igor A Yakovlev, and Steven H Strauss. Validating internal controls for quantitative plant gene expression studies. *BMC plant biology*, 4:14, 2004.
- [289] Alice Vezzaro, Sandra T. Krause, Alberto Nonis, Angelo Ramina, Jörg Degenhardt, et al. Isolation and characterization of terpene synthases potentially involved in flavor development of ripening olive (*Olea europaea*) fruits. *Journal of Plant Physiology*, 169(9):908–914, 2012.
- [290] N. Nicot. Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. *Journal of Experimental Botany*, 56(421):2907–2914, 2005.
- [291] Alberto Nonis, Alice Vezzaro, and Benedetto Ruperti. Evaluation of RNA extraction methods and identification of putative reference genes for real-time quantitative polymerase chain reaction expression studies on olive (*olea europaea* L.) fruits. *Journal of Agricultural and Food Chemistry*, 60:6855–6865, 2012.
- [292] Carla F Barsalobres-Cavallari, Fábio E Severino, Mirian P Maluf, and Ivan G Maia. Identification of suitable internal control genes for expression studies in *Coffea arabica* under different experimental conditions. *BMC molecular biology*, 10:1, 2009.
- [293] Tjasa Resetic, Natasa Stajner, Dunja Bandelj, Branka Javornik, and Jernej Jakse. Validation of candidate reference genes in RT-qPCR studies of developing olive fruit and expression analysis of four genes involved in fatty acids metabolism. *Molecular Breeding*, 32(1):211–222, 2013.
- [294] Catherine M Ronning, Svetlana S Stegalkina, Robert a Ascenzi, Oleg Bougri, Amy L Hart, et al. Comparative analyses of potato expressed sequence tag libraries. *Plant physiology*, 131:419–429, 2003.
- [295] Asad Maroufi, Erik Van Bockstaele, and Marc De Loose. Validation of reference genes for gene expression analysis in chicory (*Cichorium intybus*) using quantitative real-time PCR. *BMC molecular biology*, 11:15, 2010.
- [296] Karen E Reid, Niclas Olsson, James Schlosser, Fred Peng, and Steven T Lund. An optimized grapevine RNA isolation procedure and statistical determination of reference genes for real-time RT-PCR during berry development. *BMC plant biology*, 6:27, 2006.
- [297] José V. Die, Belén Román, Salvador Nadal, and Clara I. González-Verdejo. Evaluation of candidate reference genes for expression studies in *Pisum sativum* under different experimental conditions. *Planta*, 232:145–153, 2010.
- [298] Bornali Gohain. Rubisco-bis-phosphate oxygenase (RuBP)- A potential housekeeping gene for qPCR assays in tea. *African Journal of Biotechnology*, 11(51):11193–11199, 2012.

- [299] Karin Hoedemaekers, Jan Derksen, Suzanne W. Hoogstrate, Mieke Wolters-Arts, Sung-Aeong Oh, et al. BURSTING POLLEN is required to organize the pollen germination plaque and pollen tube tip in *Arabidopsis thaliana*. *The New phytologist*, 206(1):255–67, apr 2015.
- [300] Chunsun Gu, Sumei Chen, Zhaolei Liu, Hong Shan, Huolin Luo, et al. Reference gene selection for quantitative real-time PCR in *Chrysanthemum* subjected to biotic and abiotic stress. *Molecular biotechnology*, 49(2):192–197, 2011.
- [301] Lieven Thorrez, Katrijn Van Deun, L??on Charles Tranchevent, Leentje Van Lommel, Kristof Engelen, et al. Using ribosomal protein genes as reference: A tale of caution. *PLoS ONE*, 3(3):e1854, mar 2008.
- [302] R Carmona, A Zafra, J C Jiménez-López, P Seoane, J D Alché, et al. In the search of non described olive tree allergens by transcriptome-mining. *Inflammation Research, Conference: 44th Annual Meeting of the European-Histamine-Research-Society (EHRS)*, 64(Suppl 1:S1–S50, 2015.
- [303] Kiran Kadam, Rajiv Karbhal, V. K. Jayaraman, Sangeeta Sawant, and Urmila Kulkarni-Kale. AllerBase: a comprehensive allergen knowledgebase. *Review of Economic Studies*, 2017, jan 2017.
- [304] Payam Morakabati, Mohammad-Ali Assarehzadegan, Gholam Reza Khosravi, Bahareh Akbari, and Fatemeh Dousti. Cloning and Expression of Ama r 1, as a Novel Allergen of *Amaranthus retroflexus* Pollen. *Journal of Allergy*, 2016:1–8, jan 2016.
- [305] Takeshi Mori, Minehiko Yokoyama, Naoki Komiyama, Mitsuhiro Okano, and Kohsuke Kino. Purification, identification, and cDNA cloning of Cha o 2, the second major allergen of Japanese cypress pollen. *Biochemical and Biophysical Research Communications*, 263(1):166–171, sep 1999.
- [306] Rodrigo Barderas, Mayte Villalba, Manuel Lombardero, and Rosalía Rodríguez. Identification and characterization of Che a 1 allergen from *Chenopodium album* pollen. *International Archives of Allergy and Immunology*, 127(1):47–54, jan 2002.
- [307] R Rodríguez, M Villalba, E Batanero, O Palomares, J Quiralte, et al. Olive pollen recombinant allergens: value in diagnosis and immunotherapy., 2007.
- [308] Abdol-Reza Varasteh, Mojtaba Sankian, Terumi Midoro-Horiuti, Malihe Moghadam, Mohammad Taghi Shakeri, et al. Molecular cloning and expression of Cro s 1: an occupational allergen from saffron pollen (*Crocus sativus*). *Reports of biochemistry & molecular biology*, 1(1):1–8, oct 2012.
- [309] M. Lombardero, T. Obispo, B. Calabozo, A Lezaún, F. Polo, et al. Cross-reactivity between olive and other species. Role of Ole e 1-related proteins. *Allergy*, 57 Suppl 7(s71):29–34, apr 2002.
- [310] Rodrigo Barderas, Ashok Purohit, Rosalía Rodríguez, Gabrielle Pauli, and Mayte Villalba. Isolation of the main allergen Fra e 1 from ash (*Fraxinus excelsior*) pollen: Comparison of the natural and recombinant forms. *Annals of Allergy, Asthma and Immunology*, 96(4):557–563, apr 2006.
- [311] Michael Hauser, Gabriele Gadermaier, Sabrina Wildner, Lisa Pointner, Michael Wallner, et al. Weed pollen allergens. In Richard F. Lockey and Dennis K. Ledford, editors, *Allergens and Allergen Immunotherapy: Subcutaneous, Sublingual, and Oral*, chapter 12. CRC PRESS, sixth edition, 2020.
- [312] Brian A Baldo, Raphael C Panzani, Diana Bass, and R Zerboni. Olive (*Olea europea*) and privet (*Ligustrum vulgare*) pollen allergens. Identification and cross-reactivity with grass pollen proteins. *Molecular Immunology*, 29(10):1209–1218, oct 1992.
- [313] Oscar Palomares, Ines Swoboda, Mayte Villalba, Nadia Balic, Susanne Spitzauer, et al. The major allergen of olive pollen Ole e 1 is a diagnostic marker for sensitization to Oleaceae. *International Archives of Allergy and Immunology*, 141(2):110–118, 2006.

- [314] Fang Chao Zhu, Rui Zong Jia, Lin Xu, Hua Kong, Yun Ling Guo, et al. Genome-wide analysis of potential cross-reactive endogenous allergens in rice (*Oryza sativa* L.). *Toxicology Reports*, 2:1233–1245, jul 2015.
- [315] A J Castro, J D Alché, B Calabozo, M I Rodríguez-García, and F Polo. Pla l 1 and Ole e 1 pollen allergens share common epitopes and similar ultrastructural localization. *Journal of investigational allergology & clinical immunology : official organ of the International Association of Asthmology (INTERASMA) and Sociedad Latinoamericana de Alergia e Inmunología*, 17 Suppl 1:41–47, 2007.
- [316] Fatemeh Dousti, Mohammad Ali Assarehzadegan, Payam Morakabati, Gholam Reza Khosravi, and Bahareh Akbari. Molecular Cloning and Expression of Pro j 1: A new Allergen of Prosopis juliflora Pollen. *Iranian Journal of Allergy, Asthma and Immunology*, 15(2):122–131, 2016.
- [317] L. Castro, S. Mas, R. Barderas, C. Colás, J. García-Sellés, et al. Sal k 5, a member of the widespread ole e 1-like protein family, is a new allergen of russian thistle (*Salsola kali*) pollen. *International Archives of Allergy and Immunology*, 163(2):142–153, 2014.
- [318] Gholam Reza Khosravi, Mohammad Ali Assarehzadegan, Payam Morakabati, Bahareh Akbari, and Fatemeh Dousti. Cloning and expression of Aca f 1: A new allergen of Acacia farnesiana pollen. *Central European Journal of Immunology*, 41(3):273–281, 2016.
- [319] Mohsen Tehrani, Mojtaba Sankian, Mohammad Ali Assarehzadegan, Reza Falak, Reihaneh Noorbakhsh, et al. Identification of a new allergen from *Amaranthus retroflexus* pollen, *Ama r 2.*, sep 2011.
- [320] R. Asero, N. Wopfner, P. Gruber, G. Gadermaier, and F. Ferreira. Artemisia and Ambrosia hypersensitivity: Co-sensitization or co-recognition? *Clinical and Experimental Allergy*, 36(5):658–665, may 2006.
- [321] J. Reindl, H. P. Rihs, S. Scheurer, A. Wangorsch, D. Haustein, et al. IgE reactivity to profilin in pollen-sensitized subjects with adverse reactions to banana and pineapple. *International Archives of Allergy and Immunology*, 128(2):105–114, 2002.
- [322] S. Scheurer, A. Wangorsch, D. Haustein, and S. Vieths. Cloning of the minor allergen Api g 4 profilin from celery (*Apium graveolens*) and its cross-reactivity with birch pollen profilin Bet v 2. *Clinical and Experimental Allergy*, 30(7):962–971, 2000.
- [323] Christiane K. Fæste, Karen R. Jonscher, S. I.T. Louis, Jelena Klawitter, Kjersti E. Løvberg, et al. Differentiating cross-reacting allergens in the immunological analysis of celery (*apium graveolens*) by mass spectrometry. *Journal of AOAC International*, 93(2):451–461, 2010.
- [324] Cerrone Cabanos, Mary Rose Tandang-Silvas, Van Odijk, Peter Brostedt, Akira Tanaka, et al. Expression, purification, cross-reactivity and homology modeling of peanut profilin. *Protein Expression and Purification*, 73(1):36–45, 2010.
- [325] Won Lee Yong, Young Choi Soo, Kyung Lee Eun, Ho Sohn Jung, Jung Won Park, et al. Cross-allergenicity of pollens from the Compositae family: *Artemisia vulgaris*, *Dendranthema grandiflorum*, and *Taraxacum officinale*. *Annals of Allergy, Asthma and Immunology*, 99(6):526–533, 2007.
- [326] Lesa R. Offermann, Caleb R. Schlachter, Makenzie L. Perdue, Karolina A. Majorek, John Z. He, et al. Structural, functional, and immunological characterization of profilin panallergens Amb a 8, Art v 4, and Bet v 2. *Journal of Biological Chemistry*, 291(30):15447–15459, jul 2016.
- [327] Susanne Luoto, Wietske Lambert, Anna Blomqvist, and Cecilia Emanuelsson. The identification of allergen proteins in sugar beet (*Beta vulgaris*) pollen causing occupational allergy in greenhouses. *Clinical and Molecular Allergy*, 6:7, aug 2008.

- [328] Florin-Dan Popescu. Cross-reactivity between aeroallergens and food allergens. *World Journal of Methodology*, 5(2):31, jun 2015.
- [329] Alessandro Fiocchi, Lamia Dahdah, Carla Riccardi, Oscar Mazzina, and Vincenzo Fierro. Pre-acutionary labelling of cross-reactive foods: The case of rapeseed. *Asthma Research and Practice*, 2(1):13, dec 2016.
- [330] M. Willerroider, H. Fuchs, B. K. Ballmer-Weber, M. Focke, M. Susani, et al. Cloning and molecular and immunological characterisation of two new food allergens, Cap a 2 and Lyc e 1, profilins from bell pepper (*Capsicum annuum*) and tomato (*Lycopersicon esculentum*). *International Archives of Allergy and Immunology*, 131(4):245–255, 2003.
- [331] M. Ángeles López-Matas, Ángel Ferrer, Carlos H. Larramendi, Ángel Julio Huertas, Juan Antonio Pagán, et al. In vitro cross-reactivity between tomato and other plant allergens. *Annals of Allergy, Asthma and Immunology*, 103(5):425–431, 2009.
- [332] Xiaoyu Liu, Xiaojun Xiao, Lixin Xia, Ping Chang Yang, Zhigang Liu, et al. Identification and characterization of a pan-allergen profilin, a major allergen from *Caryota mitis* pollen. *Asian Pacific Journal of Allergy and Immunology*, 32(1):39–45, 2014.
- [333] Akram Amini, Mojtaba Sankian, Mohammad Ali Assarehzadegan, Fatemeh Vahedi, and Abdolreza Varasteh. *Chenopodium album* pollen profilin (Che a 2): Homology modeling and evaluation of cross-reactivity with allergenic profilins based on predicted potential IgE epitopes and IgE reactivity analysis. *Molecular Biology Reports*, 38(4):2579–2587, apr 2011.
- [334] Rodrigo Barderas, Mayte Villalba, and Rosalía Rodríguez. Recombinant expression, purification and cross-reactivity of chenopod profilin: rChe a 2 as a good marker for profilin sensitization. *Biological Chemistry*, 385(8):731–737, aug 2004.
- [335] Rodrigo Barderas, Mayte Villalba, Cristina Y. Pascual, Eva Batanero, and Rosalía Rodríguez. Profilin (Che a 2) and polcalcin (Che a 3) are relevant allergens of *Chenopodium album* pollen: Isolation, amino acid sequences, and immunologic properties. *Journal of Allergy and Clinical Immunology*, 113(6):1192–1198, jun 2004.
- [336] Bárbara Cases, Carlos Pastor-Vargas, Félix Gil Dones, Marina Perez-Gordo, Aroa S. Maroto, et al. Watermelon profilin: Characterization of a major allergen as a model for plant-derived food profilins. *International Archives of Allergy and Immunology*, 153(3):215–222, oct 2010.
- [337] Jesus F. Crespo, Mechthild Retzek, Kay Foetisch, Elena Sierra-Maestro, Ana B. Cid-Sanchez, et al. Germin-like protein Cit s 1 and profilin Cit s 2 are major allergens in orange (*Citrus sinensis*) fruits. *Molecular Nutrition and Food Research*, 50(3):282–290, mar 2006.
- [338] Nobue Takamatsu, Yasuto Kondo, Ikuya Tsuge, Yoichi Nakajima, Norihiko Naruse, et al. A study of cross-reactivity between citrus fruit and pollen allergens in oral allergy syndrome and food-dependent exercise-induced anaphylaxis in Japan. *Fujita Medical Journal*, 2(1):6–11, 2016.
- [339] Iris Lauer, Stefano Alessandri, Sven Pokoj, Andreas Reuter, Amedeo Conti, et al. Expression and characterization of three important panallergens from hazelnut. *Molecular Nutrition and Food Research*, 52(SUPPL. 2), nov 2008.
- [340] Abdol Reza Varasteh, Maliheh Moghadam, Fatemeh Vahedi, Tayyeb Kermani, and Mojtaba Sankian. Cloning and expression of the allergen Cro s 2 profilin from saffron (*Crocus sativus*). *Allergy International*, 58(3):429–435, jan 2009.
- [341] Mojtaba Sankian, Mahmoud Mahmoudi, and Abdol-Reza Varasteh. Molecular cloning, characterization, and expression of Cuc m 2, a major allergen in *Cucumis melo*. *Reports of biochemistry & molecular biology*, 1(2):49–63, 2013.

- [342] G. López-Torrejón, J. F. Crespo, R. Sánchez-Monge, M. Sánchez-Jiménez, J. Alvarez, et al. Allergenic reactivity of the melon profilin Cuc m 2 and its identification as major allergen. *Clinical and Experimental Allergy*, 35(8):1065–1072, 2005.
- [343] J. A. Asturias, M. C. Arilla, N. Gomez-Bayon, J. Martinez, A. Martinez, et al. Cloning and high level expression of Cynodon dactylon (Bermuda grass) pollen profilin (Cyn d 12) in Escherichia coli: Purification and characterization of the allergen. *Clinical and Experimental Allergy*, 27(11):1307–1313, 1997.
- [344] J A Asturias, M C Arilla, N Gómez-Bayón, M Aguirre, A. Martínez, et al. Cloning and immunological characterization of the allergen Hel a 2 (profilin) from sunflower pollen. *Molecular Immunology*, 35(8):469–478, jun 1998.
- [345] Kay Foetisch, Lotte Dahl, Baerbel Jansen, Wolf Meinhard Becker, Jonas Lidholm, et al. Development and in-house validation of allergen-specific ELISA tests for the quantification of Dau c 1.01, Dau c 1.02 and Dau c 4 in carrot extracts (Daucus carota). *Analytical and Bioanalytical Chemistry*, 399(2):935–943, jan 2011.
- [346] Véronique Bordas-Le Floch, Maxime Le Mignon, Laetitia Bussi eres, Karine Jain, Armelle Martelet, et al. A combined transcriptome and proteome analysis extends the allergome of house dust mite Dermatophagoides species. *PLoS ONE*, 12(10):1–15, 2017.
- [347] A. L. Karlsson, R. Aim, B. Ekstrand, S. Fj elkner-Modig, A. Schi ott, et al. Bet v 1 homologues in strawberry identified as IgE-binding proteins and presumptive allergens. *Allergy: European Journal of Allergy and Clinical Immunology*, 59(12):1277–1284, dec 2004.
- [348] Plaimlein Amnuaycheewa and Elvira Gonzalez de Mejia. Purification, characterisation, and quantification of the soy allergen profilin (Gly m 3) in soy products. *Food Chemistry*, 119(4):1671–1680, apr 2010.
- [349] J Quiralte, L Palacios, R Rodr guez, B C rdaba, J. M. Arias de Saavedra, et al. Modelling diseases: the allergens of Olea europaea pollen., 2007.
- [350] Olivia E. McKenna, Claudia Asam, Galber R. Araujo, Anargyros Roulias, Luiz R. Goulart, et al. How relevant is panallergen sensitization in the development of allergies?, sep 2016.
- [351] Virginia Bellido Linares. *Determinaci n de los perfiles de sensibilizaci n pol nica y relevancia cl nica de los al rgenos de alimentos en pacientes pol nicos sensibilizados a profilina*. PhD thesis, Universidad de Sevilla, 2020.
- [352] Farnoosh Zarinhadideh, Akram Amini, Mohammad Ali Assarehzadegan, Seyed Hamid Borsi, Najmeh Sepahi, et al. Immunochemical and molecular characterization of allergenic profilin (Koc s 2) from Kochia scoparia pollen. *Journal of the Korean Society for Applied Biological Chemistry*, 58(3):443–451, jun 2015.
- [353] Juan Juan Song, Hong Yun Zhang, Zhi Gang Liu, and Pi Xin Ran. Cloning of the panallergen profilin from lychee fruit and its cross-reactivity with birch pollen profilin Bet v 2. *Food and Agricultural Immunology*, 18(2):129–138, jun 2007.
- [354] Y. Ma, L. Zuidmeer, B. Bohle, S. T.H. Bolhaar, G. Gadermaier, et al. Characterization of recombinant Mal d 4 and its application for component-resolved diagnosis of apple allergy. *Clinical and Experimental Allergy*, 36(8):1087–1096, aug 2006.
- [355] Montserrat Fern ndez-Rivas, Suzanne Bolhaar, Eloina Gonz lez-Mancebo, Riccardo Asero, Astrid van Leeuwen, et al. Apple allergy across Europe: How allergen sensitization profiles determine the clinical expression of allergies to plant foods. *Journal of Allergy and Clinical Immunology*, 118(2):481–488, aug 2006.

- [356] A. Paschke, H. Kinder, K. Zunker, M. Wigotzki, H. Steinhart, et al. Characterization of cross-reacting allergens in mango fruit. *Allergy: European Journal of Allergy and Clinical Immunology*, 56(3):237–242, 2001.
- [357] Natalia Ukleja-Sokołowska, Ewa Gawrońska-Ukleja, Kinga Lis, Magdalena Zbikowska-Gotz, Łukasz Sokołowski, et al. Anaphylactic reaction in patient allergic to mango. *Allergy, Asthma and Clinical Immunology*, 14(1), oct 2018.
- [358] Timo Palosuo, Raphaël C. Panzani, Anand B. Singh, Renato Ariano, Harri Alenius, et al. Allergen cross-reactivity between proteins of the latex from *Hevea brasiliensis*, seeds and pollen of *Ricinus communis*, and pollen of *Mercurialis annua*, members of the Euphorbiaceae family. *Allergy and Asthma Proceedings*, 23(2):141–147, 2002.
- [359] Angel Vallverdú, Juan A Asturias, M Carmen Arilla, Nuria Gómez-Bayón, Alberto Martínez, et al. Characterization of recombinant *Mercurialis annua* major allergen Mer a 1 (profilin). *Journal of Allergy and Clinical Immunology*, 101(3):363–370, mar 1998.
- [360] J. A. Asturias, I. Ibarrola, J. Fernández, M. C. Arilla, R. González-Rioja, et al. Phod 2, a major allergen from date palm pollen, is a profilin: Cloning, sequencing, and immunoglobulin E cross-reactivity with other profilins. *Clinical and Experimental Allergy*, 35(3):374–381, mar 2005.
- [361] Irene Mittermann, Ines Swoboda, Elisabeth Pierson, Norbert Eller, Dietrich Kraft, et al. Molecular cloning and characterization of profilin from tobacco (*Nicotiana tabacum*): increased profilin expression during pollen maturation. *Plant Molecular Biology*, 27(1):137–146, jan 1995.
- [362] Scott D Russell, Prem L Bhalla, and Mohan B Singh. Transcriptome-based examination of putative pollen allergens of rice (*Oryza sativa* ssp. *japonica*). *Molecular plant*, 1(5):751–9, sep 2008.
- [363] C. Radauer, M. Willeroider, H. Fuchs, K. Hoffmann-Sommergruber, J. Thalhamer, et al. Cross-reactive and species-specific immunoglobulin E epitopes of plant profilins: An experimental and structure-based analysis. *Clinical and Experimental Allergy*, 36(7):920–929, 2006.
- [364] Juan A. Asturias, I. Ibarrola, J. L. Eseverri, M. Carmen Arilla, R. González-Rioja, et al. PCR-based cloning and immunological characterization of *Parietaria judaica* pollen profilin. *Journal of Investigational Allergology and Clinical Immunology*, 14(1):43–48, 2004.
- [365] Florin-Dan Popescu. Molecular biomarkers for grass pollen immunotherapy. *World Journal of Methodology*, 4(1):26, 2014.
- [366] G. Lund, S. Brand, T. Ramos, L. Jimeno, P. Boissy, et al. Strong and frequent T-cell responses to the minor allergen Phl p 12 in Spanish patients IgE-sensitized to Profilins. *Allergy: European Journal of Allergy and Clinical Immunology*, 73(5):1013–1021, may 2018.
- [367] Hosein Ali-Sadeghi, Ali Khodadadi, Akram Amini, Mohammad Ali Assarehzadegan, Najmeh Sepahi, et al. Pro j 2 is mesquite profilin: Molecular characteristics and specific IgE binding activity. *Asian Pacific Journal of Allergy and Immunology*, 33(2):90–98, jun 2015.
- [368] Gorji Marzban, Agata Mansfeld, Wolfgang Hemmer, Elitza Stoyanova, Hermann Katinger, et al. Fruit cross-reactive allergens: A theme of uprising interest for consumers'health. In *BioFactors*, volume 23, pages 235–241. Biofactors, 2005.
- [369] Pallavi Tawde, Yeldur P. Venkatesh, Fang Wang, Suzanne S. Teuber, Shridhar K. Sathe, et al. Cloning and characterization of profilin (Pru du 4), a cross-reactive almond (*Prunus dulcis*) allergen. *Journal of Allergy and Clinical Immunology*, 118(4):915–922, oct 2006.

- [370] Mohammad Ali Assarehzadegan, Akram Amini, Mojtaba Sankian, Mohsen Tehrani, Farahzad Jabbari, et al. Sal k 4, a new allergen of *Salsola kali*, is profilin: A predictive value of conserved conformational regions in cross-reactivity with other plant-derived profilins. *Bioscience, Biotechnology and Biochemistry*, 74(7):1441–1446, 2010.
- [371] S. Sirvent, O. Palomares, A. Vereda, M. Villalba, J. Cuesta-Herranz, et al. NsLTP and profilin are allergens in mustard seeds: Cloning, sequencing and recombinant production of Sin a 3 and Sin a 4. *Clinical and Experimental Allergy*, 39(12):1929–1936, dec 2009.
- [372] Kyohei Kiyota, Masato Yoshimitsu, Taro Satsuki-Murakami, Kazuhiko Akutsu, Keiji Kajimura, et al. Detection of the tomato allergen Sola l 1 and evaluation of its reactivity after heat and papain treatment. *Food and Agricultural Immunology*, 28(6):1450–1459, nov 2017.
- [373] Margit Cichna-Markl, Peter Forstenlechner, Karin Hoffmann-Sommergruber, Dalibor Milić, Du-bravka Matković-Čalogović, et al. Expression, purification and crystallization of wheat profilin (Tri a 12). *Clinical and Translational Allergy*, 1(S1):P4, dec 2011.
- [374] Y. Cui, L. Yu, F. Teng, C. Zhang, N. Wang, et al. Transcriptomic/proteomic identification of allergens in the mite *Tyrophagus putrescentiae*. *Allergy: European Journal of Allergy and Clinical Immunology*, 71(11):1635–1639, nov 2016.
- [375] Ching Hsiang Yu, Jaw Ji Tsai, Yi Hsueh Lin, Sheng Jie Yu, and En Chih Liao. Identification the Cross-Reactive or Species-Specific Allergens of *Tyrophagus putrescentiae* and Development Molecular Diagnostic Kits for Allergic Diseases. *Diagnostics*, 10(9), sep 2020.
- [376] E. Gadermaier, L. K. James, M. H. Shamji, K. Blatt, K. Fauland, et al. Epitope specificity determines cross-protection of a SIT-induced IgG4 antibody. *Allergy: European Journal of Allergy and Clinical Immunology*, 71(1):36–46, jan 2016.
- [377] Alkistis N. Mitropoulou, Holly Bowen, Tihomir S. Dodev, Anna M. Davies, Heather J. Bax, et al. Structure of a patient-derived antibody in complex with allergen reveals simultaneous conventional and superantigen-like recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 115(37):E8707–E8716, sep 2018.
- [378] M. Focke, W. Hemmer, B. Hayek, M. Götz, and R. Jarisch. Identification of allergens in oilseed rape (*Brassica napus*) pollen. *International Archives of Allergy and Immunology*, 117(2):105–112, oct 1998.
- [379] Nicole Wopfner, Oliver Dissertori, Fatima Ferreira, and Peter Lackner. Calcium-Binding Proteins and Their Role in Allergic Diseases, feb 2007.
- [380] Claudia Alessandri, Danila Zennaro, Alessandra Zaffiro, and Adriano Mari. Molecular allergology approach to allergic diseases in the paediatric age, oct 2009.
- [381] A. Ledesma, R. Barderas, K. Westritschnig, J. Quiralte, C. Y. Pascual, et al. A comparative analysis of the cross-reactivity in the polcalcin family including Syr v 3, a new member from lilac pollen. *Allergy: European Journal of Allergy and Clinical Immunology*, 61(4):477–484, apr 2006.
- [382] Yago Pico de Coaña, Nuria Parody, Miguel Ángel Fuertes, Jerónimo Carnés, Daniela Roncarolo, et al. Molecular cloning and characterization of Cup a 4, a new allergen from *Cupressus arizonica*. *Biochemical and Biophysical Research Communications*, 401(3):451–457, oct 2010.
- [383] S Mas, M Garrido-Arandia, E Batanero, A Purohit, G Pauli, et al. Characterization of profilin and polcalcin panallergens from ash pollen. *Journal of Investigational Allergology and Clinical Immunology*, 24(4):257–266, 2014.

- [384] Raffaella Tinghino, Anna Twardosz, Bianca Barletta, Eleonora M.R. Puggioni, Patrizia Iacovacci, et al. Molecular, structural, and immunologic relationships between different families of recombinant calcium-binding pollen allergens. *Journal of Allergy and Clinical Immunology*, 109(2):314–320, feb 2002.
- [385] A. Bonura, L. Gulino, A. Trapani, G. Di Felice, R. Tinghino, et al. Isolation, expression and immunological characterization of a calcium-binding protein from *Parietaria* pollen. *Molecular Immunology*, 45(9):2465–2473, may 2008.
- [386] Randall M. Goldblum, Bo Ning, Barbara M. Judy, Luis Marcelo F. Holthauzen, Julius van Bavel, et al. A single mouse monoclonal antibody, E58 modulates multiple IgE epitopes on group 1 cedar pollen allergens. *Molecular Immunology*, 74:106–112, jun 2016.
- [387] O. Palomares, M. Villalba, J. Quiralte, F. Polo, and Rosalía Rodríguez. 1,3- $\beta$ -glucanases as candidates in latex-pollen-vegetable food cross-reactivity. *Clinical and Experimental Allergy*, 35(3):345–351, mar 2005.
- [388] Mau Sinha, Rashmi Prabha Singh, Gajraj Singh Kushwaha, Naseer Iqbal, Avinash Singh, et al. Current overview of allergens of plant pathogenesis related protein families, 2014.
- [389] Karin Hoffmann-Sommergruber. Pathogenesis-related (PR)-proteins identified as allergens. In *Biochemical Society Transactions*, volume 30, pages 930–935. Biochem Soc Trans, nov 2002.
- [390] Salvador Mas, María Torres, María Garrido-Arandia, Guillermo Salamanca, Lourdes Castro, et al. Ash pollen immunoproteomics: Identification, immunologic characterization, and sequencing of 6 new allergens. *Journal of Allergy and Clinical Immunology*, 133(3), 2014.
- [391] María Torres, Oscar Palomares, Joaquín Quiralte, Gabrielle Pauli, Rosalía Rodríguez, et al. An enzymatically active  $\beta$ -1,3-glucanase from ash pollen with allergenic properties: A particular member in the oleaceae family. *PLoS ONE*, 10(7), jul 2015.
- [392] Annick Barre, Raphaël Culerrier, Claude Granier, Laetitia Selmán, Willy J. Peumans, et al. Mapping of IgE-binding epitopes on the major latex allergen Hev b 2 and the cross-reacting 1,3 $\beta$ -glucanase fruit allergens as a molecular basis for the latex-fruit syndrome, may 2009.
- [393] Ivana Aleksic, Milica Popovic, Rajna Dimitrijevic, Uros Andjelkovic, Emilia Vassilopoulou, et al. Molecular and immunological characterization of Mus a 5 allergen from banana fruit. *Molecular Nutrition and Food Research*, 56(3):446–453, mar 2012.
- [394] Marta Słowianek, Marta Skorupa, Ewelina Hallmann, Ewa Rembiałkowska, and Joanna Leszczyńska. Allergenic Potential of Tomatoes Cultivated in Organic and Conventional Systems. *Plant Foods for Human Nutrition*, 71(1):35–41, mar 2016.
- [395] Stefan Vieths, Stephan Scheurer, and Barbara Ballmer-Weber. Current understanding of cross-reactivity of food allergens and pollen. In *Annals of the New York Academy of Sciences*, volume 964, pages 47–68. New York Academy of Sciences, 2002.
- [396] Joana Costa, Isabel Mafra, Isabel Carrapatoso, and Maria Beatriz P.P. Oliveira. Hazelnut allergens: Molecular characterization, detection, and clinical relevance. *Critical Reviews in Food Science and Nutrition*, 56(15):2579–2605, 2016.
- [397] S. Kawamoto, T. Fujimura, M. Nishida, T. Tanaka, T. Aki, et al. Molecular cloning and characterization of a new Japanese cedar pollen allergen homologous to plant isoflavone reductase family. *Clinical and Experimental Allergy*, 32(7):1064–1070, 2002.

- [398] B. K. Ballmer-Weber, K. Skamstrup Hansen, J. Sastre, K. Andersson, I. Bätcher, et al. Component-resolved in vitro diagnosis of carrot allergy in three different regions of Europe. *Allergy: European Journal of Allergy and Clinical Immunology*, 67(6):758–766, jun 2012.
- [399] Rosario Carmona, María José Jimenez-Quesada, Elena Lima-Cabello, José Ángel Traverso, Antonio Jesús Castro, et al. S-nitroso- and nitro- proteomes in the olive (*Olea europaea* L.) pollen. Predictive versus experimental data by nano-LC-MS. *Data in Brief*, 15:474–477, 2017.
- [400] Yi Wang, Wen-Zheng Zhang, Lian-Fen Song, Jun-Jie Zou, Zhen Su, et al. Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in *Arabidopsis*. *Plant physiology*, 148(3):1201–1211, 2008.
- [401] Li Q Wei, Wen Y Xu, Zhu Y Deng, Zhen Su, Yongbiao Xue, et al. Genome-scale analysis and comparison of gene expression profiles in developing and germinated pollen in *Oryza sativa*. *BMC genomics*, 11:338, 2010.
- [402] Hanako Ono, Kazuo Ishii, Toshinori Kozaki, Isao Ogiwara, Motoki Kanekatsu, et al. Removal of redundant contigs from de novo RNA-Seq assemblies via homology search improves accurate detection of differentially expressed genes. *BMC genomics*, 16(1):1031, 2015.
- [403] María José Jiménez-Quesada, Jose Ángel Traverso, Adoración Zafra, José C Jimenez-Lopez, Rosario Carmona, et al. Identification and in silico Analysis of NADPH Oxidase Homologues Involved in Allergy from an Olive Pollen Transcriptome. In *Bioinformatics and Biomedical Engineering: Third International Conference, IWBBIO 2015*, volume 9043, pages 450–459. 2015.
- [404] Estefanía García-Quirós, Rosario Carmona, Adoración Zafra, M. Gonzalo Claros, and Juan de Dios Alché. Identification and in silico analysis of glutathione reductase transcripts expressed in olive (*Olea europaea* L.) pollen and pistil. In *Bioinformatics and Biomedical Engineering Fifth International Conference, IWBBIO Granada.*, volume Proceeding, pages 185–195. 2017.
- [405] M. Luisa Hernández, Elena Lima-Cabello, Juan D. de Alché, José M. Martínez-Rivas, and Antonio J. Castro. Lipid composition and associated gene expression patterns during pollen germination and pollen tube growth in olive (*olea europaea* L.). *Plant and Cell Physiology*, 61(7):1348–1364, 2021.
- [406] Adoración Zafra, José Carlos Jimenez-Lopez, Rosario Carmona, Gonzalo Claros, and Juan de Dios Alché. Transcriptome-based identification of a seed olive legumin (11S globulin). Characterization of subunits, 3D modelling and molecular assessment of allergenicity. In *Bioinformatics and Biomedical Engineering Fourth International Conference, IWBBIO Granada*, volume 9656, pages 59–70. 2016.
- [407] Willard M. Freeman, Stephen J. Walker, and Kent E. Vrana. Quantitative RT-PCR: Pitfalls and potential. *BioTechniques*, 26(1):112–125, 1999.
- [408] Guiqing Liu, Xuehong Qiu, Li Cao, Yi Zhang, Zubing Zhan, et al. Evaluation of Reference Genes for Reverse Transcription Quantitative PCR Studies of Physiological Responses in the Ghost Moth, *Thitarodes armoricanus* (Lepidoptera, Hepialidae). *PloS one*, 11(7):e0159060, jan 2016.
- [409] Jianyang Wu, Hongna Zhang, Liqin Liu, Weicai Li, Yongzan Wei, et al. Validation of Reference Genes for RT-qPCR Studies of Gene Expression in Preharvest and Postharvest Longan Fruits under Different Experimental Conditions. *Frontiers in plant science*, 7:780, jan 2016.
- [410] Xinjian Peng and David L McCormick. Identification of reliable reference genes for quantitative gene expression studies in oral squamous cell carcinomas compared to adjacent normal tissues in the F344 rat model. *Oncology reports*, 36(2):1076–84, aug 2016.

- [411] Daniela Cassol, Fernanda P Cruz, Kauê Espindola, Amanda Mangeon, Caroline Müller, et al. Identification of reference genes for quantitative RT-PCR analysis of microRNAs and mRNAs in castor bean (*Ricinus communis* L.) under drought stress. *Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale*, 106:101–107, mar 2016.
- [412] Michael W Pfaffl, Ales Tichopad, Christian Prgomet, and Tanja P Neuvians. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnology Letters*, 26(6):509–515, 2004.
- [413] Jo Vandesompele, Katleen De Preter, Filip Pattyn, Bruce Poppe, Nadine Van Roy, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology*, 3(7):RESEARCH0034, jun 2002.
- [414] Claus Lindbjerg Andersen, Jens Ledet Jensen, and Torben Falck Ørntoft. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer research*, 64(15):5245–50, aug 2004.
- [415] Peter D Lee, Robert Sladek, Celia M T Greenwood, and Thomas J Hudson. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome research*, 12(2):292–7, 2002.
- [416] T. Suzuki, P. J. Higgins, and D. R. Crawford. Control selection for RNA quantitation. *BioTechniques*, 29(2):332–337, 2000.
- [417] Debashree L Ray and Joshua C Johnson. Validation of reference genes for gene expression analysis in olive (*Olea europaea*) mesocarp tissue by quantitative real-time RT-PCR. *BMC research notes*, 7(1):304, 2014.
- [418] Jiong Ma, David S Skibbe, John Fernandes, and Virginia Walbot. Male reproductive development: gene expression profiling of maize anther and pollen ontogeny. *Genome Biology*, 9(12):R181, 2008.
- [419] Rosario Carmona, Pedro Seoane, Adoración Zafra, María José Jiménez-Quesada, Juan de Dios Alché, et al. Automatic Workflow for the Identification of Constitutively-Expressed Genes Based on Mapped NGS Reads. In *Bioinformatics and Biomedical Engineering Fourth International Conference, IWBBIO Granada*, volume 9656, pages 403–414. 2016.
- [420] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–8, jul 2008.
- [421] Kana Hirano, Shingo Hino, Kenzi Oshima, Daita Nadano, Atsuo Urisu, et al. Evaluation of allergenic potential for rice seed protein components utilizing a rice proteome database and an allergen database in combination with IgE-binding of recombinant proteins. *Bioscience, Biotechnology, and Biochemistry*, 80(3):564–573, 2016.
- [422] Bradley C Campbell, Edward K Gilding, Victoria Timbrell, Preethi Guru, Dorothy Loo, et al. Total transcriptome, proteome, and allergome of Johnson grass pollen, which is important for allergic rhinitis in subtropical regions. *The Journal of allergy and clinical immunology*, 135(1):133–42, jan 2014.
- [423] Bodhisattwa Saha, Gaurab Sircar, Naren Pandey, and Swati Gupta Bhattacharya. Mining novel allergens from coconut pollen employing manual de novo sequencing and homology-driven proteomics. *Journal of Proteome Research*, 14(11):4823–4833, nov 2015.
- [424] Véronique Bordas Le Floch, Maxime Le Mignon, Julien Bouley, Rachel Groeme, Karine Jain, et al. Identification of novel short ragweed pollen allergens using combined transcriptomic and immunoproteomic approaches. *PLoS ONE*, 10(8):e0136258, aug 2015.

- [425] R Rodríguez, M. Villalba, E. Batanero, E M González, R. I. Monsalve, et al. Allergenic diversity of the olive pollen. *Allergy*, 57 Suppl 7(s71):6–16, apr 2002.
- [426] C Esteve, C Montealegre, M L Marina, and M C García. Analysis of olive allergens. *Talanta*, 92:1–14, apr 2012.
- [427] R. B. Knox, J. Heslop-Harrison, and C. Reed. Localization of antigens associated with the pollen grain wall by immunofluorescence. *Nature*, 225:1066–1068, 1970.
- [428] R. B. Knox and J. Heslop-Harrison. Pollen-wall proteins: localization of antigenic and allergenic proteins in the pollen grain walls of *Ambrosia* spp. (ragweeds). *Cytobios*, 4:49–54, 1971.
- [429] Ettore Pacini. Cell biology of anther and pollen development. In E. G. Williams, A. E. Clarke, and R. B. Knox, editors, *Genetic control of self-incompatibility and reproductive development in flowering plants*, pages 289–308. Springer Netherlands, 1994.
- [430] Fang Liu, Xiaobo Zhang, Changming Lu, Xinhua Zeng, Yunjing Li, et al. Non-specific lipid transfer proteins in plants: Presenting new advances and an integrated functional analysis, sep 2015.
- [431] Eva Batanero, Mayte Villalba, Rafael I Monsalve, and Rosalía Rodríguez. Cross-reactivity between the major allergen from olive pollen and unrelated glycoproteins: Evidence of an epitope in the glycan moiety of the allergen. *Journal of Allergy and Clinical Immunology*, 97(6):1264–1271, jun 1996.
- [432] Gary A. Bannon and Tadashi Ogawa. Evaluation of available IgE-binding epitope data and its utility in bioinformatics. In *Molecular Nutrition and Food Research*, volume 50, pages 638–644. WILEY-VCH Verlag, jul 2006.
- [433] Mounira Chaki, Izabella Kovacs, Manuel Spannagl, and Christian Lindermayr. Computational prediction of candidate proteins for S-nitrosylation in *Arabidopsis thaliana*. *PLoS ONE*, 9(10):e110232, oct 2014.
- [434] Marcin Tatjewski, Marcin Kierczak, and Dariusz Plewczynski. Predicting post-translational modifications from local sequence fragments using machine learning algorithms: Overview and best practices. In *Methods in Molecular Biology*, volume 1484, pages 275–300. Humana Press Inc., 2017.
- [435] Thammakorn Saethang, D. Michael Payne, Yingyos Avihingsanon, and Trairak Pisitkun. A machine learning strategy for predicting localization of post-translational modification sites in protein-protein interacting regions. *BMC Bioinformatics*, 17(1):1–15, aug 2016.
- [436] Duolin Wang, Dongpeng Liu, Jiakang Yuchi, Fei He, Yuexu Jiang, et al. MusiteDeep: A deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Research*, 48(W1):W140–W146, jul 2021.
- [437] Taoying Li, Runyu Song, Qian Yin, Mingyue Gao, and Yan Chen. Identification of S-nitrosylation sites based on multiple features combination. *Scientific Reports*, 9(1), dec 2019.
- [438] Christian Lindermayr, Gerhard Saalbach, and Jörg Durner. Proteomic identification of S-nitrosylated proteins in *Arabidopsis*. *Plant Physiology*, 137(3):921–930, 2005.
- [439] Beatrice Belenghi, Maria C. Romero-Puertas, Dominique Vercammen, Anouk Brackener, Dirk Inzé, et al. Metacaspase activity of *Arabidopsis thaliana* is regulated by S-nitrosylation of a critical cysteine residue. *Journal of Biological Chemistry*, 282(2):1352–1358, jan 2007.
- [440] Jasmeet K. Abat, Autar K. Mattoo, and Renu Deswal. S-nitrosylated proteins of a medicinal CAM plant *Kalanchoe pinnata* - Ribulose-1,5-bisphosphate carboxylase/oxygenase activity targeted for inhibition. *FEBS Journal*, 275(11):2862–2872, jun 2008.

- [441] Jasmeet Kaur Abat and Renu Deswal. Differential modulation of S-nitrosoproteome of Brassica juncea by low temperature: Change in S-nitrosylation of Rubisco is responsible for the inactivation of its carboxylase activity. *Proteomics*, 9(18):4368–4380, sep 2009.
- [442] M. Cristina Palmieri, Christian Lindermayr, Hermann Bauwe, Clara Steinhauser, and Joerg Durner. Regulation of plant glycine decarboxylase by S-nitrosylation and glutathionylation. *Plant Physiology*, 152(3):1514–1528, mar 2010.
- [443] Maria Concetta de Pinto, Vittoria Locato, Alessandra Sgobba, Maria del Carmen Romero-Puertas, Cosimo Gadaleta, et al. S-nitrosylation of ascorbate peroxidase is part of programmed cell death signaling in Tobacco bright yellow-2 cells. *Plant Physiology*, 163(4):1766–1775, dec 2013.
- [444] María C. Romero-Puertas, María Rodríguez-Serrano, and Luisa M. Sandalio. Protein S-nitrosylation in plants under abiotic stress: An overview, sep 2013.
- [445] Jasmeet Kaur Abat, Pooja Saigal, and Renu Deswal. S-Nitrosylation - Another biological switch like phosphorylation?, apr 2008.
- [446] Francisco J. Corpas, José M. Palma, Luis A. del Río, and Juan B. Barroso. Protein tyrosine nitration in higher plants grown under natural and stress conditions, feb 2013.
- [447] E. Baudouin. The language of nitric oxide signalling. *Plant Biology*, 13(2):233–242, mar 2011.
- [448] Feng Zhao, Amr Elkelish, Jörg Durner, Christian Lindermayr, J. Barbro Winkler, et al. Common ragweed (*Ambrosia artemisiifolia* L.): Allergenicity and molecular characterization of pollen after plant exposure to elevated NO<sub>2</sub>. *Plant Cell and Environment*, 39(1):147–164, jan 2016.
- [449] Juan A. Roldán, Hernán J. Rojas, and Ariel Goldraij. In vitro inhibition of incompatible pollen tubes in *Nicotiana glauca* involves the uncoupling of the F-actin cytoskeleton and the endomembrane trafficking system. *Protoplasma*, 252(1):63–75, 2014.
- [450] Rosario Carmona, M. Claros, and Juan de Alché. Bioinformatic prediction of S-nitrosylation sites in large protein datasets. In *Methods in Molecular Biology*, volume 1747, pages 241–250. Humana Press, New York, NY, 2018.