

Integration of Bioinformatics to molecular research in forest species: the case of Holm oak (*Quercus ilex*)

Integración de la Bioinformática en la
investigación molecular en especies forestales: el
caso de la encina (*Quercus ilex*)



UNIVERSIDAD DE CÓRDOBA

Víctor Manuel Guerrero Sánchez

Supervisors: Prof. Jesús Valentín Jorrín Novo
(University of Córdoba)
Prof. Luis Valledor González
(University of Oviedo)

Programa de Doctorado en Ingeniería Agraria, Alimentaria, Forestal y
del Desarrollo Rural Sostenible por la Universidad de Córdoba y la
Universidad de Sevilla

March 2020

TITULO: *Integration of Bioinformatics to molecular research in forest species: the case of Holm oak (Quercus ilex)*

AUTOR: *Víctor Manuel Guerrero Sánchez*

© Edita: UCOPress. 2020
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

<https://www.uco.es/ucopress/index.php/es/>
ucopress@uco.es

**TÍTULO DE LA TESIS:**

Integración de la bioinformática en la investigación con especies forestales: el caso de la encina (*Quercus ilex*)

Integration of Bioinformatics to molecular research in forest species: the case of Holm oak: the case of Holm oak (*Quercus ilex*)

DOCTORANDO/A:

Víctor Manuel Guerrero Sánchez

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

La presente Tesis Doctoral, "Integración de la bioinformática en la investigación con especies forestales: el caso de la encina (*Quercus ilex*)" ha sido realizada por Víctor Manuel Guerrero Sánchez, durante los años 2017 a 2020, dentro del "Programa de Doctorado de la Universidad de Córdoba "Ingeniería Agraria, Forestal, Alimentaria y Desarrollo Rural Sostenible ". Ha sido dirigida por los Profesores Jesús V. Jorrín Novo, de la Universidad de Córdoba (AGR-164; Bioquímica, Proteómica, y Biología de Sistemas Vegetal y Agroforestal) y Luis Valledor González, de la Universidad de Oviedo (Dep. Biol. Organismos y Sistemas, Grupo Biotecnología de Plantas).

Cumple los requisitos exigidos para su presentación y defensa. en inglés, lo que le dará mayor visibilidad al trabajo.

La trayectoria de Víctor M. Guerrero Sánchez durante este periodo de formación ha sido muy brillante, habiendo logrado excelentes resultados. Ha demostrado una gran capacidad de trabajo, superando con creces los objetivos y plan de trabajo originalmente propuestos. Su grado de formación es óptimo, tanto a nivel científico-técnico como académico. Prueba de ello es su excelente *curriculum vitae*.

El trabajo realizado ha supuesto un hito para el Grupo de Investigación Bioquímica, Proteómica, y Biología de Sistemas Vegetal y Agroforestal, ya que ha permitido la utilización de diferentes programas bioinformáticos y algoritmos para el análisis de datos -ómicos y su integración en la dirección de la Biología de Sistemas. Ha manejado un gran número de programas de análisis, identificación y anotación de secuencias de ácidos nucleicos y proteínas a partir de datos de transcriptómica y proteómica. Ha optimizado métodos de análisis estadístico uni- y multivariante,

agrupamientos funcionales y establecimiento de redes de interacción de productos génicos. Todo ello en un sistema experimental que como la encina es huérfana de estudios moleculares y -ómicos, aparte de su recalcitrancia como sistema experimental.

El trabajo de Tesis ha resultado en la primera publicación del transcriptoma de la encina (Capítulos 3 y 4).

GUERRERO-SANCHEZ, VICTOR M; MALDONADO-ALCONADA, ANA M; AMIL-RUIZ, FRANCISCO; JORRIN-NOVO, JESUS V. 2017. Holm Oak (*Quercus ilex*) Transcriptome. De novo Sequencing and Assembly Analysis.. *Frontiers in Molecular Biosciences* 4, 70. DOI: 10.3389/fmolb.2017.00070. IF₂₀₁₈(JCR): 3.565 (Q2)

VÍCTOR MANUEL GUERRERO-SÁNCHEZ; ANA MARÍA MALDONADO-ALCONADA; FRANCISCO AMIL-RUIZ; ANDREA VERARDI; JESÚS V. JORRÍN-NOVO; MARÍA-DOLORES REY. 2019. Ion Torrent and Illumina, two complementary RNA-seq platforms for constructing the holm oak (*Quercus ilex*) transcriptome. *PLOS One* 14: e0210356. <https://doi.org/10.1371/journal.pone.0210356>. IF₂₀₁₈(JCR): 2.776(Q2)

VICTOR M. GUERRERO-SANCHEZ, LUIS VALLEDOR, MARÍA-DOLORES REY, ANA M MALDONADO-ALCONADA, JESUS V. JORRIN-NOVO. 2020. Specific protein database creation from transcriptomics data in non-model species: Holm oak (*Quercus ilex* L.). *Methods in Molecular Biology* (Clifton, N.J.), in press.

El análisis *in silico* de datos multiómicos (transcriptómica, proteómica y metabolómica) ha permitido la construcción y propuesta de rutas del metabolismo primario y secundario tal y como operarían en encina (capítulo 5)

CRISTINA LÓPEZ-HIDALGO, **VICTOR M. GUERRERO-SANCHEZ,** ISABEL GÓMEZ-GÁLVEZ, ROSA SÁNCHEZ-LUCAS, MARÍA ANGELES CASTILLEJO, ANA MARÍA MALDONADO-ALCONADA, LUIS VALLEDOR, JESUS V JORRIN NOVO. 2018. A multi-omics analysis pipeline for the metabolic pathway reconstruction in the orphan species *Quercus ilex*. *Frontiers in Plant Sciences*. 9: 935. doi: [10.3389/fpls.2018.00935](https://doi.org/10.3389/fpls.2018.00935). IF₂₀₁₈(JCR): 4.106 (Q1)

El capítulo 6 de la tesis corresponde a un manuscrito que se enviará para su publicación a una revista de alto impacto y en el que se lleva a cabo el estudio de la respuesta a sequía en encina mediante la integración de una doble plataforma experimental de, respectivamente, proteómica y transcriptómica.

El trabajo ha sido también publicado en diferentes artículos de revisión y capítulos de libro y ha sido presentado en diferentes congresos y reuniones científicas:

MARÍA-DOLORES REY, LUIS VALLEDOR, MARÍA ÁNGELES CASTILLEJO, ROSA SÁNCHEZ-LUCAS, CRISTINA LÓPEZ-HIDALGO, **VICTOR M. GUERRERO-SANCHEZ,** et al. 2019. Recent advances in MS-based Plant Proteomics. Proteomics data validation through integration with other classic and -omics approaches. *Progress in Botany*, Series Editors: Lüttge, U., Cánovas, F.M., Matyssek, R., Pretzsch, H. (2019).

MARÍA ÁNGELES CASTILLEJO, ROSA SÁNCHEZ-LUCAS, MARÍA-DOLORES REY, **VICTOR M. GUERRERO-SANCHEZ**, et al. 2019. Proteomics and the forest tree Holm oak (*Quercus ilex* L.), an orphan and recalcitrant experimental plant system: how do they see each other? *International Journal of Molecular Sciences* 20, 692; doi:10.3390/ijms20030692

JESUS V. JORRIN-NOVO, LUIS VALLEDOR GONZALEZ, MARI A. CASTILLEJO-SANCHEZ, ROSA SÁNCHEZ-LUCAS, ISABEL M. GOMEZ-GALVEZ, CRISTINA LOPEZ-HIDALGO, **VICTOR M. GUERRERO-SÁNCHEZ**, et al. 2018. Proteomics Analysis of Plant Tissues Based on Two-Dimensional Gel Electrophoresis. In: Springer International Publishing AG, part of Springer Nature. ISBN 978-3-319-93232-3. Chapter 19. Springer International Publishing AG, part of Springer Nature 2018 A. Sánchez-Moreiras, M. Reigosa-Roger (eds.), *Advances in Plant Ecophysiology Techniques*, https://doi.org/10.1007/978-3-319-93233-0_19.

SÁNCHEZ-LUCAS, R; LÓPEZ-HIDALGO, C; **GUERRERO SÁNCHEZ, V**; GÓMEZ et al. 2018. _Las aproximaciones -ómicas y la biología de sistemas en investigación agroforestal y biomédica. VIII Jornadas de Divulgación de la Investigación en Biología Molecular, Celular, Genética y Biotecnología. Universidad de Córdoba, UCOPress, ISBN-978-84-9927-388-4 <http://www.uco.es/ucopress/index.php/es/catalogo/e-books/product/669-ebook-viii-jornadas-de-divulgacio-n-de-la-investigacio-n-en-biologia-molecular-celular-gene-tica-y-biotecnologia-a>

Guerrero-Sánchez VM, Maldonado-Alconada AM, Amil-Ruiz F, Rey MD, Jorrín-Novó JV. Deciphering the *Quercus ilex* transcriptome using complementary sequencing and assembly strategies. XIV MEETING OF PLANT MOLECULAR BIOLOGY. Salamanca, España. 2018

Rey MD, **Guerrero-Sánchez VM**, Sánchez-Lucas R, López-Hidalgo C, Maldonado-Alconada AM, Jorrín-Novó JV. The use of -omics technologies to progress in the *Quercus ilex* biology. XIV MEETING OF PLANT MOLECULAR BIOLOGY. Salamanca, España. 2018

López-Hidalgo C, **Guerrero-Sánchez VM**, Gómez-Galvez I, Sánchez-Lucas R, Castillejo MA, Maldonado-Alconada AM, Villedor L, Jorrin-Novó JV. Proteomics, and its integration with transcriptomics and metabolomics, allowed the reconstruction of the metabolism as it occurs in *Quercus ilex*. EUPA CONGRESS "Translating genomes into biological functions". Santiago de Compostela, España. 2018

Guerrero-Sanchez VM. Integración de la bioinformática en la investigación molecular: el caso de la encina (*Quercus ilex*). VI Congreso Científico de Jóvenes en Formación, Universidad de Córdoba, España. 2018

López-Hidalgo C, Gómez-Gálvez I, Sánchez-Lucas R, **Guerrero-Sánchez VM**, Castillejo MA, Villedor L, Jorrín-Novó JV. Proteomic and metabolic analysis of *Quercus ilex* leaves in response to drought stress. Symposium of the Mexican

Proteomics Society "Mass Spectrometry Based OMICS", Guadalajara (Jalisco), México. 2017

El estudio y los resultados son totalmente novedosos para la encina, en particular, y para especies forestales, en general. El trabajo ha contribuido a la formación de Víctor M. Guerrero Sánchez como Bioinformático, un área para la que hay, hoy en día, una gran demanda. Es de destacar que su aprendizaje ha sido en gran medida autodidacta, lo que da doble valor a su trabajo.

El número de actividades formativas realizadas por Víctor M. Guerrero Sánchez se ha ajustado a las exigidas por el Programa de Doctorado, las sugeridas por sus directores, y las seguías por propia iniciativa.

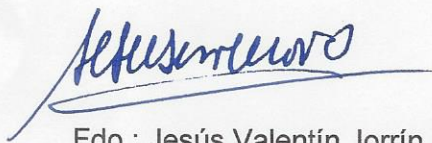
En todo momento ha demostrado enormes ganas de aprender y vocación por la actividad investigadora y la bioinformática. Siempre ha estado dispuesto, desde una posición de liderazgo en el área de la bioinformática dentro del grupo, a compartir sus conocimientos con otros estudiantes de grado, máster y doctorado, ayudándoles generosamente en la realización de las correspondientes Tesis.

Finalizaremos este informe señalando que el trabajo de Víctor M. Guerrero Sánchez durante su periodo de doctorado ha transcendido el ámbito de la presente tesis, participando en actividades docentes. Ha sido un miembro relevante del grupo de investigación Bioquímica, Proteómica, y Biología de Sistemas Vegetal y Agroforestal. Ha participado en otras investigaciones llevadas a cabo en el grupo. Ha dirigido trabajos fin de grado y tutorizado a estudiantes de otros países que realizaron estancias en el citado grupo.

Por todo ello, se autoriza la presentación de la tesis doctoral:

Córdoba, 2 de marzo de 2020

Firma del/de los director/es



Fdo.: Jesús Valentín Jorrín Novo



Fdo.: Luis Valledor González

I would like to dedicate this thesis to my loving parents.

Abstract

The term Bioinformatics, first coined by Paulien Hogeweg and Ben Hesper, back in 1970 to describe 'the study of informatic processes in biotic systems', can be defined as "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, represent, describe, store, analyze, or visualize such data" or "the development and application of data-analytical and theoretical methods, mathematical modelling, and computational simulation techniques to the study of biological, behavioural, and social systems". The first definition deals with the biological information management, and the second one with computational biology. The general objective and methodology employed in the current Thesis, "Integration of Bioinformatics to molecular research in forest species: the case of Holm oak (*Quercus ilex*)", is focused on the first definition. The use of bioinformatic tools (algorithms, programs, databases and repositories) has been used to construct the transcriptome, proteome and metabolome of Holm oak and their integration to define the metabolism and responses to drought in this species.

Since the end of the last century, biological research has moved from a reductionist to holistic paradigm, which have been possible thanks to the great technological advances, especially in the molecular biology discipline. Thus, the appearance of platforms based on the Next Generation Sequencing (NGS), for genomics and transcriptomics, and Mass Spectrometry (MS), for proteomics and metabolomics has made possible to obtain from hundreds to thousands of data in a single experiment, being impossible the management and analysis of them without the employment of informatics tools. The employment of high throughput techniques and their combination with classic approaches is what defines "Systems Biology". It do not only analyse thousands and thousands of molecular entities of an individual, but also the integration and creation of predictive models. This is quite feasible with model organisms (e.g. *Arabidopsis*), but it is a real challenge for those orphan and recalcitrant experimental systems such as *Q. ilex*. The study of this species is justified because of the environmental and economic importance in Spain and, because it faces a problem of increasing tree mortality associated to the decline syndrome, a situation that can be worsen in a climate change scenario.

Biotechnology can contribute to solve this problem through breeding programs based on markers-assisted selection of elite genotypes that are more tolerant and resistant to biotic and abiotic stresses and more resilient to climate change.

As a continuation of the work carried out since 2004 by the research group “Agroforestry and Plant Biochemistry, Proteomics, and Systems Biology”, mostly based on classic biochemistry, physiology and proteomics, and considering that neither the genome of Holm oak has been sequenced yet nor DNA or proteins sequences are available in public databases, as first objective of the Thesis was proposed the construction of the first reference transcriptome for this species. The work is presented in **chapter 3**, and has been published in *Frontiers in Molecular Bioscience*. For that purpose, the mRNA extracted from homogenized tissue from acorn embryo, leaves, and roots, was sequenced using an Illumina Hiseq 2500 platform. Three different assemblers were employed, TRINITY, RAY, and MIRA. The assemblies obtained were aligned against the most accurate and nearest phylogenetically transcriptome currently available, that of *Quercus robur* and *Quercus petraea*. MIRA generated more and longer contigs than RAY and TRINITY (MIRA>RAY>TRINITY). So, MIRA assembly was used to continue with the corresponding annotation of *Q. ilex* transcriptome, resulting in 31973 annotated sequences were obtained by Blast2GO using Swiss-Prot as reference database.

As a continuation of the previous work, and as a second objective, a new sequencing platform, Ion Torrent, was evaluated in the construction and analysis of the *Q. ilex* transcriptome. The obtained results are presented in **chapter 4** and have been already published in *PLoS ONE*. Raw sequence reads, obtained from Illumina and Ion Torrent, were assembled by three different software, MIRA, RAY and TRINITY. A hybrid transcriptome combining reads from both sequencing technologies was also assembled using RAY. The hybrid assembly generated the most complete transcriptome. The assembly of Ion Torrent reads of MIRA showed the highest number of shared sequences (84.8%) with the oak transcriptome. In addition, an *in silico* proteomic analysis was carried out using the translated assemblies as databases. Those from Ion Torrent showed more proteins compared to the Illumina and hybrid assemblies. All the assembled transcripts from the hybrid transcriptome were annotated and grouped according to the corresponding biological processes, molecular functions and cellular components (Gene Ontology). This new generated transcriptome represents a valuable tool to conduct differential gene expression studies in response to biotic and abiotic stresses and to assist and validate the ongoing *Q. ilex* whole genome sequencing.

By using the above mentioned plant sample, the transcriptomic (NGS-Illumina), proteomic (shotgun LC-MS/MS, Orbitrap), and metabolomic (GCMS) profiles were analysed. Results are presented in **chapter 5**, and have been already published in *Frontiers in Plant Science*. The annotated *Q. ilex* transcriptome was compared against the complete *in silico* proteomes of *Arabidopsis thaliana* (UP0000065489, *Oryza sativa* subsp. *Japonica* (UP00005968010), *Populus trichocarpa* (UP00000672911), and *Eucalyptus grandis* (UP00003071112) in order to elucidate the unique and shared sequences. Also, the EC numbers of each proteome were contrasted to achieve a complete picture of the metabolic pathways coverage differences among proteomes studied in previously mentioned species. The descriptive analysis and the visualization of data on a gene-by-gene basis on schematic diagrams (maps) of the biological processes described in Mapman, resulted in the identification of around 62629 transcripts, 2380 protein species, and 62 metabolites. Data were compared with those reported for model plant species, whose genome has been sequenced and well annotated, including *Arabidopsis*, *japonica* rice, poplar, and eucalyptus. The integration of the large amount of data reported using bioinformatics tools allowed the Holm oak metabolic network to be partially reconstructed. From the 127 metabolic pathways reported in KEGG pathway database, 123 metabolic pathways can be visualized when using the described methodology. They included: carbohydrate and energy metabolism, amino acid metabolism, lipid metabolism, nucleotide metabolism, and biosynthesis of secondary metabolites. The TCA cycle was the pathway most represented with 5 out of 10 metabolites, 6 out of 8 protein enzymes, and 8 out of 8 enzyme transcripts. On the other hand, gaps, missed pathways, included metabolism of terpenoids and polyketides and lipid metabolism. The multi-omics resource generated in this work will set the basis for ongoing and future studies, bringing the Holm oak closer to model species.

As a final objective of the current Thesis, an integrated transcriptomics and proteomics analysis of the response to drought in *Q. ilex* seedlings has been carried out. Seedlings were subjected to drought conditions by water withholding, and leaf tissue sampled at two times of the experiment, 20 and 25 days. RNA and proteins were extracted and analysed by using RNA-seq (Illumina), and proteomics, LC-MS/MS Orbitrap. Data are presented in **chapter 6**; it also corresponds to a manuscript to be submitted for publication.

Gene products were identified and quantified at transcript and protein levels, establishing correlations between transcript and the corresponding protein abundance. Gene ontology (GO) analysis was performed to classify identified transcripts and

proteins in terms of biological process, molecular function and cellular component. A multivariate analysis of the total and variable datasets at transcript and protein levels was performed with mixOmics. To acquire an integrated visualization of Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway maps, total transcript and protein datasets, specifying those variable transcripts and proteins, were analysed by Paintomics 3 (v0.4.5), considering *Arabidopsis thaliana* as a model reference. Pathways with $p\text{-value} < 0.05$ were considered as significantly pathways. Interaction networks were constructed using the plugin GeneMANIA under Cytoscape (v3.4.0). The interaction networks included were prediction, co-expression, co-localization, and shared protein domains. This software also finds functionally similar genes that do not exist in the input gene list. RNA-seq analysis generated 47868 transcripts corresponding to 21000 unigenes, with 3588 qualitative or quantitative differences between irrigated and droughted seedlings (1149 up, and 2439 down). From shotgun proteomics, 4008 protein species were identified, corresponding to 2767 different genes. Out of them, 640 had qualitative or quantitative differences in abundance between treatments (353 more and 287 less abundant under drought conditions). A wide gene expression reorganization was observed at the two omics levels with up and down regulation, being this transitory (observed at 20 or 25 days) or permanent (observed at 20 and 25 days). The functional groups, whose genes were most altered in response to drought, were “stress-related” and “chloroplasts”. The most affected metabolic pathways included protein translation, photosynthesis, carbohydrates, amino acids and phenolics. Variable gene products were observed at transcriptomic or proteomic levels, with a reduced number detected at both levels. This included, for example, RPS2, 4CL2, PSB28, and RIN4, among others. From the variable transcript and protein datasets, two networks were constructed, the first one included up accumulated CLPB2, CLPB3, HSP70, HSP17.4, FtsH6, AT1G23740, SMT1, and UGP3, and down accumulated ABA2, RPS1, ADK, and RPL4 genes and the second one included up accumulated CLPB2, CLPB3, HSP70, HSP17.4, FtsH6, AT1G23740, AP1, INVE, AT4G2740, CAD4, FEN1, and HIP27 and down accumulated ABA2 genes. From a biological point of view, and in terms of stress response and tolerance, *Q. ilex* seedlings were characterized by an increase in general abiotic stress related gene products, including CPLB2, CPLB3, FTSH6 and PSB28. These variable gene products overexpressed under drought conditions can be proposed as molecular markers of response and tolerance to drought stress.

As a general conclusion it is necessary to emphasize that without bioinformatics it would be impossible to analyse the huge amount of data generated by omics approaches, but also, and more important, the importance of a manual evaluation and validation of

the results before translating to the biological context, thus avoiding much speculation. Living organisms are much more complex than we can realize and, improvements in wet and *in silico* analysis will be necessary to deep in its knowledge and to shed some light to biology, to understand the mechanisms connecting genotype and phenotype, and identify gene and gene product interactions linked to different biological processes such as the plant response to stresses. So, this knowledge will allow a better progression in speed breeding programs and biotechnological-related approaches.

Resumen

El término bioinformática, acuñado por Paulien Hogeweg y Ben Hesper en 1970 para describir "el estudio de los procesos informáticos en sistemas bióticos", se puede definir como "la investigación, desarrollo o aplicación de herramientas y aproximaciones computacionales que permitan el manejo de datos biológicos, médicos o de comportamiento, incluyendo aquellas para adquirir, representar, describir, almacenar, analizar o visualizar dichos datos" o como "el desarrollo y aplicación de métodos analíticos y teóricos, modelos matemáticos y técnicas de simulación computacional para estudiar sistemas biológicos, sociales o de comportamiento". La primera definición hace referencia al manejo de información biológica, y la segunda a la biología computacional. El objetivo general y la metodología empleada en la presente tesis, "Integración de la bioinformática en la investigación molecular en especies forestales: el caso de la encina (*Quercus ilex*)" se incluyen en el primer grupo, el del uso de herramientas bioinformáticas (algoritmos, programas, bases de datos y repositorios) utilizados para el análisis de datos, principalmente ómicos, y la construcción del transcriptoma, proteoma y metaboloma de referencia en la encina, además de la integración de dichas ómicas para definir el metabolismo y la respuesta a sequía en dicha especie.

Desde el final del último siglo, la investigación biológica se ha movido desde un paradigma reduccionista a una aproximación holística, gracias al gran avance tecnológico, especialmente en la disciplina de la biología molecular. La aparición de plataformas para la secuenciación de nueva generación (NGS), en el caso de la genómica y transcriptómica, y la espectrometría de masas (MS), en el caso de la proteómica y metabolómica, ha hecho posible obtener desde cientos a miles de datos de un único experimento; el tratamiento y el análisis de los mismos es prácticamente inviable sin el empleo de herramientas informáticas. El uso de técnicas de alto rendimiento y su combinación con aproximaciones clásicas es lo que define la "Biología de Sistemas", la nueva dirección establecida en la investigación biológica. La Biología de Sistemas no solo incluye el análisis de miles de entidades moleculares, sino también su integración y el establecimiento, a partir de ellos, de modelos predictivos. Esto es bastante factible y una realidad hoy en día con organismos modelo (por ejemplo, *Arabidopsis*), sin

embargo para sistemas huérfanos de estudios moleculares y recalcitrantes, como es el caso de *Q. ilex*, constituye un auténtico desafío. El estudio de esta especie está justificado tanto por su interés medioambiental y económico para nuestra región, como por el incremento en la mortalidad del arbolado observado en las últimas décadas y asociado a estreses bióticos y abióticos, cuyo conjunto constituye el denominado síndrome de la seca. La muerte del arbolado y la pérdida de masa forestal puede verse agravada en un escenario de cambio climático. La biotecnología puede contribuir a resolver este problema a través de programas de mejora basados en la selección asistida de por marcadores moleculares para la identificación de genotipos élite que son más tolerantes a estreses bióticos y abióticos y más resilientes al cambio climático.

Como continuación al trabajo realizado desde 2004 por el grupo de investigación "Bioquímica, Proteómica y Biología de Sistemas Vegetal y Agroforestal", centrado principalmente en estudios de bioquímica clásica, fisiología y proteómica, y considerando la ausencia de secuencias de DNA y proteínas en encina, el primer objetivo de la presente tesis fue la construcción del primer transcriptoma de referencia para esta especie. Este trabajo es presentado en el **capítulo 3**, y ha sido publicado en *Frontiers in Molecular Biosciences*. Para ello, se llevó a cabo una extracción de RNAm a partir de tejido homogeneizado de embrión, hojas y raíces, y posterior secuenciación mediante la plataforma Illumina HiSeq 2500. Se emplearon tres ensambladores diferentes, TRINITY, RAY y MIRA. Las secuencias ensambladas fueron alineadas contra el transcriptoma de *Quercus robur* y *Q. petraea*, considerado como el transcriptoma filogenéticamente más preciso y cercano a *Q. ilex*. MIRA generó un mayor número de "contigs" que RAY y TRINITY (MIRA>RAY>Trinity). Por lo tanto, las secuencias ensambladas con MIRA fueron las que se usaron para continuar con la anotación correspondiente del transcriptoma *Q. ilex*, lo que resultó en 31973 secuencias anotadas obtenidas por Blast2GO utilizando Swiss-Prot como base de datos de referencia.

Como continuación del trabajo descrito en el **capítulo 4**, y como segundo objetivo, se evaluó una nueva plataforma de secuenciación, Ion Torrent, para la construcción y análisis del transcriptoma de *Q. ilex*. Los resultados obtenidos han sido publicados en *PLoS ONE*. Como en el capítulo anterior, las lecturas obtenidas a partir de Illumina y Ion Torrent se ensamblaron utilizando tres programas diferentes, MIRA, RAY y TRINITY. En el ensamblado de MIRA con Illumina y el de TRINITY con Ion Torrent generaron el mayor número de transcritos anotados (62628 y 74058 respectivamente). El ensamblado de MIRA con Ion Torrent generó el mayor número de secuencias compartidas con el transcriptoma del roble (84.8%). RAY generó los mejores resultados atendiendo al número de contigs y longitud de los mismos, con valores de E90N50 de 1122bp. Todos

los transcritos del nuevo transcriptoma de referencia fueron anotados y agrupados en términos de Gene Ontology ("Biological Process", "Cellular Component" y "Molecular Function"). Dicho transcriptoma se tradujo in silico, obteniéndose una base de datos de proteínas que será utilizada en experimentos de proteómica para la identificación de productos génicos. El uso de dicha base de datos incrementó notablemente el número de especies proteicas identificadas y los parámetros de confianza de la identificación.

A partir de las bases de datos generadas y los datos multiómicos obtenidos cuando se utilizó una muestra de encina consistente en un pool de extractos de diferentes tejidos (embrión, hoja y raíz) se reconstruyeron diferentes rutas metabólicas tal y como ocurren en *Q. ilex*. Los resultados se presentan en el *capítulo 5* y han sido publicados en *Frontiers in Plant Science*.

Se llevó a cabo la extracción independiente a partir de la misma muestra del RNA, proteínas y metabolitos, estableciéndose el perfil ómico mediante NGS-Illumina (RNA), shotgun LC-MS/MS, Orbitrap (proteínas) y GC-MS (metabolitos). Se identificaron 62629 transcritos, 2380 especies proteicas y 62 metabolitos. Se llevó a cabo la identificación de productos génicos correspondientes a enzimas mediante la comparación con genomas de referencia incluyendo *Arabidopsis thaliana* (UP0000065489), *Oryza sativa* subsp. *japonica* (UP00005968010), *Populus trichocarpa* (UP00000672911), and *Eucalyptus grandis* (UP00003071112). De las 127 rutas metabólicas descritas en KEGG, y mediante el empleo de Mapman, se visualizaron 123, entre ellas, las del metabolismo energético, de carbohidratos, de aminoácidos, lípidos, nucleótidos y secundario. El ciclo de los ácidos tricarboxílicos (TCA) fue la ruta mejor representada con 5 de 10 metabolitos, 6 de 8 proteínas enzimáticas y 8 de 8 transcritos. Por otro lado, hay rutas que no se observaron o estaban muy poco representadas, como por ejemplo las del metabolismo de lípidos, terpenoides y policétidos.

Como objetivo final de la presente tesis, se llevó a cabo un análisis transcriptómico y proteómico integrado de la respuesta a sequía en plántulas de *Q. ilex*. Los resultados se presentan en el **capítulo 6**, correspondiente a un manuscrito que será enviado para su publicación.

Las plántulas de *Q. ilex* crecieron en macetas con perlita, siendo sometidas a condiciones de sequía por falta de riego durante 30 días. Se tomaron muestras de hojas a dos tiempos, cuando la fluorescencia de las hojas disminuyó en un 30% y un 50% (20 y 25 días). Tras la extracción de RNA y proteínas se llevó a cabo su análisis mediante RNA-Seq (Illumina) y proteómica "shotgun" (LS-MS/MS, Orbitrap). El análisis de RNA-seq generó 47868 transcritos correspondientes a 21000 unigenes, con 3588 diferencias cualitativas o cuantitativas entre plántulas irrigadas y no irrigadas

(1149 sobreexpresados y 2439 reprimidos). A partir de la proteómica “shotgun” se identificaron 4008 proteoformas, productos de 2767 genes diferentes; de ellos, 640 presentaron diferencias cualitativas o cuantitativas en abundancia entre tratamientos (353 más y 287 menos abundantes en condiciones de sequía). Los productos genéticos variables se clasificaron en términos de Gene Ontology (proceso biológico, función molecular y componente celular) y en rutas metabólicas de KEGG en el caso de las enzimas. El conjunto de datos variables se sometió a análisis estadístico multivariante, PCA y sPLS. Finalmente, se usó GeneMANIA para la construcción de redes de interacción. Hubo cambios importantes en el patrón de expresión génica siendo los grupos de respuesta a estrés y cloroplastos lo más afectados. Respecto a rutas metabólicas, se detectaron cambios en la síntesis de proteínas, fotosíntesis, carbohidratos, aminoácidos y fenólicos. Hubo cambios transitorios (observado a un solo tiempo) o permanentes (comunes a los dos tiempos) detectados a nivel de transcrito y/o proteína. El número de productos génicos variables detectados por ambas plataformas fue mínimo, entre ellos RPS2, 4CL2, PSB28 y RIN4. A partir del conjunto de datos de transcritos y proteínas variables, se construyeron dos redes de interacción: la primera incluía los genes sobreexpresados CLPB2, CLPB3, HSP70, HSP17.4, FtsH6, AT1G23740, SMT1 y UGP3, y los genes reprimidos ABA2, RPS1, ADK y RPL4, y la segunda red incluía los genes sobreexpresados CLPB2, CLPB3, HSP70, HSP17.4, FtsH6, AT1G23740, AP1, INVE, AT4G2740, CAD4, FEN1 y HIP27 y el gen reprimido ABA2. Se proponen como genes marcadores de respuesta y tolerancia a sequía en encina a aquellos sobreexpresados a los dos tiempos y detectados a nivel de transcrito y proteína. Solo un número de genes cumplen dichas características entre los que se incluyen posibles proteínas de respuesta a choque térmico, CLPB2 y CLPB3, a una metaloproteasa cloroplástica, FTSH6, y la proteína del centro de reacción del fotosistema II, PSB28.

Como conclusión general, es necesario hacer énfasis en la necesidad del empleo de herramientas bioinformáticas para el análisis de la gran cantidad de datos generados por las técnicas ómicas, a la vez que, en la necesidad de la revisión y validación manual de los resultados de cara a una correcta, no especulativa, interpretación biológica. Los seres vivos son mucho más complejos de lo que podríamos imaginar, y el conocimiento de su biología requiere mejoras en las técnicas de laboratorio y análisis *in silico*, con el fin de profundizar en el conocimiento de los mecanismos que conectan el genotipo con el fenotipo y la identificación de productos génicos y sus interacciones asociados a diferentes procesos biológicos como son el de la respuesta y tolerancia/resistencia a estreses en plantas. Dicho conocimiento permitirá abordar programas de mejora mediante aproximaciones biotecnológicas.

Contents

Abstract	ix
Resumen	xv
List of Figures	xxiii
List of Tables	xxix
1 Introduction	1
1.1 Bioinformatics, definitions and history	3
1.2 Omics technologies	5
1.3 Bioinformatics methods and tools	8
1.3.1 Genomics	8
1.3.2 Transcriptomics	10
1.3.3 Proteomics	11
1.3.4 Metabolomics	11
1.3.5 Interactomics	12
1.3.6 Other approaches	13
1.4 Multi-omics study of orphan species: the case of Holm oak (<i>Quercus ilex</i>)	13
2 Objectives	15
3 Holm oak (<i>Quercus ilex</i>) transcriptome. <i>De novo</i> sequencing and assembly analysis	19
3.1 Introduction	21
3.2 Materials and Methods	22
3.2.1 Plant material	22
3.2.2 RNA extraction	22

3.2.3	Enrichment of mRNA, cDNA synthesis, and library generation for Illumina HiSeq 2500 platform. paired-end sequencing	23
3.2.4	<i>De novo</i> assembly and analysis of high throughput RNA sequencing data	23
3.3	Results	25
3.3.1	Evaluation and annotation of the assembled transcriptomes . . .	25
3.4	Direct link to deposited data	26
4	Ion Torrent and Illumina, two complementary RNA-seq platforms for constructing the Holm oak (<i>Quercus ilex</i>) transcriptome	27
4.1	Introduction	29
4.2	Materials and methods	31
4.2.1	Plant material	31
4.2.2	RNA extraction	31
4.2.3	RNA-Seq Library Construction, Illumina sequencing and de novo assembly	32
4.2.4	RNA-Seq Library Construction, Ion Torrent sequencing and de novo assembly	32
4.2.5	Development of a hybrid transcriptome	33
4.2.6	Assembly quality and completeness evaluation	34
4.2.7	De novo transcriptome alignment with <i>Quercus robur</i> and <i>Quercus petraea</i> transcriptomes	34
4.2.8	Identification of proteins from translated assemblies used as databases	35
4.3	Results	35
4.3.1	Sequencing platforms and de novo assembly structure analysis .	35
4.3.2	<i>Quercus ilex</i> <i>de novo</i> transcriptome alignment with <i>Q. robur</i> and <i>Q. petraea</i> transcriptomes	38
4.3.3	Transcriptome completeness evaluation	40
4.3.4	Gene ontology classification of <i>Quercus ilex</i> transcripts	41
4.3.5	Protein annotation in Holm oak	43
4.4	Discussion	45
4.5	Conclusions	47
4.6	Data Availability	49

5	A multi-omics analysis pipeline for the metabolic pathway reconstruction in the orphan species (<i>Quercus ilex</i>)	51
5.1	Introduction	54
5.2	Materials and methods	55
5.2.1	Plant material	55
5.2.2	Transcriptomics Analysis	56
5.2.3	Proteomics Analysis	57
5.2.4	Metabolomics analysis	59
5.2.5	Interspecies comparison	60
5.2.6	Integrated Pathway	61
5.3	Results and Discussion	61
5.4	Conclusions	75
5.5	Data Availability	75
6	Decoding drought tolerance mechanisms in Holm oak (<i>Quercus ilex</i>) through a combined transcriptomics and proteomics analysis	77
6.1	Introduction	80
6.2	Materials and methods	81
6.2.1	Plant material and drought treatment	81
6.2.2	RNA extraction	82
6.2.3	RNA-Seq Library construction, Illumina sequencing and <i>de novo</i> re-assembly	82
6.2.4	mRNA differential expression	83
6.2.5	Protein extraction and digestion	84
6.2.6	Shotgun (LC-MS/MS) protein analysis	84
6.2.7	Protein identification and quantification	85
6.2.8	Multivariate Analysis	86
6.2.9	Pathway mapping of omics data	86
6.2.10	Interaction network	86
6.2.11	Data availability	86
6.3	Results	87
6.3.1	Transcriptomic and Proteomic Profile analysis	87
6.3.2	Gene Ontology analysis	90
6.3.3	Multivariate analysis of omics data	93
6.3.4	Integrated visualization of omics data in metabolic pathways	96
6.3.5	Interaction network analysis	97
6.4	Discussion	99

6.4.1	Environmental stress, drought and climate change, biodiversity, and tolerance	99
6.4.2	Drought responses in forest tree species: the genus <i>Quercus</i> , <i>Quercus ilex</i> . Breeding for drought tolerance based on the selection of elite genotypes	101
6.4.3	Research on <i>Quercus ilex</i> at the Agroforestry and Plant Biochemistry, Proteomics, and Systems Biology Group, from classic biochemistry to -omics and systems biology approaches	102
6.4.4	Integrated proteomics and transcriptomics analysis of responses to drought in <i>Quercus ilex</i> . Identified transcripts and proteins, and functional grouping	103
7	General Conclusions	115
7.1	Conclusions	117
7.2	Conclusiones	118
	References	121
	Appendix A Supplementary Material	143
A.1	Holm oak transcriptome re-assembly. Supporting information	144
A.2	Multi-Omics data integration of <i>Quercus ilex</i> . Supporting information	150
A.3	Drought tolerance mechanisms of Holm oak. Supporting information	159

List of Figures

1.1	Number of publications per year appearing in PubMed from 1984 to 2018, using 'Bioinformatics' as searching term.	3
3.1	Evaluation of <i>Q.ilex</i> transcriptomes generated. Contig (longer than 400 nucleotides = $L > 400$ nt) length distribution and comparative evaluation against oak transcriptome (BlastN e-value = 10^{-30}). (A) Trinity; (B) Ray; (C) MIRA.	24
4.1	Alignment between <i>Q. robur</i> and <i>Q. petraea</i> transcriptomes (oak transcriptome) and <i>Q. ilex</i> (holm oak) transcriptome using MIRA, RAY, TRINITY and RAY hybrid assemblies from Illumina (a) and Ion Torrent (b) reads. Distribution of percent sequence identity between oak and <i>Q. ilex</i> (MIRA, RAY, TRINITY, RAY hybrids) transcriptomes (c).	39
4.2	Results of BUSCO analysis of the holm oak transcriptome. All the transcriptomes are organized depending on their completeness: RAY-hybrid assembly, RAY-partial hybrid assembly, MIRA-Illumina assembly, RAY-Illumina assembly, MIRA-Ion Torrent assembly; RAY-Ion Torrent assembly; TRINITY-Ion Torrent assembly; and TRINITY-Illumina assembly. Blue: complete and single-copy genes; orange: complete and duplicated genes; grey: fragmented genes; yellow: missing genes.	40
4.3	Histogram of GO classification of assembled <i>Quercus ilex</i> transcripts. Horizontal bar charts of the distribution of GO associated with the holm oak transcripts represented in the three main GO categories: biological processes (a), molecular functions (b) and cellular components (c). The first twelve transcripts assigned to each GO category are shown and the remaining transcripts assigned to each GO category are shown in Table S1.	42

4.4	Experimental work flow showing the steps carried out and bioinformatic utilities used for a transcriptome analysis.	48
5.1	Functional categorization and distribution in percentage of the identified metabolites, proteins and transcripts, according to the categories establish by MERCATOR. (A) Metabolome. (B) Transcriptome. (C) Proteome. The pie charts show different functional categories: PS (Photosynthesis), major CHO metabolism, minor CHO metabolism, glycolysis, fermentation, gluconeogenesis/glyoxylate cycle, OPP (Oxidative Pentose Phosphate), TCA/org transformation, mitochondrial electron transport/ATP synthesis, cell wall, lipid metabolism, N-metabolism, amino acid metabolism, S-assimilation, metal handling, secondary metabolism, hormone metabolism, co-factor and vitamin metabolism, tetrapyrrole synthesis, stress, redox, polyamine metabolism, nucleotide metabolism, biodegradation of xenobiotics, C1-metabolism, miscellanea, RNA, DNA, protein, signaling, cell, micro RNA, natural antisense, etc., development, transport, and not assigned.	64
5.2	Phylogenetic tree of angiosperms. The tree shows the five-species compared (<i>Arabidopsis thaliana</i> , <i>Eucalyptus grandis</i> , <i>Oryza sativa</i> subsp. <i>japonica</i> , <i>Populus trichocarpa</i> , and <i>Quercus ilex</i>). The sequence similarity of species coincides with the classification in the phylogenetic tree. Species are ranked from highest to lowest similar to <i>Q. ilex</i> : <i>P. trichocarpa</i> (91.7%), <i>E. grandis</i> (88.4%), <i>A. thaliana</i> (85.6%), and <i>O. sativa</i> subsp. <i>japonica</i> (77.8%).	65
5.3	Venn diagram for the comparison of enzymes in <i>Arabidopsis thaliana</i> , <i>Eucalyptus grandis</i> , <i>Oryza sativa</i> subsp. <i>japonica</i> , <i>Populus trichocarpa</i> <i>in silico</i> proteomes, and <i>Quercus ilex</i> proteome. The Venn diagram shows the overlap of enzymes detected.	67

- 5.4 Metabolites and enzymes (protein or transcript level) assigned to the citrate cycle (TCA cycle). Omics data are highlighted in red (metabolites), blue (proteins), yellow (transcripts), and green (both proteins and transcript). The enzymes (proteins and transcripts) are named by their EC number. EC numbers and respective detected TCA cycle enzymes: 2.3.3.1 (Citrate synthase), 4.2.1.3 (Aconitate hydratase), 1.1.1.42 [Isocitrate dehydrogenase (NADP+)], 1.2.4.2 (alpha-ketoglutarate dehydrogenase), 6.2.1.4 (Succinyl coenzyme A synthetase), 1.3.5.1 (Succinate dehydrogenase), 4.2.1.2 (Fumarate hydratase), 1.1.1.37 (Malate dehydrogenase). There are two full reactions (metabolite, protein and transcript level) This figure was adapted from KEGG reference pathway. 71
- 5.5 MapMan overview of general metabolism for the metabolites and proteins/transcripts of *Quercus ilex*. (A) Visualization of 58 metabolites in the context of general metabolism using the using MapMan software. (B) Visualization of 58 metabolites in different MapMan pathways. Each red square represents a metabolite and each gray circle represents a protein or transcript. More details can be found in (Usadel et al., 2009) 73
- 6.1 Venn diagram of the number of variable transcripts and proteins between treatments, droughted and well-watered seedlings, found at the two sampling, days 20 and 25. Intercepts show common differences at the two times and/or two platforms. 88
- 6.2 Gene Ontology analysis of the variable gene products. (a) Biological process; (b) Molecular function; (c) Cellular location. The X axes contain the categories and the Y one the number of identified gene products. Blue: transcript at day 20, Red: protein at day 20; Light blue: transcripts at day 25, Orange: proteins at day 25. Original data are included in Supplementary Table S18. 91
- 6.3 GeneMANIA network analysis. It was performed with variable gene products detected at the two -omics level. The analysis was performed at days 20 (a) and 25 (b). A red background circle indicates upregulated gene and protein, a blue background circle indicates downregulated gene and protein, a purple background circle indicates contrary gene and protein changes, and a grey background circle indicates GeneMania predicted interactions. 98
- S1 Distribution of sequence lengths over all sequences used in this study. . 146

S2	Efficiency in the use of computational resources in each assembler used in this study (RAY, MIRA and TRINITY) from Illumina clean raw data. Ncpus indicates how many central processing units (CPUs) are used by the software, Ncpus_sys indicates how many CPUs are used by the system, Mem indicates RAM memory and Process_creation indicates how many files are created.	147
S3	Efficiency in the use of computational resources in each assembler used in this study (RAY, MIRA and TRINITY) from Ion Torrent clean raw data. Ncpus indicates how many central processing units (CPUs) are used by the software, Ncpus_sys indicates how many CPUs are used by the system, Mem indicates RAM memory and Process_creation indicates how many files are created.	148
S4	Efficiency in the use of computational resources in the RAY assembler from hybrid transcriptome (a) and partial hybrid transcriptome clean raw data (b). Ncpus indicates how many central processing units (CPUs) are used by the software, Ncpus_sys indicates how many CPUs are used by the system, Mem indicates RAM memory and Process_creation indicates how many files are created.	149
S5	Density histogram for proteins in the different biological processes of <i>Q. ilex</i> annotated transcriptome.	151
S6	Density histogram for proteins in the different cellular components of <i>Q. ilex</i> annotated transcriptome	152
S7	Density histogram for proteins in the different molecular functions of <i>Q. ilex</i> annotated transcriptome	153
S8	Enzymes (transcript level and protein level) assigned to the glycolysis/gluconeogenesis.	154
S9	Correlation between transcripts and proteins abundance. Only gene products detected at both transcriptomic and proteomic levels were considered. (a) and (c) correspond to, respectively, the total and the variable gene product datasets. (b) and (d) correspond to the low abundant gene products.	163

S10	PCA and sPLS analysis of the data. PCA (on the right) and sPLS (on the left) plots based on the two first components, PC1, and PC 2. Different datasets were employed for the analysis: (A) total transcriptome and proteome, (B) total transcriptome, (C) total proteome, (D) total variable transcriptome and proteome, (E) variable transcriptome, and (F) variable proteome. (Blue) Control at day 20, (Orange) Control at day 25, (Grey) Drought at day 20, (Green) Drought at day 25. The three replicates per sample are shown.	164
S11	PCA and sPLS analysis of the data. (D) total variable transcriptome and proteome, (E) variable transcriptome, and (F) variable proteome. (Color) Control at day 20, (Color) Control at day 25, (Color) Drought at day 20, (Color) Drought at day 25. The three replicates per sample are shown.	165
S12	KEGG metabolic charts of the twelve pathways showing statistically significant differences between treatments, control and drought. In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right).	166
S13	Paintomics KEGG differential pathway analysis: Ribosome (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	167
S14	Paintomics KEGG differential pathway analysis: Glyoxylate and dicarboxylate metabolism (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	168
S15	Paintomics KEGG differential pathway analysis: Phenylpropanoid biosynthesis (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	169
S16	Paintomics KEGG differential pathway analysis: Phenylalanine metabolism (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right) . . .	170
S17	Paintomics KEGG differential pathway analysis: Flavonoid Biosynthesis (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right) . . .	171

S18	Paintomics KEGG differential pathway analysis: Stilbenoid, diarylheptanoid and gingerol biosynthesis (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	172
S19	Paintomics KEGG differential pathway analysis: Biosynthesis of Secondary Metabolites (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	173
S20	Paintomics KEGG differential pathway analysis: Photosynthesis (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	174
S21	Paintomics KEGG differential pathway analysis: Carbon metabolism (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	175
S22	Paintomics KEGG differential pathway analysis: Plant-pathogen interaction (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	176
S23	Paintomics KEGG differential pathway analysis: Cysteine and Methionine metabolism (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	177
S24	Paintomics KEGG differential pathway analysis: Ubiquinone and other terpenoid-quinone biosynthesis (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)	178

List of Tables

1.1	Milestones in Bioinformatics	4
1.2	List of some of the principal biological databases of molecular data. . .	5
1.3	The 15 Best Free Linux Bioinformatics Tools (www.linuxlinks.com) . .	8
3.1	Comparison of <i>Q. ilex</i> transcriptome assembly using Trinity, RAY, and MIRA assemblers. Statistics and structure of the transcriptome assembly are indicated, including the number of contigs obtained of a minimum length (QUAST output data). Comparative hits with oak transcriptome are shown indicating the number of genes shared with oak and those newly found in <i>Q. ilex</i> . *Oak total transcripts = 87016; **BlastN with e-value = 10^{-30}	26
4.1	Summary of the structure of the holm oak assembly. *Data from the Illumina platform were previously published in (Guerrero-Sanchez et al., 2017).	36
4.2	Blast percentage matrix of all the transcriptomes built in holm oak. Each cell in the matrix represents the overlap between two assemblers-platforms.	43
4.3	Summary of the total number of proteins annotated in Holm oak. . . .	44
5.1	Metabolite families from GC-MS data of <i>Quercus ilex</i> . Six main chemical families of metabolites are represented. Carbohydrates (19), organic acids (19), amino acids (11), fatty acids (4), polyols (2), phenolic compounds (2) and four unique compound classes (others). Data in the brackets are KEGG compound identifier of each metabolite.	68
5.2	Number of metabolites and enzymes (proteomic and transcriptomic level) in KEGG pathways. Pathways according to the KEGG pathway maps based on <i>Arabidopsis thaliana</i> . The <i>Arabidopsis</i> pathway identifiers are in brackets. The table shows the most representative pathways. The complete list of pathways is in the Supplementary Material Table S12. .	69

6.1	Features of the transcriptome and proteome analysis at the two sampling days. The total number of identified transcripts and proteins, its sequence length, as well the number of them showing qualitative or quantitative differences between treatments are indicated. Newly appeared/disappeared indicates transcripts and proteins showing qualitative changes, being only present in drought/control treatments. Up/down indicates transcripts and proteins showing quantitative changes, being more abundant in drought/control treatments. a-e correspond to minimum (a) and mean (b) bp values; number of sequenced amino-acids with at least 1 unique peptide (c); sequenced amino acids mean value (d), and number of unigenes (e-f).	87
6.2	Main features of the sPLS and PCA analysis based on the fifth first components. Values correspond to the percentage of variability explained by each component, and the number of components explaining 50% of the variability.	94
6.3	List of gene products exhibiting the highest loadings to component 1 in the sPLS analysis of the variable transcripts and proteins. Contig ID, gene3 acronyms and description, loading parameter and quantitative relative values, drought/control.	95
6.4	KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked and showed statistically significant differences (Fischer test, $p < 0,05$) between treatments. Columns include the pathway name, number of gene products identified within each pathway, significance p-values at the transcript, protein or combined levels, changes in response to drought (up or down) and identified transcript or protein enzymes.	96
6.5	List of transcripts and proteins linked to the variable pathways. The columns include the name of the pathway, and gene products up and down accumulated at days 20 and 25.	96
S1	Total number of transcripts included in the GO and Uniprot classification in holm oak. Accessible in this link:	145
S2	List of transcripts related to drought stress in the holm oak transcriptome.	145
S3	Total number of transcripts included in the GO and Uniprot classification in holm oak. Accessible in this link:	145
S4	Metabolite features.	155

S5	GC-MS metabolomic data. Mean values of normalized peak areas, as well as SD (standard deviation) and CV% (percentage of coefficient of variation) were determined for replicates of the metabolite extract. CV% = (SD/mean)*100. Range of CV% (0.7 - 40.0). CV% mean (13.70).	156
S6	KEGG pathways with metabolites, proteins, and transcripts.	157
S7	Comparison of KEGG pathways of different species.	157
S8	Bins of transcripts.	157
S9	Bins of proteins.	157
S10	Comparison of <i>in silico</i> proteomes.	157
S11	Shotgun LC-MS/MS proteomic data. Mean values of normalized peak areas, as well as SD (standard deviation) and CV% (percentage of coefficient of variation) were determined for replicates of protein extract (Jorge et al., 2005; Jorge et al., 2006). The area values correspond to replicate 1 (0.6 μ g of protein), replicate 2 (0.8 μ g of protein), and replicate 3 (1 μ g of protein). Accessible in this link:	157
S12	Enzymes (transcripts).	157
S13	Enzymes (proteins).	157
S14	Omics overview.	158
S15	List of transcripts. Columns A to M correspond to contig ID, corresponding gene acronyms, and annotations. Columns N to AG contain quantitative values for each gene for each sample within each sample and the statistical parameters, FDR and p-value:	160
S16	List of proteins. Columns A to D correspond to contig and protein IDs, corresponding gene and description. Columns E to AA contain quantitative values for each gene for each proteoform within each sample and the statistical p-value.	160
S17	Correlation between transcripts and proteins abundance. Two different datasets were employed, the total and the group containing variable ones. The first and second columns correspond to the quantitative values for, respectively, transcripts and proteins. The Pearson correlation coefficient for each dataset is indicated.	160
S21	KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked. Columns include the pathway name, number of gene products per pathway corresponding to <i>Arabidopsis thaliana</i> , and significance p-values at the transcript, protein or combined levels.	160

S21	KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked. Columns include the pathway name, number of gene products per pathway corresponding to <i>Arabidopsis thaliana</i> , and significance p-values at the transcript, protein or combined levels.	161
S21	KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked. Columns include the pathway name, number of gene products per pathway corresponding to <i>Arabidopsis thaliana</i> , and significance p-values at the transcript, protein or combined levels.	162
S18	Gene Ontology analysis of the variable gene products. Tabs correspond to biological process, molecular function, and cellular component. Each tab contains the categories, number of total and identified genes, acronyms of the identified genes, and the enrichment FDR. Data are organized according to the -omics platform (from transcriptomics to proteomics), up/down regulated in droughted seedlings, and sampling time, days 20 and 25.	163
S19	KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked. Columns include the pathway name, number of gene products per pathway corresponding to <i>Arabidopsis thaliana</i> , and significance p-values at the transcript, protein or combined levels.	163
S20	GeneMANIA network analysis. Columns correspond to the gene acronyms, functional annotation, GO code, log score, <i>Arabidopsis thaliana</i> ortholog, node type (query = interrogated gene products; result= predicted gene products; unknown= contrary gene and protein changes).	163

Chapter 1

Introduction

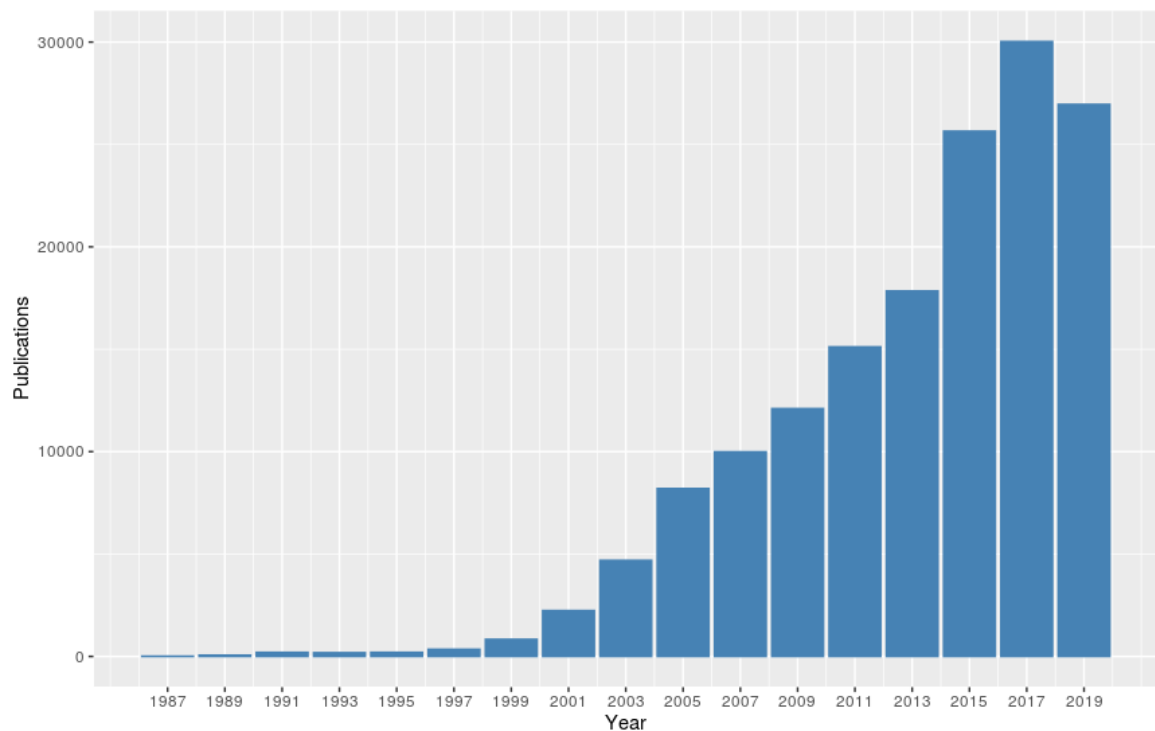


Figure 1.1 Number of publications per year appearing in PubMed from 1984 to 2018, using 'Bioinformatics' as searching term.

1.1 Bioinformatics, definitions and history

Bioinformatics, according to its etymology, is the application of informatics to biological research and knowledge. The Oxford dictionary defines Bioinformatics as "The science of collecting and analysing complex biological data such as genetic codes". A more scientific definitions are: defines Bioinformatics as the science of storing, retrieving and analysing large amounts of biological information (European Bioinformatics Institute (EMBL-EBI))or "The study of biological information using concepts and methods in computer science, statistics, and engineering (Rhee, Dickerson, and Xu, 2006). It is a highly interdisciplinary field involving many different types of specialists, including biologists, molecular life scientists, computer scientists and mathematicians. In summary, is the application to biology of mathematically based computational tools and statistics.

The term Bioinformatics was coined by Paulien Hogeweg and Ben Hesper to describe 'the study of informatic processes in biotic systems' and firstly appeared in a publication (Hesper and Hogeweg, 1970). The number of publications in which reference is made to bioinformatics has grown exponentially over the thirty last years (Figure 1.1).

Bioinformatics have been applied in different fields, from structural and functional genomics, physiology, morphometry, phenotyping, data integration, and biotic and

abiotic interactions in specific ecosystems, among others. It has been used in basic and translational research, including Biomedicine, Agriculture, Climate change and other environmental concerns.

As reviewed in (Rhee, Dickerson, and Xu, 2006), bioinformatic tools can be divided in two categories, although with no clear limits between them: biological information management and computational biology,. The first one is defined by the The National Institutes of Health (NIH) (<https://www.nih.gov/>) “research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, represent, describe, store, analyze, or visualize such data”. The second one is defined as “the development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral, and social systems”.

The methodology employed in the present Thesis belongs to the first category, as bioinformatic tools (algorithms, programs, databases and repositories) have been used to construct from wet data, mostly -omics one, the transcriptome, proteome, and metabolome of Holm oak and its integration, it aimed at defining the metabolism and responses to stresses (drought) in this species.

Some key milestones in the development of Bioinformatics as an independent discipline are mentioned in Table 1.1. In successive sections we will comment on the specific tools and software used in this doctoral thesis.

Table 1.1 Milestones in Bioinformatics

1962	First Bioinformatic software (COMPROTEIN)	(Dayhoff and Ledley, 1962)
1970	First Algorithm for DNA sequence alignment	(Needleman and Wunsch, 1970)
1971	Establishment of the Protein Data Bank	(Bernstein et al., 1977)
1974	First algorithm for predicting protein structures	(Chou and Pasman, 1974)
1978	First probabilistic model of aminoacid substitution	(Dayhoff, Schwartz, and Orcutt, 1978)
1979	First software for analyzing Sanger sequencing reads	(Staden, 1979)
1981	Creation of GCG and DNASTAR software	(Devereux, Haeberli, and Smithies, 1984)
1985	Creation of a journal specialized in bioinformatics	(Beynon, 1985)
1986	EMBL and GenBank databases are unified	https://www.insdc.org/
1987	DDBJ joins to the EMBL and GenBank union	https://www.insdc.org/
1988	Development of FASTA algorithm	(Pearson and Lipman, 1988)
1990	Release of Blast algorithm	(Altschul et al., 1990)
1991	First version of Linux	(Torvalds, 1991)
1991	First version of Python programming language	(Rossum and Boer, 1991)
1993	Release of R programming language	(Ihaka and Gentleman, 1996)
2005	454 sequencing. Next Generation Sequencing	(Margulies et al., 2005a).
2005	Release of the functional annotator, Blast2GO.	(Conesa et al., 2005)
2006	First Solexa sequencer	(Bennett, 2004)
2011	Pacific Biosciences commercialized SMRT sequencing	(Eid et al., 2009)
2015	Oxford Nanopore announces MinION sequencer	(Mikheyev and Tin, 2014)

Table 1.2 List of some of the principal biological databases of molecular data.

DNA databases	DNA Data Bank of Japan (DDBJ)
	European Nucleotide Archive (EMBL-EBI)
	GenBank (NCBI)
Gene expression databases	Gene Expression Omnibus (GEO)
	Expression Atlas
Protein databases	UniProt
	InterPro
	ProteomeXchange
	Pfam
	Protein Data Bank
Metabolite database	MetabolomeXchange
	KEGG
	MetaCyc
	Plant Metabolic Network (PMN)

With the advancement of the omics techniques, including Transcriptomics, Proteomics and Metabolomics Bioinformatics is a key tool for interpreting experimental molecular biology studies which due to their large volume of data and complexity, would be difficult to approach manually.

The increasing amount of information available on genes, proteins or metabolites, must be compiled and arranged in databases. To this end, there are multiple computer resources that make basically any pieces of information available to researchers (Table 1.2). However, most of the data available in databases correspond to model species or species that have been extensively studied for different reasons. Orphan species from molecular studies, such as forest species including Holm oak (*Quercus ilex*), European oak (*Quercus robur*) or Maritime pine (*Pinus pinaster*) are poorly represented in these databases. This fact makes a challenge to work with non-model organisms at first, but as new molecular, physiological or even behavioural data are generated, Bioinformatics to build an integrated interpretation of the functioning of a cell, tissue or organism.

1.2 Omics technologies

The technological advances and the development of computational algorithms, that allows the use of the information available in the databases have made possible the emergence of the omic disciplines (such as Genomics, Transcriptomics, Proteomics, Metabolomics and Phenomics). By omic techniques we mean those that allow the massive study of the molecules belonging to the different cellular functional levels from

genes, transcripts, proteins to metabolites. Bioinformatics, is a mandatory discipline for the interpretation of the vast amounts of data, such as those generated from the omics analyses; which in turn requires the use of techniques that can handle with high sensitivity and specificity large quantities of extremely complex biological samples, such as the recalcitrant forest species. (Schneider and Orchard, 2011). The omic approaches are briefly commented below following the flow of the biological information according to the Dogma of Molecular Biology.

Genomics is the discipline whose objective is the study and cataloguing of all the genes that an organism possesses and the organization, structure and function of each of them. Nowadays, the emergence of the Next Generation Sequencing (NGS) platforms (Illumina[®], Oxford Nanopore[®], Pacific Biosciences[®]) allows to sequence novel genomes of any species, as well as improve comprehensive sequencing from previously annotated in a fast and economic way (Weirather et al., 2017; Jung et al., 2019). In recent years the sequencing and annotation of some forest species of great importance of the genus *Quercus*, such as the common oak (*Quercus robur*) or the cork oak (*Quercus suber*) (Plomion et al., 2018; Ramos et al., 2018), have been carried out using these latest sequencing techniques and the information available for these species of great environmental importance has increased exponentially.

Transcriptomics aims to identify and quantify gene expression under certain conditions. Monitoring individual gene expression profile is routinely conducted by quantitative real time PCR (RT-PCR), while untargeted global profile of gene expression were done by microarrays. Recently Transcriptomics global analyses have taken advantage of new emerged NGS technologies, which make possible to sequence and quantify at once all the transcripts present in any biological system. Hence, these newly developed methodologies allow accomplishing the ultimate goal of Transcriptomics, to quantify, as precisely as possible, the greatest possible number of transcripts and their variants. Clustering methods are used to order and visualize the underlying patterns in large scale expression datasets showing similar patterns that can therefore be grouped according to their co-regulation/co-expression (e.g. specific developmental times or cellular/tissue locations) (Vidman, Källberg, and Rydén, 2019). In this doctoral thesis, Transcriptomics is used to quantify the expression levels of the genes of holm oak to decipher which of them are involved in drought stress tolerance.

Proteomics is the large-scale study of proteins, including their expression patterns, structures, modifications, interactions and functions (Anderson and Anderson, 1998). The development of Proteomics have been possible thanks to the advances made in mass spectrometry (MS), that allows sensitive and comprehensively analyses for identifying

proteins, separating them by their mass/load ratio. The identification of parent proteins from derived peptides now relies almost entirely on the software of search engines, which can perform *in silico* digests of protein sequence to generate peptides. Their molecular mass is then matched to the mass of the experimentally derived protein fragments (Angel et al., 2012). In the present doctoral thesis, a proteomic profiling of holm oak seedlings is performed, which together with the quantification data of transcripts, allow obtaining a holistic vision of the biology of the species.

Metabolomics is the arrange of techniques intended for the comprehensive and quantitative analyses of all the metabolites present in a biological sample. The two most common analytical approaches for the generation of metabolomics data are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) (Turi et al. 2018), but other technologies like near infrared spectroscopy (NIRS) are often used. MS based metabolomics is generally preceded by a separation step, which reduces the complexity of the biological sample and allows the MS analysis of different sets of molecules at different times. The most common separation techniques in MS technology are liquid chromatography (LC), Capillary electrophoresis (CE) and gas chromatography (GC) columns (Zhang et al., 2016b). Together with the integration of transcriptomic and proteomic data, a qualitative integration of metabolomic data corresponding to the identification of some metabolites in holm oak has been carried out (López-Hidalgo et al., 2018).

Phenotype can be any characteristic of the species, such as: the growth rate of a tree, its fruit production capacity, the proteins that constitute a certain tissue, etc. In this way, a phenomic study consists in the acquisition of datasets on the characteristics of an organism (Houle, Govindaraju, and Omholt, 2010). The phenomenon allows the understanding of the genotype-phenotype relationship and the effects of environmental factors on the development and expression of the phenome.

The independent analysis of the data generated from each of the omics technologies provides valuable information for the knowledge of the organism on study and its application in various biological areas. However, in these individualized analyses some crucial information is missing that can only come from the integration of all set of data. In this way, the development of tools that allow the analysis of the different omic data together, by means of multivariate analysis, can provide a more realistic view of what happens in the system biological.

1.3 Bioinformatics methods and tools

A high number of softwares available for analysing biological data. Those freely accessible have been very popular in recent years with the promotion of open source science. Besides being its source code available, these softwares can be permanently updated and improved according to the needs. In Table 1.3 some of the most popular free bioinformatics software are listed. Below it will be described the most important

Table 1.3 The 15 Best Free Linux Bioinformatics Tools (www.linuxlinks.com)

Bioconductor	Analysis and comprehension of high-throughput genomic data
Biopython	Tools for biological computation written in Python
BioPerl	Perl tools for computational molecular biology
InterMine	Integrate biological data sources
UGENE	Set of integrated bioinformatics software
IGV	High-performance visualization genome browser tool
BioJava	Provides Java tools for processing biological data
GROMACS	Versatile package to perform molecular dynamics
Taverna Workbench	For designing and executing bioinformatics workflows
EMBOSS	The European Molecular Biology Open Software Suite
Clustal Omega	Multiple sequence alignment program
BLAST	Algorithm for comparing primary biological sequence information
bedtools	Powerful toolset for genome arithmetic
geWorkbench	Software platform for integrated genomic data analysis
Bioclipse	Rich-client platform chemistry and biology workbench

methodologies, softwares, databases and algorithms that have helped to achieve the objectives of this doctoral thesis. As mentioned above, Bioinformatics is a multidisciplinary area that draws on knowledge from other scientific fields. In this particular case, of non-model species, software related to the assembly and annotation of sequences, algorithms for the quantification of molecular species and platforms aimed at carrying out a consensus evaluation of the results have been essential.

1.3.1 Genomics

The most widely used sequencing technologies to date are capable of sequencing small fragments of an organism's DNA (Illumina®, Oxford Nanopore®, Pacific Biosciences®). Each of them has different chemical base, which vary in the amount of data generated, the length of base pairs capable of sequencing and the amount and types of errors they generate. The millions of short sequences generated, called 'reads', are then used by different computer software to reconstruct or assemble the complete DNA sequences. However, before proceeding to the assembly of sequences, it is necessary a previous

step to evaluate the quality of these raw reads and eliminate low quality sequences and errors. Tools such as FastQC (Andrews, 2010) provide phred score information, which is a standard parameter for measuring quality in the identification of the bases generated by the sequencing platforms. Quality phred value is defined by:

$$q = -10 \log_{10} p$$

where p is the estimated error probability for that base-call. Thus, a base-call having a probability of 1/1000 of being incorrect is assigned a quality value of 30. Note that high quality values correspond to low error probabilities, and conversely (Ewing and Green, 1998). The %GC distribution should be adjusted to a normal distribution with the maximum percentage of mean GC of the particular species on study, otherwise it would mean that the DNA extract could be contaminated with genetic material from another organism. The sequences or poor quality regions of raw reads are filtered and removed using software packages such as FASTX-Toolkit (Gordon and Hannon, 2010), Trimmomatic (Bolger, Lohse, and Usadel, 2014) or Cutadapt (Martin, 2011). At present there are numerous software for the assembly of sequences, being most of them are based either on the De Bruijn graph method (Trinity, Ray, Velvet) (Grabherr et al., 2011; Boisvert, Laviolette, and Corbeil, 2010; Zerbino and Birney, 2008) or on the Overlap-Layout-consensus method (MIRA, Newbler, Edena) (Chevreux, Wetter, and Suhai, 1999; Margulies et al., 2005b; Hernandez et al., 2008). In addition, assemblers can also be used to reconstruct transcriptomes, that is, all those messenger RNAs that are being transcribed at any given time. For this purpose, Illumina technology is the most frequent choice due to the best quality/price ratio. Once the gene (Whole Genome Assembly) or transcript (Whole Transcriptome Assembly) sequences have been obtained, it is necessary to perform a gene annotation, to associate biological information to each gene. Nowadays, with the number of organisms being sequenced growing exponentially, it is necessary a nucleotide-level annotation high enough that allows the integration of genome sequence with other genetic and physical maps of the genome (Stein, 2001). Understanding the function of genes and their products in the context of cellular and organismal physiology is the goal of process-level annotation. The two most common approaches used for annotation are: Gene predictors *ab initio*. These softwares search, in the DNA strands, for certain structures and sequences of protein-coding genes. Sequence similarity gene finding methods. These softwares look for sequences that are similar to others already described in other works and deposited in gene or protein databases. Once the sequences have been identified, it is necessary to carry out a functional annotation, which consists in relating genes to biological

processes through Gene Ontology (GO) terms. The terms describe the function of genes in three classes: Biological processes, Molecular function and Cellular components. Additionally, it can be assigned the pathways to which each gene or protein belongs. The most popular annotation software is Blast2GO (Conesa and Götz, 2008), although there are other alternatives such as Sma3s, PANNZER2 or Trinotate (Munoz-Mérida et al., 2014; Törönen, Medlar, and Holm, 2018; Bryant et al., 2017).

1.3.2 Transcriptomics

When a reference genome or transcriptome has been assembled and annotated, a differential gene expression study can be performed. Sequencing platforms are used to reveal the presence and amount of each of the genes being transcribed under different experimental conditions. This process is known as RNA-seq and is currently widely used for a multitude of biological studies. The global quantification of the expression of all genes is done by mapping the reads obtained against a reference genome or transcriptome as the case may be. In other words, the raw reads are aligned against the complete sequences of the genes and the quantification of the number of readings aligned for each gene is carried out. In the case of organisms that do not contain introns in its genome, it is recommended the use of contiguous aligners such as Bowtie2 or BWA (Langmead and Salzberg, 2012; Li and Durbin, 2009) that were designed to align DNA. However, in the case of genomes with introns, the best option is to use aligners such as Hisat2, STAR or Kallisto (Kim, Langmead, and Salzberg, 2015; Dobin et al., 2013; Bray et al., 2016).

Once the mapping process is complete, a report will be generated containing the number of aligned readings and the normalized value for each gene. There are several standardization options such as CPM (Counts per Million), RPKM (Reads Per Kilobase Million) FPKM (Fragments Per Kilobase Million) or TPM (Transcripts Per Kilobase Million). Depending on the statistical package used to analyse the results, this may require a certain type of data normalization. Many of the statistical packages for this task are written in the programming language R and are deposited in the bioconductor database (<https://www.bioconductor.org/>). EdgeR and DEseq2 (Robinson, McCarthy, and Smyth, 2010; Love, Huber, and Anders, 2014) are two packages that compute the differential expression based on a model using the negative binomial distribution but DESeq can use a Wald test or a Likelihood Test while EdgeR performs a Fisher's exact test. Another well-known statistical package is limma that use of linear models for analysing designed experiments and the assessment of differential expression (Ritchie et al., 2015). Depending on the variability of the sample, the threshold for the Fold change,

the p-value or the ratio of false positives (FDR) can be adjusted. Once the statistical analysis of differential expression data has been carried out, a list of overexpressed or repressed genes will be obtained. To explore the functional processes in which a given set of genes are involved, functional enrichment can be carried out using tools such as ShinyGO, FunRich, FatiGO or Blast2GO (Al-Shahrour, Díaz-Uriarte, and Dopazo, 2004; Conesa and Götz, 2008; Ge and Jung, 2018; Pathan et al., 2015). Finally, differential expression data must be validated by quantifying the expression levels for some selected genes under the experimental condition by qPCR. Ultimately, this would indicate the usefulness of certain genes as markers for a specific characteristic.

1.3.3 Proteomics

Proteomic analyses provides complementary and equivalent information to the data commented above, but in order to carry out a proteomic study, it is necessary to have the sequences of the proteins, which in the case of model or well-studied species (i.e. Human, Arabidopsis, yeast) can be found in databases such as UniProt (<https://www.uniprot.org/>). However, when conducting proteomic analysis with orphan species there is no sequence information in the repositories, making it mandatory to use a proteome of a nearby species. However, provided a reference transcriptome is available, it is possible to translate *in silico* of the transcribed sequences to extrapolate the amino acid sequences that encode for their corresponding protein. The approach is known as Proteogenomics. With a reference proteome, it is now possible to process the peptide data analyzed by Shotgun LC-MS. Proteome DiscovererTM (ThermoFisher Scientific), Peaks and MaxQuant (Cox and Mann, 2008; Ma et al., 2003) are some of the most popular software for this type of tasks. Once protein identifications are obtained for a given species, a quantitative analysis can be performed by either, AUC (area under the curve) and spectral counting (Neilson et al., 2011). Nowadays, the normalization of the AUC ratio is the most popular and allows statistically analyse of changes in protein levels in a similar way as is done with transcripts.

1.3.4 Metabolomics

The gene expression and protein level analysis reveals the set of gene products that are being produced in the cell and represents a single and incomplete facet of cell function. In contrast, metabolic profiling goes a step further to the cell's physiology.

Metabolites are quite diverse molecules, differing in polarities, molecular weights, functional groups, stability and chemical reactivity, among other properties. Hence, its

study requires the use of multiple platforms and analytical configurations to maximize the coverage of the analysed metabolome, which is something that does not occur in genomics and proteomics experiments. The retention time for each metabolite is used together with the values of the mass/charge ratio to be obtained in the mass spectrometer by generating the necessary information for the identification of metabolites using a number of available databases such as LipidBank (<http://lipidbank.jp/>), LIPID MAPS (<https://www.lipidmaps.org/>), Metlin Database (<https://metlin.scripps.edu/>), NIST (<https://webbook.nist.gov/chemistry/>), MetaCyc (<https://metacyc.org/>), Golm Metabolome Database (gmd.mpimp-golm.mpg.de/), etc.

1.3.5 Interactomics

One of the main challenges of systems biology and functional genomics is to integrate information from proteomics, transcriptomics and metabolomics to provide a better understanding of cell biology. Interaction networks are one of the key elements within systems biology along with the differential equations that allow to define the changes in concentration of the different compounds over time. Nevertheless the latter is difficult to apply, as they involve knowledge of the specific concentrations and kinetics of numerous enzymes, substrates and products, the networks can be used easily. These networks are structures made up of nodes and connections that allow us to know the relationships established between the different nodes. In order to make sense of complex data networks, Bioinformatics helps us to organise and structure this information. Bioinformatics has become an essential instrument to perform global analyse and visualize these interactions, while databases such as KEGG (Kyoto Encyclopaedia of Genes and Genomes)(<https://www.genome.jp/kegg/>) or Reactome(<https://reactome.org/>), allow to establish maps of regulated enzymes/metabolites within different metabolic networks. Protein-protein interactions can also be determined using the STRING database (Szklarczyk et al., 2015).

The combination of proteomics and metabolomics provides a complementary source of information that improves the reliability of data interpretation. A first approximation when interpreting all previously generated data is to make an integrated visualization of each of the dataset from each omics. For this task there are tools such as Paintomics (Garcia-Alcalde et al., 2011), which is an application programmed in Python (<https://www.python.org>) and Perl (<https://www.perl.org/>), which processes quantitative data from various omics and reconstructs metabolic networks from the KEGG database. Paintomics performs a fisher test to calculate the combined p-value for each pathway from the individual data of each omic for any of its components.

1.3.6 Other approaches

The complexity of the study can still be increased if attempts are made to create predictive models with machine learning techniques from each omic dataset. Machine learning is a part of Artificial Intelligence (AI) whose objective is that a machine could learn from examples, like datasets, based on statistical methods. Nowadays there is a great availability of data, particularly in the field of bioinformatics. Machine learning techniques are increasingly being applied (Li, Wu, and Ngom, 2018), for example for automatic genome annotation and the analysis of omics data obtained in experiments with high-performance technologies. Some of the most popular methodologies used for automatic learning are: neural networks, random forest, decision trees, Naïve Bayes classification or Algorithm Clustering. A multi-layered omic approach based in learning patterns, allows such systems to make quite complex predictions when training with large datasets. This latter approach, along with those explained above, allows non-model species to be studied at almost the same level as their counterparts, generating predictive models from experimental data.

1.4 Multi-omics study of orphan species: the case of Holm oak (*Quercus ilex*)

Studying orphan species at the molecular level is a major challenge, but there are necessary to understand mechanisms that regulate physiology and response to environment of any organism, as is the case of the holm oak. Holm oak (*Quercus ilex*) is one of the predominant tree species in the western Mediterranean forest, and in the Iberian Peninsula it covers around 4 million hectares (Joffre, Rambal, and Ratte, 1999). It presents characteristics of sclerophyll species, such as small and coriaceous leaves, adaptations that allow it to tolerate drought and high temperatures (De Rigo and Caudullo, 2016). In addition, holm oak also has a high resistance to cold (Morin et al., 2007), which has allowed it to survive and dominate continental climates with cold winters followed by long, hot and dried summers. Holm oak germinates from seeds (acorns), although they are also reproduced by root and vine shoots. Its fruit, the acorn, is used in the agrosilvopastoral system of the 'dehesa', where it is used for the feeding the pigs destined to the production of hams of maximum quality (Cantos et al., 2003). Holm oak therefore has both economic and ecological value. However, these ecosystems have been threatened in recent decades, due to the presence of elderly individuals, over-exploitation, poor regeneration, inappropriate livestock management, and the severe

effect of forest decline attributed to pathogen attack (*Phytophthora cinnamoni* and *Hypoxyylon mediterraneum*). In addition to those threatens, the increase in temperatures and the extension of the periods of drought that have been developing in a scenario of climate change, endanger the survival of Holm oak (Sanchez et al., 2002). In this context, the works carried out by the Research Group 'Agroforestry and Plant Biochemistry, Proteomics and Systems Biology' (AGR-164) (<https://www.uco.es/probiveag/>), from the University of Cordoba (Spain), is focused on the study of different aspects of Holm oak biology, from physiological to molecular approaches, combining classical biochemistry with new -omics approaches (transcriptomics, proteomics and metabolomics) (Guerrero-Sanchez et al., 2017; Guerrero-Sanchez et al., 2019; López-Hidalgo et al., 2018).

Holm oak has an extreme complexity due to its biological characteristics, like recalcitrance, allogamy and a long-life cycle. In addition, like other forest species, classical breeding programmes are not viable and genetic engineering is neither feasible nor accepted within the European Union. The only possible alternative is the exploitation of biodiversity, which involves, as a preliminary step, the characterisation of biodiversity at morphological, phenological, physiological and molecular levels, including modern omics approximations. The final objective would be the identification of 'elite' or 'plus' trees with phenotypic characteristics based on molecular markers.

In the context of this doctoral thesis, it has been carried out a de novo assembly and annotation of Holm oak transcriptome. A multi-omic analysis protocol has been developed, using data from holm oak as a working model. As a result, a web repository (<http://www.uco.es/probiveag/holm-oak-database.html>) has been created where the databases of transcripts, proteins and metabolites of the *Q. ilex* are housed. Finally, a multi-omic study has made it possible to characterize the molecular response to drought in the holm oak at the level of transcripts, proteins and metabolic pathways.

Chapter 2

Objectives

- Perform a Transcriptomic, proteomic and metabolomic analysis of a representative sample of Holm oak, consisting of a mixture of different organs (root, leaf, seed and embryos). Identification and functional and structural cataloguing of molecular entities.
- Create a species-specific database of genes, proteins and metabolites of Holm oak. This database will be enriched with those data generated in the group or that, being freely accessible, appear in databases or publications.
- Generate a knowledge base of the molecular responses to drought in Holm oak by combining multiple omic analyses (transcriptomic, proteomic and metabolomic). Identification of genes and gene products responsible of drought response and tolerance.

Chapter 3

Holm oak (*Quercus ilex*) transcriptome. *De novo* sequencing and assembly analysis

Chapter published in Frontiers in Molecular Biosciences:

Guerrero-Sanchez, V. M., Maldonado-Alconada, A. M., Amil-Ruiz, F., and Jorrin-Novo, J. V. (2017). Holm Oak (*Quercus ilex*) Transcriptome. *De novo* Sequencing and Assembly Analysis. Frontiers in Molecular Biosciences, 4.

3.1 Introduction

Holm oak (*Quercus ilex* L. subsp. *ballota* [Desf.] Samp.) is the dominant tree species in the Mediterranean forest with great ecological and economic value (Pulido, Diaz, and Hidalgo de Trucios, 2001). It constitutes, together with cork oak (*Quercus suber*), the 'dehesa', a typical Mediterranean agro-forestry-pastoral ecosystem, covering almost four million hectares in the western Iberian Península (Joffre, Rambal, and Ratte, 1999). Besides, Holm oak is widely used in reforestation programs and silvicultural practices, being their seeds, acorns, used for feed, and fatten the exclusive Iberian race pigs, whose meat is the basis of a high-quality food industry (Vicente and Alés, 2006; Cañellas et al., 2007).

Nowadays, *Quercus ilex* forest maintenance and sustainability are facing severe problems and challenges. Those are related to agricultural practices, low natural regeneration, seed viability, which may be due to their non-orthodox seed character (Doody and O'Reilly, 2008), plant mortality in both adult trees and young plants after field transplantation resulting from adverse environmental conditions like drought, the so-called decline syndrome (Gallego, Algaba, and Fernandez-Escobar, 1999), especially considering the current and future climate change scenario (Plieninger, Pulido, and Schaich, 2004; Bates et al., 2008; Corcobado et al., 2013). Overcoming those threats could be greatly facilitated if Holm oak ecophysiological behavior was better understood at the molecular level. Nowadays, multidisciplinary approaches by integrating the so called omic studies transcriptomics, proteomics and metabolomics have become indispensable to shed light on the fine-tuned molecular regulation in many biological systems/species. Thus, system biology aims to describe and interpret the full complexity of cells, tissues, organs, and organisms.

In this context, our research group has been investigating different aspects of *Quercus ilex* biology such as natural variation, seed germination and seedling growth, physiology, biotic and abiotic stress-responses, combining classical biochemistry, and integrating those multidisciplinary 'omics' analysis (Echevarría-Zomeño et al., 2009; Echevarría-Zomeño et al., 2012; Jorrín-Novo et al., 2009; Valero-Galván et al., 2011; Valero-Galván et al., 2012; Valero-Galván et al., 2013; Sghaier-Hammami et al., 2013; Sghaier-Hammami et al., 2016; Romero-Rodríguez et al., 2014). Nevertheless, the scarce genomic information (to date) available for *Quercus ilex*, supposes, such as for other orphan tree species (Abril et al., 2011; Jorrín-Novo et al., 2015), a notable obstacle to successfully carry out these global studies at molecular level. Driven by that need, our main aim has been to generate a reference transcriptome of *Quercus ilex* which will support and complement future research within this species. For that purpose, as

a first approach we sequenced the mRNA of a pooled plant sample containing equal amounts of homogenized tissue from acorn embryo, leaves, and roots, using an Illumina HiSeq 2500 platform. Contrasting different assembly strategies and algorithms, we present here the first *de novo* assembled transcriptome of the non-conventional plant *Quercus ilex*.

The pre-processed raw reads generated by the sequencing platform, and used for the *de novo* assembly, have been deposited at the NCBI SRA database with accession number SRR5815058.

This new genomic resource will set the stage for ongoing and future studies to obtain a better understanding of molecular mechanisms involved in physiological processes such as seed germination, seedling establishment, drought, which are essential for selection of superior phenotypes or Candidate Plus for restoration and reforestation programs under the impending climate change in Mediterranean regions.

3.2 Materials and Methods

3.2.1 Plant material

Mature acorns from Holm oak (*Quercus ilex* L. subsp. *ballota* [Desf.] Samp.) were collected from a tree located in Aldea de Cuenca (province of Córdoba, Andalusia, Spain). Acorns were germinated and seedlings grew in a chamber under controlled conditions (a 12 h photoperiod, a temperature of 21 ± 1 °C, a relative humidity of $60 \pm 5\%$ and an irradiance of $200 \mu\text{mol m}^{-2}\text{s}^{-1}$, (Echevarría-Zomeño et al., 2009)). Germinated embryo, leaves and roots from 1 year plantlets were collected separately, weighted, and individually frozen in liquid nitrogen. The plant material used for RNA sequencing experiments consisted in a pool generated by mixing equal amounts of homogenized tissue from acorn embryo, leaves, and roots.

3.2.2 RNA extraction

Total RNA was extracted from 50 mg pooled plant sample according the procedures previously set up in our laboratory for *Quercus ilex* samples (Echevarría-Zomeño et al., 2012). Contaminating genomic DNA was removed by DNase I (Ambion) treatment. Total RNA was quantified spectrophotometrically (DU 228800 Spectrophotometer, Beckman Coulter, TrayCell Hellma GmbH & Co. KG). The high quality and integrity of the RNA preparation was tested electrophoretically (Agilent 2100 Bioanalyzer).

Only high-quality RNAs with RIN values > 8 and A260:A280 ratios near 2.0 were used for subsequent experiments.

3.2.3 Enrichment of mRNA, cDNA synthesis, and library generation for Illumina HiSeq 2500 platform. paired-end sequencing

The library construction of cDNA molecules was carried out using Illumina TruSeq Stranded mRNA Library Preparation Kit according to manufacturer instructions using 2 μ g of total RNA followed by poly-A mRNA enrichment using streptavidin coated magnetic beads and thermal mRNA fragmentation. The cDNA was synthesized, followed by a chemical fragmentation (DNA library) and sequenced in the Illumina HiSeq 2500 platform, using 100 bp paired-end sequencing (De Wit et al., 2012).

3.2.4 *De novo* assembly and analysis of high throughput RNA sequencing data

The raw reads obtained from the sequencing platform were pre-processed in order to retain only high-quality sequences to be subsequently used in the assembly. Thus, each original sequence was quality trimmed considering several parameters (quality trimming based on minimum quality scores, ambiguity trimming to trim off e.g., stretches of Ns, base trim to remove specified number of bases at either 3' or 5' end of the reads). The preprocessing parameters used were selected as following: trimming sequences by maximum 2 ambiguous nucleotides), minimum mean quality assuming error probability < 0.01 , and filtering out those sequences shorter than 30 nucleotides. Three different assemblers were employed to *de novo* assemble the *Quercus ilex* transcriptome, considering there is not a reference genome available, and further evaluated to contrast the results obtained (Figure 3.1).

Trinity 2.4.0. performs a *de novo* assembly using an algorithm based on Bruijn graphs (Grabherr et al., 2011). For the assembly, Trinity 2.4.0 was launched with a k-mer value ($k = 25$). Ray 2.3.1. assembly uses de Bruijn graphs but its framework is not based on the Eulerian steps. Specific subsequences, seeds, are defined, and for each of them, the algorithm extends it to a contig. Heuristics are defined that control the extension process in such a way that the process stops if, at some point, the readings family does not clearly indicate the address of the extension (Boisvert, Laviolette, and Corbeil, 2010). In this case we selected a k-mer value of 31.

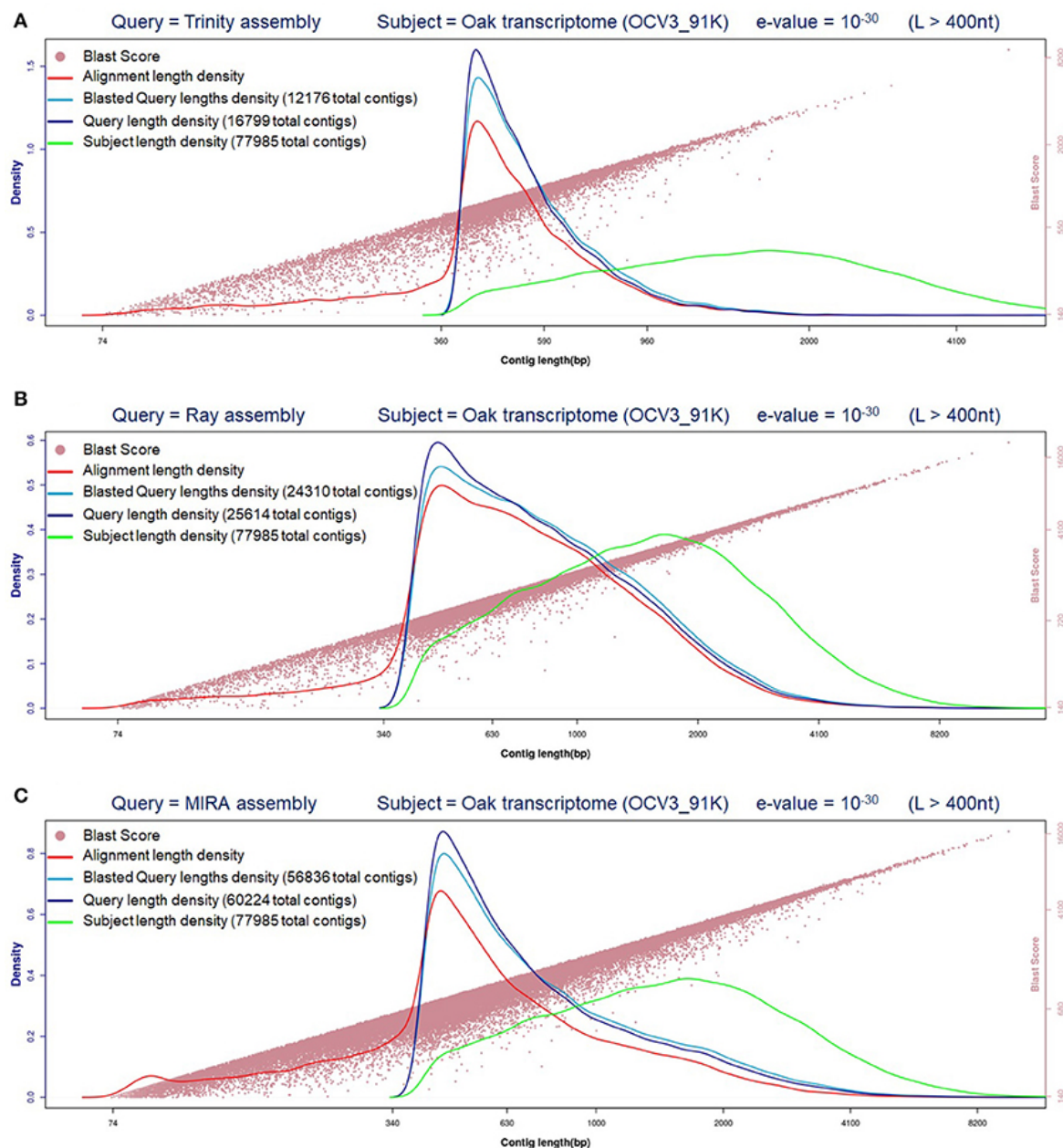


Figure 3.1 Evaluation of *Q. ilex* transcriptomes generated. Contig (longer than 400 nucleotides = L > 400 nt) length distribution and comparative evaluation against oak transcriptome (BlastN e-value = 10^{-30}). (A) Trinity; (B) Ray; (C) MIRA.

MIRA 4.9.6 software (Chevreux, Wetter, and Suhai, 1999), unlike Trinity and Ray, is based on the strategy known as Overlap /Layout/ Consensus. Following the author guidelines/recommendations for Illumina data, we used the complete raw data without a filtering process like we described previously.

Evaluation of the structure of the generated assemblies was done with the QUAST software (Gurevich et al., 2013).

The assemblies obtained using the three aforementioned softwares were blasted (e-value of 10^{-30}) against the most accurate and nearest phylogenetic transcriptome currently available, the oak transcriptome (containing *Quercus robur* and *Quercus petraea* sequences) (Lesur et al., 2015). That transcriptome database is divided in two files OCV3_91K and OCV3_101K but OCV3_91K has a larger amount of valuable information of *Quercus* spp. transcriptome. So, we chose OCV3_91K as a general oak transcriptome database.

3.3 Results

3.3.1 Evaluation and annotation of the assembled transcriptomes

There are differences between the three assembled transcriptomes in terms of transcriptome architecture/structure. Thus, the N50 value, number of contigs and the average length of the sequences generated by each algorithm differ (Table 3.1).

Considering these results, we can state that MIRA generated more and longer contigs than RAY and Trinity (MIRA>RAY>Trinity), suggesting that a more robust architecture/structure is obtained by MIRA for the *Q. ilex* transcriptome assembly. Upon the continuous development of NGS methods, data processing, and transcript assembly remains a main challenge. Several studies have been published devoted to evaluate different *de novo* assemblers varying in performance and quality in terms of number and length of transcripts and computational speed (Clarke et al., 2013). Besides, it has been reported that the quality of the assembly using a given software depends on the biological sample on study (Bradnam et al., 2013). Thus, these aspects should be taken into consideration when comparing different softwares. The comparison between the sequences generated from *Quercus ilex* and those available from the close species, oak transcriptome, reveals that MIRA assembly was the one which shared the higher number of transcripts (73073), followed by RAY assembler (Table 3.1). Besides, MIRA assembly sequences blasted against oak transcriptome

Table 3.1 Comparison of *Q. ilex* transcriptome assembly using Trinity, RAY, and MIRA assemblers. Statistics and structure of the transcriptome assembly are indicated, including the number of contigs obtained of a minimum length (QUAST output data). Comparative hits with oak transcriptome are shown indicating the number of genes shared with oak and those newly found in *Q. ilex*. *Oak total transcripts = 87016; **BlastN with e-value = 10^{-30} .

Number of original raw reads	55275472		
	MIRA	Ray	Trinity
# contigs (≥ 0 bp)	169449	107487	77159
# contigs (≥ 500 bp)	43014	20495	8803
# contigs ($\geq 1,000$ bp)	15445	8773	696
# contigs ($\geq 5,000$ bp)	155	73	1
# contigs ($\geq 10,000$ bp)	2	3	0
Largest contig	11254	12220	5916
Total length (≥ 0 bp)	83639406	41292773	26286544
Total length ($\geq 1,000$ bp)	27409911	14778197	904440
Total length ($\geq 5,000$ bp)	941227	471829	5916
Total length ($\geq 10,000$ bp)	21731	34168	0
N50	1211	1260	661
N75	742	827	563
L50	11473	5863	3428
L75	23813	11529	5931
GC (%)	41.69	42.47	39.14
Oak transcripts* present in <i>Q. ilex</i> **	73073	63950	49679
Oak transcripts* absent in <i>Q. ilex</i> **	13943	23066	37337
% of oak* transcripts in <i>Q. ilex</i> **	83,98	73,49	57,09

render the longest alignment lengths and better blast scores (Figure 3.1). Taking into consideration the data and parameters evaluated (Table 3.1 and Figure 3.1), we decided to use the MIRA assembly to continue with the corresponding annotation of *Quercus ilex* transcriptome. After blastX was completed against UniProt(Swiss-Prot) curated database (e-value of 10^{-5}), followed by the corresponding mapping process, 31973 annotated sequences were obtained by Blast2GO (Conesa and Götz, 2008).

3.4 Direct link to deposited data

The pre-processed raw reads of the transcriptome assembly generated by the sequencing platform, and used for the *de novo* assembly, have been deposited at the NCBI SRA database with the following accession number SRX2993508 and direct link: <ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR581/SRR5815058/SRR5815058.sra>

Chapter 4

Ion Torrent and Illumina, two complementary RNA-seq platforms for constructing the Holm oak (*Quercus ilex*) transcriptome

Chapter published in PLoS ONE:

Guerrero-Sanchez, V. M., Maldonado-Alconada, A. M., Amil-Ruiz, F., Verardi, A., Jorrín-Novo, J. V., and Rey, M.D. (2019). Ion Torrent and Illumina, two complementary RNA-seq platforms for constructing the Holm oak (*Quercus ilex*) transcriptome. PLoS ONE, 14(1):e0210356.

Abstract

Transcriptome analysis is widely used in plant biology research to explore gene expression across a large variety of biological contexts such as those related to environmental stress and plant-pathogen interaction. Currently, next generation sequencing platforms are used to obtain a high amount of raw data to build the transcriptome of any plant. Here, we compare Illumina and Ion Torrent sequencing platforms for the construction and analysis of the Holm oak (*Quercus ilex*) transcriptome. Genomic analysis of this forest tree species is a major challenge considering its recalcitrant character and the absence of previous molecular studies. In this study, *Quercus ilex* raw sequencing reads were obtained from Illumina and Ion Torrent and assembled by three different algorithms, MIRA, RAY and TRINITY. A hybrid transcriptome combining both sequencing technologies was also obtained in this study. The RAY-hybrid assembly generated the most complete transcriptome (1,116 complete sequences of which 1,085 were single copy) with a E90N50 of 1,122 bp. The MIRA-Illumina and TRINITY-Ion Torrent assemblies annotated the highest number of total transcripts (62,628 and 74,058 respectively). MIRA-Ion Torrent showed the highest number of shared sequences (84.8%) with the oak transcriptome. All the assembled transcripts from the hybrid transcriptome were annotated with gene ontology grouping them in terms of biological processes, molecular functions and cellular components. In addition, an *in silico* proteomic analysis was carried out using the translated assemblies as databases. Those from Ion Torrent showed more proteins compared to the Illumina and hybrid assemblies. This new generated transcriptome represents a valuable tool to conduct differential gene expression studies in response to biotic and abiotic stresses and to assist and validate the ongoing *Q. ilex* whole genome sequencing.

4.1 Introduction

Holm oak (*Quercus ilex* L.) forms natural forests or 'dehesa' ecosystems, playing an important role from an environmental and socio-economic point of view (Patón

et al., 2009). Holm oak, as with other forest tree species, can be defined as an orphan and recalcitrant experimental system, whose study at the molecular and genomic level represents a challenge. To date, partial studies using classical biochemical and proteomics approaches have shed some light on different aspects of *Q. ilex* biology such as natural variation, seed germination, seedling growth, physiology and, biotic and abiotic stress-responses (Echevarría-Zomeño et al., 2009; Echevarría-Zomeño et al., 2012; Jorrín-Novó et al., 2009; Valero-Galván et al., 2011; Valero-Galván et al., 2012; Valero-Galván et al., 2013; Sghaier-Hammami et al., 2016; Romero-Rodríguez et al., 2014).

The holm oak genome has not yet been sequenced, however, transcriptome analysis, using RNA-sequencing (RNA-Seq), offers an alternative technology now widely used to identify and characterize gene sequences (Wang, Gerstein, and Snyder, 2009; Li et al., 2014). In order to generate transcriptomes, a set of read sequences are obtained first by next generation sequencing (NGS) technologies. Of these, Illumina is the most commonly used. However, an alternative technology is provided by Ion Torrent instruments. The raw read data obtained using both platforms differ in some parameters such as fragment length, probability of base substitutions or insertion/deletion alterations in homopolymeric regions (Quail et al., 2012). Once generated, these reads must be de novo assembled to produce a transcriptome. Several de novo transcriptome assemblers are currently available (El-Metwally, Ouda, and Helmy, 2014; Biswas et al., 2014) that, combined with user-tunable parameters, enable the generation of a large figure of candidate assemblies for a single data set.

Recent studies have shown that the evaluation of de novo transcriptome assemblies remains a challenge (Li et al., 2014; Bradnam et al., 2013), and there is not a universal accepted optimal assembler identified for de novo generation.

Recently, a de novo transcriptome assembly of *Q. ilex* was published using an Illumina Hiseq 2500 platform (Guerrero-Sanchez et al., 2017; López-Hidalgo et al., 2018). Initially, 31,973 total sequences were annotated using the Blast2Go software (Guerrero-Sanchez et al., 2017) and later, the total number of transcripts was increased to 62,628 total sequences using the Sma3s v2 software (López-Hidalgo et al., 2018). To improve the amount of annotated sequences, in this work, we compare the resulting assembled sequences from two sequencing platforms, the new Ion Torrent reads against the Illumina transcriptome previously described by our group (Guerrero-Sanchez et al., 2017; López-Hidalgo et al., 2018). In addition, a hybrid transcriptome obtained from Illumina and Ion Torrent combined reads is discussed. It should be noted that the data obtained from each sequencing platform depends on the organism on study. Every

species has a different number of genes which requires a tailored sequence yield for an effective transcriptome (Lowe et al., 2017). Moreover, a comparison of three assemblers (MIRA, TRINITY and RAY), each using different algorithms, for the construction of a new de novo transcriptome of holm oak is carried out in each platform and then compared to each other. The assemblies to provide a transcriptome are highly variable in the contigs and scaffold lengths, and in the total assembly size (Bradnam et al., 2013), (Clooney et al., 2016).

4.2 Materials and methods

4.2.1 Plant material

Mature acorns from holm oak (*Quercus ilex* L. subsp. *ballota* [Desf.] Samp.) were collected from a tree located in Aldea de Cuenca (province of Cordoba, Andalusia, Spain). Acorns were germinated, and seedlings grown in a chamber under controlled conditions previously described in (Guerrero-Sanchez et al., 2017). Germinated embryos, leaves and roots from 6-months plantlets were collected and individually frozen in liquid nitrogen. The plant material used for RNA sequencing experiments consisted of a pool generated by mixing equal amounts of homogenized tissue from acorn embryos, leaves and roots.

4.2.2 RNA extraction

Total RNA was extracted from the frozen homogenized pool tissue following the procedure previously reported by (Guerrero-Sanchez et al., 2017). A total of 50 mg pooled fresh tissue was used following the protocol previously described by (Echevarría-Zomeño et al., 2012). Contaminating genomic DNA was removed by DNase I treatment (Ambion, Austin, TX). Total RNA was quantified spectrophotometrically (DU 228800 Spectrophotometer, Beckman Coulter, TrayCell Hellma GmbH & Co. KG), and the integrity of the isolated RNA was assessed using a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, Calif.). Only high-quality RNAs with RIN values > 8 and A260:A280 ratios near 2.0 were used for subsequent experiments.

4.2.3 RNA-Seq Library Construction, Illumina sequencing and de novo assembly

The holm oak Illumina transcriptome was previously described in (Guerrero-Sanchez et al., 2017; López-Hidalgo et al., 2018). Briefly, the library construction of cDNA molecules for Illumina sequencing was carried out by Illumina TruSeq Stranded mRNA library preparation kit using 2 μ g of total RNA. The cDNA was synthesized and sequenced in the Illumina Hiseq 2500 platform and three different assemblers (TRINITY 2.5.1 (Grabherr et al., 2011), RAY 2.3.1 (Boisvert, Laviolette, and Corbeil, 2010) and MIRA 4.9.6 (Chevreux, Wetter, and Suhai, 1999) algorithms) were employed to *de novo* assemble the *Q. ilex* transcriptome. Both the length and distribution of Illumina reads are shown in Figure S1

4.2.4 RNA-Seq Library Construction, Ion Torrent sequencing and de novo assembly

The cDNA library was built using the Ion Total RNA-Seq Kit v2 for whole transcriptome libraries (Life Technologies Corporation, California, USA), using an aliquot from the same RNA used for Illumina. Thus, 10 ng and 50 ng of total RNA were employed to generate in parallel two cDNA libraries that were loaded by an Ion Chef System in two Ion 540 sequencing chips and then, further sequenced by an Ion S5 System. Raw reads with length up to 372 nucleotides (mean of 112 nucleotides) from each sequencing chip were processed to filter out poor quality sequences (Cutadapt version 1.9 (-m 100) and BBDuk version 35.43 (qtrim = rt trimq = 20)). Sequencing adapters were first clipped, and low-quality bases (with phred score below a threshold) were trimmed in raw sequences. A phred score value was selected as thresholds (20) and reads shorter than 100 nucleotides were filtered out. Both the length and distribution of Ion Torrent reads are shown in Figure S1 The processed reads were assembled into contigs using the same assemblers (TRINITY version 2.5.1, RAY version 2.3.1 and MIRA version 4.9.6) used to obtain the Illumina transcriptome described in the previous section, but the parameterizations were:

TRINITY chosen parameters:

```
-max_memory 1000G -CPU 20  
-SS_lib_type F  
-bflyCalculateCPU  
-normalize_max_read_cov 20  
-KMER_SIZE 25
```

```
-min_kmer_cov 2
```

RAY chosen parameters were:

```
-n 22
```

```
-k 31
```

MIRA chosen parameters were:

```
job = denovo,est,accurate
```

```
COMMON_SETTINGS
```

```
-GENERAL:number_of_threads = 12
```

```
-KMERSTATISTICS:lossless_digital_normalisation = yes
```

```
IONTOR_SETTINGS
```

```
-ALIGN:min_relative_score = 70
```

```
-ASSEMBLY:minimum_read_length = 100
```

```
-CLIPPING:quality_clip = no
```

```
-CLIPPING:qc_window_length = 20
```

```
-CLIPPING:qc_minimum_quality = 15
```

```
-CLIPPING:clip_polyat = yes
```

```
-CLIPPING:cp_min_sequence_len = 12
```

```
technology = iontor
```

As with the Illumina transcriptome, the assembly calculations were run in the Computations Cluster of CICA (Centro de Información Científica de Andalucía, Spain) (<https://www.cica.es/servicios/supercomputacion/>), the supercomputing and bioinnovation center service of the University of Malaga (Spain) (<http://www.scbi.uma.es/site/>), and the supercomputing facilities of the Research, Technological Innovation and Supercomputing Center of Extremadura, Spain (<http://www.cenits.es/>).

4.2.5 Development of a hybrid transcriptome

A de novo hybrid transcriptome was also built using both Ion Torrent single-end and Illumina paired-end reads. Considering tested computational requirements and performance in the tests carried out in the de novo hybrid transcriptome, the RAY assembler was selected to carry out the hybrid assembly using raw data from both sequencing platforms, with the parameter k-mer = 31. In addition, we built a partial hybrid transcriptome using a random-selection of half of the Illumina reads, and half of the Ion Torrent reads, with the aim of checking if the good quality of the hybrid transcriptome was only due to the read depth of using two sequencing platforms. The partial hybrid transcriptome, using randomly-selected halves of the Illumina and Ion Torrent reads is designed as partial hybrid transcriptome in the manuscript.

4.2.6 Assembly quality and completeness evaluation

The evaluation of the structure of the generated assemblies from both sequencing platforms was performed using QUAST (version 5.0.0). The QUAST software (Gurevich et al., 2013) generates an overview of the sizes distribution (including largest contig, total length, N50, L50, N75, L75, and GC (%)) of the contigs contained in every *de novo* transcriptome. Moreover, a re-alignment of all the assemblies was carried out to obtain more transcriptome-specific metrics such as E90N50 transcript contig length, DETONATE score values, number of alignable reads and alignments in total using DETONATE (version 1.11) (Li et al., 2014) in each assembly. DETONATE (DE novo TranscriptOme rNa-seq Assembly with or without the Truth Evaluation) evaluates *de novo* transcriptome assemblies by two component packages, RSEM-EVAL and REF-Eval, providing a rigorous computational assessment of the quality of a transcriptome assembly and the best assembly is the one with the highest DETONATE score (<http://deweylab.biostat.wisc.edu/detonate/>). The assembly quality for Illumina assemblies was previously reported in (Guerrero-Sanchez et al., 2017), so it was omitted.

The completeness of all the transcriptomes obtained from Illumina, Ion Torrent and hybrid transcriptome data was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) following the BUSCO v3 user guide (version 3.0.2) using as commands (Simão et al., 2015; Waterhouse et al., 2017):

```
$ Python run_BUSCO.py -i sequence_file -o output_name -l lineage -m tran
$ Python generate_plot.py -wd working_directory
```

. A complete annotation of the *Q. ilex* transcriptome assembled from both Ion Torrent and hybrid transcriptome data (both whole and partial hybrid transcriptomes) was carried out by using the Sma3s v2 annotator (Munoz-Mérida et al., 2014; Casimiro-Soriguer, Muñoz-Mérida, and Pérez-Pulido, 2017).

4.2.7 De novo transcriptome alignment with *Quercus robur* and *Quercus petraea* transcriptomes

All the assemblies obtained in this work were aligned with the most complete and annotated transcriptome sequences of *Q. robur* and *Q. petraea* (<http://www.oakgenome.fr>) (OCV4 transcriptome version), both species being phylogenetically close to holm oak. *Quercus robur* and *Q. petraea* transcriptomes are designated as oak transcriptomes in the manuscript [30]. The alignment software used was blastN (Altschul et al., 1990) with an e-value cutoff of 10^{-30} . Alignment blast outputs were graphically and statistically analyzed using R 3.5.0 and RStudio 1.1.447 (Team, 2018; RStudio, 2016).

4.2.8 Identification of proteins from translated assemblies used as databases

A protein identification using a holm oak peptide spectra sample previously described in (López-Hidalgo et al., 2018) was used in this study. A 6-frame translation for each sequence, in all the transcriptomes generated, was performed using EMBOSS (version 6.6.0) (Rice, Longden, and Bleasby, 2000), filtering and keeping peptides longer than 50 amino acids using the R package Biostrings (version 2.48.0) (Team, 2018; RStudio, 2016; Pages et al., 2009). The resulting FASTA files were used individually as a custom holm oak protein database for the protein identification. Spectra were processed using the SEQUEST algorithm available in Proteome Discoverer 2.1 (Thermo-Scientific, Massachusetts, USA). The following settings were used as previously described in Romero-Rodríguez et al. (Romero-Rodríguez et al., 2014): precursor mass tolerance was set to 10 ppm and fragment ion mass tolerance to 0.8 Da. Only charge states +2 or greater were used. Identification confidence was set to a 5% FDR, the variable modifications were set to: oxidation of methionine, and the fixed modifications were set to carbamidomethyl cysteine formation. A maximum of two missed cleavages were set for all searches.

4.3 Results

4.3.1 Sequencing platforms and de novo assembly structure analysis

To compare the transcriptome features obtained from two different sequencing platforms, equal quantities of total RNA from three tissues, acorn embryos, leaves and roots of holm oak were mixed and used to construct a cDNA library for sequencing based on the Illumina HiSeq2500 and Ion Torrent S5 platforms. A total of 55,275,472 Illumina paired-end reads and 55,161,453 (10 ng of total RNA) and 84,364,256 (50 ng of total RNA) Ion Torrent single-end reads were generated in this study. The raw reads were preprocessed to eliminate primer/adaptor contamination and low-quality section of reads, generating a total of 50,870,724 and 46,334,832 (both RNA concentrations were preprocessed together) clean raw data in Illumina and Ion Torrent, respectively. In each sequencing platform used, the assembly was performed by three different assemblers (MIRA, RAY and TRINITY) and compared to each other (Table 4.1). However, the hybrid assemblies were built using only the RAY assembler, since TRINITY does not

allow the construction of a hybrid assembly and MIRA requires many computational resources when a hybrid assembly is built (Table 4.1).

Table 4.1 Summary of the structure of the holm oak assembly. *Data from the Illumina platform were previously published in (Guerrero-Sanchez et al., 2017).

	Assembly structure							
	Illumina*			Ion Torrent			Hybrid	Hybrid half
	MIRA	RAY	TRINITY	MIRA	RAY	TRINITY	RAY	
# contigs (≥ 0 bp)	169449	107487	77159	710041	107497	303541	132720	104640
# contigs (≥ 500 bp)	43014	20495	8803	22879	18551	118726	26670	21041
# contigs (≥ 1000 bp)	15445	8773	696	5017	5233	49190	13779	11715
# contigs (≥ 5000 bp)	155	73	1	1	4	118	185	173
# contigs (≥ 10000 bp)	2	3	0	0	0	2	9	7
Largest contig	11254	12220	5916	5273	5533	11940	15329	15043
Total length (≥ 0 bp)	83639406	41292773	26286544	145717222	35361128	185129754	56442863	45257060
Total length (≥ 1000 bp)	27409911	14778197	904440	7040671	7467041	79149878	25612168	22023591
Total length (≥ 5000 bp)	941227	471829	5916	5273	21202	710782	1206376	1107152
Total length (≥ 10000 bp)	21731	34168	0	0	0	22544	112633	82952
N50	1211	1260	661	839	930	1206	1558	1630
E90 number of transcripts	127958	65285	64150	584912	66454	224685	71023	63138
E90N50	673	806	361	215	579	946	1122	1188
Score	-2334943804	-3400761031	-6756877372	-6686768444	-5727910347	-4259931488	-7602101330	-1455877920
Number of alignable reads	48681788	39297987	9787481	35784571	27414374	42141854	82372290	22202091
Number of alignments in total	169413628	48341674	15563083	267150454	34964535	631869894	109250751	29749610
N75	742	827	563	628	685	797	972	1042
L50	11473	5863	3428	7718	6149	35219	7174	5731
L75	23813	11529	5931	14324	11404	67779	14209	11200
GC (%)	41.69	42.47	39.14	42.30	42.76	42.04	41.44	42.07

The assembly structure analysis was carried out by the QUAST software, which provided an overview of the number of contigs longer than a concrete base pairs length (from ≥ 0 bp to $\geq 10,000$ bp) (Table 4.1), together with other statistical parameters such as N50, N75, L50, L75 and % GC (Table 4.1). Moreover, the assembly structure analysis was complemented with other transcriptome-specific metrics (E90N50, overall score values, length of alignable reads and number of alignments in total) obtained by using the DETONATE software (Table 4.1). In the case of contigs $\geq 10,000$ bp, both the Illumina and hybrid assemblies resulted in a low number of contigs using MIRA (Illumina, 2 contigs), RAY (Illumina, 3 contigs), TRINITY (Ion Torrent, 2 contigs) and RAY (hybrid assembly, 9 contigs and partial hybrid assembly, 7 contigs). The number of contigs between 1,000 and $\geq 5,000$ bp was much higher in the TRINITY-Ion Torrent assembly (118 and 49,190 contigs, respectively) and the MIRA-Illumina assembly (155 and 15,445 contigs, respectively) than when the other assemblers were used (Table 4.1). The highest number of contigs in holm oak was observed in those contigs between 0 bp and ≥ 500 bp. Both the MIRA-Illumina assembly (169,449 and 43,014 contigs, respectively) and the MIRA-Ion Torrent assembly (710,041 and

22,879, respectively) showed the highest number of these contigs (Table 4.1). The largest contig was constructed by RAY using the hybrid assembly reads (15,329 bp) (Table 4.1). However, from Illumina reads, the largest contig was obtained by RAY (12,220 bp), while from Ion Torrent reads, the largest contig was obtained by TRINITY (11,940 bp) (Table 4.1). The maximum total length of annotated sequences ($\geq 10,000$ bp) was yielded in the RAY hybrid assembly (112,633 bp). Neither the TRINITY (Illumina) assembly nor MIRA and RAY (Ion torrent) assemblies showed sequence lengths higher than 10,000 bp. For $\geq 5,000$ bp total lengths of annotated sequences, RAY hybrid assembly showed more annotated sequences (1,206,376 bp) and for $\geq 1,000$ bp total length of annotated sequences, MIRA-Illumina (27,409,911 bp) and TRINITY-Ion Torrent (79,149,878 bp) assemblies showed more annotated sequences than in the remaining assemblies (Table 4.1). For annotated sequences of a total length of ≥ 0 bp, MIRA-Illumina (83,639,406 bp) and TRINITY-Ion Torrent (185,129,754 bp) assemblies showed the highest number of annotated sequences in holm oak (Table 4.1). The contig N50, in the Ion Torrent platform, was higher in TRINITY (1,206 bp) than in MIRA (930 bp) and RAY (839 bp) and, in the Illumina platform, was practically equal using MIRA (1,260 bp) and RAY (1,211 bp) (Table 4.1). The N50 value was 1,558 bp in the hybrid transcriptome and 1630 bp in the partial hybrid transcriptome (Table 4.1). The GC % content was quite similar in all the assemblers (Table 4.1). In addition, we analysed the transcriptome-specific measurement E90N50 because it is a preferable parameter over the original N50 when evaluating transcriptome assemblies [36]. Both hybrid assemblies (1,122 bp in the hybrid transcriptome and 1,188 bp in the partial hybrid transcriptome) showed the highest E90N50 values in this study, followed by RAY-Illumina (806 bp) and TRINITY-Ion Torrent (946 bp) (Table 4.1). The best DETONATE score values were observed in the partial hybrid transcriptome (-1,455,877,920 bp) and MIRA-Illumina (-2,334,943,804 bp) (Table 4.1). With regard to the number of alignable reads and total alignments, both the hybrid assembly (82,372,290) and TRINITY-Ion Torrent (109,250,751) were higher than the rest of assemblies, respectively (Table 4.1).

The efficiency of the use of resources of each assembler should be considered in a transcriptome analysis; therefore we monitored this for MIRA, TRINITY and RAY in the Illumina, Ion Torrent and hybrid transcriptomes. The MIRA-Illumina assembler used a higher amount of resources, more than 40 central processing units (CPUs) in some points and a mean of 174.80 GB of RAM memory (Figure S2b). The TRINITY-Illumina assembler used many resources during the first minutes of the assembly process, but later, only one core and a mean of 0.55 GB of RAM were used for the final process

of the assembly (Figure S2c). However, this assembler created an immense amount of files. Finally, the RAY-Illumina assembler was the most efficient in the use of resources from the Illumina reads, considering that a mean of 10.73 GB of RAM was used (Figure S2a). In addition, RAY did not generate weighty temporary files, and only used a few MB necessary for the assembly and the logs of the process. The MIRA-Ion Torrent assembler used a mean of 95.85 GB of RAM memory (Figure S3b). The TRINITY-Ion Torrent assembler used, as TRINITY-Illumina, many resources at the beginning of the assembly process, and a mean of 0.90 GB of RAM (Figure S3c). From the Ion Torrent reads, the RAY assembler was also the most convenient in terms of computational resources compared to the other assemblers analyzed in this study (15.61 GB of RAM) (Figure S3a). Regarding the RAY-hybrid assemblers, a mean of 13.55 GB of RAM was used in the hybrid transcriptome assembly and a mean of 15.62 GB of RAM was used in the partial hybrid transcriptome assembly (Figure S4a and Figure S4b).

4.3.2 *Quercus ilex* de novo transcriptome alignment with *Q. robur* and *Q. petraea* transcriptomes

An alignment between the holm oak transcriptome and the *Q. robur* and *Q. petraea* transcriptomes was carried out through a local alignment using blastN with the oak transcriptome as a database and the new assemblies obtained in this work as queries. As a result, a density graph was generated with the length of the oak transcriptome and the *Q. ilex* transcriptome built by all the assemblers used (Figure 4.1). From Illumina reads, MIRA built the best assembly (Figure 4.1a), as previously described (Guerrero-Sanchez et al., 2017). From Ion Torrent, TRINITY-Ion Torrent built the best assembly (Figure 4.1b). The oak transcriptome and *Q. ilex* (MIRA-Illumina) transcriptome showed 82.1% of shared sequences (Figure 4.1c), followed by RAY-Illumina (77.0%) and TRINITY-Illumina (55.1%) (Figure 4.1c). From Ion Torrent reads, MIRA built the best assembly with 84.8% of shared sequences between oak and *Q. ilex* transcriptomes, followed by TRINITY (84.6%) and RAY (74.7%) (Figure 4.1c). The *Q. ilex* hybrid transcriptome and the *Q. ilex* partial hybrid transcriptome showed 82.3% and 78.9% of shared sequences with oak transcriptome, respectively (Figure 4.1c). The distribution of percentage sequence identity between oak and *Q. ilex* (MIRA, RAY and TRINITY) transcriptomes from Illumina, Ion Torrent and hybrid reads was also analyzed (Figure 4.1c). The highest percentage of identity was observed in the RAY-Ion Torrent assembly (96.1%), followed by the RAY-Illumina assembly (95.8%) (Figure 4.1c).

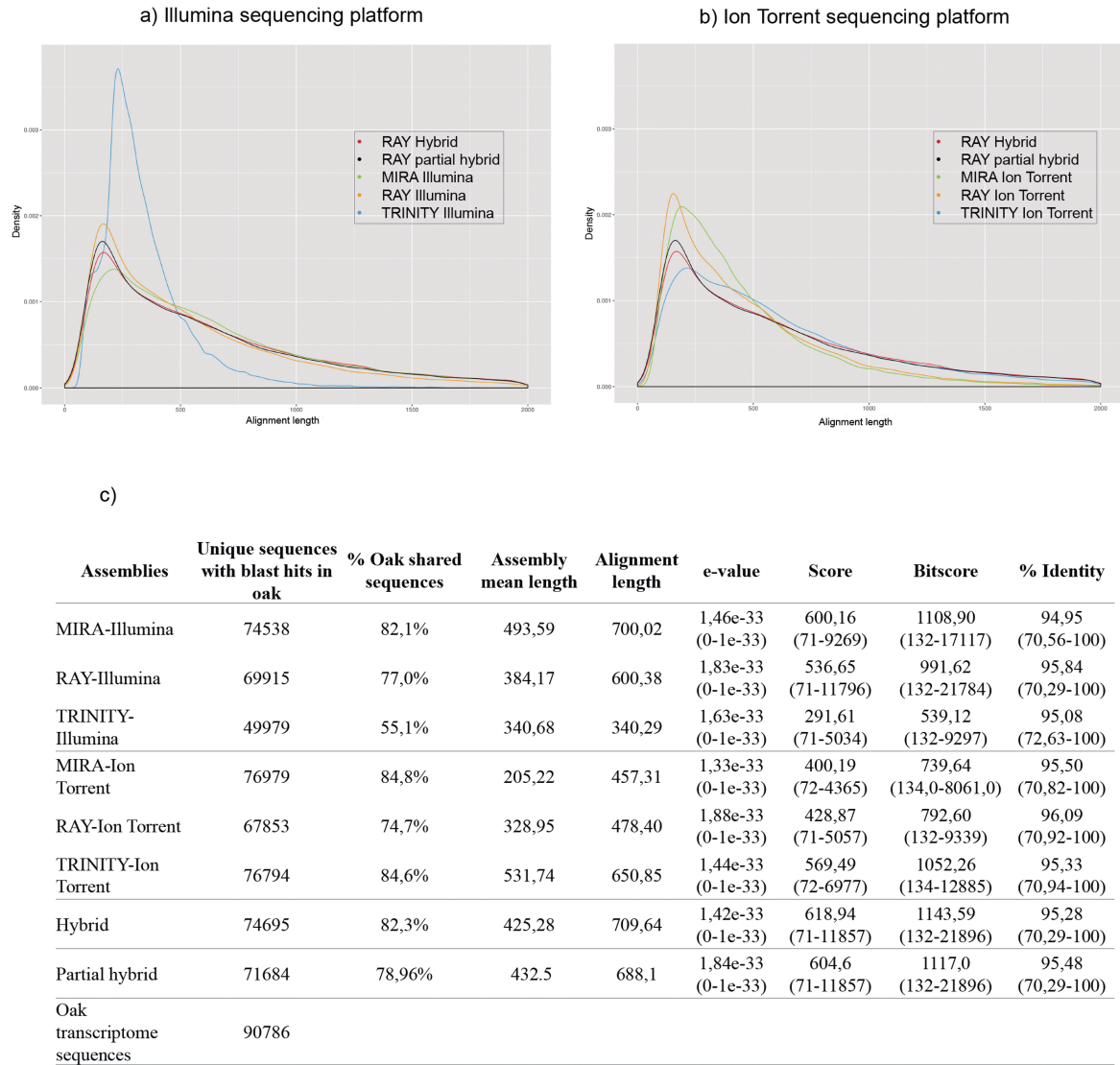


Figure 4.1 Alignment between *Q. robur* and *Q. petraea* transcriptomes (oak transcriptome) and *Q. ilex* (holm oak) transcriptome using MIRA, RAY, TRINITY and RAY hybrid assemblies from Illumina (a) and Ion Torrent (b) reads. Distribution of percent sequence identity between oak and *Q. ilex* (MIRA, RAY, TRINITY, RAY hybrids) transcriptomes (c).

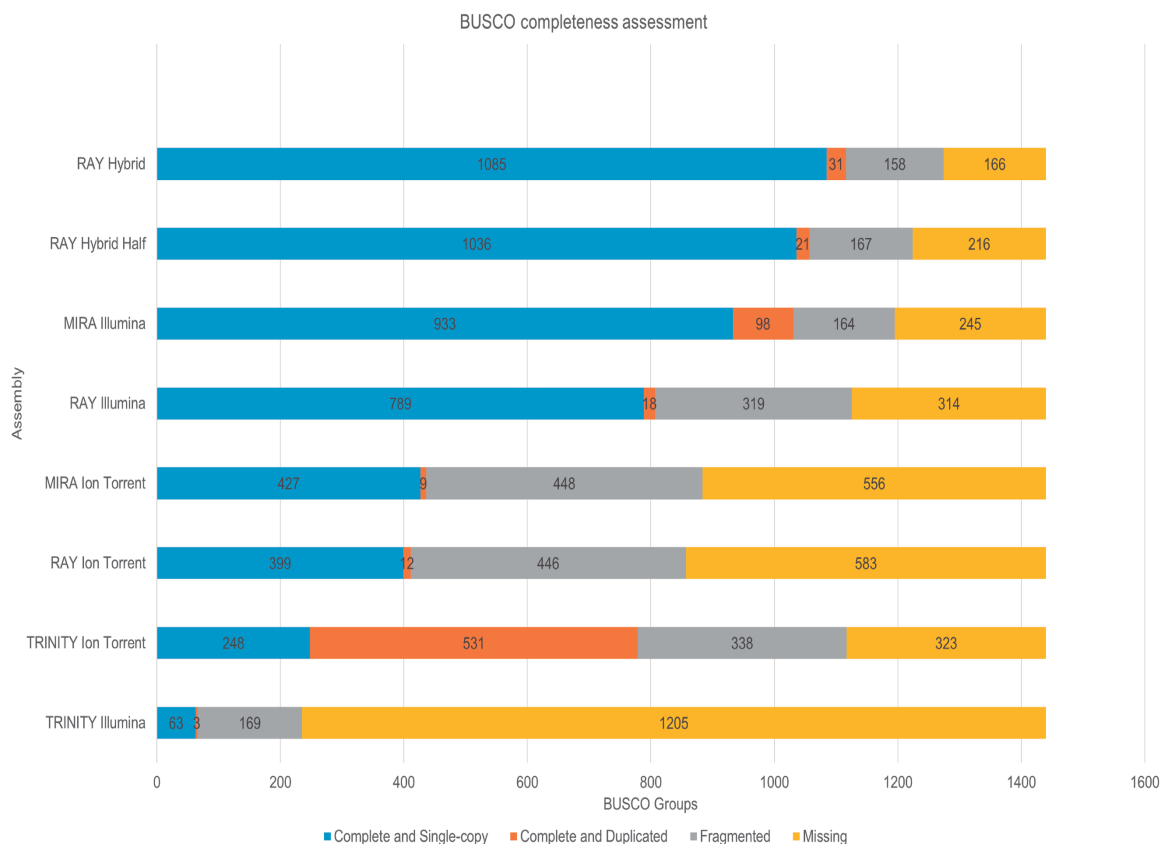


Figure 4.2 Results of BUSCO analysis of the holm oak transcriptome. All the transcriptomes are organized depending on their completeness: RAY-hybrid assembly, RAY-partial hybrid assembly, MIRA-Illumina assembly, RAY-Illumina assembly, MIRA-Ion Torrent assembly; RAY-Ion Torrent assembly; TRINITY-Ion Torrent assembly; and TRINITY-Illumina assembly. Blue: complete and single-copy genes; orange: complete and duplicated genes; grey: fragmented genes; yellow: missing genes.

4.3.3 Transcriptome completeness evaluation

The use of the BUSCO software facilitated an overview of the completeness of the assemblies obtained in this work. In BUSCO, the *embryophyta_odb9* orthologous database for Magnoliophyta plants (flowering plants) has a total of 1,440 BUSCO orthologs groups whose completeness will depend on the assembly of holm oak. According to BUSCO analysis, the RAY hybrid assemblies generated the most complete transcriptomes with 1,116 and 1,057 complete sequences of which 1,085 and 1,036 were single copy sequences in holm oak, respectively (Figure 4.2). From Illumina reads, MIRA (1,031) generated a more complete transcriptome than RAY (807 bp) and TRINITY (66) (Figure 4.2). From Ion Torrent reads, TRINITY generated the most complete transcriptome (779), followed by MIRA (436) and RAY (411) (Figure 4.2).

Annotation of the best *Q. ilex* transcriptome from each sequencing platform

The annotation was performed by the Sma3s v2 algorithm. It is worth mentioning that Blast2GO, rather than Sma3s v2, was previously used in (Guerrero-Sanchez et al., 2017). However, the annotation increased from 31,972 total transcripts annotated by Blast2GO to 62,628 total transcripts recently annotated by Sma3s v2 using the MIRA assembly (López-Hidalgo et al., 2018), both from Illumina reads. From Ion Torrent reads, 74,058 total transcripts were annotated by the TRINITY assembly while from the hybrid transcriptome assembly, 34,360 transcripts were annotated using the RAY assembly. Regarding the partial hybrid transcriptome, around 33,694 transcripts were annotated using the same assembly as for the hybrid transcriptome.

In order to facilitate the access and use of the *Q. ilex* transcriptome sequencing data, the raw data in the FASTQ format was deposited in the Sequence Read Archive (SRA-NCBI) database with accession numbers: SRR7456533 and SRR7454228 (Ion Torrent sequencing platform using 10 ng and 50 ng of total RNA, respectively) and SRR5815058 (Illumina sequencing platform), and the whole transcriptome was uploaded to the holm oak database (<http://www.uco.es/probiveag/holm-oak-database.html>; section 'data').

4.3.4 Gene ontology classification of *Quercus ilex* transcripts

Gene ontology (GO) for the *Q. ilex* transcripts obtained from the hybrid assembly were analyzed by Sma3s v2 to classify the functions of the assembled transcripts in terms of biological process, molecular function and cellular component (Figure 4.3; Table S1). Within the biological processes, more transcripts were assigned to response to stress and biosynthetic processes, followed by anatomical structure development and cellular nitrogen compound metabolic processes (Figure 4.3a; Table S1). In the case of the molecular functions, many transcripts were associated with ion binding, kinase activity, oxidoreductase activity and DNA binding (Figure 4.3b; Table S1). Finally, in the cellular component category, the transcripts were mainly classified in terms of nucleus, plastid and plasma membrane (Figure 4.3c; Table S1). A high number of transcripts (5,405 transcripts) of holm oak were assigned to response to stress (Figure 4.3a; Table S1), of which 46 (0.85%) transcripts were directly included in the drought stress category, according to our annotation (Table S2). Some of the transcripts related to drought stress were: UDP-Glucosyltransferase; TCTP (Translationally Controlled Tumor Protein); NACs (82-77-53-46) transcription factors; DICEP (Drought Inducible Cysteine Proteinase); PUF (Pumilio/Fem-3-binding factor), APUM (Arabidopsis

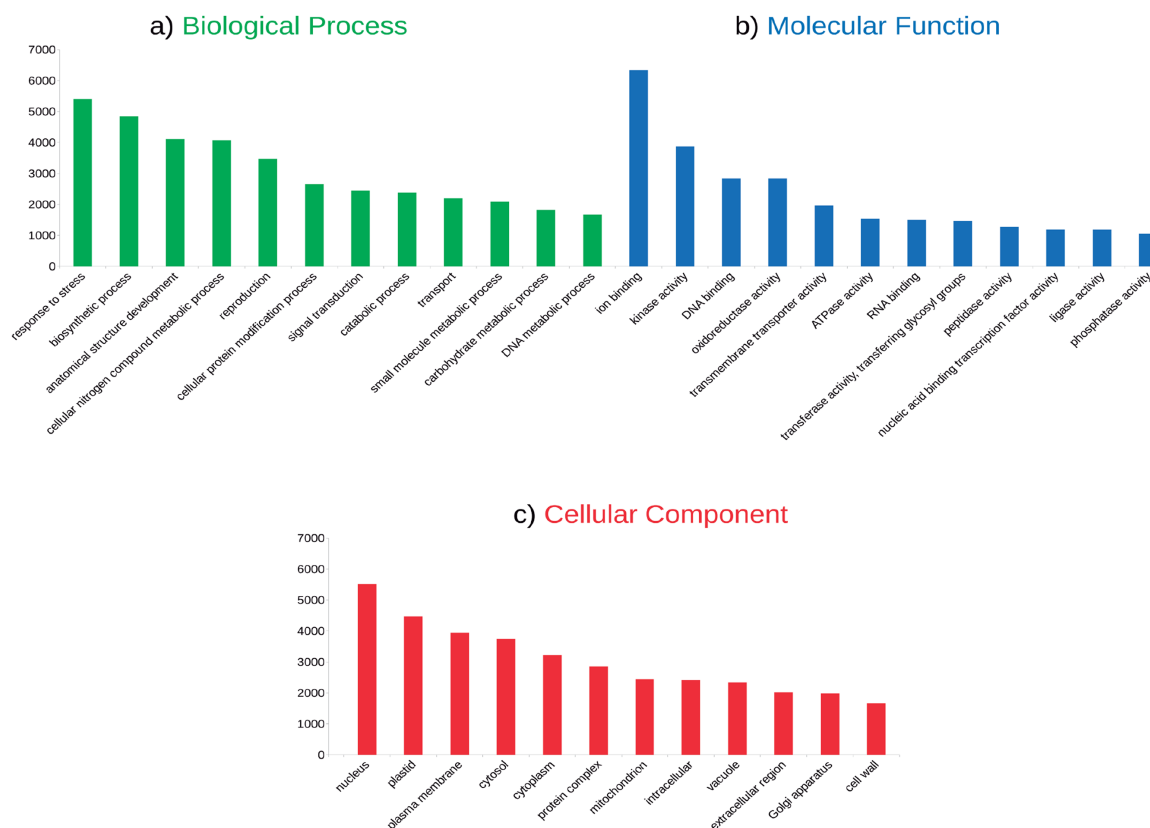


Figure 4.3 Histogram of GO classification of assembled *Quercus ilex* transcripts. Horizontal bar charts of the distribution of GO associated with the holm oak transcripts represented in the three main GO categories: biological processes (a), molecular functions (b) and cellular components (c). The first twelve transcripts assigned to each GO category are shown and the remaining transcripts assigned to each GO category are shown in Table S1.

Pumilio RNA binding protein) and PUM (Pumilio) RNA-binding proteins; PXG4 (Peroxygenase 4); PAL and PAL5 (Phenylalanine Ammonia Lyase); NH2 and NH8 (Nam Line Protein); DRS1 (Drought Sensitive 1 protein); Drought-induced protein RDI; At3g62550-drought responsive ATP-binding motif containing protein; UGT7G3 Anthocyanidin 3-O-glucosyltransferase 2; and TCM_034302 (Chloroplastic drought-induced stress protein) (Table S2).

On the other hand, we also considered a representative sample of 2,000 random transcripts to be classified by GO terms (biological process, molecular function and cellular component) (Table S3). Within the biological processes, more transcripts were assigned to response to stress, biosynthetic processes and anatomical structure development in all the transcriptome assemblies built in this study (Table S3). Within the molecular functions, the majority of transcripts in all the assemblies were grouped into ion binding, DNA binding and kinase activity (Figure 4.3b; Table S1). Finally,

in the cellular component category, the transcripts were mainly classified in terms of nucleus, plastid and cytosol (Figure 4.3c; Table S1).

In addition to the GO classification, the *Q. ilex* transcripts were classified in terms of biological process, pathway and cellular component at the Universal Protein Resource (UniProt). Within the biological processes, more transcripts were assigned to plant defense, followed by transport and transcription (Table S1). Within the pathway category, the majority of transcripts were associated with the response to stress and biosynthetic processes (Table S1). Finally, in the cellular component category, the transcripts were mainly classified in terms of the membrane and nucleus (Table S1).

To further understand the degree of transcript overlap between each of the assemblers-platforms, we created a matrix in which each cell represents the overlap between two assemblers-platforms used in this study (Table 4.2). The highest percentage overlap was observed when TRINITY-Ion Torrent was blasted with MIRA-Ion Torrent (96.58%), followed by MIRA-Illumina blasted with MIRA-Ion Torrent (95.02%) and RAY-Ion Torrent blasted with hybrid assembly (90.36%) (Table 4.2). As expected, the lowest percentage overlaps were observed when all the assemblies were blasted with TRINITY-Ion Torrent, obtaining the lowest overlap between MIRA-Ion Torrent and TRINITY-Illumina (19.51%) (Table 4.2).

Table 4.2 Blast percentage matrix of all the transcriptomes built in holm oak. Each cell in the matrix represents the overlap between two assemblers-platforms.

		RAY		MIRA		RAY		TRINITY	
		Hybrid	Partial Hybrid	Illumina	Ion Torrent	Illumina	Ion Torrent	Illumina	Ion Torrent
RAY	Hybrid	99,94	61,83	50,03	67,25	56,52	54,28	46,89	64,48
	Partial Hybrid	86,95	99,95	63,20	78,65	63,49	60,83	48,48	76,55
MIRA	Illumina	89,96	86,80	99,94	95,02	83,84	72,02	34,45	94,79
	Ion Torrent	71,85	68,62	74,22	99,97	61,34	55,52	19,51	84,27
RAY	Illumina	88,86	72,68	76,16	74,76	99,92	52,24	50,64	75,33
	Ion Torrent	90,36	77,93	64,26	88,56	61,27	99,98	36,24	85,30
TRINITY	Illumina	73,59	57,55	54,25	68,58	59,57	43,06	99,98	66,55
	Ion Torrent	87,15	82,97	86,85	96,58	77,23	73,52	40,70	100,00

4.3.5 Protein annotation in Holm oak

The protein identification carried out with Proteome Discoverer 2.1 by using a translated version of *Q. ilex* transcriptome assemblies gave a successful result (Table 4.3). In terms of total number of proteins from the Illumina translated transcriptome, 1,878, 1,930 and 565 proteins were identified after using the MIRA, RAY and TRINITY assemblers, respectively, while from the Ion Torrent translated transcriptome, 2,242, 2,356 and 2,395 proteins were identified after using the MIRA, RAY and TRINITY

assemblers, respectively. Both hybrid and the partial hybrid assemblies to obtain the holm oak proteome were also carried out in this work, giving rise to a total of 1,899 and 1,801 proteins after using the RAY assembler, respectively (Table 4.3).

Table 4.3 Summary of the total number of proteins annotated in Holm oak.

Protein identification								
	Illumina			Ion Torrent			Hybrid	Partial hybrid
	MIRA	RAY	TRINITY	MIRA	RAY	TRINITY	RAY	
Total	1878	1930	565	2242	2356	2395	1899	1801
Mean length	440,12	277,42	130,14	136,91	164,69	242,00	321,53	351,77
Annotated	1818	1881	547	1972	2303	2373	1841	1753
	-97%	-97%	-97%	-88%	-98%	-99%	-97%	-97%
Unique genes	1492	1508	460	1365	1523	1284	1522	1492
	-82%	-80%	-84%	-69%	-66%	-54%	-83%	-85%
With at least 1 unique peptide	1878	1930	565	2242	2356	2395	1899	1801
	-100%	-100%	-100%	-100%	-100%	-100%	-100%	-100%
With at least 2 unique peptides	995	1111	257	681	1153	1258	1128	1100
	-53%	-58%	-45%	-30%	-49%	-53%	-59%	-61%
With at least 3 unique peptides	620	762	133	244	629	776	795	804
	-33%	-39%	-24%	-11%	-27%	-32%	-42%	-45%
With at least 7 unique peptides	172	192	18	17	67	159	212	251
	-9%	-10%	-3%	-1%	-3%	-6%	-11%	-14%

The total number of annotated proteins was quite similar to the data described in the total number of proteins identified from each translated transcriptome (Table 4.3). A total of 1,818 (97%) (MIRA), 1,881 (97%) (RAY) and 547 (97%) (TRINITY) annotated proteins were identified from the Illumina translated transcriptome, while a total of 1,972 (88%) (MIRA), 2,303 (98%) (RAY) and 2,373 (99%) (TRINITY) annotated proteins were identified from the Ion Torrents translated transcriptome (Table 4.3). The highest number of unique genes (or unique translated protein sequences) was identified in TRINITY-Illumina (84%) and RAY-hybrid (83%) (Table 4.3). The hybrid assembly showed 1,899 proteins, of which 1,841 (97%) were annotated proteins and 1,522 unique genes (83%), and the partial hybrid assembly showed 1,801 proteins, of which 1,753 (97%) were annotated proteins and 1,492 unique genes (85%) (Table 4.3).

4.4 Discussion

In the present work, we evaluate several procedures to build an accurate *de novo* transcriptome for *Q. ilex* from a mixture of experimental raw sequence read data and statistical approaches. An accurate holm oak transcriptome has already been described by this research group (Guerrero-Sanchez et al., 2017), and therefore this present study is now focused on a comparative analysis of two sequencing platforms, Illumina and Ion Torrent, and three different assemblers (TRINITY, MIRA and RAY) used to assemble all the clean raw data obtained in the holm oak transcriptome analysis. Moreover, a *de novo* hybrid transcriptome using both sequencing platforms was built and compared to the transcriptomes obtained through Illumina and Ion Torrent alone. The *de novo* hybrid transcriptome was only assembled using RAY, as mentioned above, as neither the TRINITY nor MIRA assemblers are recommended for the assembly of a hybrid transcriptome using such a large amount of sequences. A *de novo* hybrid assembly is a setting up process of sequences by using two or more sequencing platform data. This kind of assembly was developed due to the limitations of each sequencing platform. The Illumina technology produces low percentage substitution errors (0.3–3.8%) (Dohm et al., 2008; Sleep, Schreiber, and Baumann, 2013), and the Ion Torrent technology presents indels (insertion/deletion error types) at a raw rate of 2.84% (Bragg et al., 2013). By using a hybrid assembly algorithm, we attempted to correct those errors generated in both technologies. This strategy is currently used to correct the elevated rate of errors in third generation sequencing reads (Koren et al., 2012), using high quality short reads from second generation sequencing platforms. Moreover, the use of a partial hybrid transcriptome helped in the estimation of the good quality of the hybrid transcriptome, due mainly to the correction of errors commented above rather than the read depth of using both sequencing platforms. Guerrero-Sanchez et al. (Guerrero-Sanchez et al., 2017) previously annotated 31,972 total transcripts by Blast2GO from Illumina reads assembled by MIRA, which increased the genetic information available at that time in the databases of holm oak (659 sequences on nucleotide database and 88 EST databases annotated by NCBI, (<https://www.ncbi.nlm.nih.gov/>)). The genetic information of holm oak was increased to 62,628 total transcripts annotated using Sma3s v2, rather than Blast2GO (López-Hidalgo et al., 2018), from Illumina reads assembled by MIRA. Additionally, 74,058 and 34,360 total transcripts were obtained in this work using Sma3s v2 from Ion Torrent reads assembled by TRINITY and the hybrid transcriptome assembled by RAY, respectively.

Both sequencing platforms and the assemblers available should be considered carefully, when looking for the best option, especially when there is scarce information

about the species under study, as in holm oak. Bradnam et al. (Bradnam et al., 2013) reported in the Assemblathon 2 context that more than a single assembly or a single metric should be carried out to assess the quality of an assembly. This is due to the read lengths, read counts and error profiles that are produced by different NGS technologies (Bradnam et al., 2013). So, we compared the *de novo* holm oak Illumina transcriptome previously described by (López-Hidalgo et al., 2018) to the *de novo* Ion Torrent transcriptome and *de novo* hybrid transcriptome, with the aim of building a more complete *de novo* Holm oak transcriptome. Moreover, the efficiency in the use of computational resources should be considered in a transcriptome analysis. The assembler should be chosen according to the computational resources required to process the clean raw data, since the computer resources needed represent a clear limitation for performing the assembly. In this study, the RAY assembler proved more convenient in all the transcriptomes built due to the efficient use of computational resources (Figure S2). Regarding the assembly structure, the TRINITY-Ion Torrent assembly annotated a higher number of sequences, while the MIRA-Ion Torrent assembly shared more sequences with *Q. robur* and *Q. petraea* transcriptomes (Figure 4.1).

With regard to completeness assessment, the hybrid transcriptome yielded the most complete sequences in relation to the ortholog alignment, followed by MIRA-Illumina and TRINITY-Ion Torrent assemblies (Figure 4.2). The Ion Torrent assemblies contain more duplicated and fragmented sequences than Illumina and hybrid assemblies (Figure 4.2), which may be due to both the structure of the reads and, single-end in Ion Torrent and paired-end in Illumina. Despite these differences, the Ion Torrent technology gave better assembly structure and protein identification, in addition to being quicker and cheaper than the paired-end sequencing commonly used in the Illumina platform (Lowe et al., 2017). On the other hand, the hybrid transcriptome was used to carry out the GO ontology classification as this transcriptome built the most complete sequence in relation to the ortholog alignment (Figure 4.2), identifying the highest number of unique peptides with more than 3 (Table ??) and being the most efficient in the use of resources during the assembly (Figure S2).

It was remarkable that the higher number of transcripts observed in the GO biological processes was related to the stress response (46 out of 5,405; 0.85%). Conversely, *Q. robur* did not show any stress response related transcripts (Casimiro-Soriguer, Muñoz-Mérida, and Pérez-Pulido, 2017), while they have been observed for other related species such as *Castanea dentata* and *Eucalyptus grandis* (Casimiro-Soriguer, Muñoz-Mérida, and Pérez-Pulido, 2017). The *Q. ilex* transcriptome annotations revealed interesting information about its biology, which can be used in a genetic study devoted

to investigating one of the major problems that threaten this species, drought (Giorgi and Lionello, 2008). A previous study has assessed the effect of the drought in holm oak by a proteomic analysis, reporting a large list of proteins whose levels changed under drought conditions (Simova-Stoilova et al., 2015). Interestingly, in this study, an overview of drought-resistant genes in holm oak is provided from a transcriptomic approach. Although the number of transcripts related to drought stress identified in this work was lower than the number of proteins identified previously (Simova-Stoilova et al., 2015), those transcripts are directly related to drought rather than to general stress response. Nevertheless, all the proteins identified by (Simova-Stoilova et al., 2015) were also identified in our annotations but some of them were not included in the drought stress classification.

Regarding the identification of proteins by Proteome Discovered 2.1, RAY translated assembly from Illumina reads identified more proteins than TRINITY and MIRA, and from Ion Torrent reads, the three assemblers used in this study identified similar numbers of proteins (Table 4.3). However, as a general tendency, all the Ion Torrent translated assemblies showed more proteins than the Illumina assemblies. The hybrid assemblies showed quite similar number of proteins as the Illumina translated assemblies.

4.5 Conclusions

To obtain genetic information in a non-model species, such as Holm oak whose genome has not been yet sequenced, remains a challenge. The comparison between Illumina and Ion Torrent sequencing platforms using different assemblers was carried out to further our knowledge of the *de novo* Holm oak transcriptome previously described (Guerrero-Sanchez et al., 2017; López-Hidalgo et al., 2018). We found that an increase of genetic information could be obtained when the Ion Torrent transcriptome and the hybrid (Illumina and Ion Torrent together) transcriptome were used. This work sheds light on *Q. ilex* biology. Besides, the optimized workflow described here for the Holm oak transcriptome will help to progress on other non-model species (Figure 4.4). The annotated transcripts and proteins could be used to carry out differential expression studies of different biotic and abiotic stresses such as drought or resistance to *Phytophthora cinnamomi*, which seriously affect the biology of holm oak, and as a tool of validation for the whole genome sequencing of Holm oak.

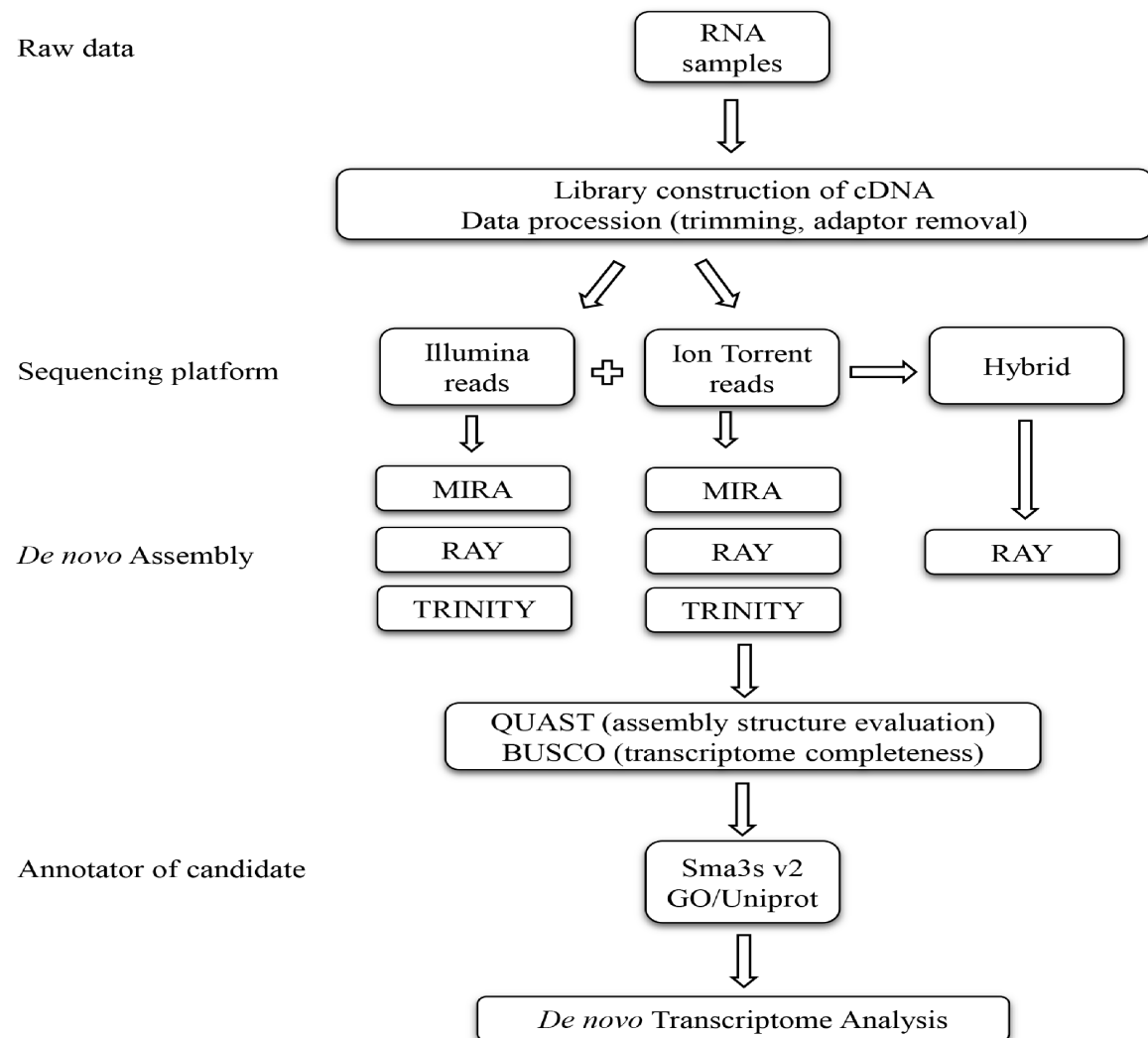


Figure 4.4 Experimental work flow showing the steps carried out and bioinformatic utilities used for a transcriptome analysis.

4.6 Data Availability

In order to facilitate the access and use of the *Q. ilex* transcriptome sequencing data, the raw data in the FASTQ format was deposited in the Sequence Read Archive (SRA-NCBI) database with accession numbers: SRR7456533 and SRR7454228 (Ion Torrent sequencing platform using 10 ng and 50 ng of total RNA, respectively) and SRR5815058 (Illumina sequencing platform), and the whole transcriptome was uploaded to the holm oak database (<http://www.uco.es/probiveag/holm-oak-database.html>; section 'data').

Chapter 5

A multi-omics analysis pipeline for the metabolic pathway reconstruction in the orphan species (*Quercus ilex*)

Chapter published in Frontiers in Plant Science:

López-Hidalgo, C^{*}., Guerrero-Sánchez, V. M^{*}., Gómez-Gálvez, I., Sánchez-Lucas, R., Castillejo-Sánchez, M. A., Maldonado-Alconada, A. M., Valledor, L., and Jorrín-Novo, J. V. (2018). A Multi-Omics Analysis Pipeline for the Metabolic Pathway Reconstruction in the Orphan Species *Quercus ilex*. Frontiers in Plant Science, 9:935.

Abstract

Holm oak (*Quercus ilex*) is the most important and representative species of the Mediterranean forest and of the Spanish agrosilvo-pastoral “dehesa” ecosystem. Despite its environmental and economic interest, Holm oak is an orphan species whose biology is very little known, especially at the molecular level. In order to increase the knowledge on the chemical composition and metabolism of this tree species, the employment of a holistic and multi-omics approach, in the Systems Biology direction would be necessary. However, for orphan and recalcitrant plant species, specific analytical and bioinformatics tools have to be developed in order to obtain adequate quality and data-density before to coping with the study of its biology. By using a plant sample consisting of a pool generated by mixing equal amounts of homogenized tissue from acorn embryo, leaves, and roots, protocols for transcriptome (NGS-Illumina), proteome (shotgun LC-MS/MS), and metabolome (GC- MS) studies have been optimized. These analyses resulted in the identification of around 62629 transcripts, 2380 protein species, and 62 metabolites. Data are compared with those reported for model plant species, whose genome has been sequenced and is well annotated, including *Arabidopsis*, japonica rice, poplar, and eucalyptus. RNA and protein sequencing favored each other, increasing the number and confidence of the proteins identified and correcting erroneous RNA sequences. The integration of the large amount of data reported using bioinformatics tools allows the Holm oak metabolic network to be partially reconstructed: from the 127 metabolic pathways reported in KEGG pathway database, 123 metabolic pathways can be visualized when using the described methodology. They included: carbohydrate and energy metabolism, amino acid metabolism, lipid metabolism, nucleotide metabolism, and biosynthesis of secondary metabolites. The TCA cycle was the pathway most represented with 5 out of 10 metabolites, 6 out of 8 protein enzymes, and 8 out of 8 enzyme transcripts. On the other hand, gaps, missed pathways, included metabolism of terpenoids and polyketides and lipid metabolism. The multi-omics resource generated in this work will set the basis for ongoing and future studies, bringing the Holm oak closer to model species, to obtain a better understanding of the molecular mechanisms underlying phenotypes of interest (productive, tolerant to environmental

cues, nutraceutical value) and to select elite genotypes to be used in restoration and reforestation programs, especially in a future climate change scenario.

5.1 Introduction

Holm oak (*Quercus ilex*) is the most representative species of the Mediterranean forest, of great importance from an environmental and economic point of view (De Rigo and Caudullo, 2016). Being the key element of the Spanish agro-forestry-pastoral ecosystem “Dehesa,” its fruit, the acorn, is the basis of the staple food of the renowned ‘black leg’ pork (Cantos et al., 2003). *Quercus* spp. have been used in the construction of wine barrels, contributing to the organoleptic properties of the maturing wine (Chira and Teissedre, 2014). The use of acorns in human nutrition and for pharmaceutical purposes has a long history. Employed in ancient civilizations, mainly in Italy and Spain, as food or beverage, nowadays it is far from being consumed like other common nuts (Rakić et al., 2006; Al-Rousan et al., 2013; Meijón et al., 2016). As a nutritionally rich product, and because of its high nutraceutical value, the interest of integrating acorns into the human diet or as a functional food has been raised (Vinha et al., 2016b; Hadidi et al., 2017).

Despite its environmental and economic interest, Holm oak is still an orphan species whose biology is almost unknown, especially at the molecular level. Nevertheless, the work of our group and others, has contributed to acquiring the knowledge on this species, focusing on natural variability (Valero-Galván et al., 2011; Akcan et al., 2017), seed germination and seedling growth (Echevarría-Zomeño et al., 2009; Romero-Rodríguez, 2015), physiology (Valero-Galván et al., 2012), and biotic and abiotic stress-responses (Echevarría-Zomeño et al., 2009; Sghaier-Hammami et al., 2013; Sardans et al., 2013; Simova-Stoilova et al., 2015). The above publications, provide fragmented information, mostly derived from classical biochemical approaches and, to a much lesser extent, those of proteomics (Valero-Galván et al., 2011; Romero-Rodríguez et al., 2014; Romero-Rodríguez, 2015) transcriptomics (Guerrero-Sanchez et al., 2017), or metabolomics (Rakić et al., 2006; Rabhi et al., 2016; Vinha et al., 2016b; López-Hidalgo, 2017), but lacking a validation and effective integration of the different molecular multilevels.

In spite of their difficulty as orphan, recalcitrant plant species, forest trees, like other experimental plant systems, deserve to be considered at the wide system level, that implicates the use of multidisciplinary approaches, from visual phenotype, to molecular -omics, through physiological and biochemical approaches (Correia et al.,

2016; Meijón et al., 2016; Escandón et al., 2017). Systems Biology approaches require the optimization of protocols for both wet and in silico analysis.

In this direction, trying to fill this gap with the use of the available high-throughput -omics, its combination and also the implementation of required methodology, we hoped to gain knowledge on the chemical composition and metabolism of the *Q. ilex* tree species, its variability among and within populations, the effect on endogenous ones and their environmental factors, and the search for molecular markers to select elite genotypes. The lack of information available in public databases on the Holm oak genome, transcriptome (Guerrero-Sanchez et al., 2017), or proteome (Romero-Rodríguez et al., 2014) and the absence of standardized laboratory and analytical protocols make this approach a real challenge.

In this work, we employed a wide range of *in silico* techniques allowing a system biology approach for a non-sequenced species. To obtain the maximum level of biochemical complexity the plant sample employed were multi-organ pools, generated by mixing equal amounts of homogenized tissue from acorn embryo, leaves, and roots. In setting up protocols for transcriptome (NGS-Illumina), proteome (shotgun LC-MS/MS) and metabolome (GC-MS) analysis, and bioinformatic pipelines for annotating transcripts, proteins and metabolites, the Holm oak metabolic pathways were partially reconstructed. This research constitutes the basis for ongoing and future studies to obtain a better understanding of the molecular bases underlying phenotypes of interest (productive, tolerant to environmental cues, nutraceutical value) and the selection of elite genotypes to be used in restoration and reforestation programs, especially in the current climate change scenario. In order to reveal the particularities of the species under study, data have been compared with those reported for model plant species, including *Arabidopsis*, rice, poplar, and eucalyptus.

5.2 Materials and methods

5.2.1 Plant material

Mature acorns from Holm oak (*Quercus ilex* L. subsp. *ballota* [Desf.] Samp.) were collected on December 2015 from a tree located in Aldea de Cuenca (province of Córdoba, Andalusia, Spain). Acorns were transported to the lab, sterilized, and germinated as previously reported (Simova-Stoilova et al., 2015). Germinated seeds were sown in pots (500 mL) with perlite and grown in a greenhouse under natural conditions for 4 months up to the 10-leaves stage. Plants were periodically watered at

field capacity and once a week with a Hoagland nutrient (Hoagland and Arnon, 1950) solution after the second month. Germinated embryos, cotyledons, leaves, and roots were collected separately, washed with distilled water and frozen in liquid nitrogen. Then, each tissue was separately homogenized in a mortar until a fine powder was obtained and finally stored at -80°C . The experiments were performed with a pool of fresh weight equivalents of the homogenized tissue from acorn embryo, cotyledons, leaves, and roots. Depending on the organ, samples from individual trees or plantlets in number of 18 (roots and leaves) to 50 (seed embryos and cotyledons) were collected and mixed. Three independent extractions were performed and only consistent proteins or metabolites, those present in the three replicates, were considered.

5.2.2 Transcriptomics Analysis

RNA Extraction and Sequencing

Total RNA was extracted from the frozen homogenized pool tissue following the procedure previously reported by (Guerrero-Sanchez et al., 2017). 50 mg pooled fresh tissue according the procedures previously set up in our laboratory for *Quercus ilex* samples was employed (Echevarría-Zomeño et al., 2012). Contaminating genomic DNA was removed by DNase I (Ambion) treatment. Total RNA was quantified spectrophotometrically (DU 228800 Spectrophotometer, Beckman Coulter, TrayCell Hellma GmbH & Co., KG). The high quality and integrity of the RNA preparation were tested electrophoretically (Agilent 2100 Bioanalyzer). Only high-quality RNAs with RIN values >8 and A260:A280 ratios near 2.0 were used for subsequent experiments.

The library construction of cDNA molecules was carried out using Illumina TruSeq Stranded mRNA Library Preparation Kit according to the manufacturer's instructions using 2 μg of total RNA followed by poly-A mRNA enrichment using streptavidin coated magnetic beads and thermal mRNA fragmentation. The cDNA was synthesized, followed by a chemical fragmentation (DNA library) and sequenced in the Illumina HiSeq 2500 platform, using 100 bp paired-end sequencing (De Wit et al., 2012).

Data processing

The raw reads obtained from the sequencing platform were preprocessed to retain only high-quality sequences to be subsequently used in the assembly. Each original sequence was quality trimmed considering several parameters (quality trimming based on minimum quality scores, ambiguity trimming to trim off, for example, stretches of Ns, base trim to remove specified number of bases at either 3' or 5' end of the

reads). The processed reads were assembled *de novo* using the assembly software MIRA 4.9.6 (Chevreux, Wetter, and Suhai, 1999). Redundancy reduction of the assembled sequenced was carried out by using the CD-HIT 4.6 clustering algorithm (Li, Jaroszewski, and Godzik, 2001; Li, Jaroszewski, and Godzik, 2002).

Gene Ontology

Assembled sequences were blasted against UniRef90 (UniProt) using the software Sma3s (Casimiro-Soriguer, Muñoz-Mérida, and Pérez-Pulido, 2017) in order to obtain the annotated sequences with the most probable gene name and protein description, EC numbers for enzymes, GO terms, and UniProt keywords and pathways. In addition, their functions were identified using MERCATOR (<http://www.plabipd.de/portal/mercator-sequence-annotation/>).

5.2.3 Proteomics Analysis

Protein Extraction and Digestion

Proteins were extracted from the frozen homogenized pool tissue by using the TCA-acetone-phenol protocol as reported in (Jorri n-Novo, 2014). Protein extracts [600–1000 ng BSA equivalents quantified with Bradford assay (Bradford, 1976)] were subjected to Orbitrap analysis after SDS–PAGE (12%) prefractionation. Electrophoresis was stopped when the sample entered the resolving gel, so that a unique protein band was revealed after Coomassie staining (Pascual et al., 2017).

Protein bands were manually excised, destained, and digested with trypsin Sequencing grade (Roche) as is described in (Castillejo, Bani, and Rubiales, 2015) with minor modifications. Briefly, gel plugs were destained by incubation (twice for 30 min) with a solution containing 100 mM ammonium bicarbonate (AmBic)/50% acetonitrile (AcN) at 37°C. Then, they were dehydrated with AcN and incubated in 100 mM AmBic containing first 20 mM DTT for 30 min, and then in the same solution containing 55 mM Iodoacetamide instead DTT for 30 min. They were washed with 25 mM AmBic and 25 mM AmBic/50% AcN two times each. After dehydration in AcN, the trypsin at a concentration of 12.5 ng/ μ l was added in a buffer containing 25 mM NH₄HCO₃, 10% AcN and 5 mM CaCl₂, and the digestion proceeded at 37 C for 12 h. Digestion was stopped, and peptides were extracted from gel plugs by adding 10 μ L of 1% (v/v) trifluoroacetic acid (TFA) and incubating for 15 min.

Shotgun LC-MS Analysis

Nano-LC was performed in a Dionex Ultimate 3000 nano UPLC (Thermo Scientific) with a C18 75 $\mu\text{m} \times 50$ Acclaim Pepmap column (Thermo Scientific). The peptide mix was previously loaded on a 300 $\mu\text{m} \times 5$ mm Acclaim Pepmap precolumn (Thermo Scientific) in 2% AcN/0.05% TFA for 5 min at 5 $\mu\text{L}/\text{min}$. Peptide separation was performed at 40°C for all runs. Mobile phase buffer A was composed of water, 0.1% formic acid. Mobile phase B was composed of 80% AcN, 0.1% formic acid. Samples were separated during a 60-min gradient ranging from 96% solvent A to 90% solvent B and a flow rate of 300 nL/min.

Eluted peptides were converted into gas-phase ions by nano electrospray ionization and analyzed on a Thermo Orbitrap Fusion (Q-OT-qIT, Thermo Scientific) mass spectrometer operated in positive mode. Survey scans of peptide precursors from 400 to 1500 m/z were performed at 120K resolution (at 200 m/z) with a 4×10^5 ion count target. Tandem MS was performed by isolation at 1.2 Da with the quadrupole, CID fragmentation with normalized collision energy of 35, and rapid scan MS analysis in the ion trap. The AGC ion count target was set to 2×10^3 and the maximum injection time was 300 ms. Only those precursors with charge state 2–5 were sampled for MS2. The dynamic exclusion duration was set to 15 s with a 10 ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on. The instrument was run in top 30 mode with 3 s cycles, meaning that the instrument would continuously perform MS2 events until a maximum of top 30 non-excluded precursors or 3 s, whichever was shorter.

Protein identification

Spectra were processed using the SEQUEST algorithm available in Proteome Discoverer Pacific Bios 1.4 (Thermo Scientific, United States). The following settings (Romero-Rodríguez et al., 2014) were used: precursor mass tolerance was set to 10 ppm and fragment ion mass tolerance to 0.8 Da. Only charge states + 2 or greater were used. Identification confidence was set to a 5% FDR and the variable modifications were set to: oxidation of methionine and the fixed modifications were set to carbamidomethyl cysteine formation. A maximum of two missed cleavages were set for all searches. The protein identification, was carried out against the annotated *Q. ilex* transcriptome, previously described. A six-frame translation for each sequence in the transcriptome was performed by using EMBOSS (Rice, Longden, and Bleasby, 2000), filtering and keeping peptides longer than 50 amino acids. Considering the identified proteins, the

protein peak areas were normalized and missing values corrected. Mean values and standard deviation (SD), as well as the coefficient of variation (CV) of the peak areas of protein species were determined for three independent analysis (Supplementary Table S11). The remaining sequences were used as a database for the protein identifications and their functions were identified using MERCATOR (Lohse et al., 2014).

5.2.4 Metabolomics analysis

Metabolite extraction

Metabolites were extracted from plant tissue as described by (Valledor et al., 2014), with three independent extractions. A buffer containing 600 μL of cold methanol: chloroform: water (5:2:2) was added to 15 mg of frozen tissue, vortexed (10 s), and the mixture sonicated (ultrasonic bath, 40 kHz for 10 min). After centrifugation (4°C , 4 min, $20,000 \times g$) the supernatant was transferred to new tubes containing 400 μL of cold chloroform: water (1:1). For phase separation, the tubes were centrifuged (4°C , 4 min, $20,000 \times g$). The upper (polar) and the lower (apolar) phases were re-extracted with 200 μL of cold chloroform (upper) and water (lower), respectively. After combining on one hand the water: methanol (upper) and, on the other the chloroform (lower) phases, they were vacuum dried at 25°C (Speedvac, Eppendorf Vacuum Concentrator Plus/5301).

GC-MS Analysis

GC-MS analysis was performed as reported (Furuhashi et al., 2012) and (Meijón et al., 2016) with some modifications. Polar (water: methanol dissolved) metabolites were derivatized by re-suspending the dried extract in 20 μL of anhydrous pyridine containing 40 mg/mL of methoxyamine hydrochloride. The mixture was incubated at 30°C for 30 min under agitation. Next, 60 μL of N-methyl-N-trimethylsilyl trifluoroacetamide (MSTFA) was added, samples incubated at 60°C for 30 min, centrifuged (3 min, $20,000 \times g$), and cooled to room temperature. Then, 80 μL of the supernatant was transferred to GC-microvials. Apolar (chloroform solubilized) metabolites were methylesterified with 295 μL tert-methyl-Butyl-Ether (MTBE), and 5 μL of trimethylsulfonium hydroxide solution (TMSH) for 30 min at room temperature. The tubes were centrifuged (3 min, $20,000 \times g$) to remove insoluble particles before transferring the supernatants to GC-microvials.

Polar metabolites were resolved and analyzed with a Gas Chromatograph/Mass Spectrometer Agilent 5975B GC/MSD. Inlet temperature was set at 230°C . Samples

were injected in discrete randomized blocks with a 1.2 mL/min flow rate. GC separation was performed splitless on a HP-5MS capillary column (30 m \times 0.25 mm \times 0.25 mm) (Agilent 19091J-433) over a 70–76°C gradient at 0.75°C/min, 76–180°C gradient at 6°C/min, 180–200°C gradient at 3.5°C/min, and then to 310°C at 6°C/min. The mass spectrometer operated in electron-impact (EI) mode at 70 eV in a scan range of m/z 40–800. For apolar metabolites a different temperature gradient was employed: 80–190°C at 8°C/min, 190–220°C at 5°C/min, and then to 270°C at 5°C/min. The mass spectrometer was operated in EI mode at 70 eV in a scan range of m/z 40–600.

Metabolite identification

Metabolites were 'tentatively assigned' based on GC retention times (RT) and m/z values (Supplementary Tables S4, S5) through searches in different databases, including the Gölm Metabolome Database (Nielsen and Jewett, 2007), Alkane, Fiehn library 1 y 2 (Kind et al., 2010), GC-TSQ, MoSys, and NIST/EPA/NIH Mass Spectral Library. Three different softwares were used for metabolite identification: MZmine 2 (2.24 version) (Pluskal et al., 2010), AMDIS software (2.66 version), and NIST.MS Search (2.01 version). Mean values and SD, as well as the CV of the peak areas of metabolites were determined for three independent extraction (Supplementary Table S5). Moreover, the metabolites were annotated using the KEGG compound reference database⁶. Metabolomics pathways of each metabolite (Supplementary Table S6) were searched against KEGG pathway maps. For other general biological networks, we employed MapMan (3.5.1 version).

5.2.5 Interspecies comparison

The annotated *Quercus ilex* transcriptome was compared against the complete *in silico* proteomes of *Arabidopsis thaliana* (UP0000065489, *Oryza sativa* subsp. *japonica* (UP00005968010), *Populus trichocarpa* (UP00000672911), and *Eucalyptus grandis* (UP00003071112) in order to elucidate the unique and shared sequences. This comparison was performed by using BLAST(<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) with blastX alignment with an e-value of 10^{-10} . Also, the EC numbers of each proteome were contrasted to achieve a complete picture of the metabolic pathways coverage differences among proteomes studied in previously mentioned species (Supplementary Table S7). For the comparison, we represented a Venn diagram plotted using VennDiagram R package (Chen and Boutros, 2011).

5.2.6 Integrated Pathway

By using MERCATOR web application (Supplementary Tables S8, S9) (Lohse et al., 2014), we could assign MapMan “Bins” to arbitrary transcript or protein input sequences (Usadel et al., 2009). The output was a text file mapping each input (proteins or transcripts) identifier to one or more Bins by searching a variety of reference databases (TAIR Release 10, SwissProt/UniProt Plant Proteins, Clusters of Orthologous Eukaryotic Genes Database (KOG), Conserved Domain Database (CDD), and InterProScan). The functional predictions generated could directly be used as a ‘mapping file’ for the high-throughput data visualization and meta-analysis software MapMan (3.5.1 version). The ImageAnnotator module allowed us to visualize the data on a gene-by-gene basis on schematic diagrams (maps) of the biological processes described.

5.3 Results and Discussion

This paper reports the study and view of the metabolism as it occurs in Holm oak, the most representative and valuable forest tree species in the Mediterranean region. For that purpose, a biological sample containing equal fresh weight amount of the different organs as starting plant material and a combination of high-throughput, -omics approaches (transcriptomics, proteomics, and metabolomics) as analytical tools were used. As each analytical platform has its own limitations (Schrimpe-Rutledge et al., 2016; Tian, Lam, and Shui, 2016; Viant et al., 2017), is their integration that will provided more confident biological knowledge of them.

The Systems Biology approach for research with species that, like Holm oak are orphan and recalcitrant is very challenging (Abril et al., 2011), and it required the optimization of experimental protocols and, more limitative, the creation of custom-made databases, and pipelines. Beyond the reconstruction of different metabolic pathways as they may occur in Holm oak, and the comparison with model plant species (*A. thaliana*, *O. sativa* subsp. *japonica*, *P. trichocarpa*, and *E. grandis*) we aimed to prove that employing state-of-the-art instrumentation and a similar workflow to those employed in model species is feasible, even though quite uncommon in the current literature.

Transcriptome Analysis

The first transcriptome of *Q. ilex* has recently been reported. For that reason, the Illumina Hiseq 2500 platform was employed to analyze the tissue mix sample, resulting in 119889 contigs, and 31973 Blast2GO annotated transcripts (Guerrero-Sanchez et al., 2017). The number of annotated sequences have been increased to 62628 after a UniRef90 database search through Sma3s software (Munoz-Mérida et al., 2014; Casimiro-Soriguer, Muñoz-Mérida, and Pérez-Pulido, 2017). Among them, 27089 sequences corresponded to unique genes. Comparatively, Sma3s performed faster than Blast2GO and allowed more elaborated results, including functional categories, such as biological processes, cellular components or molecular functions (Supplementary Figures S5, S6 and S7). The total transcriptome sequences were categorized in 35 MERCATOR functional plant categories. The result of this categorization showed a high percentage (41.8%) of non-assigned transcripts (Figure 5.2). Response to stress and biosynthetic process, and the nucleus and plastids, were, respectively, the biological processes and organelles most represented (Supplementary Figures S5, S6). With respect to molecular functions, ion binding and kinase activity were those most abundant, with around 11225 and 6372 sequences, respectively (Supplementary Figure S7).

The number of annotated transcripts, 62628, is double that previously found for the close relative *Q. robur* (38292 sequences; (Tarkka et al., 2013)), similar to the figure of 27655 protein-coding genes in *Arabidopsis* (35386 identified proteins; Araport11(<https://www.arabidopsis.org/>)), and below the 82190 unique transcripts corresponding to 34212 genes also reported in *Arabidopsis* by (Zhang et al., 2017).

The annotated sequences in *Q. ilex* transcriptome were compared with the *in silico* proteomes of *A. thaliana*, *O. sativa* subsp. *japonica*, *P. trichocarpa*, and *E. grandis* (UniProt) to elucidate the unique and shared sequences. The comparative results are shown in Supplementary Table S5. The highest percentage of similarity corresponded to *P. trichocarpa* (91.7%), and lowest to *O. sativa* subsp. *japonica* (77.8%), with intermediate values for *E. grandis* (88.5%) and *A. thaliana* (85.6%). The percentage of similarity correlated with the phylogenetic distances among the compared species as reported by The Angiosperm Phylogeny Group III (2009) (Figure 5.2).

Among the annotated transcripts, 2103 corresponded to enzyme transcript products. These enzymes were assigned to 123 KEGG metabolic pathways (Supplementary Table S6). The most represented pathways (Table 5.2) were: the carbohydrate metabolism (starch and sucrose metabolism and glycolysis/gluconeogenesis, with 26 and 30 enzyme transcripts, respectively). Also, the amino acids metabolism, primarily the cysteine and

methionine metabolism, where 37 enzyme transcripts were detected. This pathway has an important role in plants. Cysteine constitutes the sulfur donor for the biosynthesis of methionine, phytochelatins, sulfhydryl compounds, glutathione, and coenzymes. The homeostasis of sulfur metabolism in trees is more robust than in herbaceous plants. Also, a greater change in conditions to initiate a response in trees is required (Rennenberg et al., 2007). This fact is coherent with the requirement for highly flexible defense strategies in woody plant species because of longevity. In addition, the lipid metabolism (glycerophospholipid metabolism with 32 enzyme transcripts) has an important function as a mediator in hormone signal transduction in plants (Janda et al., 2013).

Proteome Analysis

The protein profile of the *Q. ilex* tissue mix sample was analyzed using a shotgun proteomics platform. Protein extracts were obtained by using a TCA-acetone-phenol protocol. After trypsin digestion, peptides were subjected to UPLC-Q-OT-qIT MS. The resulting peptides and corresponding proteins were identified by matching MS and MS/MS m/z data against the protein database resulting from the six-frame translation of the *Q. ilex* transcriptome. The employment of species specific databases instead of generic Viridiplantae ones improved the number and confidence of the identifications, as previously published (Romero-Rodríguez et al., 2014). By using Viridiplantae (SwissProt), 891 proteins were identified. Nevertheless, with our custom-built specific database, 58584 peptides were detected corresponding to 2830 proteins (with at least one unique peptide (Supplementary Tables S10, S11). Mean, SD, and CV (%) values of normalized identified protein peak areas were determined for three replicates (Supplementary Table S11). The mean of the CV obtained was 36.75% (Supplementary Table S11), which was slightly higher than the CV mean previously described using a 2-DE gel analysis (28.9%) (Jorge et al., 2005, 2006). This is due to the number proteins, considering that this number is much lower in a 2-DE gel analysis and usually highly represented than in a shotgun LC-MS/MS. However, despite having a slightly higher value of CV, the shotgun LC-MS/MS shows greater sensitivity and wide dynamic range. Proteins were categorized in 34 MERCATOR functional plant categories (Figure 5.1). 21.2% of the proteins was not assigned to a functional plant category. Up-to 18.1% proteins were related to protein fate (assembly, folding, degradation, and protein posttranslational modifications), this group being the one most represented.

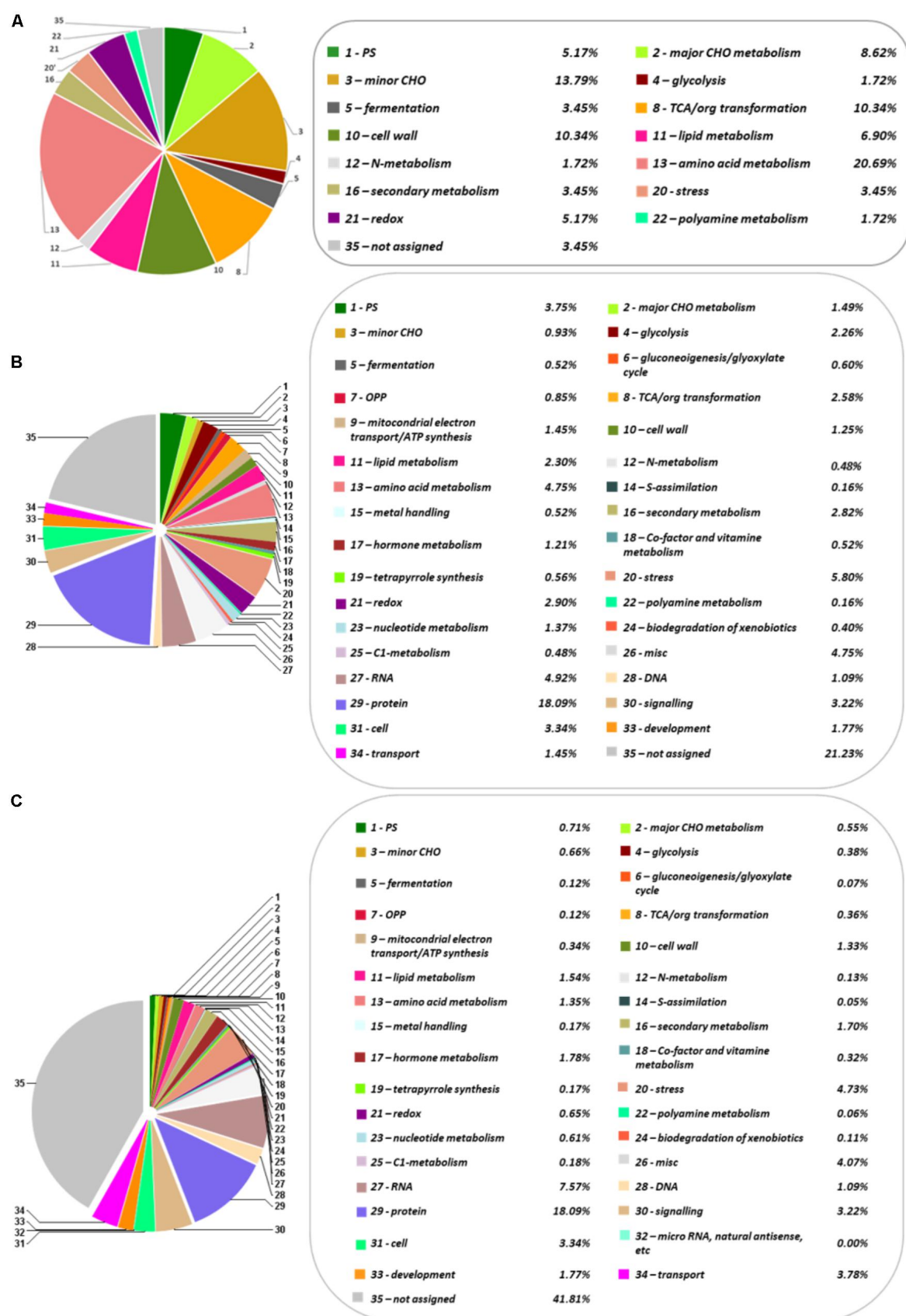


Figure 5.1 Functional categorization and distribution in percentage of the identified metabolites, proteins and transcripts, according to the categories established by MERCATOR. (A) Metabolome. (B) Transcriptome. (C) Proteome. The pie charts show different functional categories: PS (Photosynthesis), major CHO metabolism, minor CHO metabolism, glycolysis, fermentation, gluconeogenesis/glyoxylate cycle, OPP (Oxidative Pentose Phosphate), TCA/org transformation, mitochondrial electron transport/ATP synthesis, cell wall, lipid metabolism, N-metabolism, amino acid metabolism, S-assimilation, metal handling, secondary metabolism, hormone metabolism, co-factor and vitamin metabolism, tetrapyrrole synthesis, stress, redox, polyamine metabolism, nucleotide metabolism, biodegradation of xenobiotics, C1-metabolism, miscellanea, RNA, DNA, protein, signaling, cell, micro RNA, natural antisense, etc., development, transport, and not assigned.

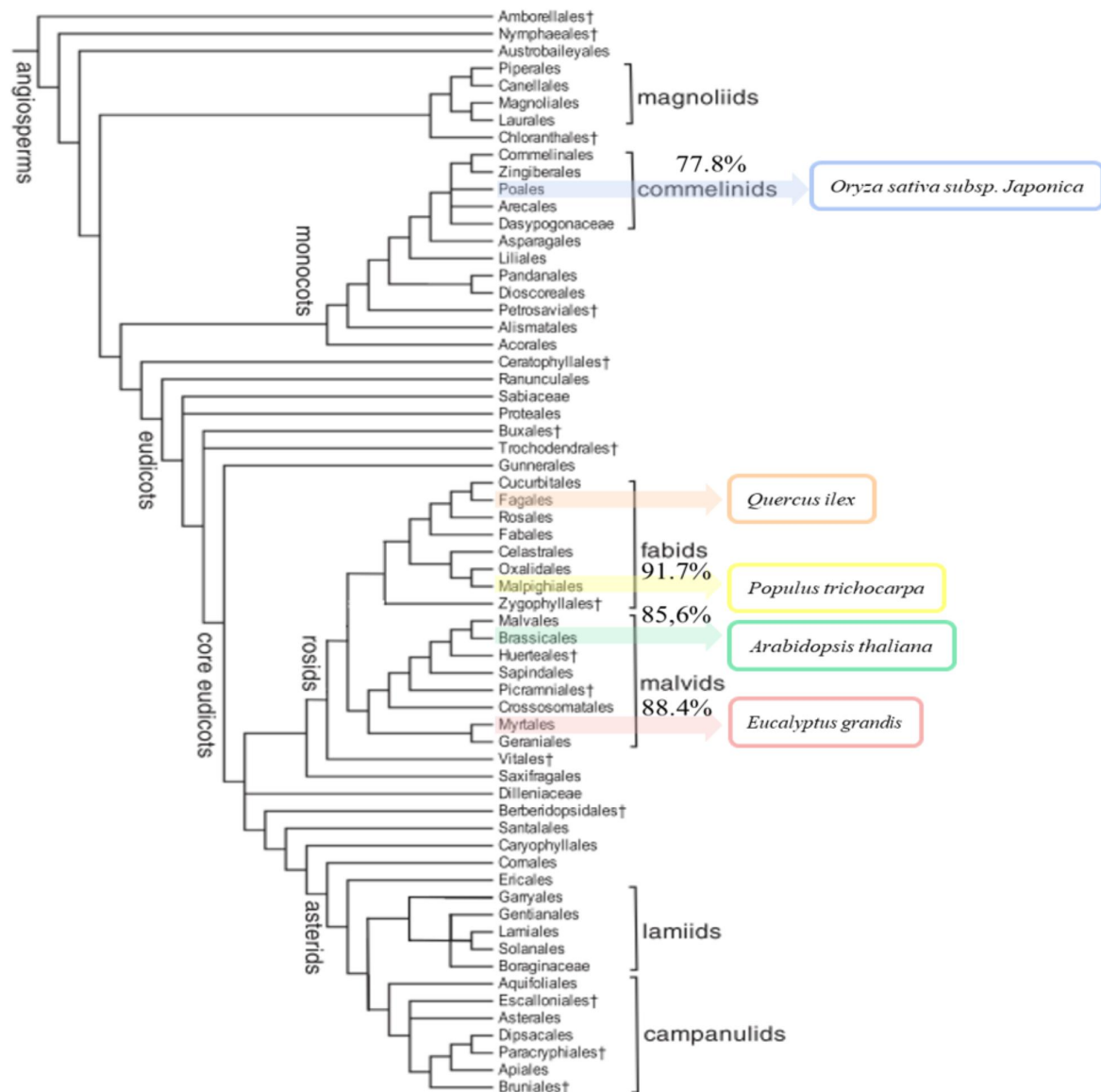


Figure 5.2 Phylogenetic tree of angiosperms. The tree shows the five-species compared (*Arabidopsis thaliana*, *Eucalyptus grandis*, *Oryza sativa* subsp. *japonica*, *Populus trichocarpa*, and *Quercus ilex*). The sequence similarity of species coincides with the classification in the phylogenetic tree. Species are ranked from highest to lowest similar to *Q. ilex*: *P. trichocarpa* (91.7%), *E. grandis* (88.4%), *A. thaliana* (85.6%), and *O. sativa* subsp. *japonica* (77.8%).

The Holm oak proteome was filtered manually looking for proteins corresponding to enzymes based on the EC number. This resulted in 228 enzyme proteins, corresponding to 10% of the protein species with EC deduced from the in silico predicted Holm oak transcriptome (2103 enzyme proteins) and around 20–50% of the enzymes predicted for the sequenced *A. thaliana* and *O. sativa* subsp. *japonica* systems at UniProt.

The proteins identified were assigned to 93 KEGG metabolic pathways (Supplementary Table S6). The most represented pathways were: the carbohydrate metabolism (starch and sucrose metabolism and glycolysis/gluconeogenesis) and the amino acids metabolism (Table 5.2). The least represented one was the enzymes related to transcription (Supplementary Table S6). These figures are much higher than those previously reported for *Q. ilex* and other forest tree species (Valero-Galván et al., 2012; Pascual et al., 2017; Szuba and Lorenc-Plucińska, 2017), maybe due to the use of the powerful LTQ-Orbitrap mass instrument (Kalli et al., 2013) and the search in custom-built specific database.

Out of the 228 enzyme proteins identified, 23 were specific for Holm oak, and 202, 157, 88, and 87, shared with, respectively, *A. thaliana*, *O. sativa* subsp. *japonica*, *P. trichocarpa*, and *E. grandis* (Figure 5.2). 84 enzymes were common to all the species, and 471, and 35 specific for *A. thaliana* and *O. sativa* subsp. *japonica*. It is worth noting that, for *P. trichocarpa* and *E. grandis* no unique enzymes were found, this proving the quality and validity of our data, with, consequently, a more complete annotated transcriptome and proteome database. Holm oak unique enzymes were related to the biosynthesis of hormones and secondary metabolites. They included those involved in the zeatin biosynthetic pathway (ath00908), such as cis-zeatin O-beta-D-glucosyltransferase (EC:2.4.1.215) and zeatin O-beta-D-xylosyltransferase (EC:2.4.2.40). Zeatin, one of the growth promoting hormones, is the predominant xylem-mobile cytokinin in many plant species (Kamboj et al., 1999). In the Holm oak unique enzymes involved in the secondary metabolism [6'-deoxychalcone synthase (EC:2.3.1.170) and prenylcysteine oxidase (EC:1.8.3.5)] were involved in flavonoid biosynthesis and terpenoid backbone biosynthesis, respectively. This is not surprising as secondary metabolites are species specific. Thus, in Holm oak, the flavonoids epicatechin gallate and epigallocatechin were found (Vinha et al., 2016a).

The 84 enzyme proteins common to the five-species corresponded mostly to pathways of the central metabolism, such as those of starch and sucrose (e.g., sucrose synthase, EC: 2.4.1.13, and glucose-6-phosphate isomerase, EC: 5.3.1.9), glycolysis and gluconeogenesis [e.g., phosphoglycerate kinase (EC:2.7.2.3) and pyruvate kinase.

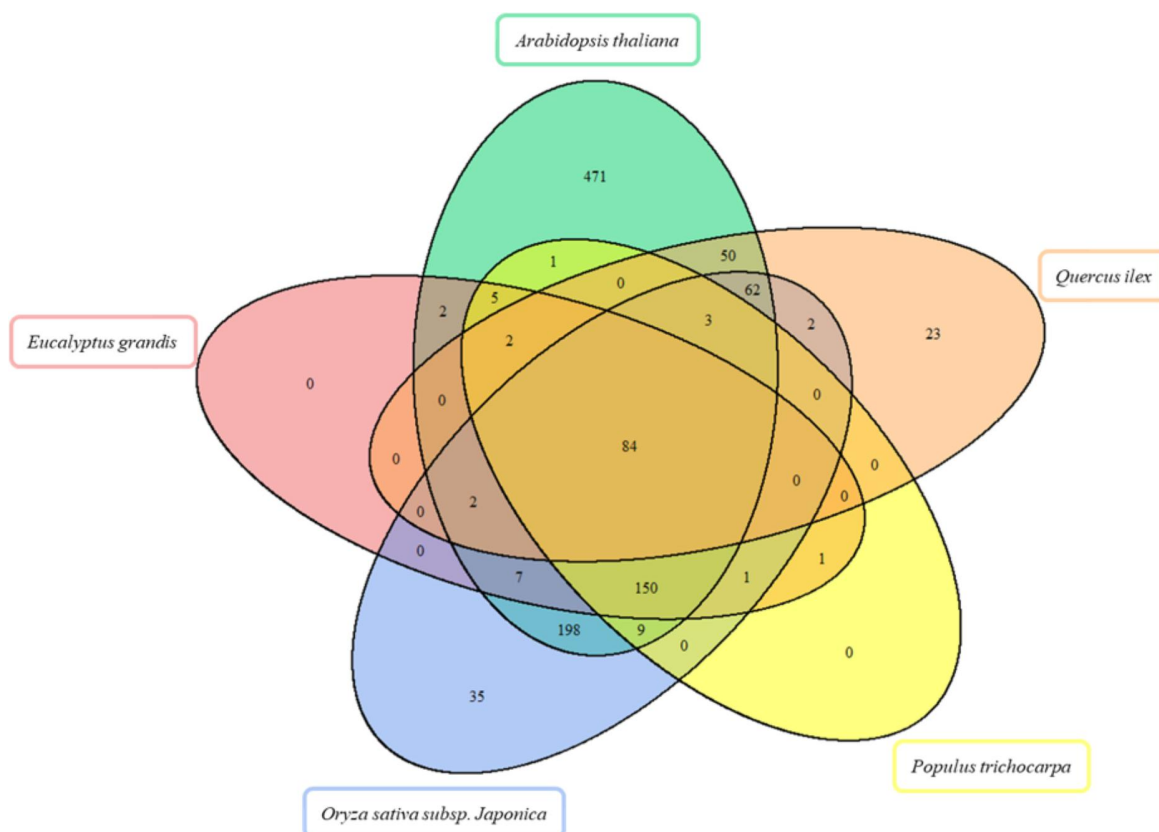


Figure 5.3 Venn diagram for the comparison of enzymes in *Arabidopsis thaliana*, *Eucalyptus grandis*, *Oryza sativa* subsp. *japonica*, *Populus trichocarpa* *in silico* proteomes, and *Quercus ilex* proteome. The Venn diagram shows the overlap of enzymes detected.

(EC:2.7.1.40)], and citrate cycle [e.g., malate dehydrogenase (EC:1.1.1.37), pyruvate dehydrogenase (EC:1.2.4.1), and aconitate hydratase (EC:4.2.1.3)].

The 228 enzyme proteins identified belonged to 109 pathways, with some of them being represented by only one enzyme [e.g., caffeine metabolism (ath00232) and arachidonic acid metabolism (ath00590)] and up to 20 enzymes [e.g., carbon fixation in photosynthetic organisms (ath00710)]. Analysis of the UniProt *in silico* enzyme proteome revealed 106 and 107 pathways for, respectively, *P. trichocarpa* and *E. grandis*, with the figure being higher for *A. thaliana* (121 pathways) and *O. sativa* subsp. *japonica* (112 pathways) (Supplementary Table S11).

The pathways most represented in Holm oak were those of the intermediate and central metabolism, including glyoxylate and dicarboxylate metabolism (ath00630) with 16 enzyme proteins and amino sugar and nucleotide sugar metabolism (ath00520) with 12 enzyme proteins (Table 5.2). For the glycolysis (Supplementary Figure S8), just as an example, there were only two enzyme proteins non-detected: phosphofructokinase (EC:2.7.1.11) and phosphoglycerate mutase (EC:5.4.2.12) (Supplementary Table S12).

Table 5.1 Metabolite families from GC-MS data of *Quercus ilex*. Six main chemical families of metabolites are represented. Carbohydrates (19), organic acids (19), amino acids (11), fatty acids (4), polyols (2), phenolic compounds (2) and four unique compound classes (others). Data in the brackets are KEGG compound identifier of each metabolite.

Nature of the compounds	Metabolite name
Amino acids	L-Glutamate (C00025), L-aspartate (C00049), L-alanine (C00041), L-asparagine (C00152), L-serine (C00065), L-threonine (C00188), L-leucine (C00123), L-valine (C00183), L-isoleucine (C00407), L-proline (C00148), L-phenylalanine (C00079)
Organic acids	Ascorbate (C00072), pyruvate (C00022), L-lactate (C00186), succinate (C00042), fumarate (C00122), malate (C00149), citrate (C00158), aconitate (C00417), gluconolactone (C00198), D-glycerate (C00258), glucarate (C00818), galactarate (C00879), maleate (C01384), salicylate (C00805), pyroglutamic acid (C01879), oxalate (C00209), gallate (C00627), quinate (C00296), D-ribonate (C01685)
Carbohydrates	D-Glucose (C00031), L-arabinose (C00259), D-xylulose (C00310), D-galacturonate (C00333), D-fructose (C00095), L-sorbose (C00247), mannitol (C00392), L-rhamnose (C00507), D-sorbitol (C00794), sucrose (C00089), D-galactose (C00124), melibiose (C05402), myo-inositol (C00137), D-glucose 6-phosphate (C00092), maltose (C00208), maltotriose (C01835), D-cellobiose (C00185), D-galactonate (C00880), D-erythrose (C01796)
Polyols	Glycerol (C00116), viburnitol (C08259)
Fatty acids	Palmitic acid (C00249), oleic acid (C00712), stearic acid (C01530), linoleic acid (C01595)
Phenolic compounds (flavonoids)	Catechin (C06562), epigallocatechin (C12136)
Others	Urea (C00086), 4-aminobutanoate (GABA) (C00334), tridecane (C13834), anthraquinone (C16207)

Table 5.2 Number of metabolites and enzymes (proteomic and transcriptomic level) in KEGG pathways. Pathways according to the KEGG pathway maps based on *Arabidopsis thaliana*. The *Arabidopsis* pathway identifiers are in brackets. The table shows the most representative pathways. The complete list of pathways is in the Supplementary Material Table S12.

Pathways	Metabolites		Proteins	Transcripts
Carbohydrate metabolism				
Glycolysis/gluconeogenesis (<i>ath00010</i>)	Pyruvate, D-glucose, L-lactate	3	20	30
Glyoxylate and dicarboxylate metabolism (<i>ath00630</i>)	Pyruvate, L-glutamate, succinate, L-serine, malate, citrate, glycolate, oxalate, glycerate, aconitate	10	16	27
Citrate cycle (TCA cycle) (<i>ath00020</i>)	Pyruvate, succinate, fumarate, malate, citrate, aconitate	6	9	16
Amino sugar and nucleotide sugar metabolism (<i>ath00520</i>)	D-Glucose, L-arabinose, D-galacturonate	3	12	38
Starch and sucrose metabolism (<i>ath00500</i>)	D-Glucose, sucrose, D-glucose 6-phosphate, D-fructose, cellobiose, maltose	6	18	26
Pentose phosphate pathway (<i>ath00030</i>)	Pyruvate, D-glucose, gluconolactone, glycerate	4	7	17
Galactose metabolism (<i>ath00052</i>)	D-Glucose, sucrose, D-fructose, glycerol, D-galactose, myo-inositol, D-sorbitol, D-galactonate, melibiose	9	9	15
Amino acid metabolism				
Alanine, aspartate, and glutamate metabolism (<i>ath00250</i>)	Pyruvate, L-glutamate, L-alanine, succinate, L-aspartate, fumarate, L-asparagine, citrate, 4-aminobutanoate (GABA)	9	9	27
Cysteine and methionine metabolism (<i>ath00270</i>)	Pyruvate, L-alanine, L-aspartate, L-serine	4	10	37
Glycine, serine, and threonine metabolism (<i>ath00260</i>)	Pyruvate, L-aspartate, L-serine, L-threonine, glycerate	5	11	31
Phenylalanine metabolism (<i>ath00360</i>)	Pyruvate, succinate, L-phenylalanine, fumarate, salicylate	5	4	14
Lipid metabolism				
Biosynthesis of unsaturated fatty acids (<i>ath01040</i>)	Palmitic acid, oleic acid, stearic acid, linoleic acid	4	3	13
Energy metabolism				
Carbon fixation in photosynthetic organisms (<i>ath00710</i>)	Pyruvate, L-alanine, L-aspartate, malate	4	20	23
Biosynthesis of other secondary metabolites				
Phenylpropanoid biosynthesis (<i>ath00940</i>)	L-Phenylalanine, gallate	2	7	17

These results are more complete than the ones found from the *in silico* analysis of the other two woody plants used for comparisons *P. trichocarpa* and *E. grandis*, with only 5 out of the 10 glycolytic enzymes.

Metabolome Analysis

The metabolites present in the pooled samples were analyzed by using GC-q-MS. Two different extraction solvents, methanol:water and chloroform, were, respectively, used for compounds of different polarities. Up to 155 and 19 peaks were resolved by gas chromatography using the above mentioned solvents. A complete list of the identified compounds with their respective RT and the mass-to-charge ratios (m/z) is included in Supplementary Tables S4, S5. From the m/z values, and after a search in seven public databases (Alkane, Fiehn library 1 and 2, Gölm Metabolome Database, GC-TSQ, MoSys, and NIST/EPA/NIH Mass Spectral Library) a total of 62 compounds were identified, 57 in the methanol:water extract and 5 in the chloroform one. The normalized peak areas of the metabolites were employed for the mean, SD, and CV determinations. The average of the CV obtained (13.70%) was lower than the obtained with proteins data (36.75%), revealing the existence of a greater variability in proteins analysis. The higher CV could be related with the higher number and diversity of identified proteins versus the metabolites identified.

Identified compounds were in the 60–500 Da and mostly belonged to the primary metabolism (59), with only three being secondary metabolites (catechin, epigallocatechin, and anthraquinone). The identified metabolites were grouped in six chemical families according to the KEGG database¹⁷, including carbohydrates (19), organic acids (19), amino acids (11), fatty acids (4), polyols (2), and phenolic compounds (2) (Table 5.1). The family most represented was that of organic acids (19) and carbohydrates (19), followed by amino acids (11). Fatty acids (4) and phenolic compounds (2) were much less represented. They were included in at least 64 different KEGG pathways (Supplementary Table S6), and in 15 functional plant categories according to MapMan classification (Figure 5.1).

These metabolites are starting metabolites or final products from primary metabolism pathways, like glyoxylate and dicarboxylate metabolism (ath00630), starch and sucrose metabolism (ath00500), citrate cycle (TCA cycle) (ath00020) of carbohydrate metabolism; alanine, aspartate, and glutamate metabolism (ath00250) of amino acid metabolism and biosynthesis of unsaturated fatty acids (ath01040) of fatty acids metabolism. Many were intermediate metabolites, with 5 (citrate, cis-Aconitate, succinate, fumarate, and malate), out of the total 8 corresponding to the Citrate

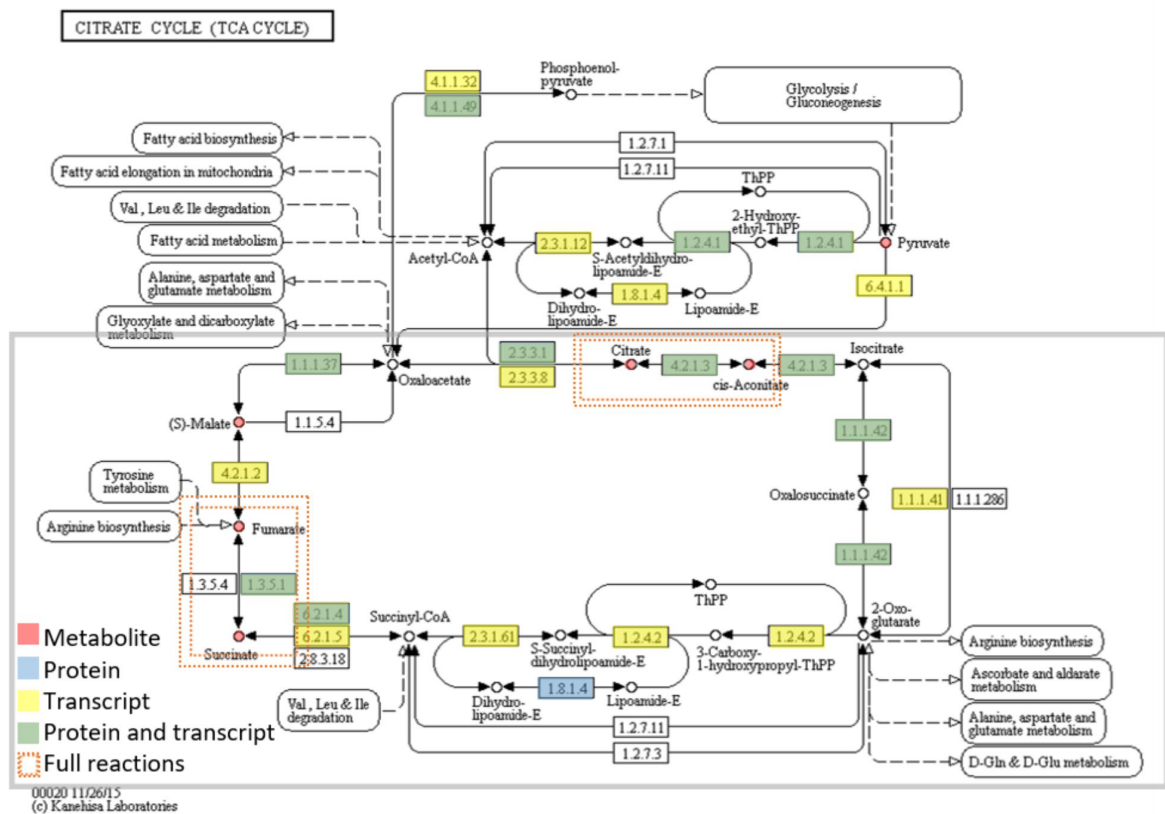


Figure 5.4 Metabolites and enzymes (protein or transcript level) assigned to the citrate cycle (TCA cycle). Omics data are highlighted in red (metabolites), blue (proteins), yellow (transcripts), and green (both proteins and transcript). The enzymes (proteins and transcripts) are named by their EC number. EC numbers and respective detected TCA cycle enzymes: 2.3.3.1 (Citrate synthase), 4.2.1.3 (Aconitate hydratase), 1.1.1.42 [Isocitrate dehydrogenase (NADP+)], 1.2.4.2 (alpha-ketoglutarate dehydrogenase), 6.2.1.4 (Succinyl coenzyme A synthetase), 1.3.5.1 (Succinate dehydrogenase), 4.2.1.2 (Fumarate hydratase), 1.1.1.37 (Malate dehydrogenase). There are two full reactions (metabolite, protein and transcript level) This figure was adapted from KEGG reference pathway.

cycle (Figure 5.4 and Table 5.2). The pathways most represented were carbohydrate and amino acid metabolisms. However, the number of secondary metabolites (catechin, epigallocatechin, and anthraquinone) was smaller than the number of secondary metabolites reported for *Quercus* spp. acorns (Vinha et al., 2016b). Due to the small number of secondary metabolites detected, the metabolic pathways related to the biosynthesis of secondary metabolites, like carotenoid biosynthesis (ath00906), anthocyanin biosynthesis (ath00942), and monoterpene biosynthesis (ath00902) are not highly represented (Supplementary Table S6). In *Arabidopsis*, the total number of secondary metabolites is still unknown due to metabolite identification being one of the bottlenecks in untargeted metabolomic studies (Wu et al., 2017). Still, in AraCyc 15.0, the total number of compounds described are 2971 and the number of metabolic pathways 610 (PMN; Plant Metabolic Network). The identification of 62 metabolites is in the order of what has been reported for non-model plant systems by using a similar

approach (Warren, Aranda, and Cano, 2012; Cadahía et al., 2015; Asai, Matsukawa, and Kajiyama, 2016; Pascual et al., 2017), but far from the figure obtained when using model systems such as *A. thaliana*, or complementary techniques such as LC-MS. The employment of complementary LC-MS strategies would increase the number of metabolites identified, as shown, for example, with *A. thaliana*, although it would greatly reduce the number of metabolites identified with no doubts. (Kim, Langmead, and Salzberg, 2015) detected 4483 distinct metabolite peaks from leaves using 11 mass spectrometric platforms, but only identifying 1348 metabolites. These results revealed that the available databases and repositories are incomplete and pointed to the need for new algorithms for elucidating structures from MSⁿ analyses.

Data Integration

To seek insights into the metabolic pathways as they occur in Holm oak, transcriptomics, proteomics, and metabolomics data have been integrated. (Table 5.1 and Supplementary Tables S11, S13). We obtained a deeper view of the metabolic pathways by implementing proteomics or transcriptomics data as the potential of these techniques is much higher than that of metabolomics. However, although technological advances and bioinformatic tools and resources for making those analyses and data interpretation have been extended to plant biology research, this has mostly been for model plants. The unique and specialized biology of such diversified species requires the adaptation of strategies conceived primarily for model organisms and the development of designed and specific methods. For their integration, we employed EC numbers (proteins and transcripts) and KEGG identifiers (metabolites). With the latter and with KEGG pathway maps we obtained the three-different level of information of 61 metabolic pathways (Supplementary Table S6). The metabolic pathways most represented are shown in Table 5.2.

In order to obtain a metabolic overview. The 'BINS' generated from the proteome/transcriptome were employed as a "mapping file," then introducing identified metabolites. The representation obtained of the general map (Figure 5.5A) for the dataset as shown from ImageAnnotator module of MapMan, showed common metabolism points between metabolites and proteins/transcripts (Figure 5.5B). From the total number of pathways reported in the plants, for example, in KEGG (127 pathways in *Arabidopsis*), we procured data from 124 of them at the metabolomic, proteomic, and transcriptomics level (Supplementary Figure S7). Table 5.2 summarizes the most representative pathways visualized, including carbohydrate metabolism [glycolysis/gluconeogenesis (ath00010), glyoxylate and dicarboxylate metabolism (ath00630),

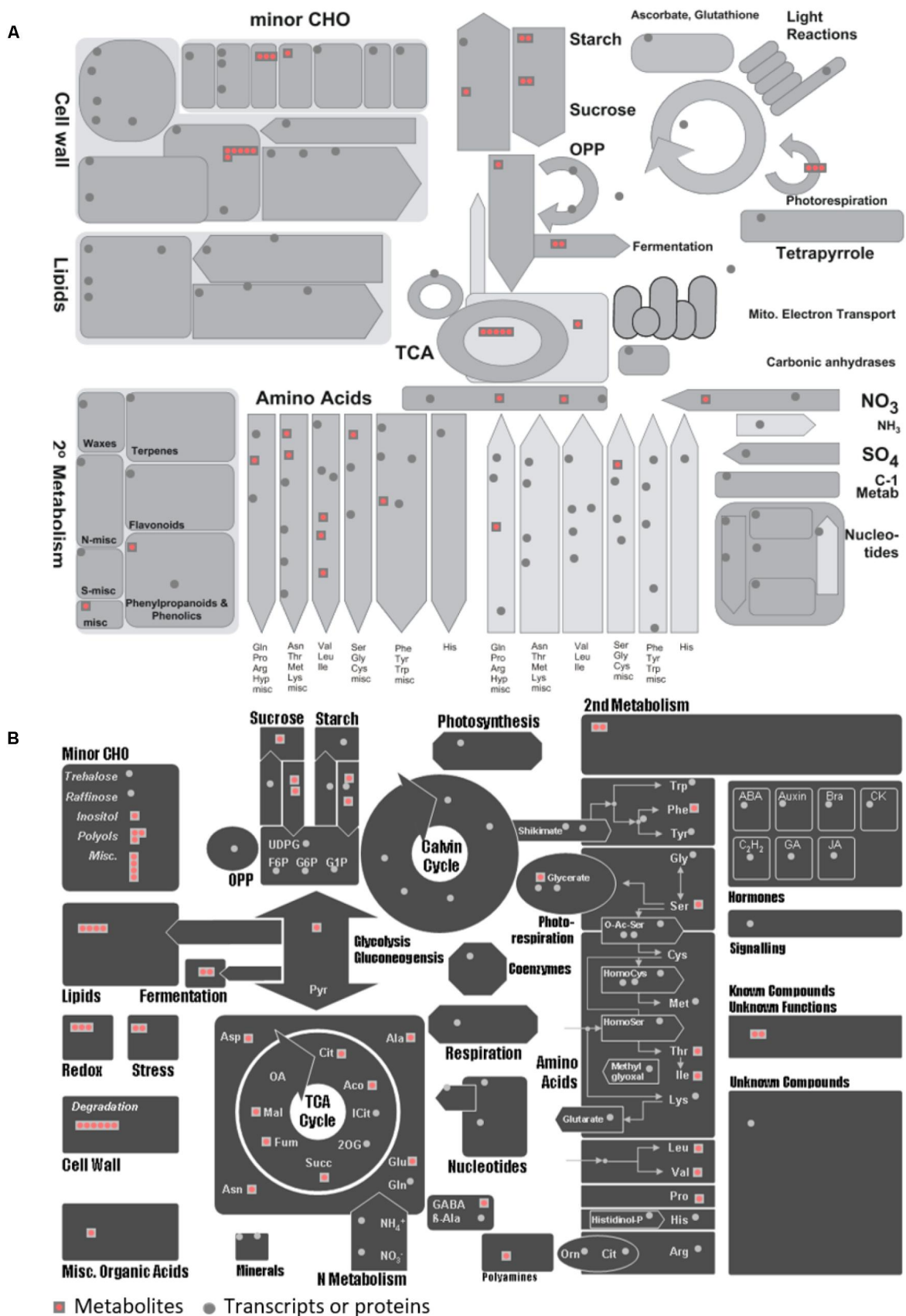


Figure 5.5 MapMan overview of general metabolism for the metabolites and proteins/transcripts of *Quercus ilex*. (A) Visualization of 58 metabolites in the context of general metabolism using the MapMan software. (B) Visualization of 58 metabolites in different MapMan pathways. Each red square represents a metabolite and each gray circle represents a protein or transcript. More details can be found in (Usadel et al., 2009)

citrate cycle (TCA cycle) (ath00020), starch and sucrose metabolism (ath00500)], amino acid metabolism [alanine, aspartate, and glutamate metabolism (ath00250) and phenylalanine metabolism (ath00360), lipid metabolism (biosynthesis of unsaturated fatty acids (ath01040)], and energy metabolism [carbon fixation in photosynthetic organisms (ath00710)]. The one most represented was the TCA, with 5 metabolites out of a total of 10, and protein and transcript corresponding to, respectively, 6 and 8 enzymes (Figure 5.2). On the other hand, there were clear gaps in the hypothetical plant metabolic chart, mainly corresponding to the secondary metabolism and hormones [anthocyanin biosynthesis (ath00942), brassinosteroid biosynthesis (ath00905)] and lipid metabolism [steroid biosynthesis (ath00100)]. For example, the brassinosteroid biosynthesis pathway, which produces plant steroidal hormones that play important roles in many stages of plant growth, has only reported 1 protein and 1 transcript (Supplementary Table S7). Also, the Figure 5.3 shows the low representation of the different metabolic pathways, also with a multi-omics data integration. From metabolomics, proteomics, and transcriptomic data we were able to identify 64, 109, and 118, pathways, respectively. The total number reported at the PMN and deduced from genome sequencing were 610 (*A. thaliana*), 519 (*E. grandis*), and 538 (*P. trichocarpa*). From these figures we can conclude that the current wet methodologies only allow the visualization of a low percentage of enzyme gene products in a single experiment.

The work and dataset generated, even considering future methodological improvements, will be the basis of ulterior studies on the particularities of the metabolism as it occurs in different organs and developmental processes, as well the changes in response to environmental cues, thus complementing our previous studies in which morphology, phenology, classical physiological and biochemical analysis, and the holistic proteomics have been employed (Echevarría-Zomeño et al., 2009; Echevarría-Zomeño et al., 2012; Valero-Galván et al., 2011; Valero-Galván et al., 2012; Sghaier-Hammami et al., 2013; Romero-Rodríguez et al., 2014; Romero-Rodríguez, 2015; Guerrero-Sanchez et al., 2017). These previously published studies provided quite fragmented and speculative biological information. Hence to go one step ahead, data validation and integration at the different molecular levels would be necessary in order to obtain an unbiased molecular interpretation of the plant biology.

5.4 Conclusions

We have proven that –omics integration, in the Systems Biology direction, is feasible not only with model organisms, but also with orphan and recalcitrant species such as the Holm oak, the most emblematic and representative tree species of the Mediterranean forest. The methodological bases, including wet protocols and *in silico* analysis, have been established, allowing the implementation of transcriptome, proteome, and metabolome databases, comprising 27089 transcripts (unigenes), 2380 protein species, and 62 metabolites (Supplementary Table S14).

Integrated analysis allowed the visualization and reconstruction of the metabolism in Holm oak. Up to 123 metabolic pathways, out of the 127-total reported in KEGG, can be visualized at the transcriptome, proteome, and metabolome level. Thus, as an example, for the Krebs cycle, six metabolites out of the eight have been detected. This route comprises eight enzymes detected at the transcriptome or proteome level. These figures are like those reported for the model plant *A. thaliana*. There is still room for improvement, and there are pathways underrepresented in the created database, including the brassinosteroid biosynthesis pathway. The *Q. ilex* genome sequencing, the use of alternative and complementary strategies such as LC-MS will improve the number of pathways visualized.

The current metabolic reconstruction achieved for this species can be considered to be sufficient to progress in the biological knowledge of this species.

5.5 Data Availability

RAW and MSF files corresponding to proteomics are available at the ProteomExchange repository; Datasets: PXD008001. The Project ID of the GC-MS *Q. ilex* metabolomic analysis is PR000618 in the Metabolomics Workbench repository.

Chapter 6

Decoding drought tolerance mechanisms in Holm oak (*Quercus ilex*) through a combined transcriptomics and proteomics analysis

In preparation

Abstract

Quercus ilex, the typical tree of the Mediterranean forest and of the “dehesa” agrosilvopastoral ecosystems, is a species well adapted to xeric conditions, being reported as one of the most drought-tolerant within the European tress and *Quercus* genus. In an attempt to identify gene products and pathways related to drought tolerance in this species, an integrated transcriptomic and proteomic analysis has been carried out. *Quercus ilex* seedlings grown on pots containing perlite were subjected to drought conditions by water withholding for 30 days. Leaves were sampled at two times, when the leaf fluorescence dropped by 30% and 50% in comparison to the irrigated seedlings (at day 20 and 25, respectively), RNA and proteins independently extracted from the same batch of samples and analysed by RNA-seq and shotgun proteomics. RNA-seq analysis generated 47868 transcripts corresponding to 21000 unigenes, with 3588 qualitative or quantitative differences between irrigated and droughted seedlings (1149 up, and 2439 down). From shotgun proteomics, 4008 protein species were identified, corresponding to 2767 different genes. Out of them, 640 had qualitative or quantitative differences in abundance between treatments (353 more and 287 less abundant under drought conditions). Variable gene products were categorized in terms of gene ontology, biological process, molecular function and cellular component, and, for enzymes, in KEGG metabolic pathways. The variable dataset was subjected to multivariate, PCA and sPLS, statistical analysis. Finally, by using GeneMANIA, interaction networks were constructed.

A wide gene expression was observed at the two omics levels with up and down regulation, being this transitory (observed at 20 or 25 days) or permanent (observed at 20 and 25 days). The functional groups, whose genes were most altered in response to drought, were “stress-related” and “chloroplasts”. The most affected metabolic pathways included protein translation, photosynthesis, carbohydrates, amino acids and phenolics. Variable gene products were observed at transcriptomic or proteomic levels, with a reduced number detected at both levels. This included, for example, RPS2, 4CL2, PSB28, and RIN4, among others.

From the variable transcript and protein datasets, two networks were constructed, the first one included up accumulated *CLPB2*, *CLPB3*, *HSP70*, *HSP17.4*, *FTSH6*, *AT1G23740*, *SMT1*, and *UGP3*, and down accumulated *ABA2*, *RPS1*, *ADK*, and *RPL4* genes and the second one included up accumulated *CLPB2*, *CLPB3*, *HSP70*, *HSP17.4*, *FTSH6*, *AT1G23740*, *AP1*, *INVE*, *AT4G2740*, *CAD4*, *FEN1*, and *HIPP27* and down accumulated *ABA2* genes.

From a biological point of view, and in terms of stress response and tolerance, *Q. ilex* seedlings were characterized by an increase in general abiotic stress related gene products, including *FTSH6*, *PSB28*, *CPLB2*, and *CPLB3*. These variable gene products overexpressed under drought conditions can be proposed as molecular markers of response and tolerance to drought stress.

6.1 Introduction

Currently, drought conditions, accompanied by high temperatures and irradiance, are considered as one of the main causes of forest decline and tree mortality (Pasho et al., 2011). Moreover, this situation could become worse in a climate change scenario (Allen, 2009; Menezes-Silva et al., 2019), considering the simulation models and predictions of an increase in both temperature and frequency of drought periods (Collins et al., 2012). Thus, high temperatures, changes in precipitation patterns, among other climatic conditions, are increasing the aridity of the Mediterranean region, strongly impacting its ecosystems composition (Peñuelas et al., 2018).

Under the current climate change conditions, the conservation of the “dehesa”, in particular, and the Mediterranean forest, in general, as well as reforestation and afforestation programmes require the employment of novel strategies for the sustainable management and conservation of these ecosystems, among which breeding for resilience should be a priority. In *Q. ilex*, considered as a non-domesticated, long-lived species, and because of its allogamous character, the only plausible alternative in a breeding programme for tolerance is the selection of elite genotypes assisted by phenotypic and molecular markers, which involve the characterization of the biodiversity and those mechanisms of tolerance from a morphological, physiological and molecular point of view (Guzmán et al., 2015; Martínez et al., 2019).

Holm oak (*Quercus ilex* L. subsp. *ballota* [Desf.] Samp.), considered as the most representative species of the Mediterranean forests and the agrosilvopastoral ecosystem “dehesa”, has suffered an increase in the mortality rate in last decades in Southern Spain an increase in the mortality rate, both in natural stands and in novel plantations

(Villar-Salvador et al., 2004; Natalini et al., 2016). Although this increased mortality is mostly associated to the root rot pathogen *Phytophthora cinnamomi*, it has also been linked to drought episodes that worsen the survival of the species. Both factors, drought and *P. cinnamomic*, are considered as the main elements of the holm oak decline syndrome (Brasier, Robredo, and Ferraz, 1993; Sanchez et al., 2002; Ruiz Gómez et al., 2018).

Quercus ilex is a sclerophyllous species well adapted to climate conditions prevailing in the Central-Western Mediterranean basin (De Rigo and Caudullo, 2016). This species is considered as one of the most drought-tolerant species (David et al., 2007; Forner, Valladares, and Aranda, 2018; Früchtenicht et al., 2018)(San Eufrazio et al., 2020) due to themorpho-functional traits and strategies developed to face conditions of low water availability (Guzmán et al., 2015; Vicente et al., 2018). However, it has been reported that both inter- and intra-population variability in the level of drought tolerance within *Q. ilex* do exist (Valero-Galván et al., 2013). To date, many studies have shown the variability in the responses to biotic (e.g. *P. cinnamomic*) and abiotic (e.g. drought) stresses in *Q. ilex* by using physiological, classic biochemistry and -omics approaches. (Jorge et al., 2006; Echevarría-Zomeño et al., 2009; Sghaier-Hammami et al., 2013; Valero-Galván et al., 2011; Valero-Galván et al., 2013; Simova-Stoilova et al., 2015; Simova-Stoilova et al., 2018; López-Hidalgo et al., 2018; Rey et al., 2019). As a continuation of the previous referenced work, a combined RNA-seq transcriptomics and shotgun, LC-MS/MS, proteomics analysis of the drought response in 6-month old *Q. ilex* seedlings was carried out, with the main objective to generate as much data as possible from each -ome. This work provides some new data and knowledge on the molecular processes and mechanisms related to drought tolerant character of this species, as well as shed some light about possible candidates to be used as molecular markers in a breeding programme.

6.2 Materials and methods

6.2.1 Plant material and drought treatment

Healthy acorns were collected from trees located in 'Almadén de la Plata', Seville, Andalusia, Spain (37° 52'N 6° 28'W). Acorns were selected and germinated as previously reported (Simova-Stoilova et al., 2015). The drought experiment was performed with 6-month-old seedlings grown in 0.5 L pots containing perlite in July 2017 in Córdoba, Andalusia (Spain) under natural conditions (44°C and 19°C mean maximum and

minimum temperatures, respectively, and 40% relative humidity), as reported in (Valero-Galván et al., 2013) and (San Eufrasio et al., 2020) (Supplementary document). Seedlings were irrigated at field capacity every two days and once a week with a Hoagland nutrient solution (Hoagland and Arnon, 1950). The drought treatment was carried out as previously described in (San Eufrasio et al., 2020). Briefly, severe drought was imposed by water withholding for 30 days. The effect of drought was determined by total pot weight, damage symptoms and leaf fluorescence. Leaf fluorescence was measured regularly with a fluorometer (FluorPen FP100, Photon Systems Instruments, Drasiv, Czech Republic). Asymptomatic leaves were collected when the leaf fluorescence dropped by 30% and 50% in the droughted seedlings with respect to the well-watered ones (at day 20 and 25, respectively). Leaves were taken from three biological replicates per treatment and time and, they were immediately shock-frozen in liquid nitrogen and kept at 80°C until RNA extraction.

6.2.2 RNA extraction

Total mRNA was extracted from plant leaves according the procedures previously set up in our laboratory for *Q. ilex* samples (Echevarría-Zomeño et al., 2012). Briefly, RNA was extracted from 50 mg of fresh tissue and DNA was removed by DNase I treatment (Ambion, Austin, TX, USA). Total RNA was quantified spectrophotometrically (DU 228800 Spectrophotometer, Fullerton, CA USA), and the integrity of the isolated RNA was assessed using a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). Only high-quality RNAs with RIN values > 8 and A260:A280 ratios near 2.0 were used for subsequent experiments.

6.2.3 RNA-Seq Library construction, Illumina sequencing and *de novo* re-assembly

Extracted total RNA was sent to Allgenetics & Biology Sl. (<https://www.allgenetics.eu/>) for library preparation and Illumina RNA sequencing. Illumina's TruSeq Stranded mRNA Library Prep Kit was used to prepare the libraries strictly following the manufacturer's instructions. Briefly, each sample was enriched in mRNA by selecting those molecules with poly-A tail at their 3' end. Captured mRNAs are then converted into cDNA, and sequencing adaptors are added to their ends in order to make the samples ready for sequencing. The fragment size distribution and concentration of the libraries were checked in the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). The libraries were quantified with Qubit dsDNA HS Assay Kit (Thermo

Scientific, Madison, WI, USA). Then, they were pooled in equimolar amounts according to the Qubit results. The library was used for high-throughput sequencing with the Illumina HiSeq 4000 platform using a paired-end sequencing system to generate raw data. A quality control of the raw reads generated to obtain only high-quality reads was carried out using FastQC (v0.11.8) (Andrews, 2010). Truseq Adapter Index 7 (ATCGGAAGAGCACACGTCTGAACTCCAGTCACCGGCTATGATCTCGTATG) and Illumina Single End PCR Primer 1 (ATCGGAAGAGCGTCGTGTAGGGAAA-GAGTGTGCCTCTATGTGTAGATCTC) adapters, ambiguous nucleotides and low quality sequences (first 12 bp of each read) were removed by using Cutadapt (v 1.9) (Martin, 2011). Previous to this study, a *de novo* transcriptome of *Q. ilex* was generated by using raw data obtained from Illumina (Guerrero-Sanchez et al., 2017) and Ion torrent (Guerrero-Sanchez et al., 2019). These raw data together the combining datasets of well-watered and droughted seedlings obtained in this study were used to assemble all the clean reads generated in *Q. ilex* into contigs using RAY (v2.3.1) (Boisvert, Laviolette, and Corbeil, 2010). The evaluation of the structure of the generated transcriptome was performed using QUASt (v5.0.0) (Gurevich et al., 2013; Guerrero-Sanchez et al., 2017; Guerrero-Sanchez et al., 2019). The new version of the *Q. ilex* transcriptome was annotated against Uniref90 (UniProt) using Sma3s (v2) (Munoz-Mérida et al., 2014; Casimiro-Soriguer, Muñoz-Mérida, and Pérez-Pulido, 2017). In addition, all the transcripts identified were subjected to a Gene Ontology term comparison and classification. A GO term was assigned to each transcript based on the GO annotations for biological process, molecular function and cellular component. GO enrichment was evaluated by Fisher's exact test with a false discovery rate (FDR) in the biological process, molecular function and cellular component categories. The assembly calculations were run in the Supercomputing and Bioinnovation Center Service of the University of Malaga (Andalusia, Spain) (<http://www.scbi.uma.es/site/>). More information on the Illumina sequencing can be found in (Guerrero-Sanchez et al., 2017; Guerrero-Sanchez et al., 2019).

6.2.4 mRNA differential expression

The transcript quantification was carried out by mapping the filtered reads to the new transcriptome generated in this study, using Expectation Maximization method (RSEM) (Li et al., 2014). Differential expression analysis were performed with the edgeR R package (Robinson, McCarthy, and Smyth, 2010) by normalizing the counts using the Trimmed Mean of M-value (TMM) method. The counts represent the total number of reads aligning to each gene/transcript. EdgeR uses a generalized linear

model (GLM) which is like a linear model but assumes that the counts (raw reads) are not normally distributed, because most genes are not differentially expressed. EdgeR fits the counts to a negative binomial distribution and estimate the expected dispersion (variance). The GLM likelihood ratio test was selected for determining differential expression patterns. All the transcripts that varied from the well-watered seedlings with an adjusted p-value Benjamini-Hochberg $FDR < 0.05$ were considered as differently expressed (Benjamini and Hochberg, 1995). Venn diagram analysis of the upregulated and downregulated transcripts under drought conditions was carried out as previously reported in (Oliveros, 2007). A Gene Ontology Enrichment analysis was performed using ShinyGO v0.61 in order to classify genes according to biological process, molecular function and cellular component (Ge and Jung, 2018).

6.2.5 Protein extraction and digestion

Protein extraction was carried out using the trichloroacetic acid (TCA)/acetone-phenol protocol previously used for Holm oak (Jorri n-Novo, 2014). The protein concentration was determined by Bradford method (Bradford, 1976) (BioRad, Hercules, CA, USA) using bovine serum albumin (BSA) as standard. 90 μ g of BSA protein equivalent from each biological replicate of all treatments were subjected to SDS-PAGE (12% acrylamide) on the Protean XL-II (20 \times 20 cm) system (Bio-Rad). The gel was run at 80 V and stopped when the bromophenol blue advanced 0.5 cm into the resolving gel, a step introduced for protein sample cleaning (Pascual et al., 2017). The gel was then stained with Coomassie Brilliant blue (CBB) (Mathesius et al., 2001), and the unique resulting band excised using a clean scalpel and digested with trypsin (Sequencing grade, Promega, Madison, WI) as described in (Castillejo, Bani, and Rubiales, 2015). Digestion was stopped, and peptides were extracted from gel plugs by adding 10 μ L of 1% (v/v) trifluoroacetic acid (TFA) and incubating for 15 min, which were later completely evaporated in speed-vac.

6.2.6 Shotgun (LC-MS/MS) protein analysis

Protein analyses were conducted at the Proteomics Facility of the Research Support Central Service (SCAI) of the University of Cordoba. Nano-LC was performed in a Dionex Ultimate 3000 nano UPLC (Thermo Scientific, Madison, WI, USA) with a C18 75 μ m \times 50 Acclaim Pepmap column (Thermo Scientific, Madison, WI, USA). The digested peptides (3 μ g in 5 μ l) was previously loaded on a 300 μ m \times 5 mm Acclaim Pepmap precolumn (Thermo Scientific, Madison, WI, USA) in 2% AcN/0.05% TFA for

5 min at 5 $\mu\text{L}/\text{min}$. Peptide separation was performed at 40°C for all runs. Samples were separated during a gradient of 120 min ranging from 95% solvent A (0.1% FA) to 80% solvent B (80% ACN, 0.1% FA) and a flow rate of 300 nL/min. LC was coupled to MS using an ESI source. Eluted peptides were converted into gas-phase ions by nano electrospray ionization and analysed on a Thermo Orbitrap Fusion (Q-OT-qIT, Thermo Scientific, Madison, WI, USA) mass spectrometer operated in positive mode. Survey scans of peptide precursors from 400 to 1500 m/z were performed at 120K resolution (at 200 m/z) with a 4×10^5 ion count target threshold. Tandem MS was performed by isolation window at 1.2 Da with the quadrupole. Monoisotopic precursor ions were CID-fragmented in the ion trap, which was set up as follows: automatic gain control, 2×10^3 , maximum injection time, 300 ms, and 35% normalized collision energy.

6.2.7 Protein identification and quantification

Spectra were processed using the SEQUEST algorithm available in Proteome DiscovererTM 2.1 (Thermo Scientific, Madison, WI, USA). The following settings (Romero-Rodríguez et al., 2014) were used: precursor mass tolerance was set to 10 ppm and fragment ion mass tolerance to 0.6 Da. Identification confidence was set to a 5% FDR and the variable modifications were set to: oxidation of methionine and the fixed modifications were set to carbamidomethyl cysteine formation. A maximum of two missed cleavages were set for all searches. The protein identification was carried out against the translated *Q. ilex* transcriptome generated in this work. A six-frame translation for each sequence in the transcriptome was performed by using transdecoder (Haas et al., 2013) filtering and keeping peptides longer than 50 amino acids. Proteome DiscovererTM filtered out those proteins groups that had no unique peptides among the considered peptides during the protein grouping process. Proteins were quantified as reported in (Silva et al., 2006), using the average MD signal response for the three most intense tryptic peptides. Values for individual proteins were then divided by the total sum of the peak area values within each sample. After the natural log transformation of these relative values, a statistical analysis using the Student's t test (p-value < 0.05) was performed to identify the differential proteins (Zybailov et al., 2006). The criteria used to consider a protein as significantly change was as variable was as follows: (a) the protein was consistently present or absent in all three replicates for a condition and (b) it exhibited statistically significant differences (t-test, p-value < 0.05; between treatments). Venn diagram analysis of variable proteins was also carried out as previously reported in (Oliveros, 2007). A Gene Ontology Enrichment analysis

was performed using ShinyGO v0.61 in order to classify proteins according to biological process, molecular function and cellular component (Ge and Jung, 2018).

6.2.8 Multivariate Analysis

A multivariate analysis of the total and variables datasets at the transcript and protein levels was performed with mixOmics (Rohart et al., 2017) using Principal Component Analysis (PCA) and sparse Partial Least Squares (sPLS). The sPLS method was used to find correlations between predictors (Transcripts matrix) and response variables (Proteins) and the PCA method was performed to corroborate the sPLS plotting.

6.2.9 Pathway mapping of omics data

To acquire an integrated visualization of Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway maps, total transcript and protein datasets, specifying those variable transcripts and proteins, were analysed by Paintomics 3 (v0.4.5) (Garcia-Alcalde et al., 2011; Diego et al., 2018). A logarithm transformation was applied to the total and variable datasets. *Arabidopsis thaliana* was considered as a model reference. Pathways with p-value > 0.05 were considered as significantly enriched pathways.

6.2.10 Interaction network

Interaction networks were constructed using GeneMANIA Cytoscape plugin (Warde-Farley et al., 2010; Shannon et al., 2003). The interaction networks included were co-expression, co-localization, shared protein domains and co-localization. This software also finds functionally similar genes that do not exist in the input gene list (Franz et al., 2018). All the variables transcripts and proteoforms at each sampling time were used to the construction of the interaction networks, and *A. thaliana* was also considered as a model reference.

6.2.11 Data availability

The transcriptome data set will be available in the Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) and the proteome data will be available in the PRoteomics IDentifications Database (PRIDE, <https://www.ebi.ac.uk/pride/>). Accession numbers will be provided after the publication of this work.

6.3 Results

The transcript and protein profiles of leaves from *Q. ilex* seedlings subjected to well-watered (control) and drought-stress (water withholding) conditions were analysed by using RNA-seq (Illumina®) and Shotgun, LC-MS/MS, proteomics platforms. The analysis was performed at two sampling times corresponding to a decrease of leaf fluorescence of 30 and 50% in droughted seedlings compared to the well-irrigated ones (at days 20 and 25, respectively) in three biological replicates per treatment.

6.3.1 Transcriptomic and Proteomic Profile analysis

Transcriptomics

Illumina RNA-seq resulted in about 40 million 150bp paired-end reads per sample (480 million in total). In this work, a deeper coverage of transcriptome was achieved from these data and previous raw data obtained recently (Guerrero-Sanchez et al., 2017; Guerrero-Sanchez et al., 2019), a more complete version of the *Q. ilex de novo* transcriptome was assembled. The assembly structure analysis provided 23826, 253 and 10 contigs that had, respectively, more than 1000, 5000, and 10000 bp. The largest contig size was 15009 bp and the N50 and L50 values were of 1044 and 14029 bp, respectively.

After mapping to the new assembled transcriptome, 71,9% of the reads generated during the drought experiment, accounted for mRNA. In total, 47868 assembled transcripts, corresponding to approximately 21000 unigenes, were obtained, having a minimum and average length of 200 and 1076 bp, respectively (Table 6.1; Supplementary Table S15).

Table 6.1 Features of the transcriptome and proteome analysis at the two sampling days. The total number of identified transcripts and proteins, its sequence length, as well the number of them showing qualitative or quantitative differences between treatments are indicated. Newly appeared/disappeared indicates transcripts and proteins showing qualitative changes, being only present in drought/control treatments. Up/down indicates transcripts and proteins showing quantitative changes, being more abundant in drought/control treatments. a-e correspond to minimum (a) and mean (b) bp values; number of sequenced amino-acids with at least 1 unique peptide (c); sequenced amino acids mean value (d), and number of unigenes (e-f).

	Total No.	Sequence length		Time (Days)	Differences			
					Qualitative		Quantitative	
					Newly appeared	Disappeared	Up	Down
Transcripts	47868	200 ^a	1076 ^b	20	155	333	362	709
	21000 ^e			25	268	360	364	1028
Proteins	4008	50 ^c	347 ^d	20	54	41	94	104
	2737 ^f			25	80	45	125	97

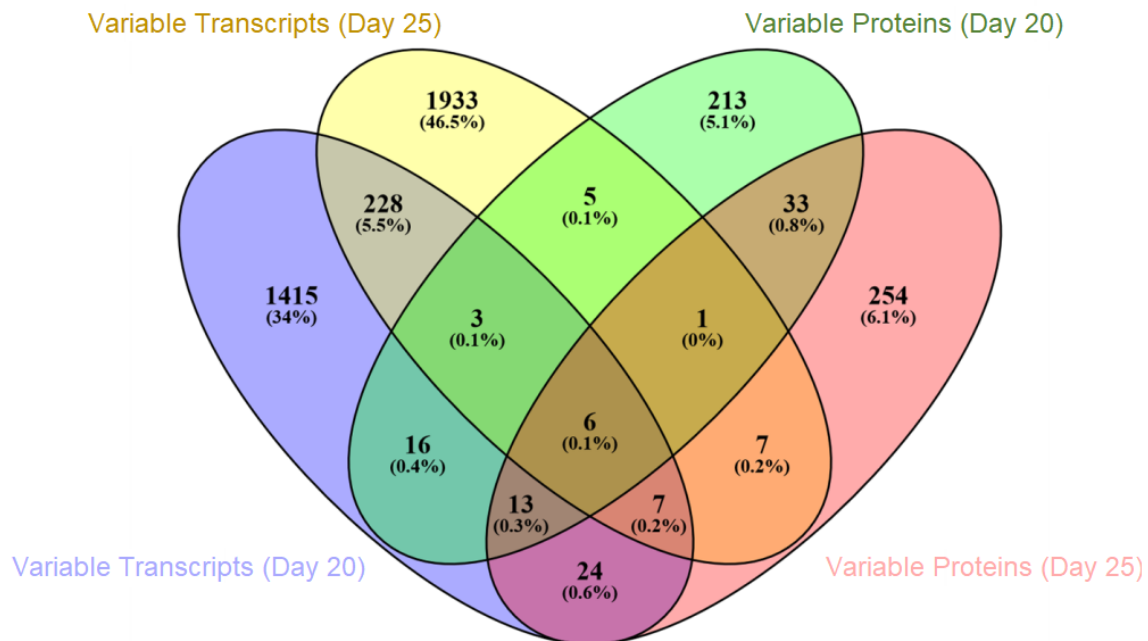


Figure 6.1 Venn diagram of the number of variable transcripts and proteins between treatments, droughted and well-watered seedlings, found at the two sampling, days 20 and 25. Intercepts show common differences at the two times and/or two platforms.

A differential gene expression analysis of the assembled transcripts was performed between well-watered and droughted seedlings at days 20 and 25 (Supplementary Table S15). The differences between treatments within each time were classified as qualitative (absence/presence) or quantitative ($p < 0.05$, t-test). Only consistent transcripts, those present in all the three replicates were considered. Out of 47868 assembled transcripts, 1116 showed qualitative differences between treatments, being 423 (155 at day 20, and 268 at day 25), and 693 (333 at day 20, and 360 at day 25) only present in droughted or well-watered seedlings, respectively (Table 6.1). The number of transcripts showing quantitative differences was much higher, 2463 in total. Out of them 726 (362 at day 20, and 364 at day 25), and 1737 (709 at day 20, and 1028 at day 25) were more and less abundant in droughted than in well-watered seedlings, respectively (Table 6.1).

When comparing transcript abundance between drought and control treatments, there were statistically significant differences common or specific to the two sampling times. Thus, while 231 variable transcripts were common to both times, 1431, and 1933, were specific of respectively, 20 and 25 days (Figure 6.1).

Proteomics

In a parallel analysis, a shotgun, LC-MS/MS, proteomic analysis was also performed in well-irrigated and droughted seedlings at days 20 and 25. A total of 4008 proteins, corresponding to 2737 unigenes, were identified satisfying the confidence parameters established (at least 1 unique peptide and $FDR < 0.05$) (Supplementary Table S16; Table 6.1). The minimum number of sequenced amino-acids with at least 1 unique peptide was 50 and the sequenced amino-acid mean number in a proteoform was 347. The number of peptides identified ranged in between 1 and 54, corresponding to sequence coverage of 1 to 93 %. Original raw data have been deposited in PRIDE Database, and the list of proteins identified, its quantitative abundance value, and statistical analysis p-value, are included in Supplementary Table S16.

After the t-test statistics analysis of the protein abundance ($p < 0.05$), 640 proteins in total were variable between treatments at the two times, 20 and 25 days, they showing qualitative, 220 proteoforms, or quantitative, 420 proteoforms, differences (Table 6.1). Within the first group, qualitative differences, 134 (54 at day 20 and 80 at day 25), and 86 (41 at day 20 and 45 at day 25) proteoforms were only detected in, respectively, droughted and well-watered seedlings. Regarding the quantitative changes, 219 (94 at day 20 and 125 at day 25) and 201 (104 at day 20 and 97 at day 25) were more abundant at, respectively, droughted and well-watered seedlings (Table 6.1).

As shown in the Venn diagram (Figure 6.1), when protein abundance in the two treatments was compared, 237, 292 and 53 proteins were identified at day 20, at day 25 and at both sampling times, respectively.

Correlations between transcript and protein abundance

The total number of gene products detected at both transcript and protein levels at the two sampling times was of 3374, with 939 out of them being variable at the transcript and/or protein level at day 20 and/or 25 (Tables S15 and S16). The maximum value for transcript abundance in the total and variable datasets was close to 11000 counts per million (CPM), with only 31 and 13 transcripts, respectively, showing values > 5000 CPM. With respect to proteoforms, the highest abundance value for the total and variable datasets was, in arbitrary units, of 0,255 and 0,05, respectively, with most of values being below 0,05.

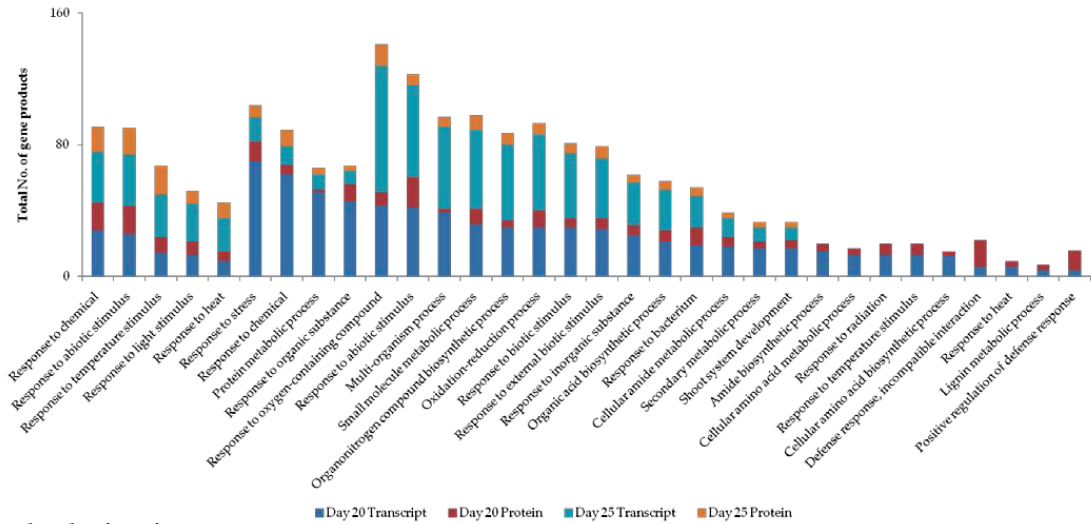
The correlation analysis between the transcriptome and proteome abundance for both datasets was carried out by using the Pearson's test. Data corresponding to gene

products detected at both levels independently of its sample and replicate origin were employed. A correlation analysis between all transcripts and proteins identified did not correlated as measured by Pearson's r (20244) = 0,1055 (Supplementary Figure S9). On the other hand, the comparison between variable transcripts and proteins did not correlated as measured by Pearson's r (5634) = 0,2139 (Supplementary Figure S9), being only a total of six common transcripts and proteins at days 20 and 25 (Figure 6.1).

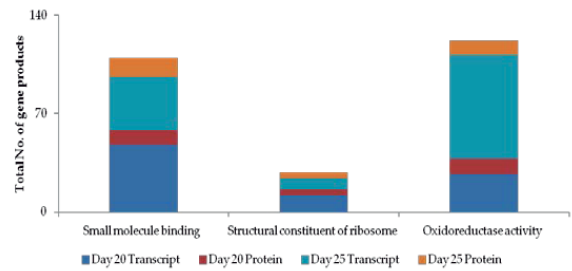
6.3.2 Gene Ontology analysis

Gene ontology (GO) analysis was performed to classify identified transcripts and proteins in terms of biological process, molecular function and cellular location. The total set of data covered most of the categories included in the GO list for the three criteria, cellular, and molecular function, and location (Supplementary Table S15). A most detailed analysis was performed with the set of variable transcripts and proteins (Figure 6.2, Supplementary Table S18). All the GO categories with an FDR < 0,05 were selected. Within the biological process, the number of functional categories ranged between 20 and 261, depending on the treatment (well-watered and droughted seedlings), sampling time (at days 20 and 25), and omics platform (transcriptome and proteome) (Supplementary Table S18). There were more functional categories with downregulated (240, and 261, at 20 and 25 days) than upregulated transcripts (20, and 59, at days 20 and 25), while the figures were more homogeneous when analysing the proteomics data, 75 and 51 (upregulated, days 20 and 25), and 72 and 31 (downregulated, days 20 and 25). These categories were filtered, firstly, to those common categories to the two -omics levels, resulting 7 and 12 categories upregulated at days 20 and 25, and, 30 and 22 categories downregulated at days 20 and 25. A second filter was established by eliminating redundant categories, keeping the final figures to 5 and 9 categories (upregulated at days 20 and 25) and, 27 and 14 categories (downregulated at days 20 and 25) (Figure 6.2). Some of the variable categories were common to both sampling times with the list of upregulated categories including "Response to chemical", "Response to abiotic stimulus", "Response to temperature stimulus", and "Response to light stimulus"; and the list of downregulated categories having "Response to chemical", "Oxidation-reduction process", "Small molecule metabolic process", "Response to oxygen-containing compound", "Response to biotic stimulus", "Response to bacterium", and "Response to heat". Some of the functional categories were included within the upregulated and downregulated categories, as for example, "Response to chemicals", one of the largest groups.

(a) Biological process



(b) Molecular function



(c) Cellular component

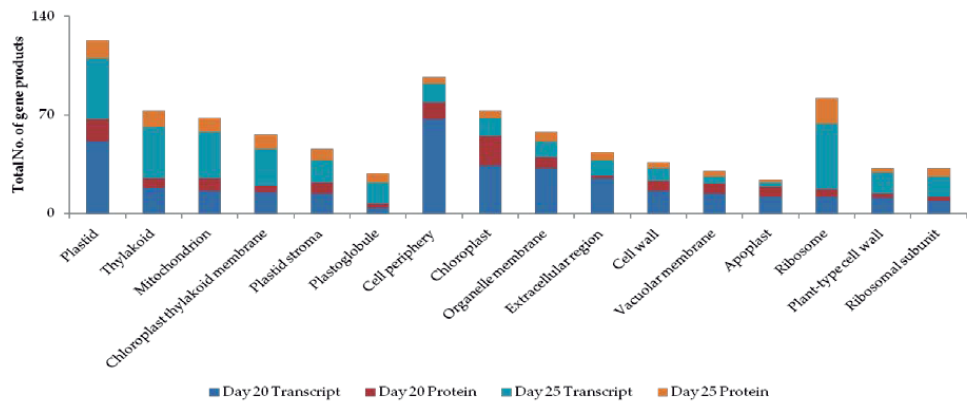


Figure 6.2 Gene Ontology analysis of the variable gene products. (a) Biological process; (b) Molecular function; (c) Cellular location. The X axes contain the categories and the Y one the number of identified gene products. Blue: transcript at day 20, Red: protein at day 20; Light blue: transcripts at day 25, Orange: proteins at day 25. Original data are included in Supplementary Table S18.

Within the molecular function, the number of functional categories identified ranged between 4-88, depending on the treatment, sampling time, and omics platform (Table S18). At day 20, the number of functional categories (both up- and downregulated categories) was quite similar in the *Q. ilex* transcriptome and proteome; however, at day 25, the number of functional categories was higher in the upregulated than downregulated transcripts and proteins. The same filters used in the biological process were applied in this function. Firstly, 0 and 22 categories in both transcripts and proteins (upregulated at days 20 and 25) and, 30 and 22 categories in transcripts and proteins (downregulated at days 20 and 25), and secondly, 0 and 2 categories (upregulated at days 20 and 25) and, 3 and 1 categories (downregulated at days 20 and 25) were identified (Figure 6.2). The most common molecular category identified in both omics platforms at days 20 and 25 between well-watered and droughted seedlings was “Small molecule binding”, in which other molecular categories were included such as “Nucleotide binding”, “Nucleoside phosphate binding”, “ATP binding”, “Cofactor binding”, “Drug binding”, among others.

Within the cellular location, annotated variable gene products were associated to 21-46 subcellular fractions, depending on the treatment, time, and omics platform, with no clear different tendencies (Supplementary Table S18). The values for transcriptomics and proteomics were 21-42 and 27-46, respectively; 21-46 and 23-42 at days 20 and 25, respectively; and 21-43 and 23-46 for up- and downregulated categories, respectively. The same filters used in the above GO categories were applied in the cellular location. Firstly, 17 and 33 categories in both transcripts and proteins (upregulated at days 20 and 25) and, 17 and 4 categories in transcripts and proteins (downregulated at days 20 and 25), and secondly, 6 and 13 categories (upregulated at days 20 and 25, respectively) and, 10 and 3 categories (downregulated at days 20 and 25, respectively) were identified (Figure 6.2). The most represented categories were “Chloroplast” and “Mitochondria” (116 and 53 upregulated transcripts and proteins, respectively, at day 20); “cell periphery”, “chloroplast”, “plasma membrane”, “cell-cell junction”, “cell wall”, “vacuole”, “peroxisome”, and others organelles (34 and 91 upregulated transcripts and proteins, respectively, at day 25), “cell periphery”, “chloroplast”, “organelle membrane”, “extracellular region”, “cell wall”, “vacuole”, “apoplast”, and “ribosome” (232 and downregulated transcripts and proteins, respectively, at day 20), “chloroplast”, “vacuole”, and “apoplast” (76 and 27 downregulated transcripts and proteins, respectively, at day 25). The most represented subcellular fraction was “chloroplast”, with proteins and transcripts changing at 20 (201) and 25 (216)

days, upregulated (297), and downregulated (120), and at the two -omics levels (284 transcripts, 123 proteins).

6.3.3 Multivariate analysis of omics data

In order to reduce the complexity of the data, to know which gene products do more contribute to the variability of the analyzed samples (well-watered and droughted seedlings at two sampling times, days 20 and 25), establishing tendencies and correlations, and which of the variables would be associated to the droughty treatment, a multivariate analysis of the variance, including sPLS and PCA tests, was carried out (Supplementary Figure S11). These analyses were performed with different datasets: (1) total transcriptome and proteome, (2) total individual transcriptome, (3) total individual proteome, (4) variable transcriptome and proteome, (5) variable transcriptome, and (6) variable proteome. The main features of the different analysis are summarized in Table 6.2).

In all the sPLS performed, component 1 separated treatments, while days were separated only by component 2 when the set of variable data were employed (Supplementary Table S15, Table 6.2). The number of components requested to account for 50 % of the variability depended on the dataset, two for the variable transcript and protein, three for the total transcriptome, and four for the total proteome dataset (Supplementary Table S16).

PCA analysis of the total transcript and proteins, whether individually or together, did not discriminate treatments or days based on PC1 and PC2. On the contrary, PC1 discriminated drought from well-watered seedlings, and PC2 20 from day 25 when, the variable protein and transcript datasets were employed (Supplementary Table S15, Table 6.2). To explain 50 % of the variability, from three to four PCs were necessary when using the whole dataset, however, the first two PCs explained 50% of variability in both variable transcript and protein datasets.

In order to identify which of the gene products, (transcript or proteins) do more contribute to the variability and that are most related to the drought treatment, the sPLS of the variable proteins and transcripts were analysed in detail. Table 6.3 shows all the gene products ordered from the ten highest and ten lowest loadings, having, respectively, positive or negative correlations.

Table 6.2 Main features of the sPLS and PCA analysis based on the fifth first components. Values correspond to the percentage of variability explained by each component, and the number of components explaining 50% of the variability.

sparse Partial Least Square (sPLS)								
DATASET	Comp1	Comp2	Comp3	Comp4	Comp5	50% Variability	Separate Treatments	Separate Times
Whole Transcriptome and Proteome	-	-	-	-	-	-	Yes	No
Whole Transcriptome	0,24	0,17	0,14	0,08	0,07	Comp1-Comp3	Yes	No
Whole Proteome	0,14	0,13	0,11	0,12	0,1	Comp1-Comp5	Yes	No
Variable Transcripts and Proteins	-	-	-	-	-	-	Yes	Yes
Variable Transcripts	0,32	0,19	0,17	0,13	0,1	Comp1-Comp2	Yes	Yes
Variable Proteins	0,34	0,18	0,12	0,06	0,06	Comp1-Comp2	Yes	Yes

Principal Component Analysis (PCA)								
DATASET	PC1	PC2	PC3	PC4	PC5	50% Variability	Separate Treatments	Separate Times
Whole Transcriptome and Proteome	0,24	0,15	0,12	0,08	0,08	PC1-PC3	No	No
Whole Transcriptome	0,25	0,16	0,12	0,08	0,07	PC1-PC3	No	No
Whole Proteome	0,17	0,13	0,12	0,1	0,09	PC1-PC4	No	Yes
Variable Transcripts and Proteins	0,33	0,18	0,11	0,07	0,06	PC1-PC3	Yes	Yes
Variable Transcripts	0,36	0,18	0,11	0,07	0,06	PC1-PC2	Yes	Yes
Variable Proteins	0,34	0,2	0,11	0,06	0,06	PC1-PC2	Yes	Yes

Table 6.3 List of gene products exhibiting the highest loadings to component 1 in the sPLS analysis of the variable transcripts and proteins. Contig ID, gene3 acronyms and description, loading parameter and quantitative relative values, drought/control.

Variable Transcriptome dataset				
Gene	Description	Loading	Day 20 Fold Change	Day 25 Fold Change
PWD	Phosphoglucan water dikinase	0.035313537	2.53608580	5.01133265
fbp1	Fructose-1,6-bisphosphatase, chloroplastic	0.035087055	6.72595130	71.90484784
PGSC0003DMG400013943	NA	0.035076491	2.77540926	2.34559807
grpE	GrpE protein homolog	0.034749614	2.85123685	2.79573939
ELIP1	Early light-induced protein 1, chloroplastic	0.034743042	2.97871283	6.96935228
CER3	Protein ECERIFERUM 3	0.034420941	4.04444142	3.85791450
PYD2	Dihydropyrimidinase	0.034304659	1.98800712	4.93674015
IAP75	Protein TOC75, chloroplastic	0.034086451	1.81732269	3.68928767
TOC159	Translocase of chloroplast 159, chloroplastic	0.033863208	2.88100706	5.51495098
GWD3	Phosphoglucan, water dikinase, chloroplastic	0.033298551	1.83614878	3.30710448
kco1b	Calcium-activated outward-rectifying potassium channel 1	-0.032773415	0.12343472	0.17303041
AMT	Ammonium transporter	-0.032832286	0.26487629	0.38353880
nl27	Disease resistance protein (TIR-NBS-LRR class)	-0.033293177	0.24637206	0.09146642
NA	TMV resistance protein N	-0.033585758	0.50673277	0.24237495
c3'h	p-coumarate 3-hydroxylase	-0.033801474	0.22347716	0.16031336
LRK10	Receptor-like kinase	-0.034174734	0.40430169	0.17128215
BI-1	Bax inhibitor 1	-0.034240492	0.32866244	0.39134065
T1.1	T1.1 protein	-0.034319035	0.10033921	0.19753233
CML19	EF hand calcium-binding family protein	-0.035269551	0.07886879	0.15301722
BVRB_4g091600	NA	-0.035278830	0.23048644	0.04465059
Variable Proteome dataset				
Gene	Description	Loading	Day 20 Fold Change	Day 25 Fold Change
sqdB	Uridine 5'-diphosphate-sulfoquinovose synthase	0.07298291	1.4001E+00	1.8899E+00
ERD7	AT3g51250/F24M12_290	0.072330095	3.8538E+00	1.7168E+09
CHRC	Probable plastid-lipid-associated protein 2, chloroplastic	0.070053381	2.6469E+09	7.5729E+00
HSP17.4B	17.4 kDa class III heat shock protein	0.069930911	4.5386E+00	4.0172E+00
HSP70	Heat shock cognate 70 kDa protein	0.069501331	3.5000E+00	4.8181E+00
SARED1	Os05g0110300 protein	0.069231708	3.9104E+00	2.1154E+00
stk	Serine/threonine-protein kinase	0.069066131	5.5881E+08	5.2851E+08
SARED1	Os05g0110300 protein	0.068555284	2.4203E+00	1.9273E+00
hspB	Heat shock protein 70	0.068053709	1.4765E+00	2.3667E+00
PGSC0003DMG400008259	Galactose-1-phosphate uridylyltransferase	0.067922253	3.2889E+09	3.4138E+09
PPH1	Protein phosphatase 2C 57	-0.064501978	9.2683E-10	3.2415E-10
GPP2	(DL)-glycerol-3-phosphatase 2	-0.065429023	8.8551E-03	3.0915E-11
IFR	Isoflavone reductase, putative	-0.065574026	3.3291E-10	4.7938E-11
ychF	Ribosome-binding ATPase YchF	-0.066309298	3.5516E-01	2.9291E-01
At3g26720	Alpha-mannosidase At3g26720	-0.066399069	3.8231E-10	4.7078E-01
hdr	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	-0.066655366	5.0349E-01	7.7475E-01
SBE3	Starch branching enzyme 3	-0.066868425	5.3121E-01	3.3846E-01
fnr	Ferredoxin--NADP reductase	-0.067204050	2.9376E+10	4.7920E-01
RPS13A	40S ribosomal protein S13-1	-0.073315490	4.8753E-10	2.1372E-10
4CLL9	4-coumarate--CoA ligase-like 9	-0.073643243	1.1418E-09	3.4723E-10

6.3.4 Integrated visualization of omics data in metabolic pathways

To obtain a more global and integrative vision of the metabolic changes in response to drought in *Q. ilex*, PaintOmics 3 was used to carry out a KEGG pathway analysis. Identified transcript and protein enzymes were linked to 125 KEGG pathways (Table S21).

Table 6.4 KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked and showed statistically significant differences (Fischer test, $p < 0,05$) between treatments. Columns include the pathway name, number of gene products identified within each pathway, significance p-values at the transcript, protein or combined levels, changes in response to drought (up or down) and identified transcript or protein enzymes.

Pathway name	Unique genes	Gene expression	Proteomics	Combined p-value (Fisher)
Ribosome	53	5.46694E-14	0.390805479	6.93874E-13
Glyoxylate and dicarboxylate metabolism	26	0.075875383	0.002218035	0.001630737
Phenylpropanoid biosynthesis	17	0.001955976	0.167367291	0.002954295
Phenylalanine metabolism	8	0.00311676	0.111185675	0.003107594
Flavonoid biosynthesis	4	0.000547595	1	0.004660021
Stilbenoid, diarylheptanoid and gingerol biosynthesis	6	0.006338102	-	0.006338102
Biosynthesis of secondary metabolites	268	0.005297628	0.263029184	0.010556612
Photosynthesis	35	0.527196172	0.002659193	0.010612389
Carbon metabolism	68	0.032768139	0.069147043	0.016064252
Plant-pathogen interaction	62	0.021472121	0.167367291	0.023821276
Cysteine and methionine metabolism	26	0.035504942	0.106842142	0.024939814
Ubiquinone and other terpenoid-quinone biosynthesis	11	0.017242761	0.280591355	0.03063151

Table 6.5 List of transcripts and proteins linked to the variable pathways. The columns include the name of the pathway, and gene products up and down accumulated at days 20 and 25.

Metabolic pathway	Transcript		Proteins	
	Up	Down	Up	Down
Ribosomal RNAs		rrn18		rrn18
Ribosomal proteins	RPS10, RPL4, RPS17, RPL18, RPL9, rps18, RPS6	Rps3, rpl14, rps8, RPS15A, RPL15, RPL34, RPS13A, RPL27, RPS1, SAG24, RPL24, RPS1, UBQ1, rps2		rps2
Glyoxylate and dicarboxylate metabolism	CSY1, HPR, LPD1, AGT, GLU1	MDH		
Phenylpropanoid biosynthesis	HCT, FAH1	4CL2		4CL2
Phenylalanine metabolism		4CL2		4CL2
Flavonoid biosynthesis	HCT			
Stilbenoid, diarylheptanoid and gingerol biosynthesis	PPL1, PSBY, PSAG, ATPC2, PSB28	psbC, psbB, psbM, psbZ, psaB, psaC, atpF	PSB28 , psbH	
Plant-pathogen interaction	NOA1, FRK1, BAK1, EFR, RPM1, MYB30	CPK1, CML11, MPK6, WRKY2, MKK5, PRB1, SHD, EDS1, RIN4		RIN4
Cysteine and methionine metabolism	CSYD1	MDH, BCAT3, AK3, ACS1, SAMDO		
Ubiquinone and other terpenoid-quinone biosynthesis	PG1, ICS2, PPT1, MENG, ABC4	HST, 4CL2		HST, 4CL2

Out of the 125 pathways, 10 showed significant differences (Fischer p-value < 0.05) between treatments (well-watered and droughted seedlings) at transcript and

protein levels (Tables 6.4 and 6.5). The number of gene products identified per variable pathway ranged between 1 (caffeine metabolism, monoterpenoid biosynthesis, riboflavin metabolism and other types of O-glycan biosynthesis) and 268 (biosynthesis of secondary metabolites) (Supplementary Table 6.5). All the significant KEGG pathways are showed in Table 6.4.

6.3.5 Interaction network analysis

A functional network analysis was performed by using GeneMANIA. The analysis included those gene products detected at the two -omics levels, transcripts and proteins, with statistically significant differences between treatments at days 20 and 25 (Figure 6.1). The list is reduced to 29 and 14 gene products at, respectively, days 20 and 25.

At day 20, a principal cluster composed by heat shock proteins and molecular chaperons was observed (e.g. *HSP70*, *HSP17.4B*, *CLPB3*, and *CLPB2*) (Figure 6.3; Supplementary Table S20). At day 25, two principal clusters composed by genes implicated in both the cellular response to DNA damage/DNA repair and response to heat (e.g. *FEN1*, *ABA2*, *EGY3* or *CLPB3*) were observed (Figure 6.3; Supplementary Table S20). Variable gene products detected at both sampling times at the transcript and protein levels, namely *ABA2*, *CLPB3*, *CLPB2*, *FTSH6*, and *AT1G23740*, were included in the three established networks. All of them, except *ABA2* that was downregulated, were upregulated under drought conditions (Figure 6.3).

Apart from these genes, other genes were included in these networks. At day 20, the genes *AtHB26*, *SMT1*, *UGP3*, *AT5G48020*, *HSP70*, *HSP17.4B* and *AT3G23600* were upregulated and *ADK*, *RPL4Z* and *RPS1* were downregulated; and at day 25, *FEN1*, *HIPP27*, *CAD4*, *AP1*, *INVE* and *EGY3* were upregulated under drought conditions (Figure 6.3).

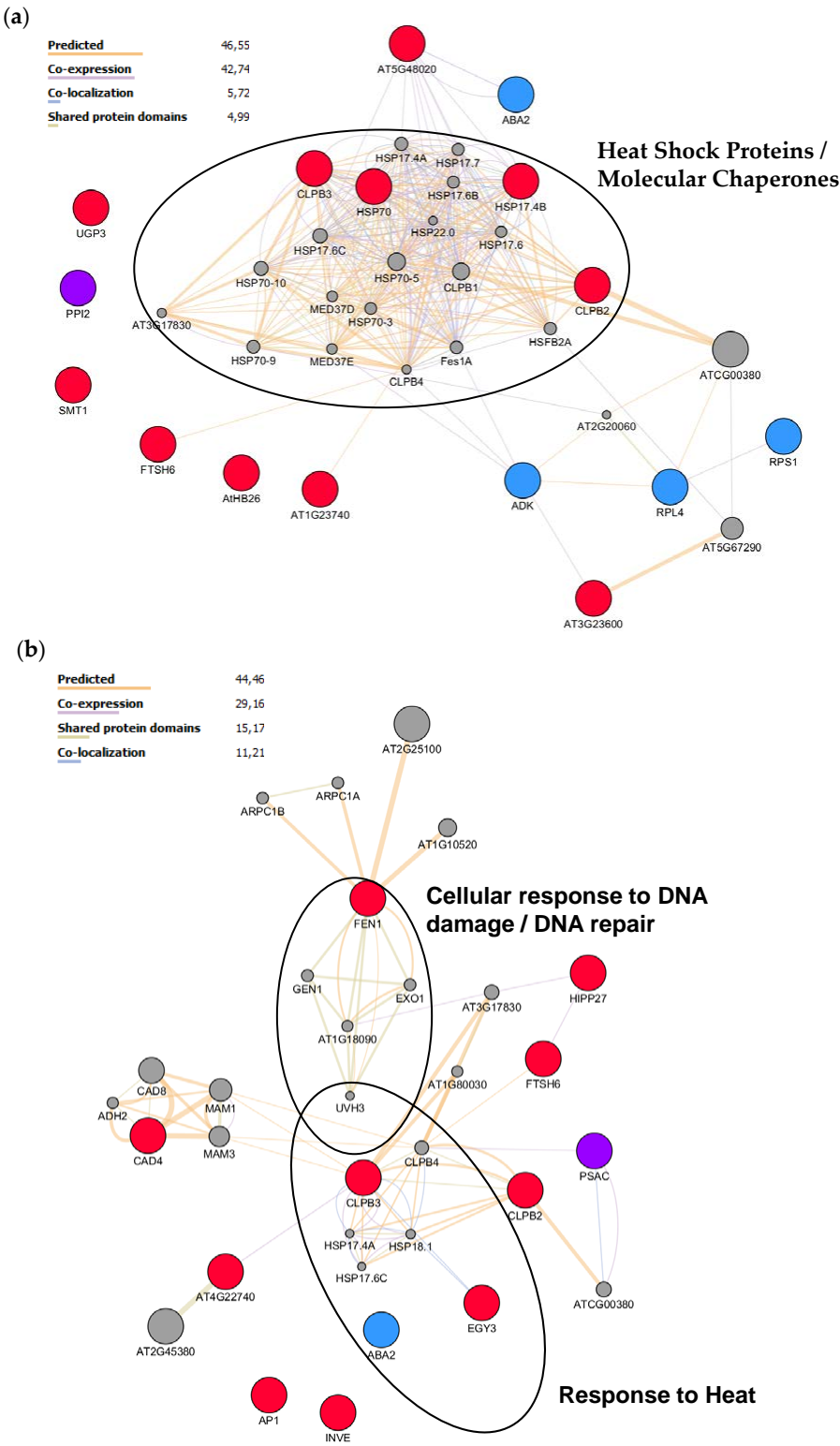


Figure 6.3 GeneMANIA network analysis. It was performed with variable gene products detected at the two -omics level. The analysis was performed at days 20 (a) and 25 (b). A red background circle indicates upregulated gene and protein, a blue background circle indicates downregulated gene and protein, a purple background circle indicates contrary gene and protein changes, and a grey background circle indicates GeneMania predicted interactions.

6.4 Discussion

6.4.1 Environmental stress, drought and climate change, biodiversity, and tolerance

The environment determines plant reproduction, growth, development and physiology, and, hence, geographical distribution. In situations of adverse environmental conditions and when the intensity and duration thresholds cause serious damages in the plant species, the mortality rate increases significantly, it determining geographical distribution (Schäfer and Dirk, 2011; Harfouche, Meilan, and Altman, 2014). Among the main factors that cause restrictions in the survival of plant species, drought is one of the most restrictive factors, predicting to be the most climatic extreme that affects terrestrial ecosystems, including natural and intervened forest, such as the Mediterranean and “dehesa” (Greenwood et al., 2017; Schwalm et al., 2017; Peñuelas et al., 2017; Ruosteenoja et al., 2018). Increased tree mortality associated to drought episodes has been observed in last 30 years, being argued as evidence of vulnerability to forest dieback. This is because of that the main objectives of the research community have been focused on the determination of processes related to tree species mortality and survival (Meir, Mencuccini, and Dewar, 2015). The relevance of this topic is closely related to the climate change issue, with models predicting extreme weather events, including severe drought episodes in the next decades, species distribution, forest composition, and biomes replacement, with the prevalence of drought-tolerant and lower growth rate species (Menezes-Silva et al., 2019). Social concerns and environmental and economic interests favour those studies carried out in plant species and genotypes well-adapted to arid conditions.

Despite being sessile organisms, plant species are able to survive under biotic and abiotic stress conditions. During the course of evolution, mechanisms have evolved to cope with these extreme environmental conditions, allowing plants to colonize different and variable ecological niches. These mechanisms, either permanent or transitory, and depending on the intensity and duration of the stress, occur and operate at different time scales and organism levels from the plant body to the subcellular fractions, from morphometry to chemical composition, and from short to long term adjustments (Polle et al., 2019). The different phenotypes behind adaptation to water shortage and drought conditions, including avoidance (e.g. water content homeostasis, leaf and root morphology) or tolerance/resistance (e.g. metabolic adjustments) mechanisms are, ultimately, the result of the genotype, including epigenetic labels, and the reprogramming

of gene expression and interaction among different gene products (Müller and Gailing, 2019).

Between and within forest tree species, there is an enormous genetic diversity for stress tolerance. Thus, and as an example, European beech (*Fagus sylvatica*) and Penduculate oak (*Quercus robur*), two related species within the genus *Fagaceae* whose genomes have been recently sequenced (Mishra et al., 2018; Plomion et al., 2018), exhibit quite different ecological behaviour (Roman et al., 2015); the same can be applied to populations within a species (Valero-Galván et al., 2013; Aranda et al., 2015). For a better management of forest populations, it is necessary to understand the relationship between tree genotype and phenotype, as well as the searching of morphometric and molecular markers. The genetic and molecular bases of tolerance are rapidly being known thanks to the advances in sequencing and -omics technologies, and its integration with phenotyping, physiological and classic biochemistry approaches (Harfouche, Meilan, and Altman, 2014). The investigation of the molecular and genetic bases of polygenic traits as is the tolerance to drought is an important goal in tree breeding and forest conservation and management.

Plant adaptation to water resources, water status and homeostasis is determined by a number of phenological, morphological traits and physiological and molecular mechanisms, such as root structure, leaf morphology, surface, and shedding, anatomical adjustment of the conducting system, hydraulic conductivity and cavitation, chemical composition of the cuticular leaf surface, stomatal conductance, and metabolic and osmotic adjustment, among the most relevant ones (Polle et al., 2019; Müller and Gailing, 2019).

Drought tolerance is associated to well-known changes in the cellular metabolism such as photosynthesis and energy production, stored non-structural carbohydrates mobilization and respiration, secondary metabolism, membrane composition, protein folding, osmotic adjustment, redox homeostasis by Reactive Oxygen Species (ROS) scavenging, aquaporins, and the induction of drought-related proteins such as Late Embryogenesis Abundant proteins (LEAs). These biochemical changes are mediated by sensors, inter and intracellular signalling, calcium and related signal transduction pathways, hormones such as abscisic acid (ABA), ethylene, as well as other intracellular and jasmonic, as main hallmarks (Harfouche, Meilan, and Altman, 2014; Polle et al., 2019; Müller and Gailing, 2019).

6.4.2 Drought responses in forest tree species: the genus *Quercus*, *Quercus ilex*. Breeding for drought tolerance based on the selection of elite genotypes

The effect, and the responses and tolerance to drought have been widely studied in model plant species (Sharma et al., 2018), crops (Swamy and Kumar, 2013; Zhu et al., 2016; Pieczynski et al., 2018) and much lesser in forest trees, among which productive species such as *Populus* (Cohen et al., 2010), *Eucalyptus* (Correia et al., 2018), and *Pinus* (Moran et al., 2017) have received more attention from the scientific community. On the contrary, little attention has been paid to other species of more environmental than economics interest such as those of the genus *Quercus* (Müller and Gailing, 2019).

In the present work, *Q. ilex* has been used as an experimental system to study responses to drought. It is the typical tree of the Mediterranean forest and of the dehesa agrosilvopastoral ecosystem (Moreno and Pulido, 2009). Its election is justified because of the environmental and economic importance in the Andalusian region, and because is one of the most drought tolerant species within the *Quercus spp.* (San Eufrasio et al., 2020). This character makes it an excellent candidate in reforestation programs. Although it is well adapted to xeric conditions (Echevarría-Zomeño et al., 2009; Valero-Galván et al., 2013; Gil-Pelegrín, Peguero-Pina, and Sancho-Knapik, 2018), drought stress is the main cause of *Q. ilex* seedling mortality in forest plantations, and one of the damaging factors of the decline syndrome (Villar-Salvador et al., 2004; Gentilesca et al., 2017; Colangelo et al., 2018; Ruiz Gómez et al., 2018). Plant responses and mechanisms of tolerance to drought in *Q. ilex* have been approached at different levels in field and greenhouse experiments, employing morphometric, physiological, classic biochemistry, and -omics approaches (Peña-Rojas, Aranda, and Fleck, 2004; Serrano et al., 2005; Limousin et al., 2010; Vaz et al., 2011; Sardans, Peñuelas, and Lope-Piedrafita, 2010; Galiano et al., 2012; Barbeta, Ogaya, and Peñuelas, 2013; Rosas et al., 2013; Valero-Galván et al., 2013; Rico et al., 2014; Simova-Stoilova et al., 2015; Chiatante et al., 2015; Sperlich et al., 2016; Salomón et al., 2017; Rodríguez-Calcerrada et al., 2018).

Breeding for resilience to drought and other stresses is a priority in tree breeding programmes (Polle et al., 2019). Within this objective, and being a long-lived, and non-domesticated species of allogamous and promiscuous properties, selection of elite or plus genotypes based on molecular markers is the almost unique and most plausible biotechnological approach in *Q. ilex*. Conventional breeding, genetic engineering, or genome edition are not possible or realistic alternatives. In this regard, genetic diversity

within *Q. ilex* and other *Quercus* species has been very well-documented, with some studies reporting variability in drought tolerance (Jorge et al., 2006; Echevarría-Zomeño et al., 2009; Valero-Galván et al., 2013; Rico et al., 2014; Müller and Gailing, 2019).

Selection of elite genotypes are based on phenotypic, physiology or molecular characteristics, with the last ones aimed at profiling the different cell biomolecules, DNA, RNA, proteins, enzymes, and metabolites, by using classic biochemistry, or modern DNA marker or -omics techniques (Porth and El-Kassaby, 2014; Gudeta, 2018).

6.4.3 Research on *Quercus ilex* at the Agroforestry and Plant Biochemistry, Proteomics, and Systems Biology Group, from classic biochemistry to -omics and systems biology approaches

In the direction presented in the previous sub-sections, the Agroforestry and Plant Biochemistry, Proteomics, and Systems Biology Group is currently investigating on different aspects of the biology of *Q. ilex* by using a molecular biology approach. Biological processes under study include development, seed maturation and germination, and responses to biotic (*Phytophthora cinnamomi*) and abiotic (drought) stresses (Valero-Galván et al., 2013; Sghaier-Hammami et al., 2013; Romero-Rodríguez et al., 2018; Simova-Stoilova et al., 2018; Romero-Rodríguez, Jorrín-Novó, and Castillejo, 2019). As methodological approaches, and beyond classic biochemistry, and physiology techniques, microsatellites, and -omics approaches have been optimized to this experimental system, mostly proteomics, and to a lesser extent, transcriptomics and metabolomics (Guerrero-Sanchez et al., 2017; Guerrero-Sanchez et al., 2019; López-Hidalgo et al., 2018; Marti et al., 2018). The objective is to integrate all of them in the Systems Biology direction (López-Hidalgo et al., 2018). Up to 2010, -omics approaches were developed and employed independently with not much integration between them, which made proteomics and transcriptomics mostly descriptive and speculative. In the 2010, papers reporting the integrated employment of the two or three -omics approaches, mostly transcriptomics and proteomics, have started to appear in plant biology research. As stated in (Rey et al., 2019), “*The logical transition from reductionists to a holistic strategy and integration of multidimensional biological information is currently accepted by the scientific community as the only way to decipher the complexity of living organisms and predict through multiscale networks and models.*” *The integrated use of the -omics approaches will not only allow us to connect the phenotype and the genotype*

but also, more importantly, to deepen the knowledge of gene expression mechanisms, including posttranscriptional (RNA splicing, micro-RNAs, small interfering RNA, long noncoding RNAs), and posttranslational (phosphorylation, glycosylation, acetylation, 286 methylation, etc.) events". This new strategy requires novel methodologies and equipment, with bioinformatics and computer skills being the real bottleneck, and because of that the present doctoral thesis was programmed and executed.

6.4.4 Integrated proteomics and transcriptomics analysis of responses to drought in *Quercus ilex*. Identified transcripts and proteins, and functional grouping

It has recently been reported that *Q. ilex* was the most tolerant species, within the genus *Quercus*, found in the Iberian Peninsula, including *Q. robur*, *Q. faginea*, *Q. pyrenaica*, and *Q. suber* (San Eufasio et al., 2020). The tolerant phenotype was established based on the appearance of damage symptoms in leaves and plant mortality. Within the *Q. ilex* species, five populations covering the Andalusian geography were surveyed, observing differences in tolerance among them. At the physiological level, it was observed that under severe drought conditions as well as high temperature and illumination, tolerant individuals kept well hydrated. Leaf fluorescence, gas exchange (stomata closure) and photosynthesis were reduced at different degrees depending on the population. Metabolic homeostasis was proven as there were not differences in the protein, sugar, amino acid, photosynthetic pigments, and phenolic contents between well-hydrated and droughted seedlings. Following these previous results, an integrated transcriptome and proteome analysis was performed on six-old-month seedlings of the Seville population at two times (days 20 and 25) corresponding to a 30 and 50 %, respectively, of leaf fluorescence decrease.

RNA and proteins were independently extracted from the same batch of samples, and transcripts and protein species were analysed by Next-Generation Sequencing (Illumina HiSeq 4000) and shotgun proteomics (LC-MS/MS, Orbitrap). After a search in the *Q. ilex* transcriptome generated during the current Thesis (Chapters 3 and 4; (Guerrero-Sanchez et al., 2017; Guerrero-Sanchez et al., 2019) and the current chapter), 47868 transcripts, corresponding to 21000 unigenes, were identified Table 6.1. From these data, a new improved version of the *Q. ilex* transcriptome has been constructed. It contains a higher number of larger contigs (> 1000, 5000 and 10000 bp) than the previously reported version (Guerrero-Sanchez et al., 2017; Guerrero-Sanchez et al., 2019). Assuming a *Q. ilex* estimated genome size of, approximately, 930 Mb/C with a

total length of 1.87 Gb as assessed by flow cytometry ($2n=2x=24$, Rey et al., 2019), we could conclude a good coverage of the whole exome. The number of proteins identified was one order of magnitude lower, 4008 proteoforms, corresponding to 2737 unigenes, around 10 % of the total expected proteome (Ramírez-Sánchez et al., 2016). These figures are similar to those reported in similar experimental systems (Madritsch et al., 2019; Gugger et al., 2017; Romero-Rodríguez, Jorrín-Novó, and Castillejo, 2019; Liu et al., 2019). The difference between whole transcriptome and proteome coverage can be due to the employed platforms, analytical techniques, and to the physico-chemical and biological properties of both type of biomolecules. Thus, PCR will ensure the detection of low abundant transcripts and differently from proteins, nucleic acids only differ in length, having similar physicochemical properties, being the number of species and the dynamic range lower (Wang et al., 2019a). The advantage of the proteome analysis is, among other considerations (Post-Translational modification, PTMs, interactomics), its close position to the phenotype, hence it is not always possible to jump from the presence of a transcript to the biological function due to posttranscriptional and posttranslational events (Vélez-Bermúdez and Schmidt, 2014).

A relative quantitation at both -omics levels was performed. The dynamic ranges were of four (100-104, in counts per million) and twelve (10-14-10-2, relative peak area) orders of magnitude, which is in the range reported in the literature for the two platforms employed (Schiess, Wollscheid, and Aebersold, 2009; Zhao et al., 2014). For both biomolecules, most of them are in the low abundant range, 0-2x10³ (transcripts), and 10-14-10-3 (proteins). Correlations between mRNA and protein abundance were analysed by the Pearson test (Supplementary Figure S9). The Pearson correlation coefficients were low both in the total transcriptome and proteome and in the variable transcript and protein datasets, indicating no correlation between both omics approaches. This agrees with other studies carried out in plant species (Pan et al., 2012; Li et al., 2016; Xing et al., 2018)). Apart from analytical explanations, biological ones can also be the cause, including, among others, variable translation efficiency of the different transcripts, the mRNA transport to distant tissues, modification and degradation of proteins (Thieme et al., 2015).

Univariate statistical analysis of the data (GLM for transcripts, and t-test for proteins) revealed the existence of significant differences between treatments (control and drought) at the two sampling times (days 20 and 25). The variable dataset was filtered to consistent transcripts and proteins, those present in all the three biological replicates performed. As summarized in Table 6.1, 3579 transcripts (8% of the total) and 640 proteoforms (14 % of the total) responded to the drought treatment, with

qualitative (1116 for RNA, and 220 for proteins) or quantitative (2463 for RNA, and 420 for proteins) changes in abundance. The number of variables at day 25 (2020 for RNA and 347 for proteins) was higher than at day 20 (1559 for RNA, and 293 for proteins). Out of the total variable transcripts and proteins, 726 (RNA) and 219 (proteins) were up accumulated in droughted seedlings, while 1737 (RNA) and 201 (proteins) were down accumulated. These data reveal important and complex gene expression reorganization, as previously reported in similar published papers. Transcript figures presented are higher and lower than those reported for *Q. ilex*, *Q. pubescens*, and *Q. robur* by (Madritsch et al., 2019), and *Q. lobata* by (Gugger et al., 2017). Protein figures are in the range of those reported for other studies on changes in the leaf protein profile in response to drought, as reviewed by (Wang et al., 2016).

Following, the variable gene products will be discussed at the category and individual levels, focusing, firstly, on permanent changes, which are common to the two sampling times, better than time specific ones and, secondly, to those observed at the transcriptomic and proteomic level. Data will be compared with those previously reported in *Q. ilex* and other *Quercus spp.* at the transcriptomic level (Madritsch et al., 2019) and in a recent review by (Wang et al., 2016) on proteomics studies of drought responses in plant leaves. However, an exact comparison is not always possible as the experimental design (plant systems, intensity and duration of the stress) and the methodological approaches employed are different.

Both variable transcripts and proteins were grouped in terms of biological process, molecular function and cellular location (Figure 6.2; Supplementary Table S18). Focusing on the GO categories included both at transcriptome and proteome levels at both sampling times, the following functional ones showed statistically significant differences: i) Up regulated gene products were mostly involved in “response to chemical”, “response to abiotic stimulus”, “response to light stimulus” and “response to temperature stimulus”. ii) Down regulated gene products were mostly involved in “small molecule metabolic process”, “oxidation reduction process”, “response to biotic stimulus”, “responses to heat”, “responses to bacterium”, “response to oxygen-containing compounds”, and “response to chemicals”.

Under the experimental conditions employed in the current work, water restriction under high temperature and illumination, some clear tendencies can be observed, such as an up regulation of the responses to abiotic stresses functional categories, and a down regulation of responses to biotic stimulus, and metabolism. However, there are components of the same functional category that are up or down regulated, as is the case of “response to chemical” (Supplementary Table S18). In similar studies

on the responses to drought in *Q. ilex* and other *Quercus spp.* (Gugger et al., 2017; Madritsch et al., 2019), also showed that differentially expressed gene dataset was enriched in response to stimulus/stress GO terms. Similar studies with other plant species also got to similar conclusions (Shan et al., 2018; Wang et al., 2019b). With respect to the cellular component categories, differential abundant transcripts and proteins, were mostly located at the different chloroplastic and mitochondrial fractions. Both organelles play an important role in the responses of plant to stresses, being the communication and crosstalk with the nucleus, the retrograde signalling, of pivotal role in the correct function of the plant cells (Zhao et al., 2018).

In order to simplify the dataset of variable transcripts and proteins and to obtain an integrative vision of the metabolic changes, a KEGG pathway analysis using PaintOmics 3 was employed (<http://www.paintomics.org/>; (Fàbregas et al., 2018)). From the confidence statistical parameters, the following significant variable pathways were identified: ribosome and protein translation, primary (photosynthesis, cysteine and methionine, glyoxylate and dicarboxylate) and secondary (shikimate, phenylpropanoid and extension to flavonoids, stilbenoid and terpenoids) metabolism. Following each pathway will be briefly discussed.

Common variable gene products belonging to the ribosome protein translation included 36 (transcripts) and 11 (proteins), with only one, down regulated *RPS2*, detected at two sampling times. *RPS2* encodes a structural component of the mitochondrial ribosome small subunit, implicated in the signalling cascade of the immune response in plant (Kim et al., 2009). Out of variable transcripts, 7 were up and 3 down accumulated. Within the first several members of the RPS and RPL family were included, which encode different ribosomal proteins. Apart from having an essential role in protein synthesis, and consequently, in metabolism, growth and development, it has been reported their participation in the response to stress with up or down regulation depending on the type of stress, intensity and duration (Wang et al., 2013). Apart from ribosomal proteins, other genes included in KEGG within the ribosome category, were identified, such as *UBQ1* that encodes for an ubiquitin extension protein. In *Arabidopsis*, it has been reported as an early-responsive to dehydration (Callis, Raasch, and Vierstra, 1990). Together with *GADPH*, *UBQ1* has been proposed as a marker of sugarcane genotypes with different tolerance to drought (Andrade et al., 2017).

It is widely known that photosynthesis is probably most affected pathway in under stress conditions. Drought causes a closure of stomata, and a decrease of available CO₂, that together with photo-oxidative damage, alters and decreases the photosynthetic

activity (Yang et al., 2006). Four (up) and 7 (down) gene products changes were common to both sampling time. Out of them, only one, up regulated *PSB28*, was detected by two platforms. *PSB28* encodes a PSII reaction centre protein. It has been implicated, together with other member of PSB family, in the response to high-light intensity (Parrine et al., 2018). The list of up accumulated transcripts included other members of the photosystem proteins (*PPL1*, *PSBY*, *PSAG*), and *ATPC2*, encoding for ATP synthase gamma chain 2, chloroplast, putative. Enhanced stability of thylakoid membrane proteins contributes to drought stress tolerance, as previously reported for some wheat mutants (Tian et al., 2013). Down accumulation of seven gene products was observed under drought conditions, most of them belonging to PSB and PSA families. The decrease of these proteins has been proven by immunoblotting analysis in *Arabidopsis* (Chen et al., 2016). A striking recurrent result is once more observed, which is how different members of the same family are up and down accumulated. This is the case of PSAs where *PSAG* is up and *PSAB* and *PSAC* are down, since it is impossible to conclude on phenotype (activity of pathways) based on descriptive omics data. So, it needed the functional validation of data using microscopic and classical biochemistry and physiology.

Amino acid metabolism is also an enriched pathway in response to drought. It is related to its role in the biosynthesis of proteins, GSH, secondary metabolism, compatible osmolytes, ethylene, and their usage as a carbohydrate alternative respiratory substrate under reduced photosynthesis activity. In the current study, we have observed changes in the metabolism of amino acids, particularly of the sulphur containing, cysteine and methionine, pathways. Common changes to the two times included one up and six down accumulated gene products only detected at the transcriptome level. The up accumulated was *CYSD1* encoding for a cysteine synthase CYSD1. Cysteine is the precursor of glutathione, the major antioxidant biomolecules whose induction is a common feature of the plant responses to oxidative damage and stresses (Ahmad et al., 2016). Within the down accumulated transcripts, *SAM1*, *SAMDC*, *ACS1*, encoding for s-adenosylmethionine synthase, decarboxylase, and 1-aminocyclopropane-1-carboxylate synthase, respectively, are implicated in the biosynthesis of the hormone ethylene. Although mostly linked to biotic stresses, it has been also implicated in responses to abiotic ones. Different members of the SAM family have organ and stress specific expression patterns, so it is difficult to conclude from the transcriptomic data to the metabolic phenotype (Wang, Oh, and Komatsu, 2016).

The glyoxylate pathway comprises those reactions occurring in plants that convert Acetyl-CoA, as a product of fatty-acid beta oxidation and amino acid catabolism

in succinyl-CoA, which is used in the biosynthesis of carbohydrates. This pathway becomes relevant in situations of photosynthesis inhibition, as a source of sugars. This could be the case of drought and other water related stresses. In our system, a number of variable transcripts common to the both sampling times were up accumulated under drought conditions. The list includes genes encoding for glyoxylate/hydroxypyruvate reductase (HPR3) and Dihydrolipoyl dehydrogenase 1, chloroplastic (LPD1) (Ebeed et al., 2018).

Included in different pathways, the malate dehydrogenase gene, encoding a chloroplastic NADP-dependent MDH (EC 1.1.1.82) was down accumulated under drought conditions. It is a key enzyme controlling the malate valve, which “plays an essential role in the regulation of catalase activity and the accumulation of a hydrogen peroxide-dependent signal by transmitting the redox state of the chloroplast to other cell compartments” (Heyno et al., 2014).

Different enzymes of the secondary metabolic pathways have been altered under drought conditions including shikimate, phenylpropanoid and its extension to flavonoids, stilbenoids, and terpenoids. While some are up (*HCT* and *FAH1*), other (*4CL2*) was down accumulated. Changes in the phenolic patterns are a common phenomenon well-reported since long in different plant systems as a response to biotic and abiotic stresses (Cheynier et al., 2013; Sharma et al., 2019). In our system, *HCT* that encodes a hydroxycinnamoyl-Coenzyme A shikimate/quinate hydroxycinnamoyltransferase both synthesizing and catabolizing the hydroxycinnamoyl esters (coumaroyl/caffeoyl shikimate and quinate) was up accumulated (Carocha et al., 2015). Previous studies have demonstrated that this gene is induced by drought and other abiotic stresses (low temperature and high salinity) (Sun, Yang, and Tzen, 2018). On the other hand, isogen 2 of the 4-coumarate: CoA ligase (*4CL2*) was down accumulated. The different expression pattern for isogenes has been reported for this enzyme. Thus, in *Bassica napus*, it has been reported the organ specific up accumulation in root and leaves of isoforms 1 and 5 of this enzyme, 4CL1 and 4CL5 (Liu et al., 2015).

With respect to the changes observed at the individual gene product level, a short discussion of the most relevant and more related to the responses to drought stress is presented, starting by the qualitative changes between treatments (control and drought). At the transcriptomic level, at day 20, droughted plants were enriched in 15 members of the gene expression group, including ribosomal proteins, transcription factors related to auxin and gibberellin hormones, a maturase K, and a RNase P (*RPL4*, *RPL34*, *AIL1*, *RPS13*, *ARF17*, *RPL5*, *PRORP1*, *NLP8*, *AT2G45640*, *TBP1*, *TPR4*, *GR1*, *NLP3*, *RPL29*, and *MATK*), plus two members of the ALA family, 3

and 8, ATPases implicated in phospholipid transfer. The synthesis of phospholipids involved in membrane stabilization and stress signalling is one of the reported responses to drought (Zhang, Xu, and Huang, 2019). At day 25, new transcripts appeared, belonging to different cellular biology processes and metabolic pathways related to drought responses, including starch, cellulose, and cell wall polysaccharides (*DBE1*, *SS4*, *DPE1*, *CSLG3*, and *XTH5*) (Thalmann and Santelia, 2017), photosynthesis (*PSBR*, *PSBO1*, *RBL*) (Pelloux et al., 2001; Suja and Parida, 2008; Pawłowicz, Kosmala, and Rapacz, 2012), chlorophyll catabolism (*PPD*, *NOL*) (Foulkes et al., 2004; Sato, Ito, and Tanaka, 2015). Droughted plants at day 20 were enriched in stress responsive gene products when proteomics data were analysed. The list of proteins included: heat-shock proteins (HSP17; (Sewelam et al., 2019)), RubisCO activase (RCA, (Ji et al., 2012)), AT3G01520 (Kim et al., 2015), tubulin (TUBA 6; TUBA genes can be used as housekeeping in gene expression analysis, (Montilla-Bascón et al., 2017)), proteasome (RPT 2A; (Stone, 2019)), thioredoxin (TRX 2; (Cha et al., 2014)), auxin-related (PIN3; (Jiang, Li, and Qu, 2017)). The enrichment of stress-related proteins in droughted plants was more evident at day 25. The list includes at least three members of the HSP family, 17, 26.5, and 83 (Mishra et al., 2018), RubisCO activase, (RCA (Chen et al., 2015)), thioredoxin, TRX 9 (Cha et al., 2014), RBG4, with a possible role in RNA-transcription or processing during stress (Kwak, Kim, and Kang, 2005), related to proteasome (CSN1), ribosomal protein RPS12 (Moin et al., 2017). Some of the proteins are related to ABA (RAB, (Alqurashi et al., 2018)) or auxin (CHY1; (Gonçalves et al., 2019)). Two proteins of the shikimate-phenolic pathways, CS1, and CAD (Kim, Bae, and Huh, 2010; Guo et al., 2018), and a metalloprotease (EGY3) were also included in the list.

Changes in gene product abundance also include those that are absent in stress, but present in control conditions, although they are not always considered in the current literature. In the present -omics study there were a few of transcripts or proteins that fit in this criterium (Tables S15 and S16). The list included also members of multigene families of the functional groups also detected in stressed seedlings, as for example cellular process, gene expression, protein metabolism. Several functional groups and genes were common at both days 20 and 25. It was noticeable that at day 20, some components of the ribosomal proteins RPL and RPS large families were absent (Moin, 2017). Also, some stress related gene products were only detected in control plants (*CRT3*, *MED15A*, *EXL1*, *YDA*, *ARK1*, *BBD1*, *ADH*, *TPS1*, *RPL13*, *FRK1*, *LAC2*, *HSI2*, *CRK1*, *ACS1*, *AT2G47470*, *CA1*, *PAP1*, *CRRSP38*, *RPL12*, *RPL24*, *ACO2*,

RAC7, CYC1, DME, CPK1, CLT3, DND1, COL1, EFR, CAS, BIP, ERF096, PK5, BZIP1, EXL3, HSP83, PP7, IAA9, and RPS12).

With respect to protein abundance, new families appeared with members within the filter absent in drought and present in control at day 20, such as cell death (RIN4, CRY1), defence response to bacterium (RCA, RIN4). Others that contained members that were only detected under stress came out, including pigment metabolic process (ACSF, CRY1), translation and other biosynthetic processes (MYB3, ABA2, RPS24A, TCP20, ACSF, GRF2, and RPS8). Two of the genes, ABA2 and CRY1, were included within the group of response to water deprivation, a result that would not be expected (Endo et al., 2014; D'Amico-Damião and Carvalho, 2018). However, it has been reported that ABA2 transcript did not change in response to ABA or dehydration in rice (Endo et al., 2014). At day 25, several proteins included in different metabolic pathways were only detected in control, being absent in drought treatment. The set include enzymes of the sugar metabolism CYT1 and PMM, implicated in the biosynthesis of mannose, L-ascorbic, and nucleotide-sugars, biosynthesis of small organic molecules, including carboxylic acids, oxylipins, (LOX3, ABA2, CYT1, PMM, 4CLL9, among others, Supplementary Tables S15 and S16).

Time dependent regulation of the gene expression may operate independently at the transcriptional or translational levels. As recurrently discussed along this section, differences may be methodological or biologically explained, so, for a confident interpretation of the data focus should be given to permanent better than transitory differences. There were not common permanent (days 20 and 25) qualitative changes observed at the two -omics levels. Both criteria should be used in searching for markers of drought tolerant related genes. The following criterium of confidence established was those genes products up or down regulated at both sampling times. Following, specific transcripts and proteins are discussed. At the mRNA level, 9 were only present in drought treatment at both times, that corresponded to accession: *glysoja_034204*, *OSIGBa0096P03.5*, *RFC1*, *PRUPE_ppa020167mg*, *PERK2*, *TCM_026400*, *contig-8784000021*, *At1g56120*, and *PGSC0003DMG400030893*. Out of them, three corresponded to receptors, including *contig-2451000013* and *contig-7688000013*, putatives LRR receptor-like serine/threonine-protein kinase, and *contig-1900000023*, a member of the proline-rich receptor-like protein kinase family. Members of the Ser/Thr PK have been reported as ABA-receptors, being implicated in drought and salt tolerance in soybean (Sun et al., 2013). In *A. thaliana* roots, PERK4 has been reported as a regulator of Ca²⁺ signalling that is required for ABA responses (Bai et al., 2009). For the other transcripts, neither annotated function nor relationship with drought

has been reported, including *contig-1473000003* that matched *OSIGBa0096P03.5*, *contig-5570000008*, that matched RFC1, a replication factor subunit 1 (“it Plays a role as mediator of transcriptional gene silencing (TGS), DNA replication, DNA repair, hypersensitive response (HR) and telomere length regulation; (Liu et al., 2010)), *contig-1259000017* that matched *PRUPE_ppa020167mg*, *contig-6159000022*, a sulphate/thiosulfate import ATP-binding protein *cysA*, and *contig-8784000021*, a putative U4/U6.U5 tri-snRNP-associated protein 1-like.

Following gene products showing quantitative differences between treatments will be discussed. The total variable dataset was filtered for those that according to the sPLS and PCA multivariate tests do more explain the variability, discriminating between well-watered and droughted seedlings (Table 6.2, Supplementary Figure S11). Among them, those common to the two times, with the highest and lowest loadings, showing positive and negative correlations will be included. There were not gene products detected at both transcriptomic and proteomics levels. The list of gene products detected at the transcriptomic level and up accumulated in droughted seedlings was *DG1* encodes a pentatricopeptide repeat containing protein that is targeted to the chloroplast (Chi et al., 2008) and *TOC159* is an integral membrane GTPase that functions as a transit-sequence receptor required for the import of proteins necessary for chloroplast biogenesis (Chang et al., 2017). On the other hand, some of the down accumulated gene products under drought conditions were genes encoding for enzymes associated to phenolic metabolism (*C3H*, *HCT* and *LAR*), and related to plant immunity (Ankyrin repeat family protein) (Yang et al., 2012). *CYP98A3* encodes coumarate 3-hydroxylase (*C3H*) involved in both lignin and flavonoids biosynthesis (Kim et al., 2006; Varbanova et al., 2011) and *LAR* encodes a pinorensinol reductase involved in lignan biosynthesis (Nakatsubo et al., 2008). Several gene products detected at the proteomic level and up accumulated under drought conditions were identified in the sPLS of the variable transcripts and proteins. Of these, two heat shock proteins (*HSP17.4* and *HSP70*) and one chaperon (*cp10*) were identified. HSP family considered as stress-inducible genes respond to abiotic stresses such as drought (Cho and Hong, 2004). *CP10* encode a chloroplast-localized chaperonin 10 (Koumoto et al., 2001). *SQDB* encode an uridine 5'-diphosphate-sulfoquinovose synthase and is related to the glycolipid biosynthesis (Shimajima and Benning, 2003) and *STIP1* encode a stress-induced-phosphoprotein and it has been identified as an up-regulated protein in alfalfa roots in response to water deficit stress (Byung-Hyun, 2016). *M6PR* encodes a NADPH-dependent mannose 6-phosphate reductase and is a key enzyme involved in mannitol biosynthesis that is affected by drought and salt stresses (Carvalho et al., 2014). On the other hand, two

gene products (*SBE3* and *HDR*) were down accumulated under drought conditions. *SBE3* encodes a starch branching enzyme similar to SBE2 in rice (Han et al., 2007) and *HDR* encodes a protein with 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase activity involved in the last step of mevalonate-independent isopentenyl biosynthesis (Hsieh et al., 2014).

GeneMANIA web site contained algorithm has been employed to establish network analysis (Franz et al., 2018). For plants, it is based on Arabidopsis and has been previously used in the multi-omics analysis in for example in the analysis of the cell wall signalling proteins (Ihsan et al., 2017). Other alternative algorithms have been more employed, such as STRING (Dai et al., 2015; Escandón et al., 2020). The election of GeneMANIA was merely empirical based on the resulting networks and its interpretation in terms of response to drought. It is based on prediction, co-expression, co-localization, and shared protein domains. Two times specific, days 20 and 25, networks were obtained from variable gene products detected at two omics levels. The first one corresponding to day 20 contained 11 up and 4 down regulated genes plus 10 predicted genes (Figure 6.3). Within this network, we have found the previously discussed drought responsive proteins including up accumulated heat shock proteins (CLPB2, CLPB3, HSP70, and HSP17.4), FTSH6, AT1G23740 (*AOR* gene), SMT1, and UGP3, and down accumulated ABA2, RPS1, ADK, and RPL4. Some of the genes have been not previously reported to drought stress; as it is the case of *SMT1* that controls the levels of cholesterol in plants and that could be a factor of the drought response at the membrane level as determine its fluidity (Hartmann, 1998). The second one corresponding to day 25 contained 10 up and 1 down regulated genes plus 20 predicted genes (Figure 6.3). The network is enriched in cellular response to DNA damage/DNA repair and response to heat. Some of them also appeared in the day 20 network (*CLPB2*, *CLPB3*, *HSP70*, *HSP17.4*, *FTSH6*, *AT1G23740*, and *ABA2*). The ones only appeared at day 25 included the up accumulated *AP1*, *INVE*, *AT4G22740*, *CAD4*, *FEN1*, and *HIPP27* genes. Details of genes that are not described above can be found below. FTSH6 encodes an ATP-dependent zinc metalloprotease, reported to be the major thylakoid membrane protease implicated in the biogenesis of thylakoid membranes, quality control in the photosystem II repair cycle, and retrograde signalling mechanism (Kato, Hyodo, and Sakamoto, 2018). CLPB (casein lytic proteinase/heat shock protein) are chaperones that act to remodel or disassemble protein complexes and/or aggregates using the energy of ATP (Lee et al., 2007). AT1G23740 (*AOR* gene) is an oxidoreductase that helps to maintain the photosynthetic process by detoxifying reactive carbonyls formed during lipid peroxidation (Yamauchi et al., 2012). *UGP3* encodes for a UDP-

glucose phosphorylase and is involved in anthocyanins biosynthesis in *Arabidopsis* (Rajpal et al., 2019). *ABA2* encodes a cytosolic short-chain dehydrogenase/reductase involved in the conversion of xanthoxin to ABA-aldehyde during ABA biosynthesis (Lin et al., 2007). *ADK1* encodes an adenosine kinase 1 and is associated with the protein and receptor kinase group of drought tolerance (Sarwar et al., 2019). *INVE* encodes a chloroplast-targeted alkaline/neutral invertase implicated in the development of the photosynthetic apparatus. This is enzyme included in the sucrose synthesis and degradation in the carbohydrate metabolism that is up regulated under drought stress (Shaar-Moshe, Hübner, and Peleg, 2015). *AT4G22740* encodes a glycine-rich protein that is involved in response to drought stress as previously described in *Arabidopsis* (Yao et al., 2016). *CAD4* encodes a catalytically active cinnamyl alcohol dehydrogenase involved in the biosynthesis of lignin (Goujon et al., 2003). *HIPP27* encodes a heavy metal transport/detoxification superfamily protein. HIPP family is induced during cold, salt and drought stress (Barth et al., 2009). So far, no clear relationships between AP1 and FEN1, and drought have been described. *AP1* and *FEN1* encode a MADS domain protein homologous to SRF transcription factors and 5'-3' exonuclease family protein, respectively (Busch, Bomblies, and Weigel, 1999; Zhang et al., 2016a).

In conclusion, because of the huge amount of variable gene products (47868 transcripts and 4008 proteins), complexity of the results and low correlation between transcripts and proteins, only the variable dataset has been analysed based on functional groups and pathways according to GO (biological process, molecular functions and cellular component) and KEGG. Even so, it is important to realise that the work presented should be considered as a descriptive analysis of difficult biological interpretation until functional validation. Variable gene products were transitory, observed in one of the two sampling times, or permanent, observed at two sampling times. As we are also searching molecular markers, the discussion was limited, as a general rule, to those observed at both days 20 and 25. For a confident biological interpretation of the data in terms of response and tolerance to drought stress and to search molecular markers, we pretended to focus on those gene products detected at the two omics levels, but it was almost impossible because the number of gene products was too low compared to the total dataset (6 gene products). The low correlation observed can be explained from a methodological or biological point of view, considering the particularities and characteristics of the two employed analytical platforms and the complexity of the gene expression regulation including transcriptional, post-transcriptional, translational or post-translational events, as well as the stability, degradation of the RNA and proteins. These conclusions can be also applied to similar studies where a multi-omics

approach is employed to study different biological processes. We have observed both up and down regulation of the different groups paying attention at those with more abundant transcripts and proteins under drought conditions, as within them, the search of molecular markers must be performed. The up and down changes in response to drought is of difficult interpretation from a biological point of view as the same category is up and down, with members of the same or close gene families included in one or the other. The situation is even more complicated for isozymes located in different subcellular fractions.

In addition to the analysis of groups, individual gene product analysis has also been performed. For that, the total variable dataset has been filtered to those consistent (present in all the three biological replicates) observed at the two sampling times, with the same up or down tendency and, if any, detected at transcriptomic and proteomic levels. Firstly, qualitative changes, with no statistical analysis, were considered. Quantitative variables were filtered based on univariate (GLM for transcripts, and t-test for proteins) and multivariate (PCA and sPLS) analysis. As it has been discussed above the biological interpretation based on the omics data is not always possible, even being quite restrictive in the confidence parameters. Finally, within the qualitative and quantitative variable gene products commonly observed at both sampling times and both omics levels, we could propose as molecular markers of response and tolerance to drought stress: CLPB2, CLPB3, FTSH6 and PSB28.

Chapter 7

General Conclusions

7.1 Conclusions

1. A reference transcriptome for *Quercus ilex* has been constructed. It contains 47868 annotated transcripts being present in different organs (leaves, roots and embryos), either constitutive or induced, in response to drought stress.
2. The employment of complementary NGS platforms (Illumina and Ion Torrent) and assembling algorithms (MIRA > TRINITY > RAY) resulted in a deeper transcriptome coverage, with longer sequences.
3. From the transcriptome a custom *Q. ilex* specific protein database has been constructed to be used in the proteomic analysis and protein identification in this species.
4. The integrated multiomic analysis resulted in the metabolic pathways reconstruction as it occurs in *Q. ilex*. Out of 127 metabolic pathways reported in KEGG, 123 could be visualized at the transcriptome, proteome and/or metabolome levels. the TCA cycle was the pathway most represented with 5 out of 10 metabolites, 6 out of 8 protein enzymes, and 8 out of 8 enzyme transcripts.
5. Drought causes changes in the leaf transcript and protein profiles, resulting in 3588 mRNA and 640 proteoforms showing qualitative or quantitative differences in abundance. Up or down changes were observed, being transitory in just sampling time, or permanent, at both sampling times.
6. A low correlation (lower than 0.2) between mRNA and protein abundances was obtained
7. Responses to stress and chloroplast were the gene ontology groups more enriched in variable gene products.
8. The KEGG pathways more affected in responses to drought were photosynthesis, protein translation, carbohydrates, amino-acid metabolism and phenolics.
9. Variable gene products detected at both omics levels, being up-accumulated at both both sampling times, can be proposed as molecular markers of responses and tolerance to drought in *Q. ilex*. The list of genes fitting in these characteristics induced CLPB2, CLPB3, FTSH6 and PSB28.

7.2 Conclusiones

1. Se ha construido un transcriptoma de referencia para *Q. ilex*, conteniendo 47868 transcritos anotados. Las secuencias se obtuvieron a partir de muestras de RNA de diferentes órganos (hojas, raíces y embriones), siendo estos expresados de manera constitutiva o inducida en respuesta a sequía.
2. El empleo de dos plataformas complementarias de secuenciación de última generación (Illumina e Ion Torrent) junto con los algoritmos de ensamblaje de “contigs” (RAY, TRINITY y MIRA) generaron secuencias más largas y una mayor cobertura del exoma.
3. A partir de dicho transcriptoma de referencia, se ha construido una base de datos de proteínas específica para *Q. ilex*, que será utilizada en la identificación de proteínas en estudios de proteómicas.
4. El uso integrado de las diferentes plataformas ómicas, transcriptómica, proteómica y metabolómica, permitió la reconstrucción de las rutas metabólicas en *Q. ilex*. Del total de las 127 rutas incluidas en la base de datos KEGG, se visualizaron, a nivel de transcrito, proteínas y/o metabolitos, 123. El ciclo de Krebs fue la ruta más representada con 5 de los 10 metabolitos, 6 de los 8 enzimas y el total de los transcritos, 8, identificados.
5. El tratamiento de sequía provocó grandes cambios en el patrón de expresión génica, detectándose 3588 mRNAs y 640 proteoformas como variables, presentado diferencias cualitativas o cuantitativas. Dichos cambios fueron transitorios (detectados en un solo tiempo) o permanentes (detectados a los dos tiempos).
6. La correlación entre la abundancia a nivel de transcritos y la correspondiente proteína fue muy baja (valores inferiores a 0.2 en el coeficiente de correlación de Pearson).
7. Los grupos génicos más alterados en condiciones de sequía fueron, desde un punto de vista funcional, los de la “respuesta a estrés”, y en cuanto a su localización celular los “cloroplásticos”.
8. Los productos génicos variables correspondientes a enzimas correspondieron a las siguientes rutas metabólicas: síntesis de proteínas, fotosíntesis, carbohidratos, aminoácidos y fenólicos.

9. Los productos génicos variables detectados a ambos niveles ómicos y sobreexpresados a ambos tiempos de muestreo pueden ser propuestos como marcadores moleculares de respuesta y tolerancia a sequía en *Q. ilex*. La lista de genes que cumplen dichos criterios fue muy corta, incluyendo hipotéticas proteínas de respuesta a choque térmico, CPLB2 y CPLB3, una metaloproteasa cloroplástica, FTSH6, y una proteína del centro de reacción del fotosistema II, PSB28.

References

- Abril, Nieves et al. (2011). “Proteomics research on forest trees, the most recalcitrant and orphan plant species”. In: *Phytochemistry* 72.10, pp. 1219–1242.
- Ahmad, Nisar et al. (2016). “Drought stress in maize causes differential acclimation responses of glutathione and sulfur metabolism in leaves and roots”. In: *BMC plant biology* 16.1, p. 247.
- Akcan, T. et al. (2017). “Acorn (*Quercus* spp.) as a novel source of oleic acid and tocopherols for livestock and humans: discrimination of selected species from Mediterranean forest”. In: *Journal of Food Science and Technology*, pp. 1–8.
- Al-Rousan, W. M. et al. (2013). “Characterization of Acorn Fruit Oils Extracted from Selected Mediterranean *Quercus* Species”. In: *Grasas y Aceites* 64.5, pp. 554–560.
- Al-Shahrour, Fátima et al. (2004). “FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes”. In: *Bioinformatics* 20.4, pp. 578–580.
- Allen, Craig D et al. (2009). “Climate-induced forest dieback: an escalating global phenomenon”. In: *Unasylva* 231.232, pp. 43–49.
- Alqurashi, May et al. (2018). “Early responses to severe drought stress in the *Arabidopsis thaliana* cell suspension culture proteome”. In: *Proteomes* 6.4, p. 38.
- Altschul, Stephen F et al. (1990). “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3, pp. 403–410.
- Anderson, N Leigh et al. (1998). “Proteome and proteomics: new technologies, new concepts, and new words”. In: *Electrophoresis* 19.11, pp. 1853–1861.
- Andrade, Larissa Mara de et al. (2017). “Reference genes for normalization of qPCR assays in sugarcane plants under water deficit”. In: *Plant Methods* 13.1, p. 28.
- Andrews, Simon (2010). *FASTQC A Quality Control tool for High Throughput Sequence Data*.
- Angel, Thomas E et al. (2012). “Mass spectrometry-based proteomics: existing capabilities and future directions”. In: *Chemical Society Reviews* 41.10, pp. 3912–3928.
- Aranda, Ismael et al. (2015). “Variation in photosynthetic performance and hydraulic architecture across European beech (*Fagus sylvatica* L.) populations supports the case for local adaptation to water stress”. In: *Tree physiology* 35.1, pp. 34–46.

- Asai, Tomonori et al. (2016). “Metabolomic analysis of primary metabolites in citrus leaf during defense responses”. In: *Journal of Bioscience and Bioengineering* 123.3, pp. 376–381.
- Bai, Ling et al. (2009). “Plasma membrane-associated proline-rich extensin-like receptor kinase 4, a novel regulator of Ca²⁺ signalling, is required for abscisic acid responses in *Arabidopsis thaliana*”. In: *The Plant Journal* 60.2, pp. 314–327.
- Barbeta, Adrià et al. (2013). “Dampening effects of long-term experimental drought on growth and mortality rates of a Holm oak forest”. In: *Global change biology* 19.10, pp. 3133–3144.
- Barth, Olaf et al. (2009). “Stress induced and nuclear localized HIP26 from *Arabidopsis thaliana* interacts via its heavy metal associated domain with the drought stress related zinc finger transcription factor ATHB29”. In: *Plant molecular biology* 69.1-2, pp. 213–226.
- Bates, B C et al. (2008). “Climate change and water Technical Paper of the Intergovernmental Panel on Climate Change (Geneva: IPCC Secretariat)”. In: *Climate Change* 95, p. 96.
- Benjamini, Yoav et al. (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.
- Bennett, Simon (2004). “Solexa ltd”. In: *Pharmacogenomics* 5.4, pp. 433–438.
- Bernstein, Frances C et al. (1977). “The Protein Data Bank: a computer-based archival file for macromolecular structures”. In: *Journal of molecular biology* 112.3, pp. 535–542.
- Beynon, RJ (1985). *CABIOS editorial*.
- Biswas, Abhishek et al. (2014). “Big data challenges for estimating genome assembler quality”. In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, pp. 653–660.
- Boisvert, Sébastien et al. (2010). “Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies”. In: *Journal of Computational Biology* 17.11, pp. 1519–1533.
- Bolger, Anthony M. et al. (2014). “Trimmomatic: A flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15, pp. 2114–2120.
- Bradford, Marion M (1976). *A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein-Dye Binding*. Tech. rep., pp. 248–254.
- Bradnam, Keith R et al. (2013). “Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species”. In: *GigaScience* 2.1, p. 10.
- Bragg, Lauren M et al. (2013). “Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data”. In: *PLoS computational biology* 9.4, e1003031.

- Brasier, CM et al. (1993). “Evidence for *Phytophthora cinnamomi* involvement in Iberian oak decline”. In: *Plant pathology* 42.1, pp. 140–145.
- Bray, Nicolas L et al. (2016). “Near-optimal probabilistic RNA-seq quantification”. In: *Nature Biotechnology* 34.5, pp. 525–527.
- Bryant, Donald M et al. (2017). “A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors.” In: *Cell reports* 18.3, pp. 762–776.
- Busch, Maximilian A et al. (1999). “Activation of a floral homeotic gene in *Arabidopsis*”. In: *Science* 285.5427, pp. 585–587.
- Byung-Hyun, Lee (2016). “Proteome analysis of alfalfa roots in response to water deficit stress”. In: *Journal of integrative agriculture* 15.6, pp. 1275–1285.
- Cadahía, Estrella et al. (2015). “Non-targeted metabolomic profile of *Fagus Sylvatica* l. leaves using liquid chromatography with mass spectrometry and gas chromatography with mass spectrometry”. In: *Phytochemical Analysis* 26.2, pp. 171–182.
- Callis, Judy et al. (1990). “Ubiquitin extension proteins of *Arabidopsis thaliana*. Structure, localization, and expression of their promoters in transgenic tobacco.” In: *Journal of Biological Chemistry* 265.21, pp. 12486–12493.
- Cañellas, I. et al. (2007). “An approach to acorn production in Iberian dehesas”. In: *Agroforestry Systems* 70.1, pp. 3–9.
- Cantos, Emma et al. (2003). “Phenolic compounds and fatty acids from acorns (*Quercus* spp.), the main dietary constituent of free-ranging Iberian pigs”. In: *Journal of Agricultural and Food Chemistry* 51.21, pp. 6248–6255.
- Carocha, Victor et al. (2015). “Genome-wide analysis of the lignin toolbox of *Eucalyptus grandis*”. In: *New Phytologist* 206.4, pp. 1297–1313.
- Carvalho, Kenia de et al. (2014). “Homeologous genes involved in mannitol synthesis reveal unequal contributions in response to abiotic stress in *Coffea arabica*”. In: *Molecular genetics and genomics* 289.5, pp. 951–963.
- Casimiro-Soriguer, Carlos S et al. (2017). “Sma3s: a universal tool for easy functional annotation of proteomes and transcriptomes”. In: *Proteomics* 17.12, p. 1700071.
- Castillejo, María Ángeles et al. (2015). “Understanding pea resistance mechanisms in response to *Fusarium oxysporum* through proteomic analysis”. In: *Phytochemistry* 115, pp. 44–58.
- Cha, Joon-Yung et al. (2014). “NADPH-dependent thioredoxin reductase A (NTRA) confers elevated tolerance to oxidative stress and drought”. In: *Plant physiology and biochemistry* 80, pp. 184–191.
- Chang, Jun-Shian et al. (2017). “Chloroplast preproteins bind to the dimer interface of the Toc159 receptor during import”. In: *Plant physiology* 173.4, pp. 2148–2162.
- Chen, Hanbo et al. (2011). “VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R”. In: *BMC bioinformatics* 12.1, p. 35.

- Chen, Yang-Er et al. (2016). “Different response of photosystem II to short and long-term drought stress in *Arabidopsis thaliana*”. In: *Physiologia plantarum* 158.2, pp. 225–235.
- Chen, Yue et al. (2015). “Rubisco activase is also a multiple responder to abiotic stresses in rice”. In: *PLoS One* 10.10.
- Chevreur et al. (1999). “Genome Sequence Assembly Using Trace Signals and Additional Sequence Information”. In: *German conference on bioinformatics* 99.
- Cheyrier, Véronique et al. (2013). “Plant phenolics: recent advances on their biosynthesis, genetics, and ecophysiology”. In: *Plant Physiology and Biochemistry* 72, pp. 1–20.
- Chi, Wei et al. (2008). “The pentatricopeptide repeat protein DELAYED GREENING1 is involved in the regulation of early chloroplast development and chloroplast gene expression in *Arabidopsis*”. In: *Plant Physiology* 147.2, pp. 573–584.
- Chiatante, D et al. (2015). “Interspecific variation in functional traits of oak seedlings (*Quercus ilex*, *Quercus trojana*, *Quercus virgiliana*) grown under artificial drought and fire conditions”. In: *Journal of plant research* 128.4, pp. 595–611.
- Chira, Kleopatra et al. (2014). “Chemical and sensory evaluation of wine matured in oak barrel: effect of oak species involved and toasting process”. In: *European Food Research and Technology* 240.3, pp. 533–547.
- Cho, Eun Kyung et al. (2004). “Molecular cloning and expression pattern analyses of heat shock protein 70 genes from *Nicotiana tabacum*”. In: *Journal of Plant Biology* 47.2, pp. 149–159.
- Chou, Peter Y. et al. (1974). “Conformational Parameters for Amino Acids in Helical, β -Sheet, and Random Coil Regions Calculated from Proteins”. In: *Biochemistry* 13.2, pp. 211–222.
- Clarke, Kaitlin et al. (2013). “Comparative analysis of de novo transcriptome assembly”. In: *Science China Life Sciences* 56.2, pp. 156–162.
- Clooney, Adam G et al. (2016). “Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis”. In: *PloS one* 11.2, e0148028.
- Cohen, David et al. (2010). “Comparative transcriptomics of drought responses in *Populus*: a meta-analysis of genome-wide expression profiling in mature leaves and root apices across two genotypes”. In: *BMC genomics* 11.1, p. 630.
- Colangelo, Michele et al. (2018). “Drought and *Phytophthora* are associated with the decline of oak species in southern Italy”. In: *Frontiers in plant science* 9, p. 1595.
- Collins, WJ et al. (2012). “Global and regional temperature-change potentials for near-term climate forcers”. In: *Atmospheric Chemistry and Physics Discussions* 12.9, pp. 23261–23290.
- Conesa, Ana et al. (2008). “Blast2GO: A comprehensive suite for functional analysis in plant genomics.” In: *International journal of plant genomics* 2008, p. 619832.

- Conesa, Ana et al. (2005). “Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research”. In: *Bioinformatics* 21.18, pp. 3674–3676.
- Corcobado, Tamara et al. (2013). “Quercus ilex forests are influenced by annual variations in water table, soil water deficit and fine root loss caused by *Phytophthora cinnamomi*”. In: *Agricultural and Forest Meteorology* 169, pp. 92–99.
- Correia, Barbara et al. (2016). “Integrated proteomics and metabolomics to unlock global and clonal responses of *Eucalyptus globulus* recovery from water deficit”. In: *Metabolomics* 12.8.
- Correia, Barbara et al. (2018). “Gene expression analysis in *Eucalyptus globulus* exposed to drought stress in a controlled and a field environment indicates different strategies for short-and longer-term acclimation”. In: *Tree physiology* 38.11, pp. 1623–1639.
- Cox, Jürgen et al. (2008). “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification”. In: *Nature Biotechnology* 26.12, pp. 1367–1372.
- Dai, Jiewen et al. (2015). “Irf6-related gene regulatory network involved in palate and lip development”. In: *Journal of Craniofacial Surgery* 26.5, pp. 1600–1605.
- David, Teresa Soares et al. (2007). “Water-use strategies in two co-occurring Mediterranean evergreen oaks: surviving the summer drought”. In: *Tree physiology* 27.6, pp. 793–803.
- Dayhoff, M et al. (1978). “22 a model of evolutionary change in proteins”. In: *Atlas of protein sequence and structure*. Vol. 5. National Biomedical Research Foundation Silver Spring MD, pp. 345–352.
- Dayhoff, Margaret Oakley et al. (1962). “Comprotein”. In: *Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall)*. New York, New York, USA: ACM Press, pp. 262–274.
- De Rigo, D et al. (2016). “*Quercus ilex* in Europe: Distribution, habitat, usage and threats”. In: *European Atlas of Forest Tree Species; European Union: Luxembourg*, pp. 130–131.
- De Wit, Pierre et al. (2012). “The simple fool’s guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis”. In: *Molecular Ecology Resources* 12.6, pp. 1058–1067.
- Devereux, John et al. (1984). “A comprehensive set of sequence analysis programs for the VAX”. In: *Nucleic acids research* 12.1Part1, pp. 387–395.
- Diego, Rafael Hernández-de et al. (2018). “PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data”. In: *Nucleic Acids Research* 46.W1, W503–W509.
- Dobin, Alexander et al. (2013). “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1, pp. 15–21.
- Dohm, Juliane C et al. (2008). “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing”. In: *Nucleic acids research* 36.16, e105.

- Doody, Colin N. et al. (2008). “Drying and soaking pretreatments affect germination in pedunculate oak”. In: *Annals of Forest Science* 65.5, pp. 509–509.
- D’Amico-Damião, Victor et al. (2018). “Cryptochrome-related abiotic stress responses in plants”. In: *Frontiers in plant science* 9, p. 1897.
- Ebeed, Heba T et al. (2018). “Conserved and differential transcriptional responses of peroxisome associated pathways to drought, dehydration and ABA”. In: *Journal of experimental botany* 69.20, pp. 4971–4985.
- Echevarría-Zomeño, Sira et al. (2009). “Changes in the protein profile of *Quercus ilex* leaves in response to drought stress and recovery”. In: *Journal of Plant Physiology* 166.3, pp. 233–245.
- Echevarría-Zomeño, Sira et al. (2012). “Simple, rapid and reliable methods to obtain high quality RNA and genomic DNA from *Quercus ilex* L. leaves suitable for molecular biology studies”. In: *Acta Physiologiae Plantarum* 34.2, pp. 793–805.
- Eid, John et al. (2009). “Real-time DNA sequencing from single polymerase molecules”. In: *Science* 323.5910, pp. 133–138.
- El-Metwally, Sara et al. (2014). *Next generation sequencing technologies and challenges in sequence assembly*. Vol. 7. Springer Science & Business.
- Endo, Akira et al. (2014). “Functional characterization of xanthoxin dehydrogenase in rice”. In: *Journal of plant physiology* 171.14, pp. 1231–1240.
- Escandón, Mónica et al. (2017). “System-wide analysis of short-term response to high temperature in *Pinus radiata*”. In: *Journal of Experimental Botany* 68.13, pp. 3629–3641.
- Escandón, Mónica et al. (2020). “Protein interaction networks: functional and statistical approaches”. In: *In press*.
- Ewing, Brent et al. (1998). “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities”. In: *Genome Research* 8.3, pp. 186–194.
- Fàbregas, Norma et al. (2018). “Overexpression of the vascular brassinosteroid receptor BRL3 confers drought resistance without penalizing plant growth”. In: *Nature communications* 9.1, pp. 1–13.
- Forner, Alicia et al. (2018). “Mediterranean trees coping with severe drought: Avoidance might not be safe”. In: *Environmental and Experimental Botany* 155, pp. 529–540.
- Foulkes, MJ et al. (2004). “Effects of a photoperiod-response gene Ppd-D1 on yield potential and drought resistance in UK winter wheat”. In: *Euphytica* 135.1, pp. 63–73.
- Franz, Max et al. (2018). “GeneMANIA update 2018”. In: *Nucleic acids research* 46.W1, W60–W64.
- Früchtenicht, Elena et al. (2018). “Response of *Quercus robur* and two potential climate change winners—*Quercus pubescens* and *Quercus ilex*—To two years summer drought in a semi-controlled competition study: I—Tree water status”. In: *Environmental and experimental botany* 152, pp. 107–117.

- Furuhashi, Takeshi et al. (2012). “Metabolite changes with induction of *Cuscuta haustorium* and translocation from host plants”. In: *Journal of plant interactions* 7.1, pp. 84–93.
- Galiano, Lucía et al. (2012). “Determinants of drought effects on crown condition and their relationship with depletion of carbon reserves in a Mediterranean holm oak forest”. In: *Tree physiology* 32.4, pp. 478–489.
- Gallego, By F. J. et al. (1999). “Etiology of oak decline in Spain”. In: *Forest Pathology* 29.1, pp. 17–27.
- Garcia-Alcalde, F. et al. (2011). “Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data”. In: *Bioinformatics* 27.1, pp. 137–139.
- Ge, Steven Xijin et al. (2018). “ShinyGO: a graphical enrichment tool for ani-mals and plants”. In: *bioRxiv*, p. 315150.
- Gentilesca, Tiziana et al. (2017). “Drought-induced oak decline in the western Mediterranean region: an overview on current evidences, mechanisms and management options to improve forest resilience”. In: *iForest-Biogeosciences and Forestry* 10.5, p. 796.
- Gil-Pelegrín, Eustaquio et al., eds. (Mar. 26, 2018). *Oaks Physiological Ecology. Exploring the Functional Diversity of Genus Quercus L.* Vol. 7. Tree Physiology. Cham: Springer International Publishing. ISBN: 978-3-319-69098-8.
- Giorgi, Filippo et al. (2008). “Climate change projections for the Mediterranean region”. In: *Global and planetary change* 63.2-3, pp. 90–104.
- Gonçalves, Luana P et al. (2019). “Rootstock-induced molecular responses associated with drought tolerance in sweet orange as revealed by RNA-Seq”. In: *BMC genomics* 20.1, p. 110.
- Gordon, A. et al. (2010). *Fastx-toolkit. Computer program distributed by the author.*
- Goujon, Thomas et al. (2003). “Genes involved in the biosynthesis of lignin precursors in *Arabidopsis thaliana*”. In: *Plant Physiology and Biochemistry* 41.8, pp. 677–687.
- Grabherr, Manfred G et al. (2011). “Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data”. In: *Nature biotechnology* 29.7, pp. 644–52.
- Greenwood, Sarah et al. (2017). “Tree mortality across biomes is promoted by drought intensity, lower wood density and higher specific leaf area”. In: *Ecology Letters* 20.4, pp. 539–553.
- Gudeta, Temesgen Bedassa (2018). “Molecular marker based genetic diversity in forest tree populations”. In: *Forestry Research and Engineering: International Journal* 2.4, pp. 176–182.
- Guerrero-Sanchez, Victor M. et al. (2017). “Holm Oak (*Quercus ilex*) Transcriptome. De novo Sequencing and Assembly Analysis”. In: *Frontiers in Molecular Biosciences* 4.

- Guerrero-Sanchez, Victor M et al. (2019). “Ion Torrent and Illumina, two complementary RNA-seq platforms for constructing the holm oak (*Quercus ilex*) transcriptome.” In: *PLoS one* 14.1. Ed. by Mukesh Jain, e0210356.
- Gugger, Paul F et al. (2017). “Whole-transcriptome response to water stress in a California endemic oak, *Quercus lobata*”. In: *Tree physiology* 37.5, pp. 632–644.
- Guo, Rui et al. (2018). “Metabolic responses to drought stress in the tissues of drought-tolerant and drought-sensitive wheat genotype seedlings”. In: *AOB Plants* 10.2, p1016.
- Gurevich, Alexey et al. (2013). “QUAST: quality assessment tool for genome assemblies”. In: *Bioinformatics* 29.8, pp. 1072–1075.
- Guzmán, Beatriz et al. (2015). “Protected areas of Spain preserve the neutral genetic diversity of *Quercus ilex* L. irrespective of glacial refugia”. In: *Tree genetics & genomes* 11.6, p. 124.
- Haas, Brian J et al. (2013). “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.” In: *Nature protocols* 8.8, pp. 1494–512.
- Hadidi, Lila et al. (2017). “*Quercus ilex* L.: How season, Plant Organ and Extraction Procedure Can Influence Chemistry and Bioactivities”. In: *Chemistry & Biodiversity* 14.1, e1600187.
- Han, Yuepeng et al. (2007). “Three orthologs in rice, *Arabidopsis*, and *Populus* encoding starch branching enzymes (SBEs) are different from other SBE gene families in plants”. In: *Gene* 401.1-2, pp. 123–130.
- Harfouche, Antoine et al. (2014). “Molecular and physiological responses to abiotic stress in forest trees and their relevance to tree improvement”. In: *Tree physiology* 34.11, pp. 1181–1198.
- Hartmann, Marie-Andrée (1998). “Plant sterols and the membrane environment”. In: *Trends in plant science* 3.5, pp. 170–175.
- Hernandez, David et al. (2008). “De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer.” In: *Genome research* 18.5, pp. 802–9.
- Hesper, B et al. (1970). “Bioinformatica: een werkconcept”. In: *Kameleon* 1.6, pp. 28–29.
- Heyno, Eiri et al. (2014). “Putative role of the malate valve enzyme NADP-malate dehydrogenase in H₂O₂ signalling in *Arabidopsis*”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1640, p. 20130228.
- Hoagland, D. R. et al. (1950). “The water-culture method for growing plants without soil.” In: *Circular. California Agricultural Experiment Station* 347.2nd edit.
- Houle, David et al. (2010). “Phenomics: the next challenge”. In: *Nature reviews genetics* 11.12, p. 855.

- Hsieh, Wei-Yu et al. (2014). “Functional evidence for the critical amino-terminal conserved domain and key amino acids of Arabidopsis 4-HYDROXY-3-METHYLBUT-2-ENYL DIPHOSPHATE REDUCTASE”. In: *Plant physiology* 166.1, pp. 57–69.
- Ihaka, Ross et al. (1996). “R: a language for data analysis and graphics”. In: *Journal of computational and graphical statistics* 5.3, pp. 299–314.
- Ihsan, Muhammad Z et al. (2017). “Gene mining for proline based signaling proteins in cell wall of Arabidopsis thaliana”. In: *Frontiers in plant science* 8, p. 233.
- Janda, Martin et al. (2013). “Phosphoglycerolipids are master players in plant hormone signal transduction”. In: *Plant cell reports* 32.6, pp. 839–851.
- Ji, Kuixian et al. (2012). “Drought-responsive mechanisms in rice genotypes with contrasting drought tolerance during reproductive stage”. In: *Journal of plant physiology* 169.4, pp. 336–344.
- Jiang, Zhaoyun et al. (2017). “Auxins”. In: *Hormone Metabolism and Signaling in Plants*. Elsevier Inc., pp. 39–76. ISBN: 9780128115633.
- Joffre, R. et al. (1999). “The dehesa system of southern Spain and Portugal as a natural ecosystem mimic”. In: *Agroforestry Systems* 45.1-3, pp. 57–79.
- Jorge, Inmaculada et al. (2005). “The Holm Oak leaf proteome: Analytical and biological variability in the protein expression level assessed by 2-DE and protein identification tandem mass spectrometry *de novo* sequencing and sequence similarity searching”. In: *PROTEOMICS* 5.1, pp. 222–234.
- Jorge, Inmaculada et al. (2006). “Variation in the holm oak leaf proteome at different plant developmental stages, between provenances and in response to drought stress”. In: *PROTEOMICS* 6.S1, S207–S214.
- Jorrin-Novo, Jesus V (2014). “Plant proteomics methods and protocols”. In: *Plant Proteomics*. Springer, pp. 3–13.
- Jorrín-Novo, Jesús V. et al. (2009). “Plant proteomics update (2007–2008): Second-generation proteomic techniques, an appropriate experimental design, and data analysis to fulfill MIAPE standards, increase plant proteome coverage and expand biological knowledge”. In: *Journal of Proteomics* 72.3, pp. 285–314.
- Jorrín-Novo, Jesus V. et al. (2015). “Fourteen years of plant proteomics reflected in *Proteomics* : Moving from model species and 2DE-based approaches to orphan species and gel-free platforms”. In: *PROTEOMICS* 15.5-6, pp. 1089–1112.
- Jung, Hyungtaek et al. (2019). “Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes”. In: *Trends in plant science*.
- Kalli, Anastasia et al. (2013). “Evaluation and Optimization of Mass Spectrometric Settings during Data-Dependent Acquisition Mode: Focus on LTQ- Orbitrap Mass Analyzers”. In: *Journal of Proteome Research* 12.7, pp. 3071–3086.
- Kamboj, JS et al. (1999). “Identification and quantitation by GC-MS of zeatin and zeatin riboside in xylem sap from rootstock and scion of grafted apple trees”. In: *Plant Growth Regulation* 28.3, pp. 199–205.

- Kato, Yusuke et al. (2018). “The photosystem II repair cycle requires FtsH turnover through the EngA GTPase”. In: *Plant physiology* 178.2, pp. 596–611.
- Kim, Daehwan et al. (2015). “HISAT: A fast spliced aligner with low memory requirements”. In: *Nature Methods* 12.4, pp. 357–360.
- Kim, Do Jin et al. (2015). “Crystal structure of the protein A t3g01520, a eukaryotic universal stress protein-like protein from arabidopsis thaliana in complex with AMP”. In: *Proteins: Structure, Function, and Bioinformatics* 83.7, pp. 1368–1373.
- Kim, Min Gab et al. (2009). “The Pseudomonas syringae type III effector AvrRpm1 induces significant defenses by activating the Arabidopsis nucleotide-binding leucine-rich repeat protein RPS2”. In: *The Plant Journal* 57.4, pp. 645–653.
- Kim, Young-Hwa et al. (2010). “Transcriptional regulation of the cinnamyl alcohol dehydrogenase gene from sweetpotato in response to plant developmental stage and environmental stress”. In: *Plant cell reports* 29.7, pp. 779–791.
- Kim, Young Jin et al. (2006). “Wound-induced expression of the ferulate 5-hydroxylase gene in *Camptotheca acuminata*”. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1760.2, pp. 182–190.
- Koren, Sergey et al. (2012). “Hybrid error correction and de novo assembly of single-molecule sequencing reads”. In: *Nature biotechnology* 30.7, p. 693.
- Koumoto, Yasuko et al. (2001). “Chloroplasts have a novel Cpn10 in addition to Cpn20 as co-chaperonins in *Arabidopsis thaliana*”. In: *Journal of Biological Chemistry* 276.32, pp. 29688–29694.
- Kwak, Kyung Jin et al. (2005). “Characterization of transgenic Arabidopsis plants overexpressing GR-RBP4 under high salinity, dehydration, or cold stress”. In: *Journal of Experimental Botany* 56.421, pp. 3007–3016.
- Langmead, Ben et al. (2012). “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4, pp. 357–359.
- Lee, Ung et al. (2007). “The Arabidopsis ClpB/Hsp100 family of proteins: chaperones for stress and chloroplast development”. In: *The Plant Journal* 49.1, pp. 115–127.
- Lesur, Isabelle et al. (2015). “The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery of genes regulated during bud dormancy release”. In: *BMC Genomics* 16.1, p. 112.
- Li, Bo et al. (2014). “Evaluation of de novo transcriptome assemblies from RNA-Seq data”. In: *Genome biology* 15.12, p. 553.
- Li, H. et al. (2009). “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14, pp. 1754–1760.
- Li, Jiajia et al. (2016). “Differential proteomics analysis to identify proteins and pathways associated with male sterility of soybean using iTRAQ-based strategy”. In: *Journal of proteomics* 138, pp. 72–82.
- Li, Weizhong et al. (2001). “Clustering of highly homologous sequences to reduce the size of large protein databases”. In: *Bioinformatics* 17.3, pp. 282–283.

- (2002). “Tolerating some redundancy significantly speeds up clustering of large protein databases”. In: *Bioinformatics* 18.1, pp. 77–82.
- Li, Yifeng et al. (2018). “A review on machine learning principles for multi-view biological data integration”. In: *Briefings in bioinformatics* 19.2, pp. 325–340.
- Limousin, Jean-Marc et al. (2010). “Do photosynthetic limitations of evergreen *Quercus ilex* leaves change with long-term increased drought severity?” In: *Plant, Cell & Environment* 33.5, pp. 863–875.
- Lin, Pei-Chi et al. (2007). “Ectopic expression of ABSCISIC ACID 2/GLUCOSE INSENSITIVE 1 in *Arabidopsis* promotes seed dormancy and stress tolerance”. In: *Plant physiology* 143.2, pp. 745–758.
- Liu, Chunqing et al. (2015). “Comparative analysis of the *Brassica napus* root and leaf transcript profiling in response to drought stress”. In: *International journal of molecular sciences* 16.8, pp. 18752–18777.
- Liu, Qian et al. (2010). “DNA replication factor C1 mediates genomic stability and transcriptional gene silencing in *Arabidopsis*”. In: *The Plant Cell* 22.7, pp. 2336–2352.
- Liu, Yan et al. (2019). “Physiological and Proteomic Responses of Mulberry Trees (*Morus alba*. L.) to Combined Salt and Drought Stress”. In: *International journal of molecular sciences* 20.10, p. 2486.
- Lohse, Marc et al. (2014). “Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data”. In: *Plant, Cell and Environment* 37.5, pp. 1250–1258.
- López-Hidalgo, Cristina (2017). “Análisis metabolómico de especies forestales, la encina (*Quercus ilex*)”. PhD thesis. Universidad de Córdoba, p. 46.
- López-Hidalgo, Cristina et al. (2018). “A Multi-Omics Analysis Pipeline for the Metabolic Pathway Reconstruction in the Orphan Species *Quercus ilex*”. In: *Frontiers in Plant Science* 9, p. 935.
- Love, Michael I. et al. (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12, p. 550.
- Lowe, Rohan et al. (2017). “Transcriptomics technologies”. In: *PLoS computational biology* 13.5, e1005457.
- Ma, Bin et al. (2003). “PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry”. In: *Rapid communications in mass spectrometry* 17.20, pp. 2337–2342.
- Madritsch, Silvia et al. (2019). “Elucidating drought stress tolerance in European oaks through cross-species transcriptomics”. In: *G3: Genes, Genomes, Genetics* 9.10, pp. 3181–3199.
- Margulies, Marcel et al. (2005a). “Genome sequencing in microfabricated high-density picolitre reactors”. In: *Nature* 437.7057, p. 376.

- Margulies, Marcel et al. (2005b). “Genome sequencing in microfabricated high-density picolitre reactors.” In: *Nature* 437.7057, pp. 376–80.
- Marti, Angel Fernández i et al. (2018). “Population genetic diversity of *Quercus ilex* subsp. *ballota* (Desf.) Samp. reveals divergence in recent and evolutionary migration rates in the Spanish dehesas”. In: *Forests* 9.6, p. 337.
- Martin, Marcel (2011). “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1, p. 10.
- Martínez, María Teresa et al. (2019). “Holm Oak Somatic Embryogenesis: Current Status and Future Perspectives”. In: *Frontiers in Plant Science* 10, p. 239.
- Mathesius, Ulrike et al. (2001). “Establishment of a root proteome reference map for the model legume *Medicago truncatula* using the expressed sequence tag database for peptide mass fingerprinting”. In: *PROTEOMICS: International Edition* 1.11, pp. 1424–1440.
- Meijón, Mónica et al. (2016). “Exploring natural variation of *Pinus pinaster* Aiton using metabolomics: Is it possible to identify the region of origin of a pine from its metabolites?” In: *Molecular Ecology* 25.4, pp. 959–976.
- Meir, Patrick et al. (2015). “Drought-related tree mortality: addressing the gaps in understanding and prediction”. In: *New Phytologist* 207.1, pp. 28–33.
- Menezes-Silva, Paulo Eduardo et al. (2019). “Different ways to die in a changing world: Consequences of climate change for tree species performance and survival through an ecophysiological perspective”. In: *Ecology and evolution* 9.20, pp. 11979–11999.
- Mikheyev, Alexander S et al. (2014). “A first look at the Oxford Nanopore MinION sequencer”. In: *Molecular ecology resources* 14.6, pp. 1097–1102.
- Mishra, Bagdevi et al. (2018). “A reference genome of the European beech (*Fagus sylvatica* L.)” In: *Gigascience* 7.6, giy063.
- Moin, Mazahar et al. (2017). “Expression profiling of ribosomal protein gene family in dehydration stress responses and characterization of transgenic rice plants overexpressing RPL23A for water-use efficiency and tolerance to drought and salt stresses”. In: *Frontiers in chemistry* 5, p. 97.
- Montilla-Bascón, Gracia et al. (2017). “Reduced nitric oxide levels during drought stress promote drought tolerance in barley and is associated with elevated polyamine biosynthesis”. In: *Scientific reports* 7.1, pp. 1–15.
- Moran, Emily et al. (2017). “The genetics of drought tolerance in conifers”. In: *New Phytologist* 216.4, pp. 1034–1048.
- Moreno, Gerardo et al. (2009). “The functioning, management and persistence of dehesas”. In: *Agroforestry in Europe*. Springer, pp. 127–160.
- Morin, X. et al. (2007). “Variation in cold hardiness and carbohydrate concentration from dormancy induction to bud burst among provenances of three European oak species”. In: *Tree Physiology* 27.6, pp. 817–825.

- Müller, Markus et al. (2019). “Abiotic genetic adaptation in the Fagaceae”. In: *Plant Biology* 21.5, pp. 783–795.
- Munoz-Mérida, ANTONIO et al. (2014). “Sma3s: a three-step modular annotator for large sequence datasets”. In: *DNA research* 21.4, pp. 341–353.
- Nakatsubo, Tomoyuki et al. (2008). “Characterization of *Arabidopsis thaliana* pinorensinol reductase, a new type of enzyme involved in lignan biosynthesis”. In: *Journal of Biological Chemistry* 283.23, pp. 15550–15557.
- Natalini, Fabio et al. (2016). “The role of climate change in the widespread mortality of holm oak in open woodlands of Southwestern Spain”. In: *Dendrochronologia* 38, pp. 51–60.
- Needleman, S B et al. (1970). “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” In: *Journal of molecular biology* 48.3, pp. 443–53.
- Neilson, Karlie A. et al. (2011). “Less label, more free: Approaches in label-free quantitative mass spectrometry”. In: *Proteomics* 11.4, pp. 535–553.
- Oliveros, Juan Carlos (2007). “VENNY. An interactive tool for comparing lists with Venn Diagrams”. In: <http://bioinfogp.cnb.csic.es/tools/venny/index.html>.
- Pages, H et al. (2009). “String objects representing biological sequences, and matching algorithms”. In: *R package version 2.2*.
- Pan, Zhiyong et al. (2012). “An integrative analysis of transcriptome and proteome provides new insights into carotenoid biosynthesis and regulation in sweet orange fruits”. In: *Journal of proteomics* 75.9, pp. 2670–2684.
- Parrine, Débora et al. (2018). “Proteome modifications on tomato under extreme high light induced-stress”. In: *Proteome science* 16.1, p. 20.
- Pascual, Jesús et al. (2017). “Integrated Physiological, Proteomic, and Metabolomic Analysis of Ultra Violet (UV) Stress Responses and Adaptation Mechanisms in *Pinus radiata*”. In: *Molecular & Cellular Proteomics* 16.3, pp. 485–501.
- Pasho, Edmond et al. (2011). “Impacts of drought at different time scales on forest growth across a wide climatic gradient in north-eastern Spain”. In: *Agricultural and Forest Meteorology* 151.12, pp. 1800–1811.
- Pathan, Mohashin et al. (2015). “FunRich: An open access standalone functional enrichment and interaction network analysis tool”. In: *Proteomics* 15.15, pp. 2597–2601.
- Patón, Daniel et al. (2009). “Influence of climate on radial growth of holm oaks (*Quercus ilex* subsp. *ballota* Desf) from SW Spain”. In: *Geochronometria* 34.1, pp. 49–56.
- Pawłowicz, Izabela et al. (2012). “Expression pattern of the psbO gene and its involvement in acclimation of the photosynthetic apparatus during abiotic stresses in *Festuca arundinacea* and *F. pratensis*”. In: *Acta physiologiae plantarum* 34.5, pp. 1915–1924.

- Pearson, W R et al. (1988). “Improved tools for biological sequence comparison.” In: *Proceedings of the National Academy of Sciences of the United States of America* 85.8, p. 2444.
- Pelloux, J et al. (2001). “Changes in Rubisco and Rubisco activase gene expression and polypeptide content in *Pinus halepensis* M. subjected to ozone and drought”. In: *Plant, Cell & Environment* 24.1, pp. 123–131.
- Peña-Rojas, Karen et al. (2004). “Stomatal limitation to CO₂ assimilation and down-regulation of photosynthesis in *Quercus ilex* resprouts in response to slowly imposed drought”. In: *Tree physiology* 24.7, pp. 813–822.
- Peñuelas, Josep et al. (2017). “Impacts of global change on Mediterranean forests and their services”. In: *Forests* 8.12, p. 463.
- (2018). “Assessment of the impacts of climate change on Mediterranean terrestrial ecosystems based on data from field experiments and long-term monitored field gradients in Catalonia”. In: *Environmental and Experimental Botany* 152, pp. 49–59.
- Pieczynski, Marcin et al. (2018). “Genomewide identification of genes involved in the potato response to drought indicates functional evolutionary conservation with *Arabidopsis* plants”. In: *Plant biotechnology journal* 16.2, pp. 603–614.
- Plieninger, Tobias et al. (2004). “Effects of land-use and landscape structure on holm oak recruitment and regeneration at farm level in *Quercus ilex* L. dehesas”. In: *Journal of Arid Environments* 57.3, pp. 345–364.
- Plomion, Christophe et al. (2018). “Oak genome reveals facets of long lifespan”. In: *Nature Plants* 4.7, p. 440.
- Pluskal, Tomáš et al. (2010). “MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data”. In: *BMC bioinformatics* 11.1, p. 395.
- Polle, Andrea et al. (2019). “Engineering drought resistance in forest trees”. In: *Frontiers in plant science* 9, p. 1875.
- Porth, Ilga et al. (2014). “Assessment of the genetic diversity in forest tree populations using molecular markers”. In: *Diversity* 6.2, pp. 283–295.
- Pulido, Fernando J et al. (2001). “Size structure and regeneration of Spanish holm oak *Quercus ilex* forests and dehesas: effects of agroforestry use on their long-term sustainability”. In: *Forest Ecology and Management* 146.1-3, pp. 1–13.
- Quail, Michael A et al. (2012). “A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers”. In: *BMC genomics* 13.1, p. 341.
- Rabhi, Faten et al. (2016). “Sterol, aliphatic alcohol and tocopherol contents of *Quercus ilex* and *Quercus suber* from different regions”. In: *Industrial Crops and Products* 83, pp. 781–786.
- Rajpal, Vijay Rani et al. (2019). *Genetic Enhancement of Crops for Tolerance to Abiotic Stress: Mechanisms and Approaches, Vol. I*. Springer.

- Rakić, S. et al. (2006). “Oak acorn, polyphenols and antioxidant activity in functional food”. In: *Journal of Food Engineering* 74.3, pp. 416–423.
- Ramírez-Sánchez, Obed et al. (2016). “Plant proteins are smaller because they are encoded by fewer exons than animal proteins”. In: *Genomics, proteomics & bioinformatics* 14.6, pp. 357–370.
- Ramos, António Marcos et al. (2018). “The draft genome sequence of cork oak”. In: *Scientific data* 5, p. 180069.
- Rennenberg, H et al. (2007). “Sulfur metabolism in plants: are trees different?” In: *Plant Biology* 9.05, pp. 620–637.
- Rey, María-Dolores et al. (2019). “Proteomics, Holm Oak (*Quercus ilex* L.) and Other Recalcitrant and Orphan Forest Tree Species: How do They See Each Other?” In: *International journal of molecular sciences* 20.3, p. 692.
- Rhee, Seung Yon et al. (2006). “Bioinformatics and its applications in plant biology”. In: *Annu. Rev. Plant Biol.* 57, pp. 335–360.
- Rice, P et al. (2000). “EMBOSS: the European Molecular Biology Open Software Suite.” In: *Trends in genetics : TIG* 16.6, pp. 276–7.
- Rico, L et al. (2014). “Changes in DNA methylation fingerprint of *Quercus ilex* trees in response to experimental field drought simulating projected climate change”. In: *Plant Biology* 16.2, pp. 419–427.
- Ritchie, Matthew E. et al. (2015). “Limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43.7, e47.
- Robinson, Mark D et al. (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” In: *Bioinformatics (Oxford, England)* 26.1, pp. 139–40.
- Rodríguez-Calcerrada, Jesús et al. (2018). “A molecular approach to drought-induced reduction in leaf CO₂ exchange in drought-resistant *Quercus ilex*”. In: *Physiologia plantarum* 162.4, pp. 394–408.
- Rohart, Florian et al. (2017). “mixOmics: An R package for ‘omics feature selection and multiple data integration”. In: *PLoS computational biology* 13.11, e1005752.
- Roman, DT et al. (2015). “The role of isohydric and anisohydric species in determining ecosystem-scale response to severe drought”. In: *Oecologia* 179.3, pp. 641–654.
- Romero-Rodríguez, M. Cristina (2015). “Integrated “ omics ” approaches to study non - orthodox seed germination : the case of holm oak (*Quercus ilex* subsp . *ballota* [Desf .] Samp) María Cristina Romero Rodríguez Doctoral Thesis”. In: p. 234.
- Romero-Rodríguez, M. Cristina et al. (2014). “Improving the quality of protein identification in non-model species. Characterization of *Quercus ilex* seed and *Pinus radiata* needle proteomes by using SEQUEST and custom databases”. In: *Journal of Proteomics* 105, pp. 85–91.

- Romero-Rodríguez, M Cristina et al. (2018). “Germination and early seedling development in *Quercus ilex* recalcitrant and non-dormant seeds: targeted transcriptional, hormonal, and sugar analysis”. In: *Frontiers in plant science* 9, p. 1508.
- Romero-Rodríguez, María Cristina et al. (2019). “Toward characterizing germination and early growth in the non-orthodox forest tree species *Quercus ilex* through complementary gel and gel-free proteomic analysis of embryo and seedlings”. In: *Journal of proteomics* 197, pp. 60–70.
- Rosas, Teresa et al. (2013). “Dynamics of non-structural carbohydrates in three Mediterranean woody species following long-term experimental drought”. In: *Frontiers in plant science* 4, p. 400.
- Rossum, Guido van et al. (1991). “Interactively testing remote servers using the Python programming language”. In: *CWi Quarterly* 4.4, pp. 283–303.
- RStudio, Team (2016). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio.
- Ruiz Gómez, Francisco et al. (2018). “Differences in the response to acute drought and *Phytophthora cinnamomi* Rands Infection in *Quercus ilex* L. seedlings”. In: *Forests* 9.10, p. 634.
- Ruosteenoja, Kimmo et al. (2018). “Seasonal soil moisture and drought occurrence in Europe in CMIP5 projections for the 21st century”. In: *Climate dynamics* 50.3-4, pp. 1177–1192.
- Salomón, Roberto L et al. (2017). “Stem hydraulic capacitance decreases with drought stress: implications for modelling tree hydraulics in the Mediterranean oak *Quercus ilex*”. In: *Plant, cell & environment* 40.8, pp. 1379–1391.
- San Eufasio, Bonoso et al. (2020). “Responses to drought stress and differences in tolerance at seedling level in *Quercus* spp., and Andalusian *Q. ilex* populations”. In: *Under review*.
- Sanchez, M. E. et al. (2002). “*Phytophthora* disease of *Quercus ilex* in south-western Spain”. In: *Forest Pathology* 32.1, pp. 5–18.
- Sardans, J. et al. (2013). “Metabolic responses of *Quercus ilex* seedlings to wounding analysed with nuclear magnetic resonance profiling”. In: *Plant Biology* 16, pp. 395–403.
- Sardans, Jordi et al. (2010). “Changes in water content and distribution in *Quercus ilex* leaves during progressive drought assessed by in vivo ¹H magnetic resonance imaging”. In: *BMC plant biology* 10.1, p. 188.
- Sarwar, Muhammad Bilal et al. (2019). “De novo assembly of *Agave sisalana* transcriptome in response to drought stress provides insight into the tolerance mechanisms”. In: *Scientific reports* 9.1, pp. 1–14.
- Sato, Rei et al. (2015). “Chlorophyll b degradation by chlorophyll b reductase under high-light conditions”. In: *Photosynthesis research* 126.2-3, pp. 249–259.

- Schäfer, KVR et al. (2011). “The physical environment within forests”. In: *Nat Educ Knowl* 2.12, p. 5.
- Schiess, Ralph et al. (2009). “Targeted proteomic strategy for clinical biomarker discovery”. In: *Molecular oncology* 3.1, pp. 33–44.
- Schneider, Maria V. et al. (2011). “Omics Technologies, Data and Bioinformatics Principles”. In: *Methods in Molecular Biology*. Humana Press, pp. 3–30.
- Schrimpe-Rutledge, Alexandra C. et al. (2016). “Untargeted Metabolomics Strategies???Challenges and Emerging Directions”. In: *Journal of the American Society for Mass Spectrometry* 27.12, pp. 1897–1905.
- Schwalm, Christopher R et al. (2017). “Global patterns of drought recovery”. In: *Nature* 548.7666, pp. 202–205.
- Serrano, Lydia et al. (2005). “Tissue-water relations of two co-occurring evergreen Mediterranean species in response to seasonal and experimental drought conditions”. In: *Journal of plant research* 118.4, pp. 263–269.
- Sewelam, Nasser et al. (2019). “The AtHSP17. 4C1 Gene Expression Is Mediated by Diverse Signals that Link Biotic and Abiotic Stress Factors with ROS and Can Be a Useful Molecular Marker for Oxidative Stress”. In: *International journal of molecular sciences* 20.13, p. 3201.
- Sghaier-Hammami, Besma et al. (2013). “Physiological and proteomics analyses of Holm oak (*Quercus ilex* subsp. *ballota* [Desf.] Samp.) responses to *Phytophthora cinnamomi*”. In: *Plant Physiology and Biochemistry* 71, pp. 191–202.
- Sghaier-Hammami, Besma et al. (2016). “Protein profile of cotyledon, tegument, and embryonic axis of mature acorns from a non-orthodox plant species: *Quercus ilex*”. In: *Planta* 243.2, pp. 369–396.
- Shaar-Moshe, Lidor et al. (2015). “Identification of conserved drought-adaptive genes using a cross-species meta-analysis approach”. In: *BMC plant biology* 15.1, p. 111.
- Shan, Zhongying et al. (2018). “Physiological and proteomic analysis on long-term drought resistance of cassava (*Manihot esculenta* Crantz)”. In: *Scientific reports* 8.1, pp. 1–12.
- Shannon, Paul et al. (2003). “Cytoscape: A software Environment for integrated models of biomolecular interaction networks”. In: *Genome Research* 13.11, pp. 2498–2504.
- Sharma, Anket et al. (2019). “Response of phenylpropanoid pathway and the role of polyphenols in plants under abiotic stress”. In: *Molecules* 24.13, p. 2452.
- Sharma, Rinku et al. (2018). “Comparative transcriptome meta-analysis of *Arabidopsis thaliana* under drought and cold stress”. In: *PloS one* 13.9.
- Shimajima, Mie et al. (2003). “Native uridine 5'-diphosphate-sulfoquinovose synthase, SQD1, from spinach purifies as a 250-kDa complex”. In: *Archives of biochemistry and biophysics* 413.1, pp. 123–130.
- Silva, Jeffrey C et al. (2006). “Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition”. In: *Molecular & Cellular Proteomics* 5.1, pp. 144–156.

- Simão, Felipe A et al. (2015). “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs”. In: *Bioinformatics* 31.19, pp. 3210–3212.
- Simova-Stoilova, Lyudmila P. et al. (2015). “2-DE proteomics analysis of drought treated seedlings of *Quercus ilex* supports a root active strategy for metabolic adaptation in response to water shortage”. In: *Frontiers in Plant Science* 6, p. 627.
- Simova-Stoilova, Lyudmila P et al. (2018). “Holm oak proteomic response to water limitation at seedling establishment stage reveals specific changes in different plant parts as well as interaction between roots and cotyledons”. In: *Plant science* 276, pp. 1–13.
- Sleep, Julie A et al. (2013). “Sequencing error correction without a reference genome”. In: *BMC bioinformatics* 14.1, p. 367.
- Sperlich, Dominik et al. (2016). “Balance between carbon gain and loss under long-term drought: impacts on foliar respiration and photosynthesis in *Quercus ilex* L”. In: *Journal of experimental botany* 67.3, pp. 821–833.
- Staden, Rodger (1979). “A strategy of DNA sequencing employing computer programs”. In: *Nucleic acids research* 6.7, pp. 2601–2610.
- Stein, Lincoln (2001). “Genome annotation: From sequence to biology”. In: *Nature Reviews Genetics* 2.7, pp. 493–503.
- Stone, Sophia L. (2019). “Role of the Ubiquitin Proteasome System in Plant Response to Abiotic Stress”. In: *International Review of Cell and Molecular Biology* 343, pp. 65–110.
- Suja, G et al. (2008). “Isolation and characterization of photosystem 2 PsbR gene and its promoter from drought-tolerant plant *Prosopis juliflora*”. In: *Photosynthetica* 46.4, pp. 525–530.
- Sun, Chi-Hui et al. (2018). “Molecular Identification and Characterization of Hydroxycinnamoyl Transferase in Tea Plants (*Camellia sinensis* L.)” In: *International journal of molecular sciences* 19.12, p. 3938.
- Sun, XiaoLi et al. (2013). “A Glycine soja ABA-responsive receptor-like cytoplasmic kinase, GsRLCK, positively controls plant tolerance to salt and drought stresses”. In: *Planta* 237.6, pp. 1527–1545.
- Swamy, BP Mallikarjuna et al. (2013). “Genomics-based precision breeding approaches to improve drought tolerance in rice”. In: *Biotechnology advances* 31.8, pp. 1308–1318.
- Szklarczyk, Damian et al. (2015). “STRING v10: Protein-protein interaction networks, integrated over the tree of life”. In: *Nucleic Acids Research* 43.D1, pp. D447–D452.
- Szuba, Agnieszka et al. (2017). “Field proteomics of *Populus alba* grown in a heavily modified environment – An example of a tannery waste landfill”. In: *Science of The Total Environment* 610.Supplement C, pp. 1557–1571.
- Tarkka, Mika T. et al. (2013). “OakContigDF159.1, a reference library for studying differential gene expression in *Quercus robur* during controlled biotic inter-

- actions: use for quantitative transcriptomic profiling of oak roots in ectomycorrhizal symbiosis”. In: *New Phytologist* 199.2, pp. 529–540.
- Team, R Core (2018). *R: A language and environment for statistical computing; 2015*.
- Thalmann, Matthias et al. (2017). “Starch as a determinant of plant fitness under abiotic stress”. In: *New Phytologist* 214.3, pp. 943–951.
- Thieme, Christoph J et al. (2015). “Endogenous Arabidopsis messenger RNAs transported to distant tissues”. In: *Nature Plants* 1.4, p. 15025.
- Tian, Fengxia et al. (2013). “Enhanced stability of thylakoid membrane proteins and antioxidant competence contribute to drought stress resistance in the *tasg1* wheat stay-green mutant”. In: *Journal of experimental botany* 64.6, pp. 1509–1520.
- Tian, He et al. (2016). “Metabolomics, a powerful tool for agricultural research”. In: *International Journal of Molecular Sciences* 17.11.
- Törönen, Petri et al. (2018). “PANNZER2: A rapid functional annotation web server”. In: *Nucleic Acids Research* 46.W1, W84–W88.
- Torvalds, Linus (1991). *Linux kernel 0.01*.
- Usadel, Björn et al. (2009). “A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize”. In: *Plant, Cell & Environment* 32.9, pp. 1211–1229.
- Valero-Galván, José et al. (2011). “Studies of variability in Holm oak (*Quercus ilex* subsp. *ballota* [Desf.] Samp.) through acorn protein profile analysis”. In: *Journal of Proteomics* 74.8, pp. 1244–1255.
- Valero-Galván, José et al. (2012). “Population variability based on the morphometry and chemical composition of the acorn in Holm oak (*Quercus ilex* subsp. *ballota* [Desf.] Samp.)” In: *European Journal of Forest Research* 131.4, pp. 893–904.
- Valero-Galván, José et al. (2012). “Proteomic analysis of Holm oak (*Quercus ilex* subsp. *ballota* [Desf.] Samp.) pollen”. In: *Journal of Proteomics* 75.9, pp. 2736–2744.
- Valero-Galván, José et al. (2013). “Physiological and Proteomic Analyses of Drought Stress Response in Holm Oak Provenances”. In: *Journal of Proteome Research* 12.11, pp. 5110–5123.
- Valledor, Luis et al. (2014). “A universal protocol for the combined isolation of metabolites, DNA, long RNAs, small RNAs, and proteins from plants and microorganisms”. In: *Plant Journal* 79.1, pp. 173–180.
- Varbanova, Marina et al. (2011). “Molecular and biochemical basis for stress-induced accumulation of free and bound p-coumaraldehyde in cucumber”. In: *Plant physiology* 157.3, pp. 1056–1066.
- Vaz, Margarida et al. (2011). “Leaf-level responses to light in two co-occurring *Quercus* (*Quercus ilex* and *Quercus suber*): leaf structure, chemical composition and photosynthesis”. In: *Agroforestry Systems* 82.2, pp. 173–181.

- Vélez-Bermúdez, Isabel Cristina et al. (2014). “The conundrum of discordant protein and mRNA expression. Are plants special?” In: *Frontiers in plant science* 5, p. 619.
- Viant, Mark R. et al. (2017). “How close are we to complete annotation of metabolomes?” In: *Current Opinion in Chemical Biology* 36, pp. 64–69.
- Vicente, Ángel Martín et al. (2006). “Long Term Persistence of Dehesas. Evidences from History”. In: *Agroforestry Systems* 67.1, pp. 19–28.
- Vicente, Eduardo et al. (2018). “Water Balance of Mediterranean *Quercus ilex* L. and *Pinus halepensis* Mill. Forests in Semiarid Climates: A Review in A Climate Change Context”. In: *Forests* 9.7, p. 426.
- Vidman, Linda et al. (2019). “Cluster analysis on high dimensional RNA-seq data with applications to cancer research-An evaluation study”. In: *BioRxiv*, p. 675041.
- Villar-Salvador, Pedro et al. (2004). “Drought tolerance and transplanting performance of holm oak (*Quercus ilex*) seedlings after drought hardening in the nursery”. In: *Tree physiology* 24.10, pp. 1147–1155.
- Vinha, A. F. et al. (2016a). “Chemical and antioxidant profiles of acorn tissues from *Quercus* spp.: Potential as new industrial raw materials”. In: *Industrial Crops and Products* 94, pp. 143–151.
- Vinha, Ana F. et al. (2016b). “A New Age for *Quercus* spp. Fruits: Review on Nutritional and Phytochemical Composition and Related Biological Activities of Acorns”. In: *Comprehensive Reviews in Food Science and Food Safety* 15.6, pp. 947–981.
- Wang, Dongxue et al. (2019a). “A deep proteome and transcriptome abundance atlas of 29 healthy human tissues”. In: *Molecular systems biology* 15.2.
- Wang, Jinyan et al. (2013). “Expression changes of ribosomal proteins in phosphate- and iron-deficient *Arabidopsis* roots predict stress-specific alterations in ribosome composition”. In: *BMC genomics* 14.1, p. 783.
- Wang, X et al. (2016). “Characterization of S-adenosylmethionine synthetases in soybean under flooding and drought stresses”. In: *Biologia Plantarum* 60.2, pp. 269–278.
- Wang, Xiaoli et al. (2016). “Drought-Responsive Mechanisms in Plant Leaves Revealed by Proteomics.” In: *International journal of molecular sciences* 17.10, p. 1706.
- Wang, Xuan et al. (2019b). “Comparative proteomics and physiological analyses reveal important maize filling-kernel drought-responsive genes and metabolic pathways”. In: *International journal of molecular sciences* 20.15, p. 3743.
- Wang, Zhong et al. (2009). “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nature reviews genetics* 10.1, p. 57.
- Warde-Farley, David et al. (2010). “The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function”. In: *Nucleic acids research* 38.suppl_2, W214–W220.

- Warren, Charles R. et al. (2012). “Metabolomics demonstrates divergent responses of two Eucalyptus species to water stress”. In: *Metabolomics* 8.2, pp. 186–200.
- Waterhouse, Robert M et al. (2017). “BUSCO applications from quality assessments to gene prediction and phylogenomics”. In: *Molecular biology and evolution* 35.3, pp. 543–548.
- Weirather, Jason L et al. (2017). “Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis”. In: *F1000Research* 6.
- Xing, Miaomiao et al. (2018). “Integrated analysis of transcriptome and proteome changes related to the Ogura cytoplasmic male sterility in cabbage”. In: *PLoS One* 13.3.
- Yamauchi, Yasuo et al. (2012). “Chloroplastic NADPH-dependent alkenal/one oxidoreductase contributes to the detoxification of reactive carbonyls produced under oxidative stress”. In: *FEBS letters* 586.8, pp. 1208–1213.
- Yang, Xinghong et al. (2006). “Tolerance of photosynthesis to photoinhibition, high temperature and drought stress in flag leaves of wheat: A comparison between a hybridization line and its parents grown under field conditions”. In: *Plant Science* 171.3, pp. 389–397.
- Yang, Yuanai et al. (2012). “The ankyrin-repeat transmembrane protein BDA1 functions downstream of the receptor-like protein SNC2 to regulate plant immunity”. In: *Plant physiology* 159.4, pp. 1857–1865.
- Yao, LM et al. (2016). “Overexpression of a glycine-rich protein gene in *Lablab purpureus* improves abiotic stress tolerance”. In: *Genet. Mol. Res* 15.
- Zerbino, Daniel R. et al. (2008). “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs”. In: *Genome Research* 18.5, pp. 821–829.
- Zhang, Jixiang et al. (2016a). “Requirement for flap endonuclease 1 (FEN 1) to maintain genomic stability and transcriptional gene silencing in Arabidopsis”. In: *The Plant Journal* 87.6, pp. 629–640.
- Zhang, Runxuan et al. (2017). “A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing”. In: *Nucleic acids research* 45.9, pp. 5061–5073.
- Zhang, Tianlei et al. (2016b). “Current trends and innovations in bioanalytical techniques of metabolomics”. In: *Critical reviews in analytical chemistry* 46.4, pp. 342–351.
- Zhang, Xiexiang et al. (2019). “Lipidomic reprogramming associated with drought stress priming-enhanced heat tolerance in tall fescue (*Festuca arundinacea*)”. In: *Plant, cell & environment* 42.3, pp. 947–958.
- Zhao, Chenchen et al. (2018). “Roles of chloroplast retrograde signals and ion transport in plant drought tolerance”. In: *International journal of molecular sciences* 19.4, p. 963.

- Zhao, Shanrong et al. (2014). “Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells”. In: *PloS one* 9.1.
- Zhu, Mengmeng et al. (2016). “Molecular and systems approaches towards drought-tolerant canola crops”. In: *New Phytologist* 210.4, pp. 1169–1189.
- Zybailov, Boris et al. (2006). “Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*”. In: *Journal of proteome research* 5.9, pp. 2339–2347.

Appendix A

Supplementary Material

A.1 Ion Torrent and Illumina, two complementary RNA-seq platforms for constructing the Holm oak (*Quercus ilex*) transcriptome. Supporting information

Table S1 Total number of transcripts included in the GO and Uniprot classification in holm oak. Accessible in this link:

<https://doi.org/10.1371/journal.pone.0210356.s001>

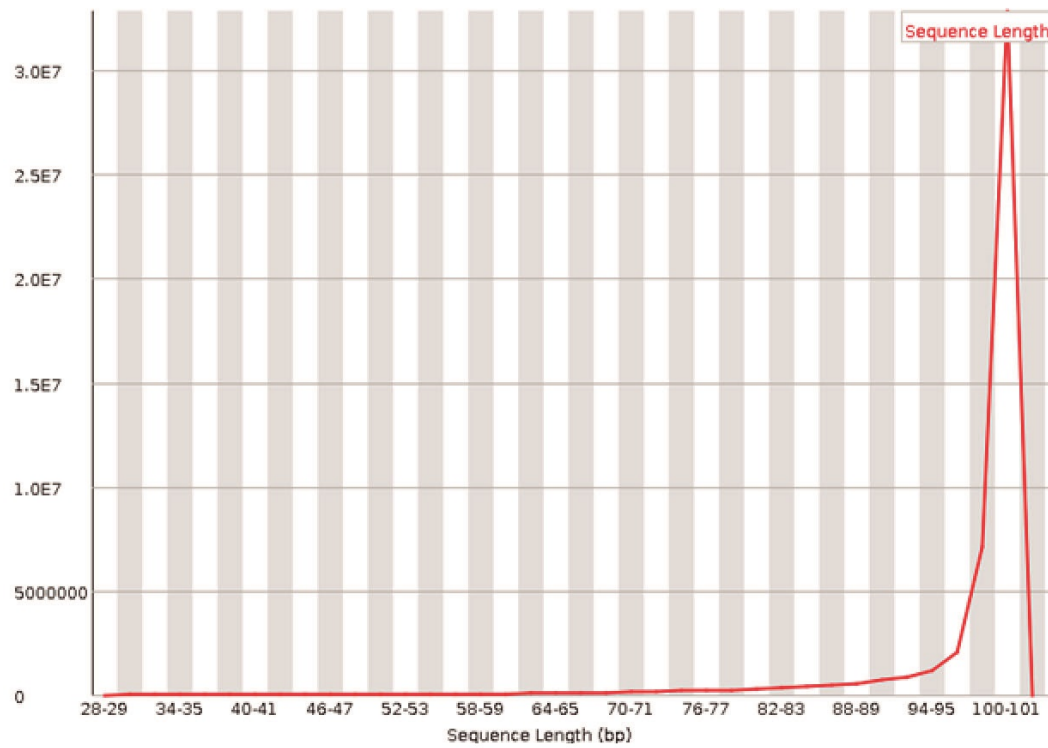
Table S2 List of transcripts related to drought stress in the holm oak transcriptome.

Abbreviation	Gene name	EC (enzyme code)
Bpm	BPM	
At3g62550	Drought responsive ATP-binding motif containing protein	
dreb5	Drought responsive element binding protein 5	
	Drought-induced protein RDI	
drs1	Drought-sensitive 1 protein	
drs1	Drought-sensitive 1 protein	
drs1	Drought-sensitive 1 protein	
drs1	Drought-sensitive 1 protein	
drs1	Drought-sensitive 1 protein	
GT1	Glycosyltransferase	2.4.1.-;2.4.1.115;2.4.1.91
GT1	Glycosyltransferase	2.4.1.-;2.4.1.115;2.4.1.91
NAC13	NAC domain class transcription factor	1.11.1.7
NAC014	NAC domain-containing protein 14	1.11.1.7
NAC078	NAC domain-containing protein 78	
NAC078	NAC domain-containing protein 78	
NAC078	NAC domain-containing protein 78	
NAC078	NAC domain-containing protein 78	
AIF	NAC domain-containing protein 78	
AIF	NAC domain-containing protein 78	4.6.1.1
NAC082	NAC domain-containing protein 82	
NAC082	NAC domain-containing protein 82	
NAC086	NAC domain-containing protein 86	1.11.1.7
NAC1	NAC transcription factor	
NTL5	NAC transcription factor NTL5	1.11.1.7
NAM	No apical meristem	1.11.1.7
Os05g0109600	Os05g0109600 protein	
pal	Phenylalanine ammonia-lyase	4.3.1.25;2.3.3.10;4.3.1.24
PAL	Phenylalanine ammonia-lyase	4.3.1.25;2.3.3.10;4.3.1.24
PAL	Phenylalanine ammonia-lyase	4.3.1.25;2.3.3.10;4.3.1.24
PXG4	Probable peroxygenase 4	1.11.2.3
NTL9	Protein NTM1-like 9	1.11.1.7
NTL9	Protein NTM1-like 9	1.11.1.7
APUM2	Pumilio homolog 2	
APUM5	Pumilio homolog 5	
Bpm	Pumilio isogeny 2	
puf2	Pumilio isogeny 2	
Bpm	Pumilio isogeny 2	
puf2	Pumilio isogeny 2	
DICP	Putative drought-inducible cysteine proteinase	3.4.22.-;3.4.22.16
NAC1	Putative membrane bound NAC transcription factor 1	
NAC77	Putative NAC domain class transcription factor	
TCTP	Translationally-controlled tumor protein homolog	
TCTP	Translationally-controlled tumor protein homolog	2.5.1.47
TCTP	Translationally-controlled tumor protein homolog	2.5.1.47
UGT71K1	UDP-glycosyltransferase 71K1	2.4.1.-;2.4.1.115;2.4.1.91
UGT71K1	UDP-glycosyltransferase 71K1	2.4.1.-;2.4.1.115;2.4.1.91

Table S3 Total number of transcripts included in the GO and Uniprot classification in holm oak. Accessible in this link:

<https://doi.org/10.1371/journal.pone.0210356.s003>

a) Illumina reads



b) Ion Torrent reads

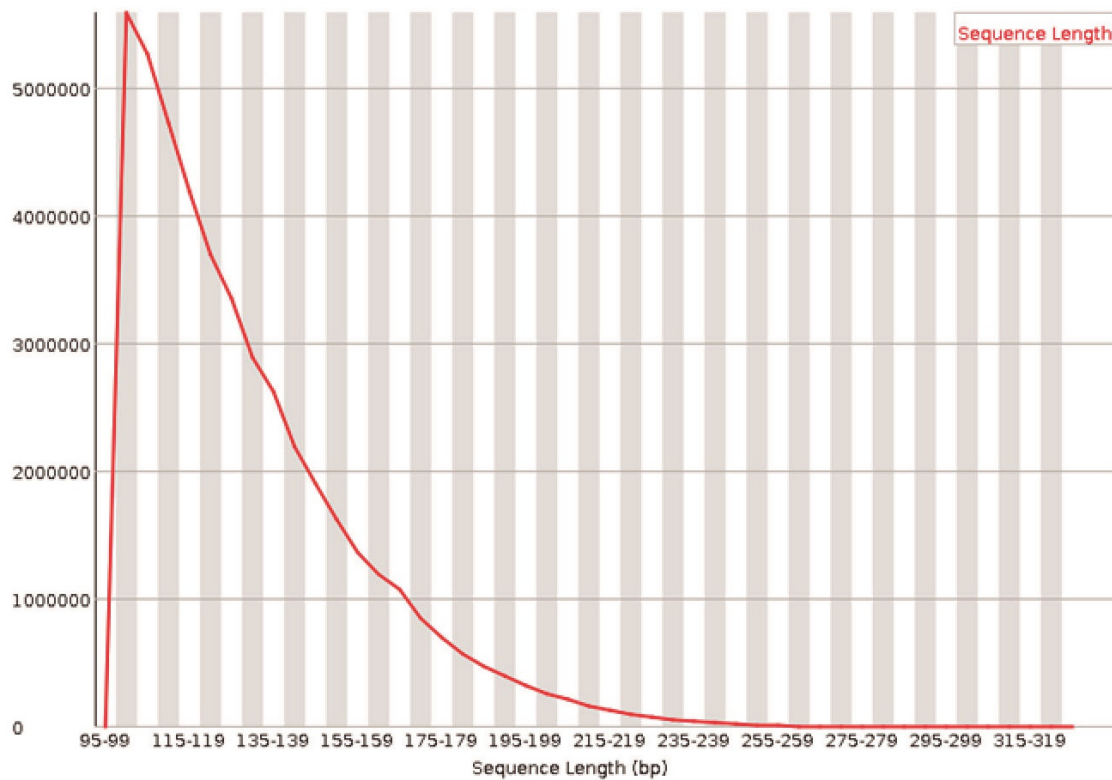


Figure S1 Distribution of sequence lengths over all sequences used in this study.

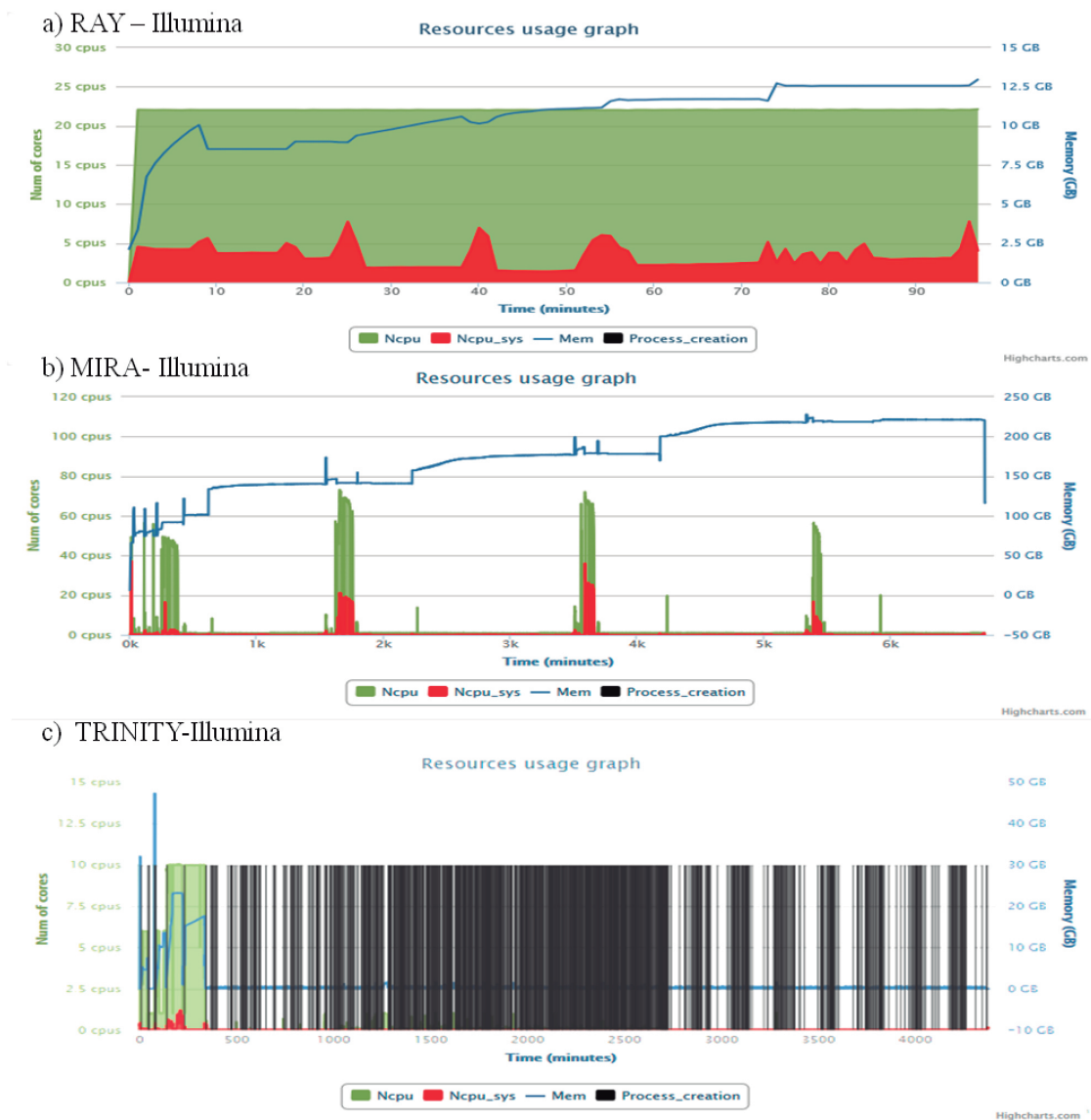
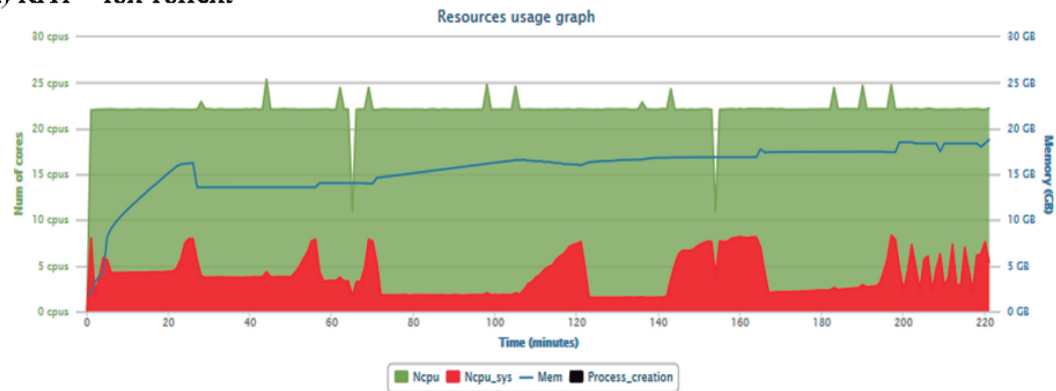
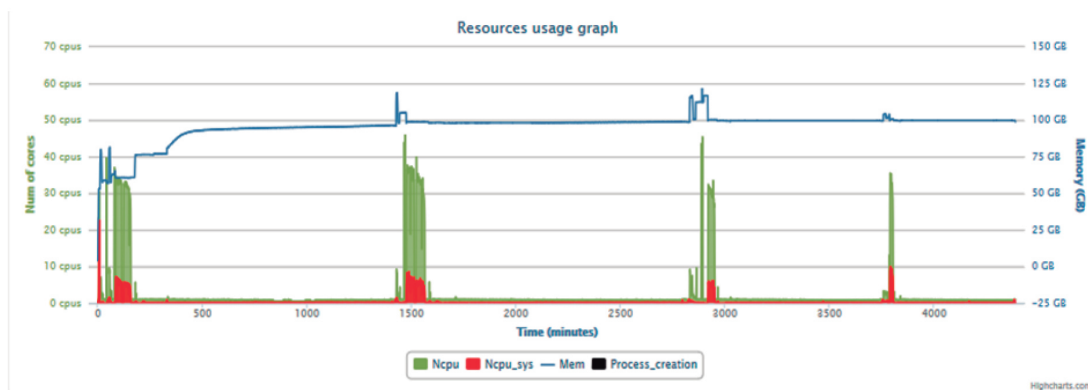


Figure S2 Efficiency in the use of computational resources in each assembler used in this study (RAY, MIRA and TRINITY) from Illumina clean raw data. Ncpus indicates how many central processing units (CPUs) are used by the software, Ncpu_sys indicates how many CPUs are used by the system, Mem indicates RAM memory and Process_creation indicates how many files are created.

a) RAY – Ion Torrent



b) MIRA – Ion Torrent



c) TRINITY – Ion Torrent

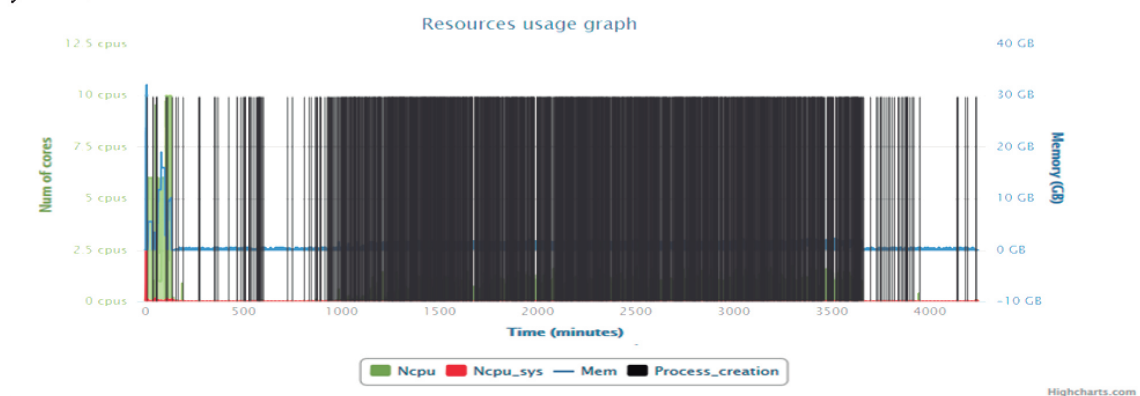
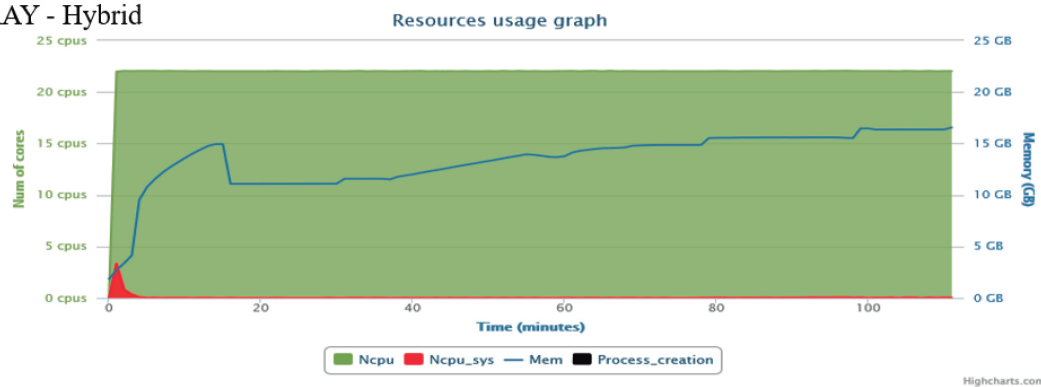


Figure S3 Efficiency in the use of computational resources in each assembler used in this study (RAY, MIRA and TRINITY) from Ion Torrent clean raw data. Ncpus indicates how many central processing units (CPUs) are used by the software, Ncpu_sys indicates how many CPUs are used by the system, Mem indicates RAM memory and Process_creation indicates how many files are created.

a) RAY - Hybrid



b) RAY - Partial hybrid

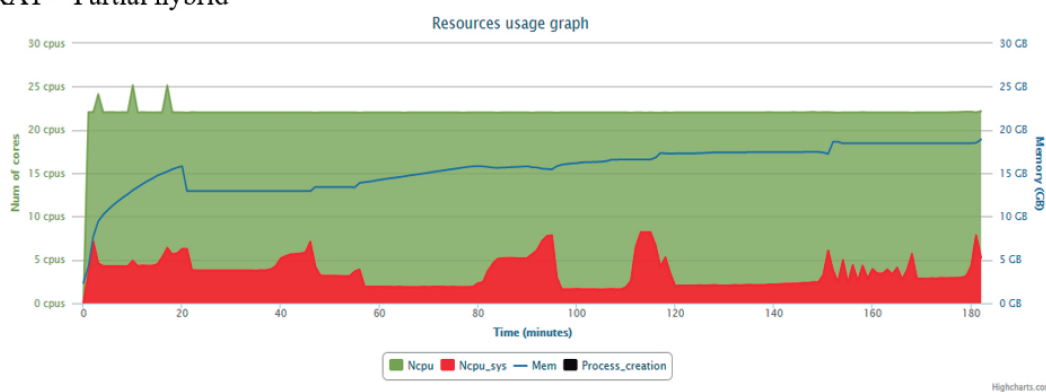


Figure S4 Efficiency in the use of computational resources in the RAY assembler from hybrid transcriptome (a) and partial hybrid transcriptome clean raw data (b). Ncpus indicates how many central processing units (CPUs) are used by the software, Ncpus_sys indicates how many CPUs are used by the system, Mem indicates RAM memory and Process_creation indicates how many files are created.

A.2 A multi-omics analysis pipeline for the metabolic pathway reconstruction in the orphan species (*Quercus ilex*). Supporting information

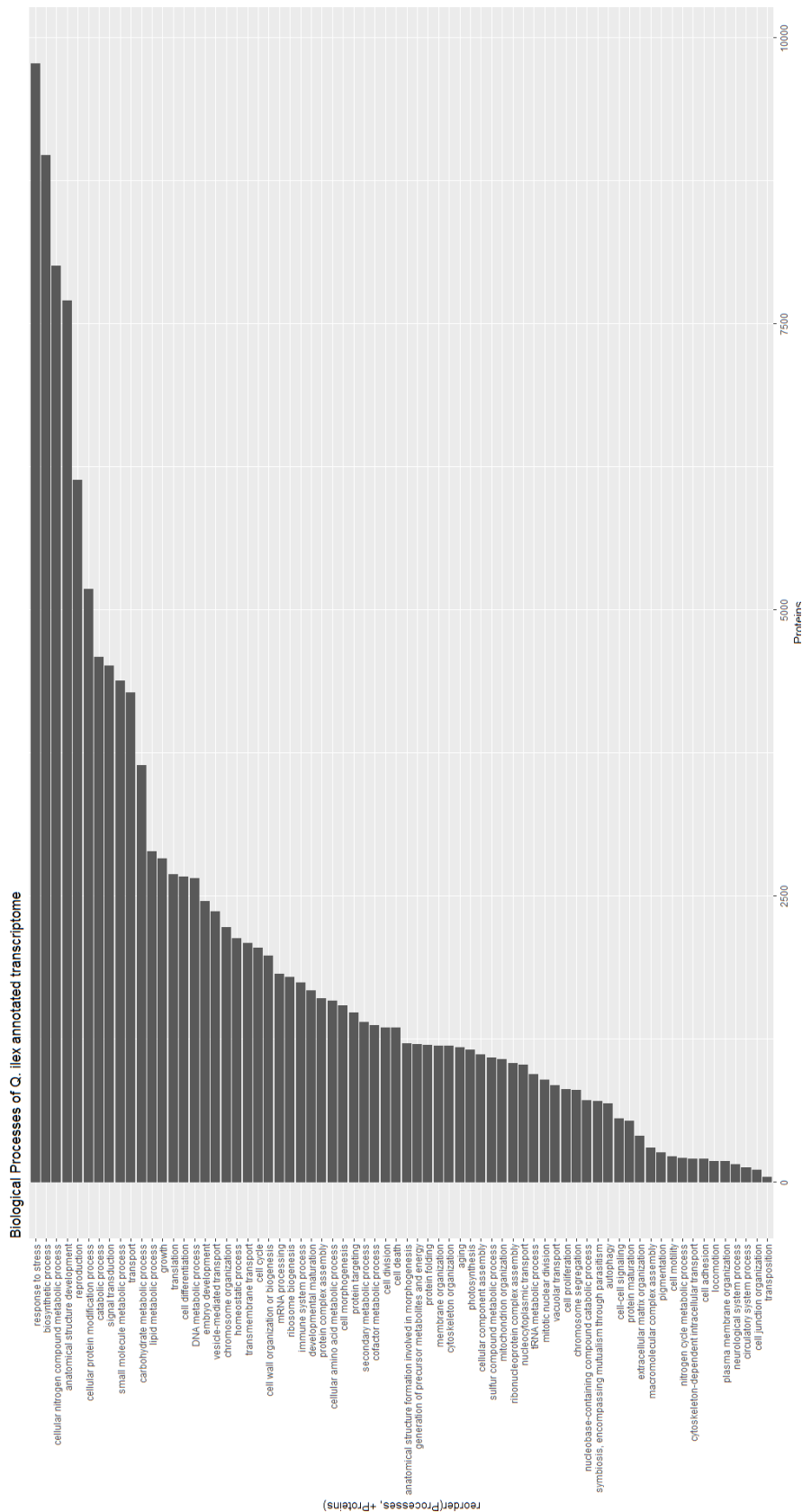


Figure S5 Density histogram for proteins in the different biological processes of *Q. ilex* annotated transcriptome.

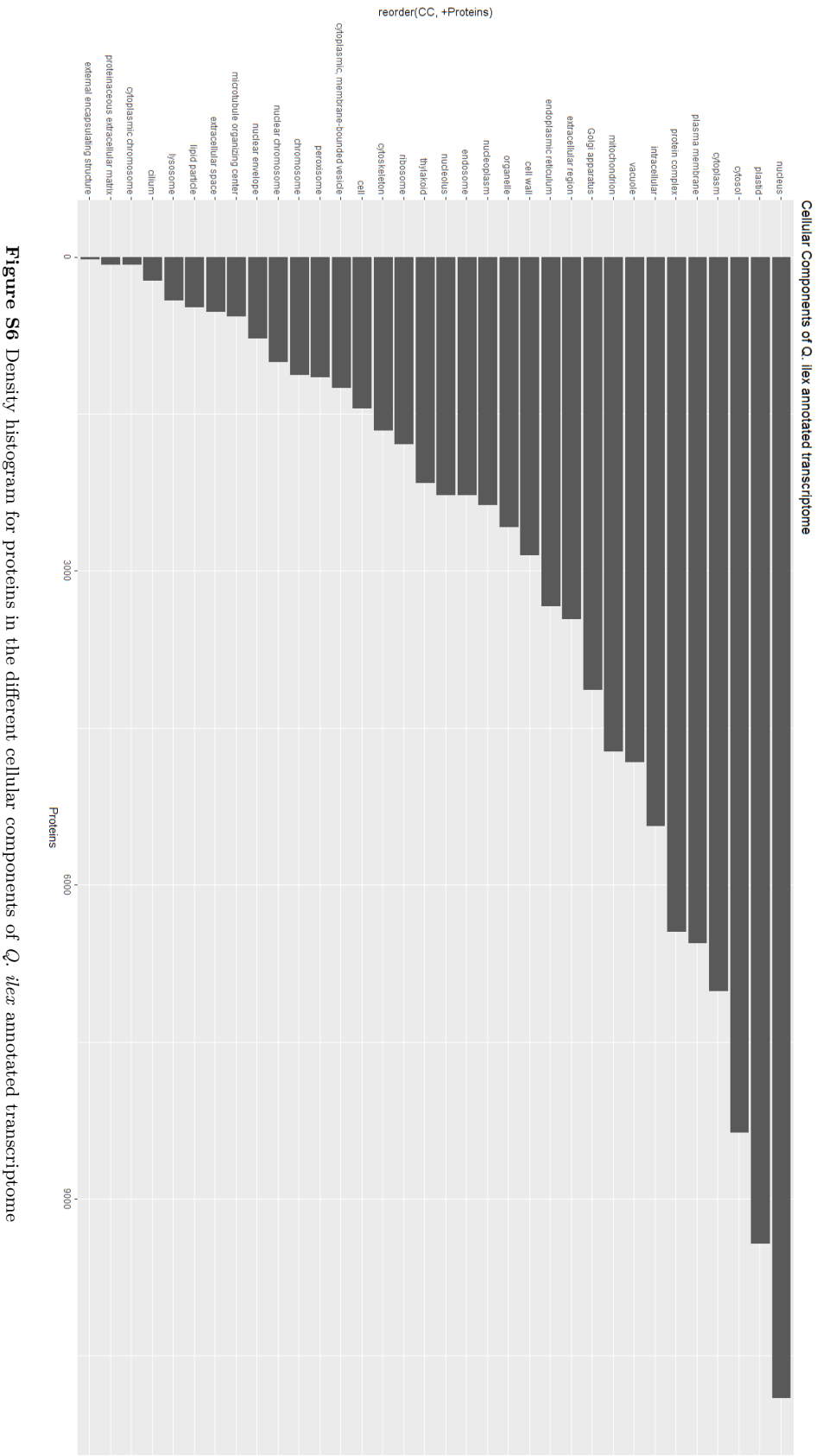


Figure S6 Density histogram for proteins in the different cellular components of *Q. ilex* annotated transcriptome

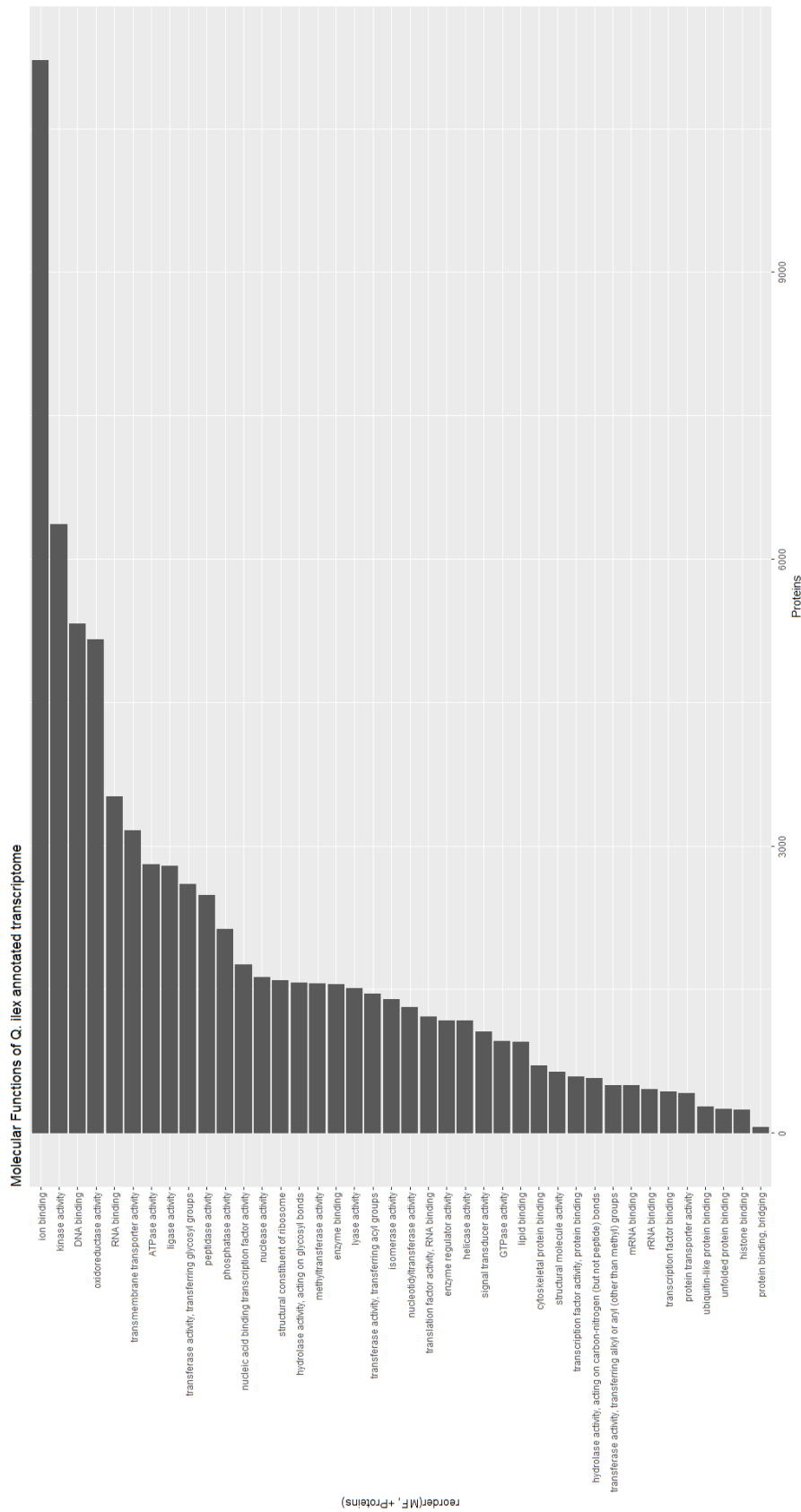


Figure S7 Density histogram for proteins in the different molecular functions of *Q. ilex* annotated transcriptome



Figure S8 Enzymes (transcript level and protein level) assigned to the glycolysis/gluconeogenesis.

Table S4 Metabolite features.

	Metabolite	m/z	RT
1	Pyruvic acid	147, 217, 133	11.59
2	L-Glutamic acid	246, 147, 128	25.23
3	Glucose	319, 205, 147	31.35
4	Alanine	116, 190, 75	12.13
5	Succinic acid	147, 247, 75	18.40
6	Aspartic acid	232, 218, 147	23.32
7	L-Serine	204, 218, 147	19.74
8	L(+)-Ascorbic acid	332, 147, 205	32.13
9	L-Phenylalanine	218, 192, 266	25.25
10	Urea	147, 189, 171	16.82
11	Sucrose	361, 217, 437	43.56
12	Glucose-6-phosphate	387, 299, 357	39.08
13	Glycerol	205, 218, 133	17.60
14	Fructose	217, 437, 147	29.33
15	Fumaric acid	156, 147, 232	23.19
16	Leucine	158, 147, 232	17.39
17	Galactose	319, 205, 147	31.20
18	myo-Inositol	305, 318, 217	31.65
19	L-Proline	142, 147, 75	17.87
20	Malic acid	147, 233, 245	22.70
21	Asparagine	231, 116, 132	26.24
22	Citric acid	273, 147, 347	29.50
23	Glycolic acid	147, 205, 177	11.21
24	Valine	144, 218, 75	15.78
25	D-Cellobiose	204, 361, 217	44.79
26	Lactic acid	147, 117, 191	10.69
27	L-Threonine	218, 291, 117	20.38
28	Gluconic acid	292, 319, 147	33.31
29	Maltose	361, 204, 217	44.89
30	Oxalic acid	147, 133, 220	13.35
31	Sorbose	217, 307, 103	31.08
32	Palmitic acid	313, 117, 129	33.50
33	Glyceric acid	292, 147, 189	19.03
34	Arabinose	307, 217, 103	26.34
35	D(-)-Quinic acid	345, 255, 147	30.37
36	Xylulose	205, 147, 263	26.62
37	D-(+)-Galacturonic acid	217, 204, 292	39.80
38	4-Aminobutyric acid (GABA)	174, 304, 147	23.37
39	Mannitol	319, 205, 217	31.91
40	L-Isoleucine	158, 147, 218	17.94
41	cis-Aconitic acid	147, 229, 375	27.69
42	L-Rhamnose	204, 319, 220	33.96
43	Oleic acid	339, 129, 117	36.60
44	Sorbitol	319, 147, 217	32.05
45	Salicylic acid	267, 370, 193	27.97
46	Glucaric acid	333, 292, 305	33.63
47	Galactaric acid	217, 204, 292	39.80
48	Galactonic acid	205, 275, 147	30.89
49	Maleic acid	147, 245, 75	18.19
50	Gallic acid	458, 281, 443	32.17
51	Stearic acid	341, 117, 129	37.03
52	Linoleic Acid	294, 263, 109	18.36
53	Ribonic acid	292, 217, 147	28.39
54	D-Erythrose	147, 201, 117	21.49
55	Maltotriose	361, 204, 217	44.19
56	Pyroglutamic acid	156, 147, 232	23.19
57	Melibiose	204, 217, 147	44.27
58	Catechin	368, 355, 204	46.10
59	Viburnitol	217, 191, 204	32.50
60	Epigallocatechin	456, 355, 345	46.65
61	Tridecane	112, 98, 85	9.82
62	Anthraquinone	369, 267, 204	46.26

Table S5 GC-MS metabolomic data. Mean values of normalized peak areas, as well as SD (standard deviation) and CV% (percentage of coefficient of variation) were determined for replicates of the metabolite extract. CV%= (SD/mean)*100. Range of CV% (0.7 - 40.0). CV% mean (13.70).

Assigned to/Identified	Replicate 1	Replicate 2	Replicate 3	mean	SD	CV%
Pyruvic acid	0.07	0.08	0.06	0.07	0.01	17.8
L-Glutamic acid	5.15	4.63	4.45	4.74	0.37	7.72
Glucose	28.39	36.31	38.59	34.43	5.35	15.5
Alanine	2.98	3.26	2.69	2.98	0.28	9.45
Succinic acid	0.98	0.89	0.77	0.88	0.11	12.5
Aspartic acid	11.27	10.10	8.57	9.98	1.36	13.6
L-Serine	0.98	0.74	0.85	0.86	0.12	14.1
L(+)-Ascorbic acid	0.71	0.97	0.80	0.83	0.13	15.8
L-Phenylalanine	0.22	0.29	0.25	0.25	0.04	14.5
Urea	0.04	0.07	0.06	0.06	0.02	30.8
Sucrose	25.66	25.14	27.52	26.11	1.26	4.81
Glucose-6-phosphate	0.16	0.21	0.17	0.18	0.02	12.8
Glycerol	1.16	1.49	1.21	1.29	0.18	13.9
Fructose	0.42	0.66	0.52	0.53	0.12	22.7
Fumaric acid	0.65	0.42	0.46	0.51	0.12	23.5
Leucine	0.17	0.34	0.19	0.23	0.09	40
Galactose	4.00	4.66	4.14	4.27	0.35	8.1
myo-Inositol	20.29	21.69	19.46	20.48	1.13	5.5
L-Proline	0.77	0.73	0.74	0.75	0.02	3.04
Malic acid	41.81	40.02	40.45	40.76	0.94	2.3
Asparagine	2.09	2.33	2.30	2.24	0.13	5.82
Citric acid	16.22	16.84	16.41	16.49	0.32	1.92
Glycolic acid	0.26	0.30	0.23	0.26	0.04	13.4
Valine	1.18	1.52	1.31	1.33	0.17	12.6
D-Cellobiose	0.23	0.36	0.33	0.31	0.07	22.5
Lactic acid	2.54	2.21	2.49	2.42	0.18	7.46
L-Threonine	0.17	0.25	0.19	0.21	0.04	20.7
Gluconic acid	0.11	0.11	0.11	0.11	0.00	2.66
Maltose	3.52	1.61	2.30	2.48	0.97	39.2
Oxalic acid	0.59	0.54	0.63	0.59	0.05	8.29
Sorbose	12.83	13.42	12.99	13.08	0.30	2.31
Palmitic acid	0.86	0.88	0.84	0.86	0.02	2.07
Glyceric acid	0.11	0.24	0.20	0.18	0.07	38
Arabinose	0.09	0.11	0.09	0.10	0.01	10
D(-)-Quinic acid	26.66	27.01	26.76	26.81	0.18	0.66
Xylulose	0.03	0.05	0.03	0.04	0.01	20
D-(+)-Galacturonic acid	3.16	2.19	2.26	2.54	0.54	21.2
4-Aminobutyric acid (GABA)	6.41	5.50	6.13	6.01	0.46	7.73
Mannitol	0.39	0.38	0.38	0.38	0.01	1.57
L-Isoleucine	0.33	0.58	0.40	0.44	0.13	29.3
cis-Aconitic acid	0.30	0.24	0.28	0.27	0.03	11.7
L-Rhamnose	7.99	7.06	7.32	7.46	0.48	6.45
Oleic acid	1.65	1.67	1.62	1.65	0.02	1.29
Sorbitol	0.13	0.11	0.12	0.12	0.01	6.78
Salicylic acid	0.09	0.08	0.08	0.09	0.01	7.87
Glucaric acid	0.12	0.14	0.14	0.14	0.01	7.78
Galactaric acid	0.37	0.31	0.33	0.34	0.03	10
Galactonic acid	1.37	1.79	1.37	1.51	0.24	16.1
Maleic acid	3.75	2.80	2.39	2.98	0.70	23.4
Gallic acid	2.72	2.25	2.31	2.43	0.26	10.6
Stearic acid	0.46	0.49	0.57	0.51	0.05	10.6
Linoleic Acid	19.33	17.29	16.26	17.63	1.56	8.85
Ribonic acid	0.26	0.25	0.23	0.25	0.02	6.74
D-Erythrose	0.05	0.05	0.05	0.05	0.00	2.31
Maltotriose	0.60	1.03	0.71	0.78	0.22	28.4
Pyroglutamic acid	1.27	1.06	1.08	1.14	0.12	10.3
Melibiose	0.74	0.54	0.77	0.68	0.12	18.2
Catechin	12.40	9.05	9.98	10.47	1.73	16.5
Viburnitol	16.88	18.60	17.86	17.78	0.86	4.86
Epigallocatechin	1.47	2.25	2.07	1.93	0.41	21.1
Tridecane	2.78	2.71	4.40	3.29	0.96	29
Anthraquinone	0.16	0.33	0.27	0.26	0.09	34.9

Table S6 KEGG pathways with metabolites, proteins, and transcripts.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050436/bin/Table_3.XLSX

Table S7 Comparison of KEGG pathways of different species.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050436/bin/Table_4.XLSX

Table S8 Bins of transcripts.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050436/bin/Table_5.XLSX

Table S9 Bins of proteins.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050436/bin/Table_6.XLSX

Table S10 Comparison of *in silico* proteomes.

Protein databases (Uniprot Proteomes)	Number of proteins in DB	<i>Q. ilex</i> blasted transcripts	% of <i>Q. ilex</i> transcripts
<i>Oryza sativa</i> subsp. <i>Japonica</i>	39947	48724	77.80
<i>Arabidopsis thaliana</i>	39179	53644	85.65
<i>Populus trichocarpa</i>	44466	57440	91.72
<i>Eucalyptus grandis</i>	44150	55392	88.45
Total annotated sequences in <i>Q. ilex</i> transcriptome = 62628			
Blastx (e-value 10^{-10})			

Table S11 Shotgun LC-MS/MS proteomic data. Mean values of normalized peak areas, as well as SD (standard deviation) and CV% (percentage of coefficient of variation) were determined for replicates of protein extract (Jorge et al., 2005; Jorge et al., 2006). The area values correspond to replicate 1 (0.6 μ g of protein), replicate 2 (0.8 μ g of protein), and replicate 3 (1 μ g of protein). Accessible in this link:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050436/bin/Table_8.XLSX

Table S12 Enzymes (transcripts).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050436/bin/Table_9.xlsx

Table S13 Enzymes (proteins).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050436/bin/Table_10.xlsx

Table S14 Omics overview.**Omics mapping**

TRANSCRIPTOME		
	<i>Q. ilex</i> annotated sequences	62628
	Unique genes	27080
	Transcripts of Enzymes	2103
PROTEOME		
	Identified proteins	2380
	Enzymes	228
METABOLOME		
	Identified metabolites	62
	Metabolites mapped to the metabolic reconstruction (Mapman)	58

A.3 Decoding through a multiomic analysis, drought tolerance determinants of Holm oak (*Quercus ilex*). Supporting Information

Table S15 List of transcripts. Columns A to M correspond to contig ID, corresponding gene acronyms, and annotations. Columns N to AG contain quantitative values for each gene for each sample within each sample and the statistical parameters, FDR and p-value:

http://www.uco.es/probiveag/files/supplementary/Supplementary_Table_1.xlsx

Table S16 List of proteins. Columns A to D correspond to contig and protein IDs, corresponding gene and description. Columns E to AA contain quantitative values for each gene for each proteoform within each sample and the statistical p-value.

http://www.uco.es/probiveag/files/supplementary/Supplementary_Table_2.xlsx

Table S17 Correlation between transcripts and proteins abundance. Two different datasets were employed, the total and the group containing variable ones. The first and second columns correspond to the quantitative values for, respectively, transcripts and proteins. The Pearson correlation coefficient for each dataset is indicated.

http://www.uco.es/probiveag/files/supplementary/Supplementary_Table_3.xlsx

Table S21 KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked. Columns include the pathway name, number of gene products per pathway corresponding to *Arabidopsis thaliana*, and significance p-values at the transcript, protein or combined levels.

Pathway name	UniGenes	Gene expression	Proteomics	Combined p-value
Ribosome	53	5.46694E-14	0.39081	6.93874E-13
Glyoxylate and dicarboxylate metabolism	26	0.07588	0.00222	0.00163
Phenylpropanoid biosynthesis	17	0.00196	0.16737	0.00295
Phenylalanine metabolism	8	0.00312	0.11119	0.00311
Flavonoid biosynthesis	4	0.00055	1	0.00466
Stilbenoid, diarylheptanoid and gingerol biosynthesis	6	0.00634	NA	0.00634
Biosynthesis of secondary metabolites	268	0.00530	0.26303	0.01056
Photosynthesis	35	0.52720	0.00266	0.01061
Carbon metabolism	68	0.03277	0.06915	0.01606
Plant-pathogen interaction	62	0.02147	0.16737	0.02382
Cysteine and methionine metabolism	26	0.03550	0.10684	0.02494
Ubiquinone and other terpenoid-quinone biosynthesis	11	0.01724	0.28059	0.03063
Nitrogen metabolism	21	0.02418	0.39007	0.05342
Monobactam biosynthesis	4	0.01271	1	0.06819
Plant hormone signal transduction	64	0.01357	1	0.07193
Biosynthesis of amino acids	68	0.01842	0.85423	0.08106
Carbon fixation in photosynthetic organisms	19	0.25709	0.07021	0.09052
Propanoate metabolism	8	0.35336	0.06164	0.10513
Glycerolipid metabolism	24	0.02208	1	0.10626
Linoleic acid metabolism	4	0.11377	NA	0.11377
Oxidative phosphorylation	38	0.03273	0.91770	0.13532
alpha-Linolenic acid metabolism	15	0.41035	0.11119	0.18648
Glycine, serine and threonine metabolism	20	0.07296	0.62891	0.18728
Phenylalanine, tyrosine and tryptophan biosynthesis	16	0.08488	0.62891	0.20982
Diterpenoid biosynthesis	6	0.23127	NA	0.23127
Citrate cycle (TCA cycle)	15	0.12518	0.62891	0.27883
Lipoic acid metabolism	2	0.28317	NA	0.28317
Tyrosine metabolism	8	0.73628	0.11119	0.28674
2-Oxocarboxylic acid metabolism	22	0.23982	0.34836	0.29093
Butanoate metabolism	5	0.56511	0.15170	0.29633
Vitamin B6 metabolism	4	0.48626	0.28059	0.40821
C5-Branched dibasic acid metabolism	3	1.00000	0.15170	0.43779
Valine, leucine and isoleucine biosynthesis	5	0.56511	0.28059	0.45058
Limonene and pinene degradation	5	0.17084	1	0.47272
Selenocompound metabolism	5	0.17084	1	0.47272
Carotenoid biosynthesis	13	0.61503	0.28059	0.47577
Spliceosome	30	0.30732	0.56194	0.47599
Taurine and hypotaurine metabolism	4	0.48626	NA	0.48626
Sulfur metabolism	10	0.18827	1	0.50266
Photosynthesis - antenna proteins	10	0.18827	1	0.50266
Isoquinoline alkaloid biosynthesis	4	0.48626	0.39007	0.50499
Tropane, piperidine and pyridine alkaloid biosynthesis	4	0.48626	0.39007	0.50499

Table S21 KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked. Columns include the pathway name, number of gene products per pathway corresponding to *Arabidopsis thaliana*, and significance p-values at the transcript, protein or combined levels.

Pathway name	UniGenes	Gene expression	Proteomics	Combined p-value
Alanine, aspartate and glutamate metabolism	20	0.57515	0.34836	0.52247
Glutathione metabolism	31	0.46489	0.46299	0.54585
Pyruvate metabolism	15	0.32060	0.68574	0.55288
Valine, leucine and isoleucine degradation	12	1.00000	0.22708	0.56371
Fatty acid biosynthesis	11	0.23113	1	0.56969
Other glycan degradation	6	0.23127	1	0.56988
Endocytosis	36	0.82492	0.28798	0.57902
beta-Alanine metabolism	13	0.61503	0.39007	0.58237
Fructose and mannose metabolism	16	0.69446	0.34836	0.58524
Mismatch repair	13	0.88554	0.28059	0.59445
Ribosome biogenesis in eukaryotes	16	1.00000	0.28059	0.63718
Glucosinolate biosynthesis	2	0.28317	1	0.64045
Ascorbate and aldarate metabolism	13	1.00000	0.28798	0.64648
One carbon pool by folate	7	0.29268	1	0.65229
Arginine and proline metabolism	13	0.61503	0.48303	0.65766
Phosphatidylinositol signaling system	22	0.64597	0.48303	0.67543
Fatty acid metabolism	21	0.40567	0.77480	0.67809
Galactose metabolism	13	0.32060	1	0.68531
Protein processing in endoplasmic reticulum	51	0.50812	0.69115	0.71868
Tryptophan metabolism	14	0.36574	1	0.73362
Arginine biosynthesis	15	0.65658	0.56194	0.73684
Fatty acid degradation	15	0.91806	0.40698	0.74147
Inositol phosphate metabolism	21	1.00000	0.39007	0.75729
Fatty acid elongation	9	0.77680	NA	0.77680
Pentose phosphate pathway	9	0.41209	1	0.77742
Amino sugar and nucleotide sugar metabolism	54	0.72459	0.60666	0.80089
Starch and sucrose metabolism	56	0.49582	0.90260	0.80735
Base excision repair	10	0.81110	NA	0.81110
RNA transport	41	0.98970	0.46299	0.81582
Phagosome	11	0.46805	1	0.82338
Lysine biosynthesis	4	0.48626	1	0.83686
Glycosphingolipid biosynthesis - globo series	4	0.48626	1	0.83686
Glycolysis / Gluconeogenesis	26	0.87272	0.61084	0.86844
Terpenoid backbone biosynthesis	19	0.53659	1	0.87063
mRNA surveillance pathway	32	1.00000	0.56194	0.88582
Glycerophospholipid metabolism	39	0.56506	1	0.88761
Brassinosteroid biosynthesis	5	0.56511	1	0.88764
Thiamine metabolism	5	0.56511	1	0.88764
RNA degradation	23	1.00000	0.68574	0.94444
DNA replication	15	0.69446	1	0.94767
Aminoacyl-tRNA biosynthesis	14	0.90315	0.77480	0.94959
Ubiquitin mediated proteolysis	49	0.99729	0.73393	0.96035
Biosynthesis of unsaturated fatty acids	8	0.73628	1	0.96169
Cyanoamino acid metabolism	17	0.75985	1	0.96853
Homologous recombination	21	0.96997	NA	0.96997
Purine metabolism	48	0.86819	0.88476	0.97076
Pyrimidine metabolism	43	0.99925	0.77480	0.97234
Nicotinate and nicotinamide metabolism	9	0.77680	1	0.97300
Sphingolipid metabolism	9	0.77680	1	0.97300
Pantothenate and CoA biosynthesis	9	0.77680	1	0.97300
Steroid biosynthesis	9	0.77680	1	0.97300
Circadian rhythm - plant	18	0.78776	1	0.97569
Basal transcription factors	13	0.88554	1	0.99318
SNARE interactions in vesicular transport	26	0.92534	1	0.99714
Nucleotide excision repair	19	0.95803	1	0.99911
Porphyrin and chlorophyll metabolism	20	0.96450	1	0.99936
Proteasome	24	0.98183	1	0.99983
Peroxisome	39	0.98808	1	0.99993
Caffeine metabolism	1	1	NA	1
Monoterpenoid biosynthesis	1	1	NA	1
Riboflavin metabolism	1	1	NA	1
Other types of O-glycan biosynthesis	1	1	NA	1
Histidine metabolism	2	1	1	1
Sesquiterpenoid and triterpenoid biosynthesis	2	1	NA	1
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	2	1	NA	1
Sulfur relay system	3	1	1	1
ABC transporters	3	1	NA	1
Glycosphingolipid biosynthesis - ganglio series	4	1	1	1
Biotin metabolism	4	1	1	1
Indole alkaloid biosynthesis	4	1	1	1

Table S21 KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked. Columns include the pathway name, number of gene products per pathway corresponding to *Arabidopsis thaliana*, and significance p-values at the transcript, protein or combined levels.

Pathway name	UniGenes	Gene expression	Proteomics	Combined p-value
Non-homologous end-joining	5	1	NA	1
Glycosaminoglycan degradation	5	1	1	1
Folate biosynthesis	6	1	NA	1
Regulation of autophagy	6	1	NA	1
Protein export	6	1	1	1
Pentose and glucuronate interconversions	6	1	1	1
Arachidonic acid metabolism	7	1	1	1
Cutin, suberine and wax biosynthesis	8	1	NA	1
Lysine degradation	8	1	1	1
Zeatin biosynthesis	8	1	NA	1
Ether lipid metabolism	8	1	NA	1
N-Glycan biosynthesis	15	1	NA	1
RNA polymerase	20	1	1	1

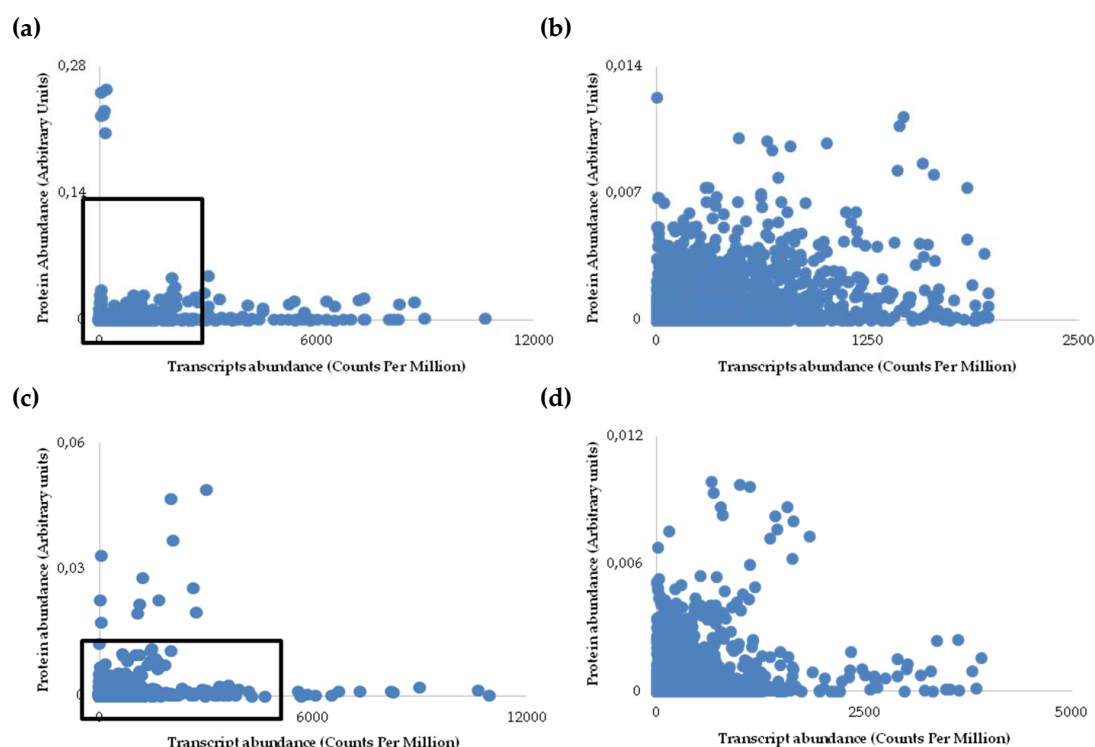


Figure S9 Correlation between transcripts and proteins abundance. Only gene products detected at both transcriptomic and proteomic levels were considered. (a) and (c) correspond to, respectively, the total and the variable gene product datasets. (b) and (d) correspond to the low abundant gene products.

Table S18 Gene Ontology analysis of the variable gene products. Tabs correspond to biological process, molecular function, and cellular component. Each tab contains the categories, number of total and identified genes, acronyms of the identified genes, and the enrichment FDR. Data are organized according to the -omics platform (from transcriptomics to proteomics), up/down regulated in droughted seedlings, and sampling time, days 20 and 25.

http://www.uco.es/probiveag/files/supplementary/Supplementary_Table_4.xlsx

Table S19 KEGG pathway analysis. List of pathways to which identified transcripts and/or protein enzymes are linked. Columns include the pathway name, number of gene products per pathway corresponding to *Arabidopsis thaliana*, and significance p-values at the transcript, protein or combined levels.

http://www.uco.es/probiveag/files/supplementary/Supplementary_Table_5.xlsx

Table S20 GeneMANIA network analysis. Columns correspond to the gene acronyms, functional annotation, GO code, log score, *Arabidopsis thaliana* ortholog, node type (query = interrogated gene products; result= predicted gene products; unknown= contrary gene and protein changes).

http://www.uco.es/probiveag/files/supplementary/Supplementary_Table_6.xlsx

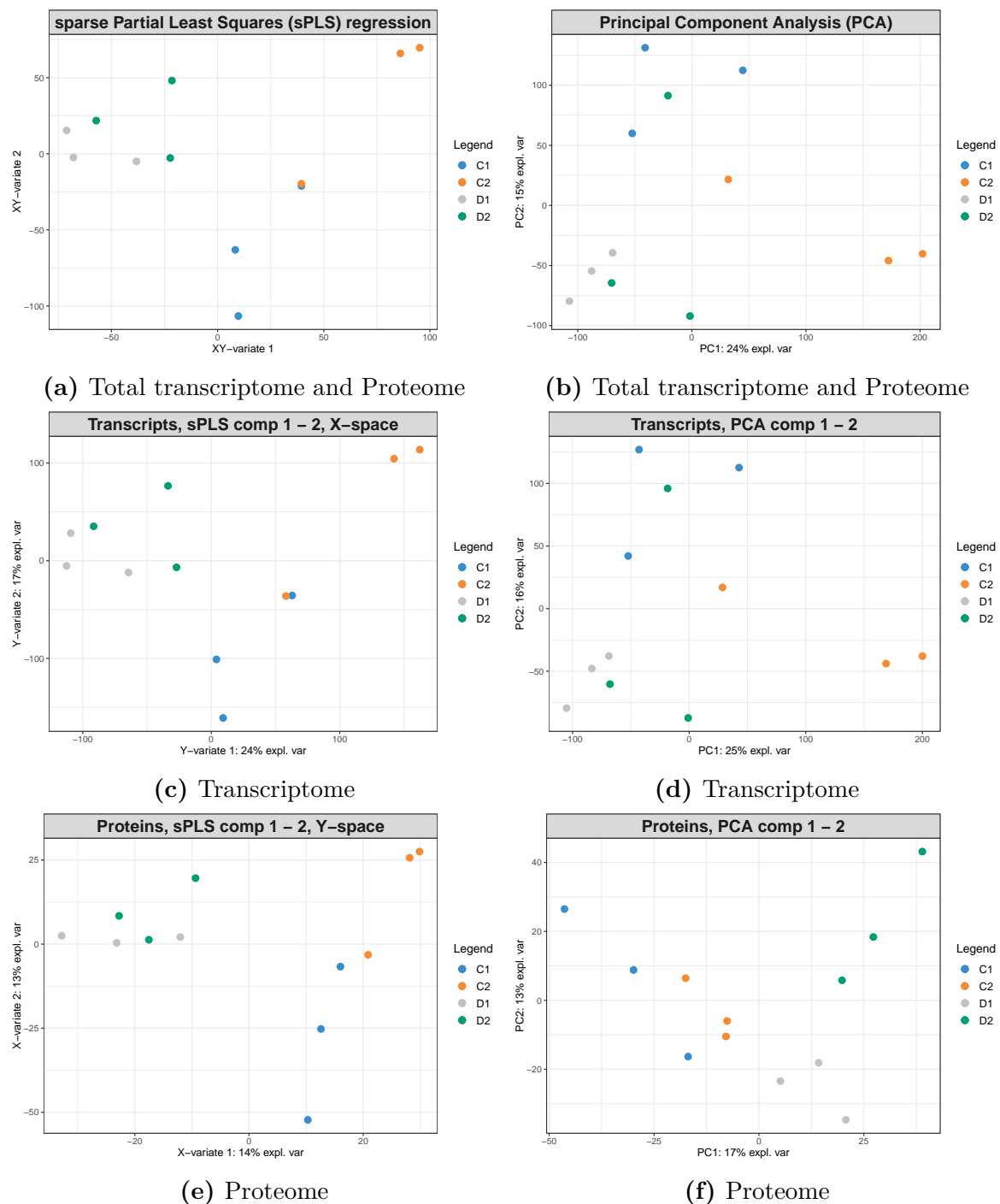


Figure S10 PCA and sPLS analysis of the data. PCA (on the right) and sPLS (on the left) plots based on the two first components, PC1, and PC 2. Different datasets were employed for the analysis: (A) total transcriptome and proteome, (B) total transcriptome, (C) total proteome, (D) total variable transcriptome and proteome, (E) variable transcriptome, and (F) variable proteome. (Blue) Control at day 20, (Orange) Control at day 25, (Grey) Drought at day 20, (Green) Drought at day 25. The three replicates per sample are shown.

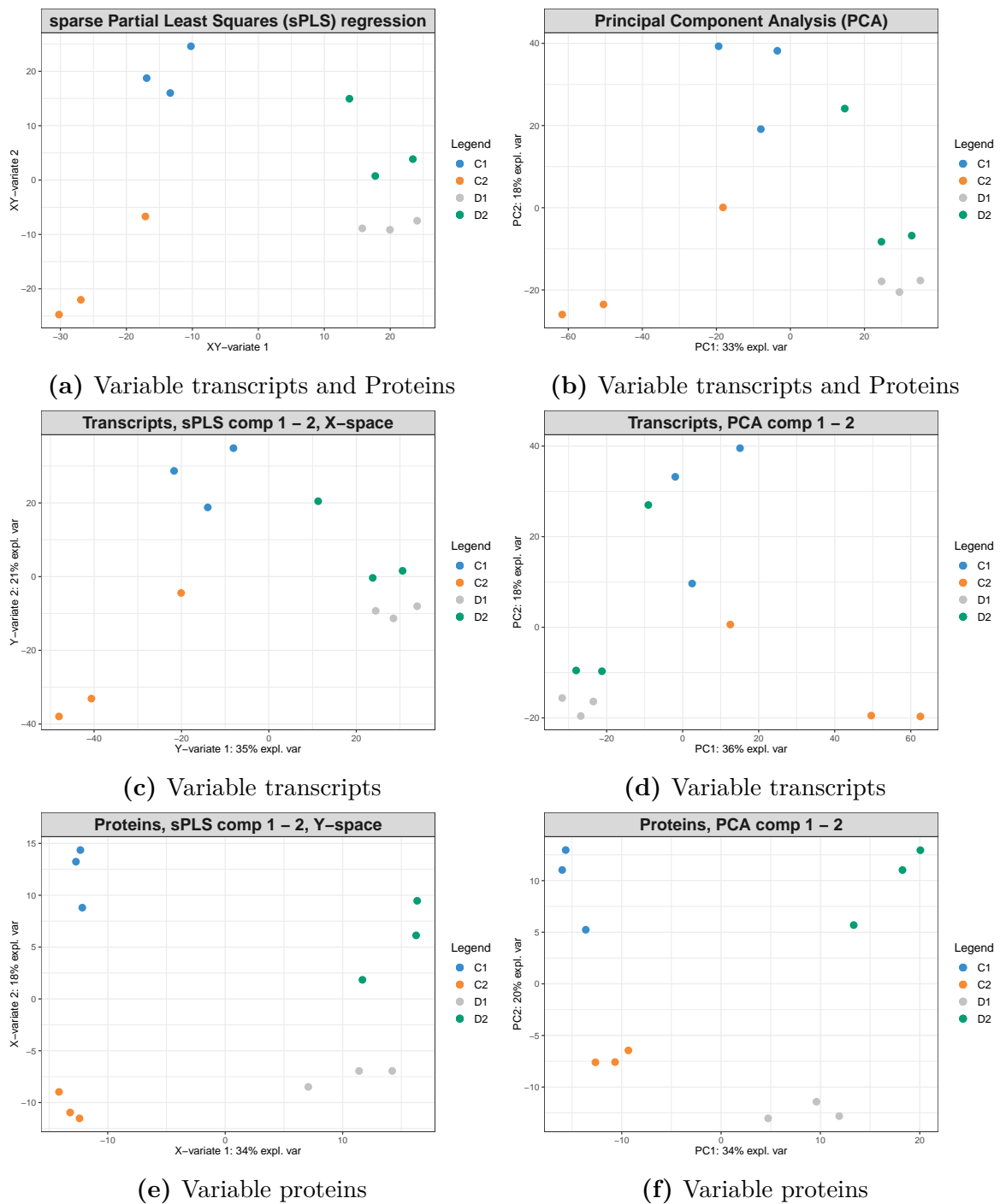


Figure S11 PCA and sPLS analysis of the data. (D) total variable transcriptome and proteome, (E) variable transcriptome, and (F) variable proteome. (Color) Control at day 20, (Color) Control at day 25, (Color) Drought at day 20, (Color) Drought at day 25. The three replicates per sample are shown.

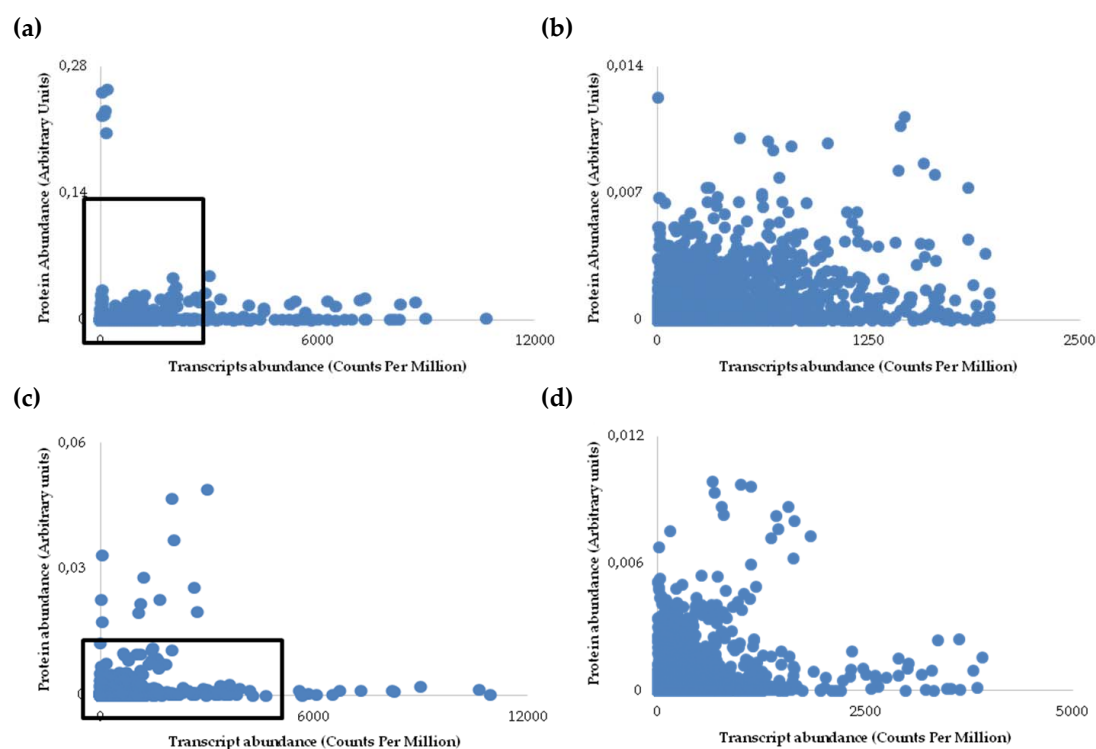


Figure S12 KEGG metabolic charts of the twelve pathways showing statistically significant differences between treatments, control and drought. In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right).

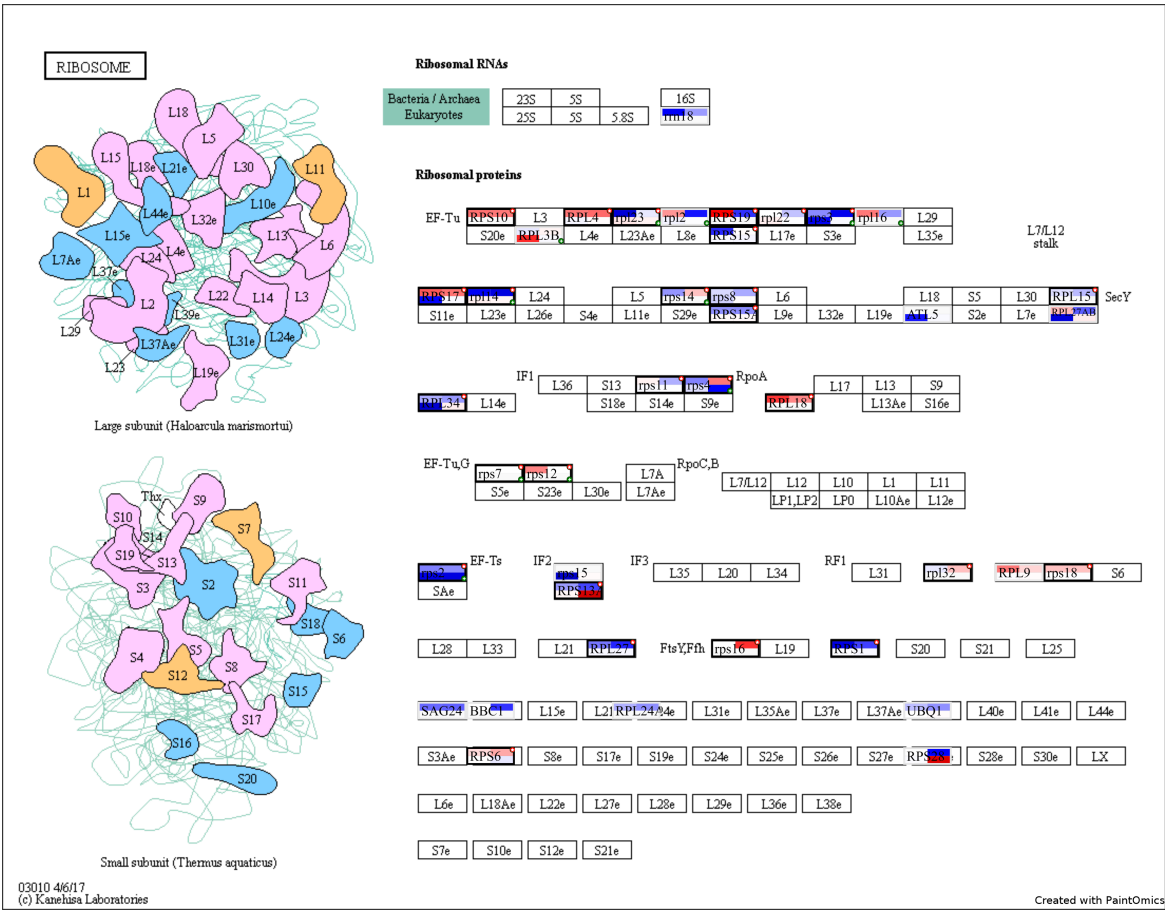


Figure S13 Paintomics KEGG differential pathway analysis: Ribosome (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)

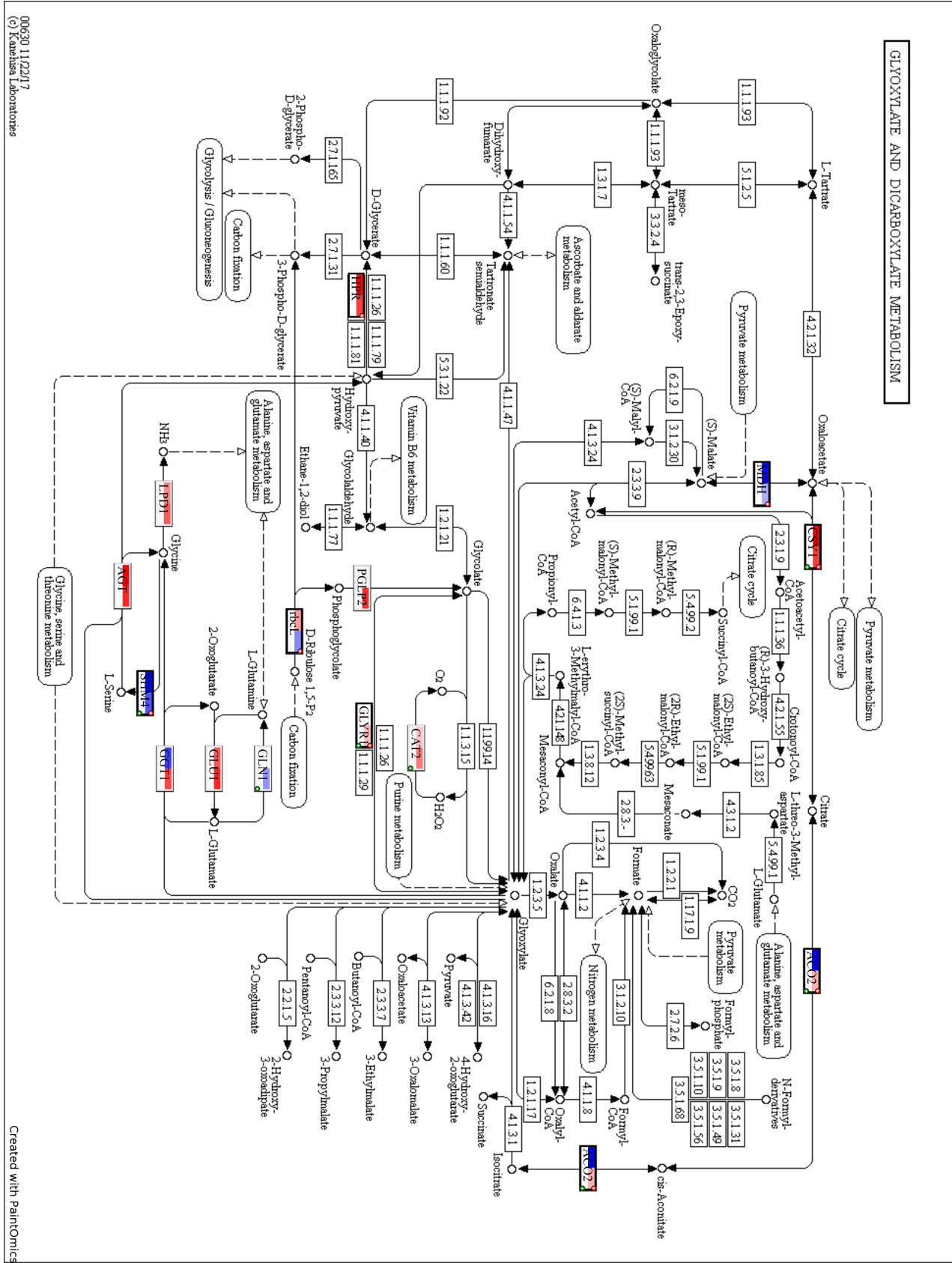


Figure S14 Paintomics KEGG differential pathway analysis: Glyoxylate and dicarboxylate metabolism (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)



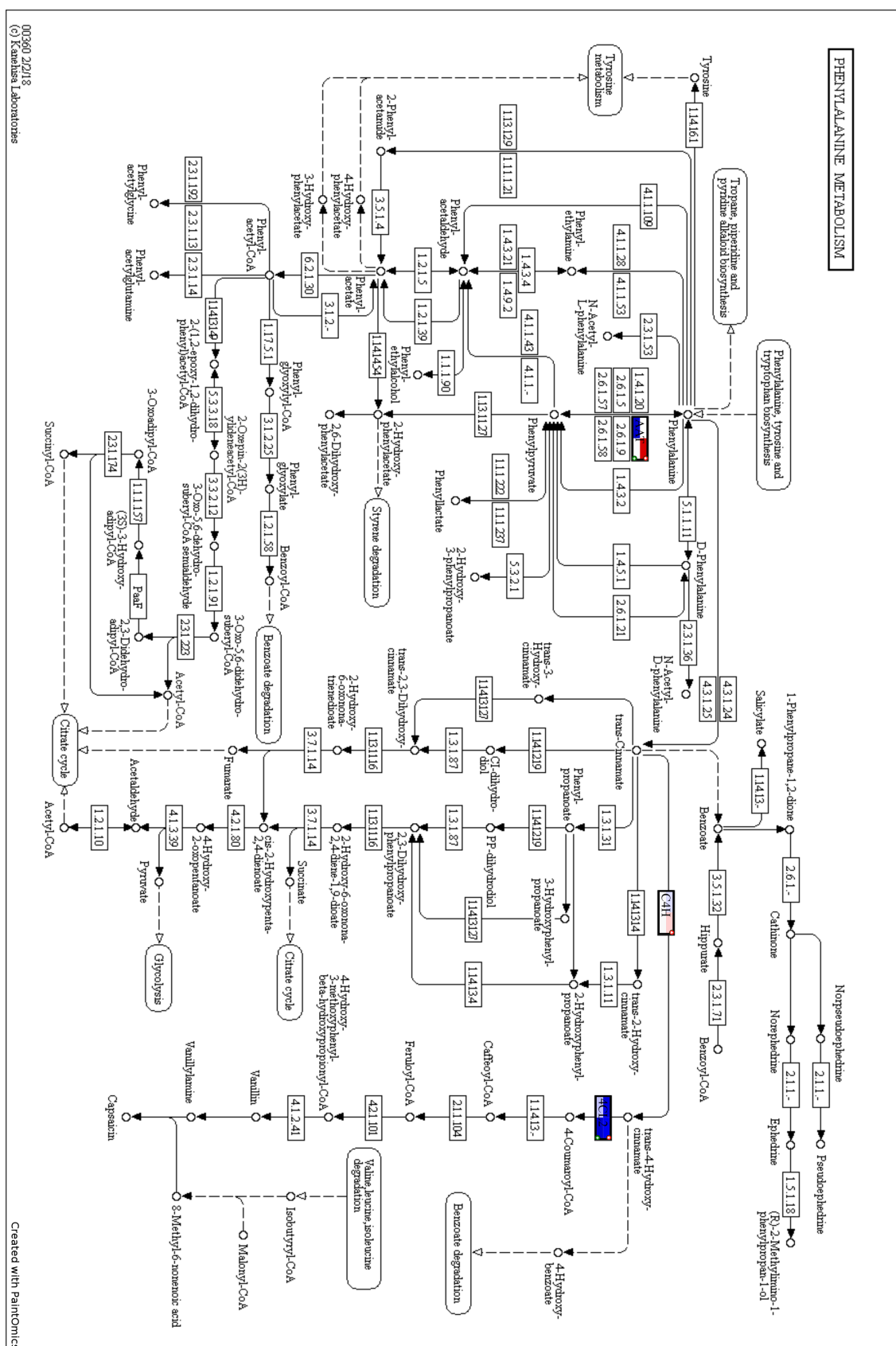


Figure S16 Paintomics KEGG differential pathway analysis: Phenylalanine metabolism (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)



Figure S17 Paintomics KEGG differential pathway analysis: Flavonoid Biosynthesis (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated droughted seedlings at days 20 (left) and 25 (right)





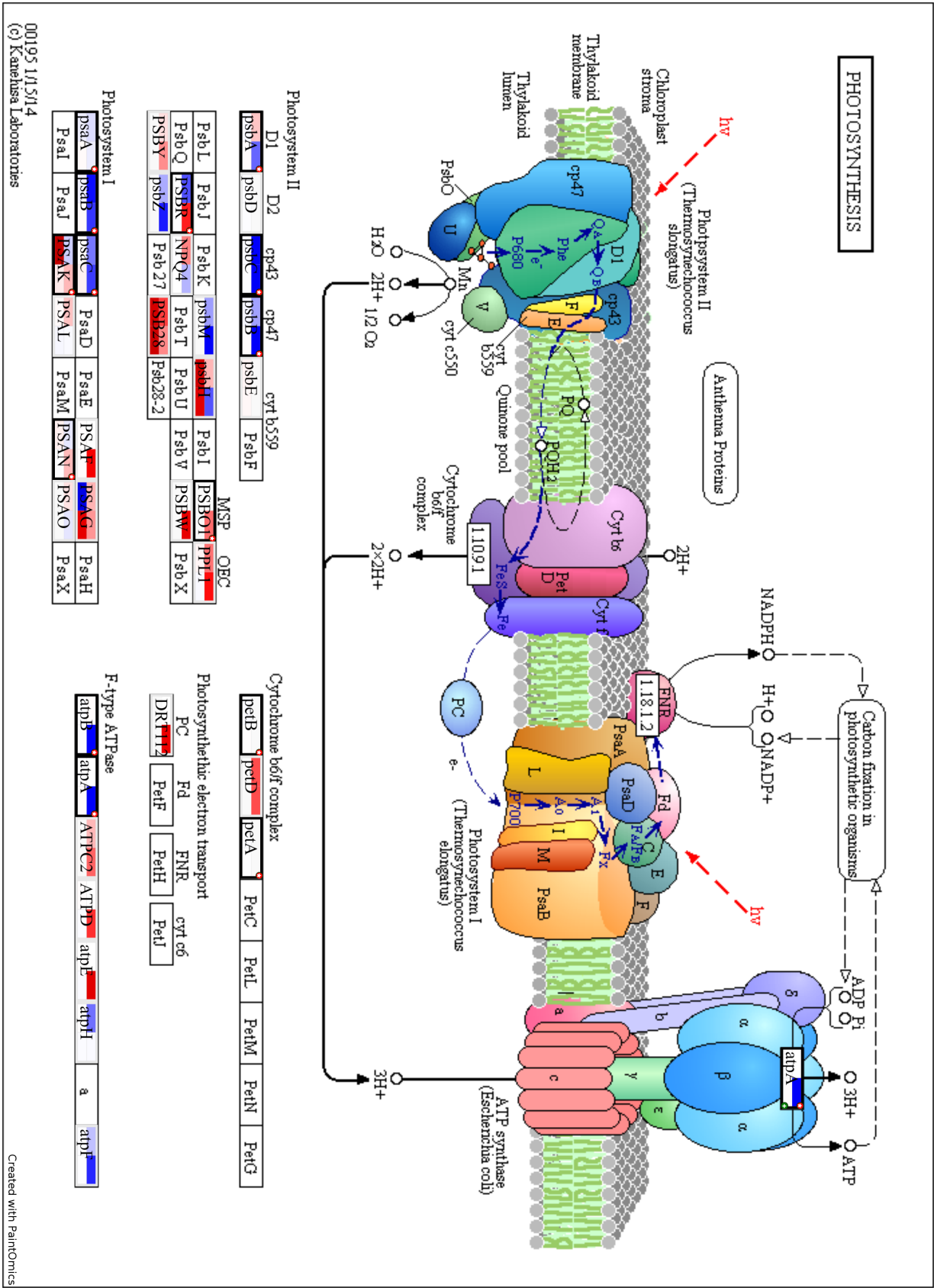


Figure S20 Paintomics KEGG differential pathway analysis: Photosynthesis (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)

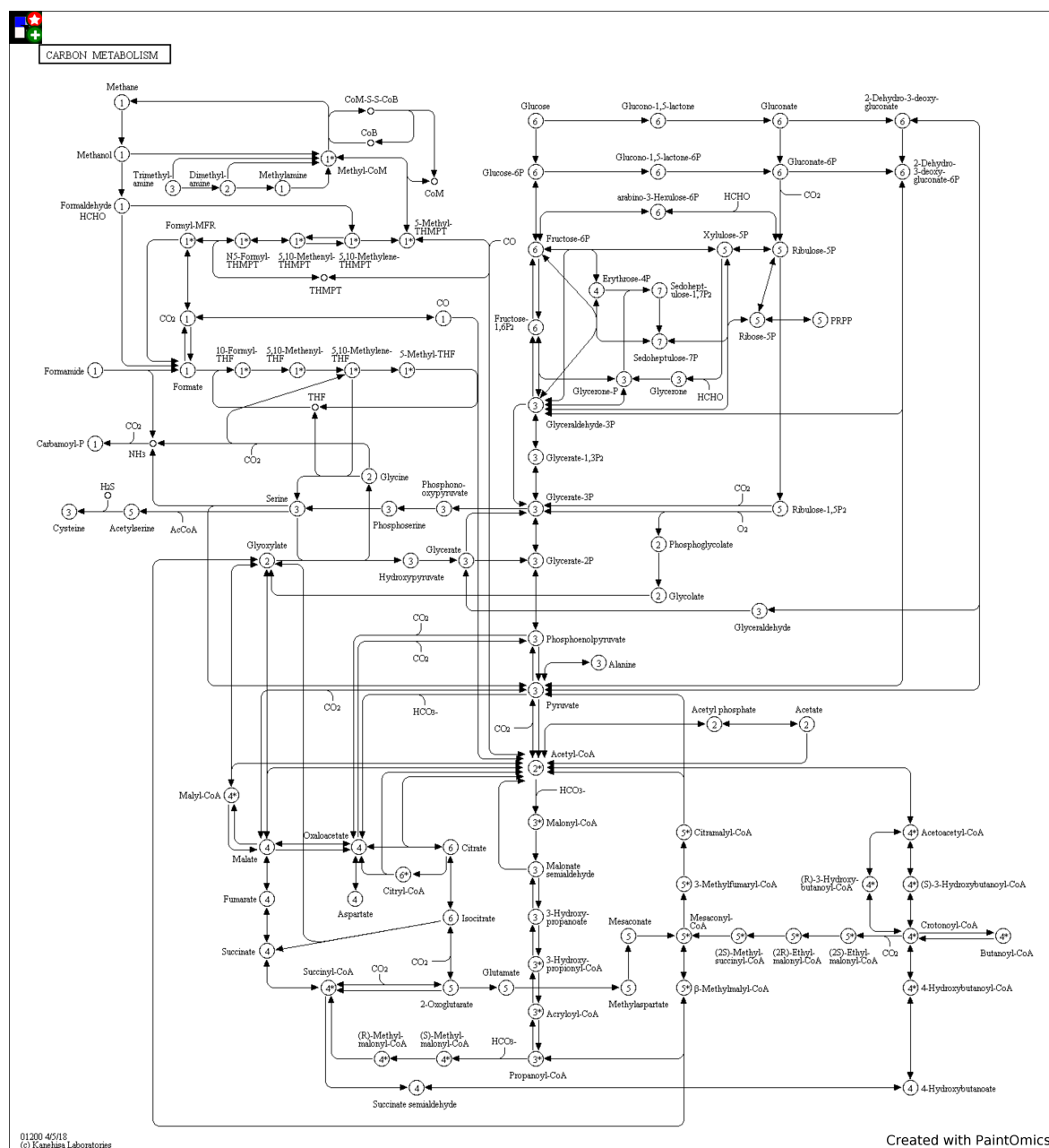


Figure S21 Paintomics KEGG differential pathway analysis: Carbon metabolism (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)

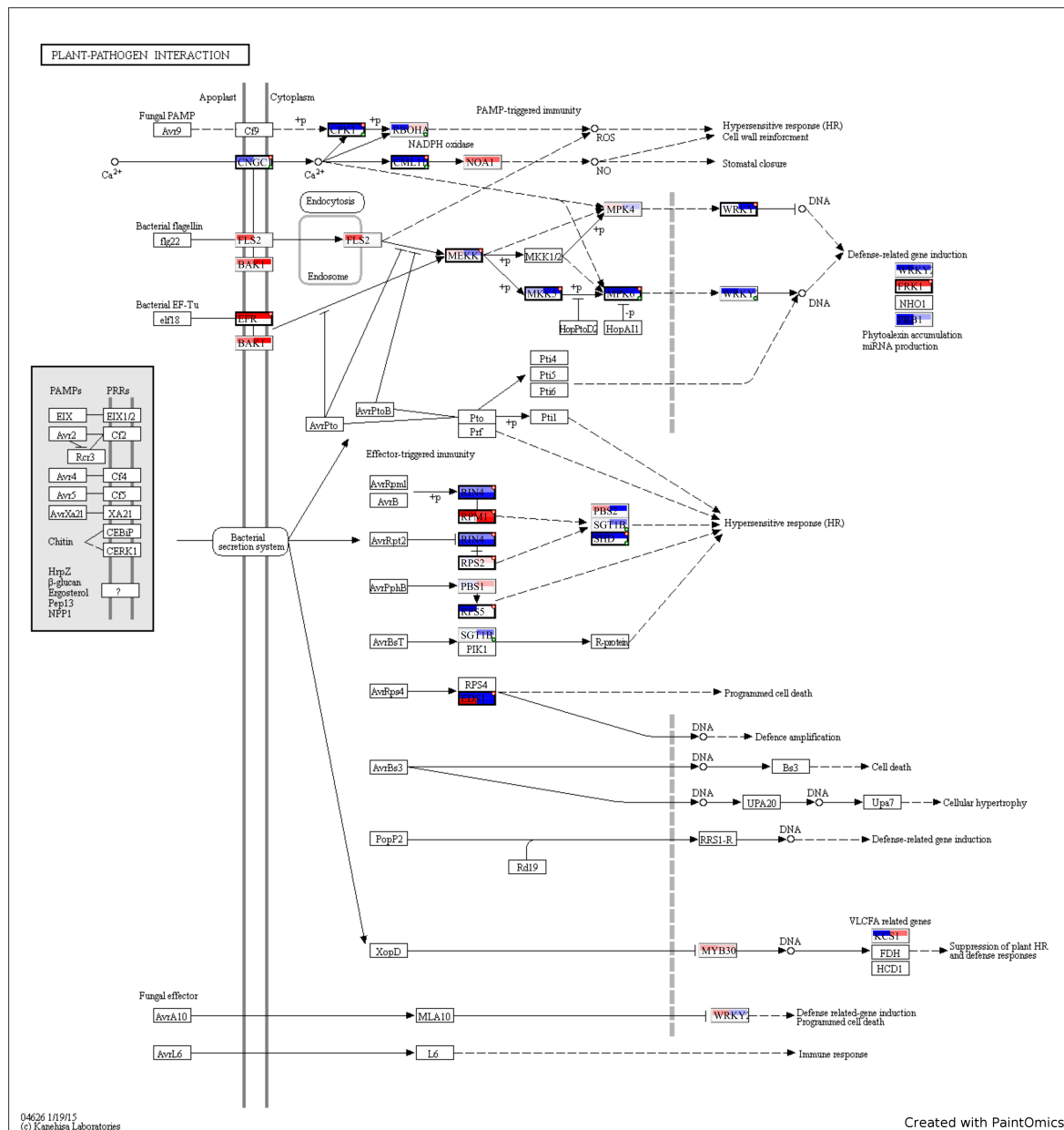
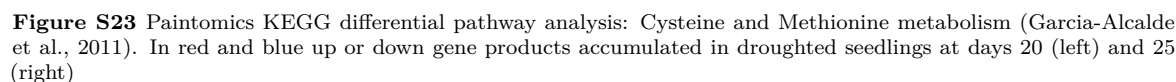


Figure S22 Paintomics KEGG differential pathway analysis: Plant-pathogen interaction(Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)



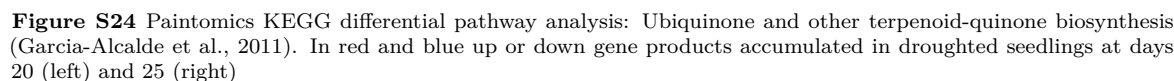


Figure S24 Paintomics KEGG differential pathway analysis: Ubiquinone and other terpenoid-quinone biosynthesis (Garcia-Alcalde et al., 2011). In red and blue up or down gene products accumulated in droughted seedlings at days 20 (left) and 25 (right)