UAB
Universitat Autònoma de Barcelona

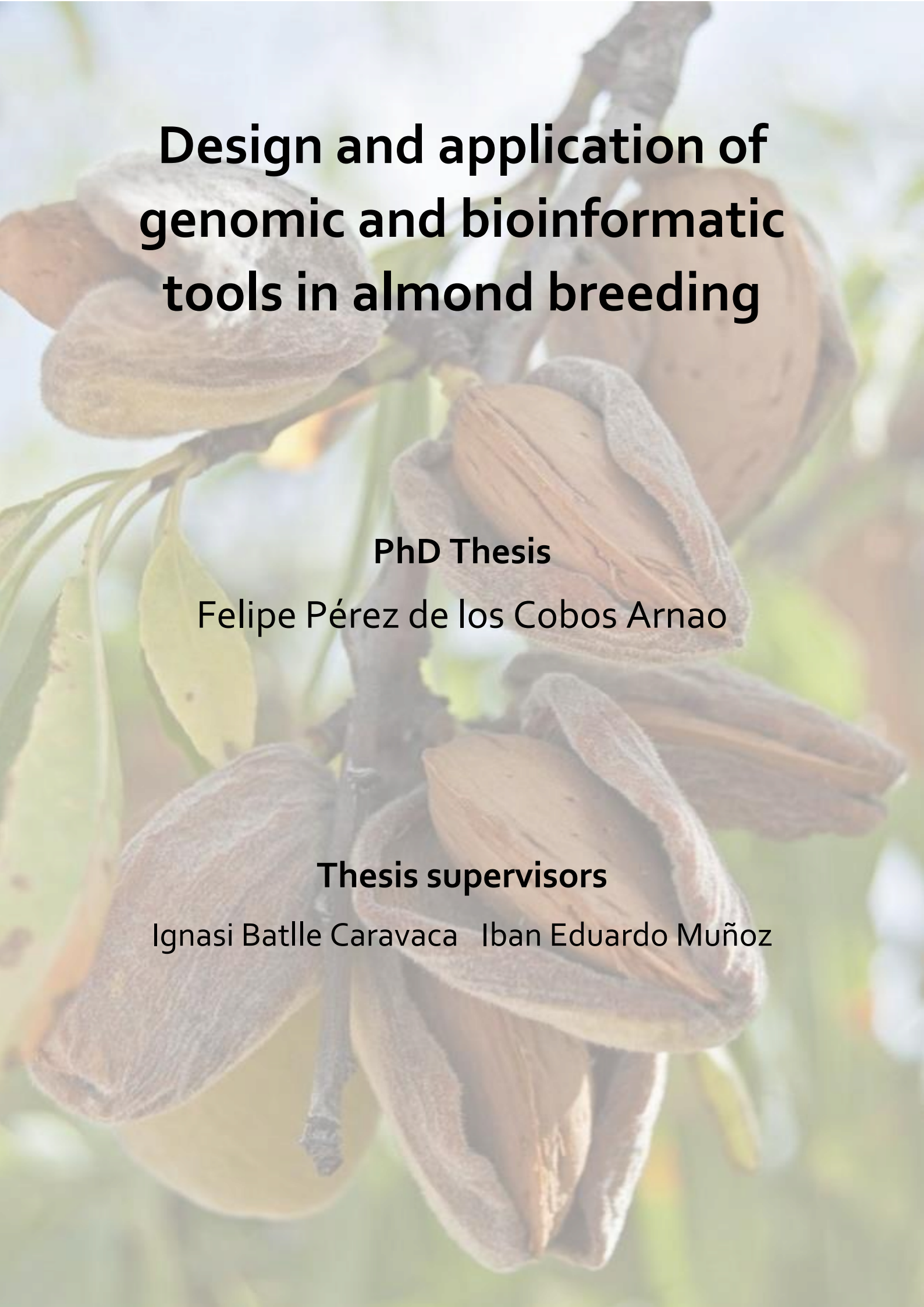
Design and application of genomic and bioinformatic tools in almond breeding

Felipe Pérez de los Cobos Arnao

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. The access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



Design and application of genomic and bioinformatic tools in almond breeding

PhD Thesis

Felipe Pérez de los Cobos Arnao

Thesis supervisors

Ignasi Batlle Caravaca Iban Eduardo Muñoz

PhD Thesis

Design and application of genomic and bioinformatic tools in almond breeding

Dissertation presented by Felipe Pérez de los Cobos Arnao for the degree of Doctor in Plant Biology and Biotechnology by Universitat Autònoma de Barcelona (UAB)

This work have been perfomed in Institut of Agrifood Research and Technology of Catalonia (IRTA) and Institut of Research of Agricultural Genomics (CRAG), Bellaterra.

PhD Candidate

Felipe Pérez de los Cobos Arnao

PhD Supervisors

Ignasi Batlle Caravaca & Iban Eduardo Muñoz

Academic Tutor

Josep Mª Casacuberta Suñer

Index

List of Figures

Figure 3.4. Top enriched terms of the most suitable candidate genes. A) Prupe.*2G196600* B) Prupe.*2G169700* C) Prupe.*7G052700* D) Prupe.*7G092200*.

Figure 4.1. Violin plot of node degree connectivity in each of the aggregated and non-aggregated networks with relaxed or stringent sparsity (COO300, COO100, HRR300 and HRR100). Boxplots of node degree connectivity were added for each violin plot.

Figure 4.2. Boxplots of the AUROC value for every subset of Bioprojects (from 2 to 26) and method used.

Figure 4.3. Scatter plot and Loess regression representation of average node degree connectivity by AUROC value for each of the GCNs (COO300, HRR300, COO100 and HRR100) in all the datasets used for network annotation (CObp, GOmf, GOcc, Pfam, KEEG, PANTHER and Mapman).

Figure 4.4. *PpPG21*, *PpPG22* and MF subnetworks. MF subnetwork is highlighted in orange.

Figure 4.5. Lollipop plot of enriched terms found in the MF subnetwork. Enriched terms were sorted by the number of genes annotated by each term.

List of Tables

List of abbreviations

MAI → Marker-assisted introgression.

MAS → Marker-assisted selection.

QTL → Quantitative trait loci.

SNP → Single nucleotide polymorphism.

GWAS → Genome-wide association analysis.

Sk → Sweet kernel.

Lb → Late blooming.

Sf → Self-compatibility.

T x E → 'Texas' x 'Early Gold'.

RAPD → Random amplification of polymorphic DNA.

RNA-Seq → Ribonucleic acid sequencing.

GCN → Gene coexpression network.

F → Inbreeding coefficient.

r → Pairwise relatedness.

GC → Genetic contribution.

OP → Open pollination.

IA → Image analysis.

GC-FID → Gas-chromatography with flame ionization detector.

FAME → Fatty acid methyl esters

LOD → Logarithm of the odds.

DNA → Deoxyribonucleic acid.

PCA → Principal component analysis.

ROH →Region of homozygosity.

NW → Nut weight.

KW → Kernel weight.

CRO → Crack-out percentage.

DK → Double kernel percentage.

GObp → Gene ontology biological process.

GOmf → Gene ontology molecular function.

GOcc → Gene ontology cellular component.

MF → Melting flesh.

COO → Co-ocurrence.

HRR → High reciprocal ranking.

AUROC → Area under the receiver operator characteristic curve.

FPKM → Fragments per Kilobase million.

PCC → Pearson correlation coefficient.

TET → Top enriched term.

LD → Linkage disequilibrium.

GBA → Guilt-by-association.

BLO → Blooming time.

Summary

Almond [*Prunus dulcis* (Miller) D.A. Webb, syn. *P. amygdalus* (L) Batsch] is the most economycaly important tree nut worldwide. In the year 2021, after duplicate its production in only 10 years, it arrived to 1.76 kernel million tons worldwide. Apart from its economic importance, almond shows a high adaptability to different environments and irrigation regimes. Additionally, its kernel has a high nutritional value, making the almond a crop with a high potential to adapt to an agriculture threatened by the climate change and the needs of feeding an increasing world population.

Modern almond breeding started in the 1920's. The first breeding programs were based on classical breeding, making controlled crosses and seedling selection to develop new almond varieties with superior performance. Currently, breeding programs are based in marker-assisted breeding, carrying on activities such as germplasm characterization, marker-assisted selection (MAS) and marker-assisted introgression (MAI). However, the number of traits included in MAS pipelines is still limited. Only three traits such as self-compatibility, sweet kernel and late blooming have molecular markers associated to specific phenotypes. This situation contrast with the progress that have been made in almond genomic research. In the last years, three reference genomes and a 60K almond SNP array have been published. At the same time, bioinformatics allow breeders and researcher to face complex biological questions with novel and powerful approaches.

The main objective of this thesis was the designing and application of genomic and bioinformatic tools and approaches applied in almond breeding. We performed a pedigree analysis of 220 accessions from 9 countries with the objective of study breeding tendencies in the last 50 years of almond breeding. Our results detected two worldwide mainstream breeding lines: one European line, based mainly in 'Tuono' and 'Cristomorto' as founders and the Californian-Australian line, based primarily in 'Nonpareil'. Indeed, the repeated use of these three founders and their related genotypes resulted in a loss of genetic variability and an increase of inbreeding in almond breeding. Additionally, we found out that the use of the cultivar 'Tuono' as a source of self-compatibility has been a common practice in most breeding programs, creating a bottleneck effect.

We also performed a QTL mapping of kernel quality traits in a F1 population coming from the cross 'Marcona' x 'Marinada'. Even if this technique have been applied in almond breeding for decades, there is still room for new approaches and improvements, as it has been proved in this thesis. The use of the almond 60K SNP array combined with novel bioinformatic protocols allowed us to build a high quality and highly saturated linkage map. Additionally, the high correlation found between traits measured with conventional and image analysis methods and the fact that the same QTLs were found indicating the accuracy of image data methods as a new phenotyping tool. We used new tools and protocols such as image analysis and the phenotypic data transformation to lsmean data. Another important innovation carried out in chapter 2 was the use of lsmean data instead of raw phenotypic data. This is an approach already mainstream in other trait-loci analyses such as GWAS, but never applied in QTL mapping before. Finally, the QTLs reported here, will allow the implementation of efficient MAS strategies applied to kernel quality traits.

Additionally, we carried out a genetic structure analysis and non-additive GWAS in a set of different almond accessions from 20 countries. Our results strongly supported the subdivision of these accessions into five ancestral groups. Each group was formed by accessions with a common geographical origin, agreeing with the archaeological and historical evidence that

separate almond dissemination into four phases: Asiatic, Mediterranean, Californian and southern hemisphere. Through a homozygosity analysis, we detected low levels of inbreeding in most of the accessions under study. However, high levels of inbreeding were detected in some breeding cultivars, agreeing with the results found in our pedigree analysis, where we concluded that breeding practices could be increasing inbreeding in almond. Also, signals of domestication were detected in chromosomes one, four and five. Among the 13 QTLs detected, only one had an additive effect. This indicated that non-additive effects could be the main source of genotype-phenotype interactions in almond and other Prunus species. Finally, the use of the peachGCN, developed in this thesis, allowed us to propose four candidate genes for the main QTLs mapped.

Finally, we created a new tool for predicting gene function that can be used for any *Prunus* breeder or researcher. For that, we constructed four GCNs from publicly available RNA-Seq data, we evaluated the performance of every GCN and finally, we validated the GCN with the best performance. To validate the performance of the GCN, we selected two well-characterized genes responsible for fruit flesh softening in peach, the endopolygalacturonases *PpPG21* and *PpPG22*. The Melting Flesh (MF) subnetwork, constituted by the genes coexpressing with *PpPG21* and *PpPG22*, was mainly formed by genes involved in cell wall organization and biogenesis, with expression regulated by ripening-related phytohormones such as ethylene, auxin and MeJA. Additionally, we found in MF subnetwork 25 genes previously reported as involved in softening, some taking part in key steps of that process. These results demonstrated that the MF subnetwork was closely related to peach fruit softening and therefore to the function of *PpPG21* and *PpPG22*.

Resumen
El almendro [Prunus dulcis (Miller) D.A. Webb, sin. P. amygdalus (L) Batsch] es el fruto seco de mayor importancia económica en todo el mundo. En el año 2021, despues de duplicar su producción en solo 10 años, su producción llegó a 1,76 millones de toneladas a nivel mundial. Aparte de su importancia económica, el almendro muestra una alta adaptabilidad a diferentes entornos y regímenes de riego. Además, su semilla tiene un alto valor nutricional, lo que convierte a la almendra en un cultivo con un alto potencial para adaptarse a una agricultura amenazada por el cambio climático y las necesidades de alimentación de una población mundial creciente.

La mejora genética moderna del almendro comenzó en la década de 1920. Los primeros programas de mejora se basaron en técnicas clásicas, realizando cruces controlados y selección de semillas para desarrollar nuevas variedades de almendro con un rendimiento superior. Actualmente, los programas de mejora se basan en la mejora asistida por marcadores, realizando actividades como caracterización de germoplasma, selección asistida por marcadores (SAM) e introgresión asistida por marcadores (IAM). Sin embargo, la cantidad de caracteres incluidos en los protocolos de MAS aún es limitada. Sólo tres caracteres, como la autocompatibilidad, la pepita dulce y la floración tardía, tienen marcadores moleculares asociados a fenotipos específicos. Esta situación contrasta con los avances que se han producido en la investigación genómica del almendro. En los últimos años se han publicado tres genomas de referencia y un chip de 60 mil SNPs de almendro. Al mismo tiempo, la bioinformática permite a los mejoradores e investigadores afrontar cuestiones biológicas complejas con enfoques novedosos y potentes.

El principal objetivo de esta tesis ha sido el diseño y aplicación de herramientas y enfoques genómicos y bioinformáticos aplicados en la mejora genética del almendro. Por ello, realizamos un análisis de pedigrí de 220 accesiones de 9 países con el objetivo de estudiar las tendencias de mejora durante los últimos 50 años. Nuestros resultados detectaron dos líneas genéticas principales a nivel mundial: una línea europea, basada principalmente en 'Tuono' y 'Cristomorto' como fundadores y la línea californiana-australiana, basada principalmente en 'Nonpareil'. De hecho, el uso repetido de estos tres fundadores y sus genotipos relacionados resultó en una pérdida de variabilidad genética y un aumento de la endogamia en el almendro. Además, descubrimos que el uso de la variedad 'Tuono' como fuente de autocompatibilidad ha sido una práctica común en la mayoría de los programas de mejoramiento, creando un efecto de cuello de botella.

También realizamos un mapeo de QTL de rasgos de calidad del grano en una población F1 proveniente del cruce 'Marcona' x 'Marinada'. Aunque esta técnica se ha aplicado en el mejoramiento del almendro durante décadas, todavía hay espacio para nuevos enfoques y mejoras, como se ha demostrado en esta tesis. El uso de la matriz de 60 mil SNPs combinada con protocolos bioinformáticos novedosos nos permitió construir un mapa de ligamiento altamente saturado y de alta calidad. Además, la alta correlación encontrada entre los caracteres medidos con métodos convencionales y de análisis de imágenes y el hecho de que se encontraran los mismos QTL indica la precisión de los métodos de análisis de imágenes como una nueva herramienta de fenotipado. Utilizamos nuevas herramientas y protocolos como el análisis de imágenes y la transformación de datos fenotípicos a datos medios. Otra innovación importante llevada a cabo en el capítulo 2 fue el uso de datos modelados en lugar de datos fenotípicos sin procesar. Este es un enfoque que ya es común en otros carácter-loci análisis como GWAS, pero que nunca antes se había aplicado en el mapeo de QTL. Finalmente, los QTL

reportados aquí permitirán la implementación de estrategias de SAM eficientes aplicadas a los caracteres de calidad de la pepita.

Además, llevamos a cabo un análisis de estructura genética y GWAS no aditivo en un conjunto de diferentes accesiones de almendras de 20 países. Nuestros resultados apoyaron firmemente la subdivisión de estas accesiones en cinco grupos ancestrales. Cada grupo estuvo formado por accesiones con un origen geográfico común, coincidiendo con la evidencia arqueológica e histórica que separa la diseminación del almendro en cuatro fases: asiática, mediterránea, californiana y hemisferio sur. A través de un análisis de homocigosidad, detectamos bajos niveles de endogamia en la mayoría de las accesiones bajo estudio. Sin embargo, se detectaron altos niveles de endogamia en algunas variedades provenientes de programas de mejora, lo que coincide con los resultados encontrados en nuestro análisis de pedigrí, donde concluimos que las prácticas de mejoramiento podrían estar aumentando la endogamia en el almendro. Además, se detectaron señales de domesticación en los cromosomas uno, cuatro y cinco. Entre los 13 QTL detectados, sólo uno tuvo un efecto aditivo. Esto indica que los efectos no aditivos podrían ser la principal fuente de interacciones genotipo-fenotipo en almendros y otras especies de *Prunus*. Finalmente, el uso de la PeachGCN, desarrollada en esta tesis, nos permitió proponer cuatro genes candidatos para los principales QTL mapeados.

Finalmente, creamos una nueva herramienta para predecir la función genética que puede ser utilizada por cualquier mejorador o investigador de *Prunus*. Para ello, construimos cuatro GCN a partir de datos de RNA-Seq disponibles públicamente, evaluamos el rendimiento de cada GCN y, finalmente, validamos la GCN con el mejor rendimiento. Para validar el rendimiento de la GCN escogida, seleccionamos dos genes bien caracterizados responsables del ablandamiento de la pulpa del melocotón, las endopoligalacturonasas *PpPG21* y *PpPG22*. La subred Melting Flesh (MF), constituida por los genes que coexpresan con *PpPG21* y *PpPG22*, estaba formada principalmente por genes implicados en la biogénesis y organización de la pared celular, cuya expresión estaba regulada por fitohormonas relacionadas con la maduración como el etileno, la auxina y los jasmonatos de metilo. Además, encontramos en la subred MF 25 genes previamente hayados involucrados en el ablandamiento, algunos de los cuales participan en pasos clave de ese proceso. Estos resultados demostraron que la subred MF estaba estrechamente relacionada con el ablandamiento del melocotón y, por lo tanto, con la función de los genes *PpPG21* y *PpPG22*.

Resum

L'ametller [*Prunus dulcis* (Miller) D.A. Webb, syn. *P. amygdalus* (L) Batsch] és, econòmicament, l'arbre productor de fruita seca més important al món. L'any 2021, després de duplicar la seva producció en només 10 anys, es van produir 1.76 millions de tones de gra a nivell mundial. A banda de la seva importancia econòmica, l'ametller presenta una gran adaptació a diferents ambients i dotacions d'aigua de regadiu. També, el seu gra té un gran valor nutricional, fent de l'ametller un conreu amb un elevat potencial d'adaptació a una agricultura amenaçada pel canvi climàtic, alhora que esdevé de gran importància, tenint present la necessitat d'alimentar una población mundial cada cop més gran.

La millora genètica moderna de l'ametller va començar als anys 1920. Els primers programes de millora es basaren en la millora clàssica mitjançant la realització d'encreuaments dirigits i la selección d'individus de llavor per a l'obtenció de noves varietats amb millor comportament que les varietats locals. Actualment, els programes de millora es basen en la millora assistida amb marcadors duent a terme activitats tals com la caracterització del germoplasma, selecció assistida amb marcadors (SAM) i introgressió assistida amb marcadors (IAM). No obstant, el nombre de caràcters inclosos en el procés de la SAM és limitat, ja que només a tres caràcters com l'auto-compatibilitat, el sabor dolç i la floració tardana, s'han pogut associar marcadors moleculars específics. Aquesta situació contrasta amb els avenços que s'han realitzat en recerca sobre la genòmica de l'ametller. En els darrers anys, tres genomes de referència i un ensamblatge de 60K han estat publicats. Tanmateix, la bioinformàtica permet als milloradors i investigadors confrontar questions complexes biològiques amb aproximacions novedoses i potents.

El principal objectiu d'aquesta tesi fou el diseny i l'aplicación d'eines genòmiques i bioinformàtiques i d'altres aproximacions, aplicades a la millora genètica de l'ametller. Inicialment es va dur a terme una anàlisi de la genealogia de 220 introduccions de 9 països diferents per tal d'estudiar les tendències de la millora genètica de l'ametller en els darrers 50 anys. Els resultats detectaren dos línies principals de millora: una la línea europea, basada principalmente en les varietats 'Tuono' i 'Cristomorto', com a fundadors i la línia Californiana-Australiana, basada principalment en la varietat 'Nonpareil'. Així, l'ús freqüent d'aquetes tres varietats i el seus genotips relacionats, va donar lloc a una pèrdua de variabilitat genètica i a l'increment de la consanguinitat en la millora genètica de l'ametller. També es va trobar que l'ús de la varietat 'Tuono', com a font d'autocompatibilitat, ha estat una pràctica comuna en la majoria dels programes de millora, originant un efecte de coll d'ampolla.

S'ha dut a terme un mapeig de QTL de trets qualitat de l'ametlla en una població F1 derivada de l'encreuament 'Marcona' x 'Marinada'. Tot i que que aquesta tècnica ja s'ha aplicat a la millora de l'ametller durant dècades, encara hi ha espai per a noves aproximacions i millores, tal com s'ha pogut comprovar en aquesta tesi. Així l'ús d'un ensamblatge de 60K de SNP combinat amb nous protocols bioinformàtics ha permès construir un mapa molt saturat d'elevada qualitat. També, l'elevada correlació trobada entre caràcters mesurats convencionalment i mitjançant mètodes d'analisi d'imatges i el fet que els mateixos QTLs es van trobar, indiquen la precissió del mètodes de dades d'imatge com una nova eina de fenotipat. Es varen utilizar noves eines i protocols tals com els d'analisis d'imatges i la seva transformació en dades Ismean. Una altra innovació important desenvolupada en el capítol 2, va ser l'ús de dades Ismean en lloc de dades fenotípiques sense processar. Aquesta és una aproximació principal en altres anàlisis de dades caràcter-loci tal com GWAS, però que mai

s'havia aplicat en mapeig de QTL. Finalment, els QTLs referits en aquest treball, permetran la implementació d'estratègies eficients SAM aplicades als caràcters de qualitat de de l'ametlla.

D'altra banda, es va realitzar una anàlisi d'estructura genètica i dades no aditives GWAS en un grup de diferents introduccions de 20 països. Els resultats argumenten la subdivisió d'aquestes introduccions en cinc grups ancestrals. Cada grup està format per introduccions amb un origen geogràfic comú, d'acord amb les evidències arqueològiques i històriques que separen l'ametller en cinc etapes l'Asiàtica, la Mediterrania, la Californiàna i la de l'hemisferi sud. A través d'una anàlisi d'homozigositat, es van detectar nivells baixos de consanguinitat en la majoria d'introduccions estudiades. En canvi, es van detectar alts nivells de consanguinitat en varietats millorades, d'acord amb els resultats obtinguts en l'anàlisi de genealogia, on es va poder concloure, que les pràctiques de millora poden augmentar la consanguinitat en l'ametller. També es van trobar senyals de domesticació als cromosomes, un, quatre i cinc. Entre els 13 QTLs detectats únicament un mostrava efectes additius. Això indica que els efectes no additius poden ser la font primària de les interaccions genotip per fenotip en l'ametller i d'altres espècies de *Prunus*. Finalment, l'ús de GCN de presseguer desenvolupades en aquesta tesi, permetrà proposar quatre gens candidats pels principals QTLs mapats.

Finalment, es va crear una nova eina per predir la funció de gens que pot utilizar qualsevol millorador o investigador de *Prunus*. Per això, es varen construir quatre GNCs a partir de dades de RNA-Seq disponibles públicament, i es va avaluar el comportament de cada GNC i finalment es varen validar les GNCs amb el millor rendiment. Per a validar el rendiment de les GNCs es seleccionaren dos gens ben caracteritzats i responsables de la tovor de la polpa del presseguer, la endopoligalacturonasa *PpPG21* i *PpPG22.* La xarxa de la Melting Flesh (MF), en presseguer està constituïda per gens coexpressant amb *PpPG21* i *PpPG22*, està formada per gens implicats en l'organització de la paret cel·lular i la seva biogènesi amb l'expressió regulada per fitohormones de la maduració, com l'etilè, l'auxina i els jasmonats de metil. També es varen trobar a la xarxa de la MF, 25 gens, prèviament citats com involucrats en la tovor de la polpa, alguns d'ells en esglaons clau del procés. Aquets resultats van demostrar que la xarxa TP està molt relacionada amb la tovor de la polpa del presseguer, i per tant amb la funció de *PpPG21* i *PpPG22*.

General Introduction

I. General Introduction

I.1. Taxonomy and botanical aspects

Almond [*Prunus dulcis* (Miller) D.A. Webb, syn. *P. amygdalus* (L) Batsch] is the most important tree nut crop in terms of commercial production. It belongs to the *Prunus* genus, included in the *Rosaceae* family. The botanical classification is as follow:

Kingdom: *Plantae*
Subkingdom: *Tracheobionta*
Division: *Magnoliophyta*
Class: *Magnoliopsida*
Subclass: *Rosidae*
Order: *Rosales*
Family: *Rosaceae*
Subfamily: *Amygdaloideae*
Genus: *Prunus*
Subgenus: *Amygdalus*
Species: *Prunus dulcis*

Rosaceae is a family with around 3,000 species. It includes species of economic importance due to their edible fruit such as apple (*Malus domestica*), pear (*Pyrus communis*) or strawberry (*Fragaria x ananassa*) or due to ornamental value such as rose (*Rosa sp*). The *Prunus* genus has over 250 species of trees and shrubs (Wen et al., 2008), some of them being important fruit tree crops as peach (*P. persica*), European plum (*P. domestica*), Japanese plum (*P. salicina*), sweet cherry (*P. avium*), sour cherry (*P. cerasus*) and apricot (*P. armeniaca*). According to the last evidence, the *Prunus* genus can be divided into five subgenera: *Amygdalus* (peaches and almonds), *Prunus* (plums and apricots), *Cerasus* (cherries), *Padus* (bird cherries) and *Laurocerasus* (Laurel cherries) (Chin et al., 2014).

Almond is a deciduous tree with different size depending on the cultivar, soil and horticultural management. Almond natural root system is characterized by a strong pivotal taproot, which reaches a very deep level in the soil. The leaves are alternate, lanceolate and have a serrated margin. The flowers are hermaphrodite, white to pale pink, with five petals and sepals, a variable number of stamens a single pistil and usually appearing before the leaves in late winter or early spring.

The almond fruit is oval to round, and it is botanically classified as a drupe. It is characterized by an outer fibrous layer or hull (pericarp and mesocarp), equivalent to the flesh of the stone fruits. The hull splits at maturity, showing the shell (endocarp) that contains the seed (kernel). In contrast to the other species of its genus, whose commercial interest lies in their fruits, almond is the only *Prunus* species cultivated exclusively for its kernels. This is due to the fact that its kernels are sweet, unlike the other species, whose kernels are usually bitter. Generally, a single kernel is present within the shell, but occasionally two kernels occur. After the fruit matures, the mesocarp splits and separates from the shell, and an abscission layer forms between the stem and the fruit so that the fruit can fall from the tree.

I.2. Origin and dissemination

I.2.1 Origin

The geographic origin of almond is not clear yet, but the last evidence places it somewhere between the Eastern Mediterranean, Southwest Asia and Central Asia. The most accepted

theory about the origin of the species is that almond originated from hybridizations with several wild relatives, including species such as *P. fenzliana*, *P. orientalis, P. bucharica*, *P. kuramica* or *P. webbi*. Over 20 wild almond species are native to Western and Central Asia, however, these species are morphologically different to almond. Probably, the domesticated almond originated by hybridization between wild almond species encouraged by their outcrossing nature (Evreinoff, 1958; Grasselly, 1976a; Browicz and Zohary, 1996; Ladizinsky, 1999).

Apart from morphology and distribution, studies using molecular markers are in agreement with the hybridization theory. Zeinalabedini et al., 2010 proposed *P. fenzliana* as the most probable ancestor of almond based on data from nuclear and chloroplast SSRs. However, this study included a limited number of species, since it did not include neither *P. kuramica* nor *P. bucharica* specimens. A more comprehensive study pointed *P. kuramica* as the closest wild relative of the cultivated almond (Delplancke et al., 2016).

Archaeobotanic evidence also supports an Asian origin of the species. Apparently, almonds were collected from the wild a long time before domestication. The oldest *Prunus sp* nuts remains found so far date back to the Epipalaeolithic (20,000-10,000 b.c.), in the Ohalo II site, and to the Pre-Pottery Neolithic A (10,000-9,000 b.c.), in the Netiv Hagdud site, Israel (Kislev et al., 1992, 1997). Other remains in similar periods of time were found in the Çayönü and Hallan Çemi Tepesi sites, in Turkey (Van Zeist and de Roller, 1992; Rosenberg et al., 1995). In Europe, the first site where pre-domesticated *Prunus sp* shells have been found is the Franchthi Cave, Southern Greece, in beds dating back to the Mesolithic and Neolithic (15,000-3,000 b.c.) (Hansen, 1991).

The first almond remains suspected of being cultivated date back to the Early Holocene (10,000-9,000 b.c.), found in Jerf el Ahmar and Tell Qaramel sites, Nothern Syria (Willcox et al., 2008a). On these sites there are several evidences of early plant cultivation: wild species occurring outside their natural habitats, weeds associated with cultivation increased with time, a gradual decrease in gathered plants such as small seeded grasses and *Polygonum*/*Rumex*, and barley grains increased in breadth and thickness with time. Also, rodent droppings were found on these sites, suggesting large-scale grain storage. In the site of Bab edh-Dhra, in Jordan, evidence of almond cultivation dating back to the Early Bronze Age (3,300-2,100 b.c.) has been found. In this site, almonds appear with numerous remains of grape vine and olive, suggesting their cultivation (McCreery, 1979).

I.2.2 Dissemination

Almond was one of the first domesticated tree crops along with figs, olives, dates and pomegranates. Delplancke et al., 2013 proposed the Fertile Crescent as the center of domestication, since they detected a pattern of decreased genetic variability towards east and west of that point, in a population of 1032 almonds from sites covering the whole Mediterranean Basin and part of western Asia. From its center of domestication, it expanded rapidly to the Western Mediterranean and Central Asia (Figure I.1). These patterns are consistent with an almond human-driven dispersal through trade routes. Greeks and Phoenicians played a major role on the almond dissemination, as they brought almonds to their colonies spread throughout the Mediterranean. In the Western Mediterranean, almond remains started to appear since the 9,000 b.c., coinciding with the establishment of Greek and Phoenician colonies in that region (Pérez-Jordà et al., 2021).

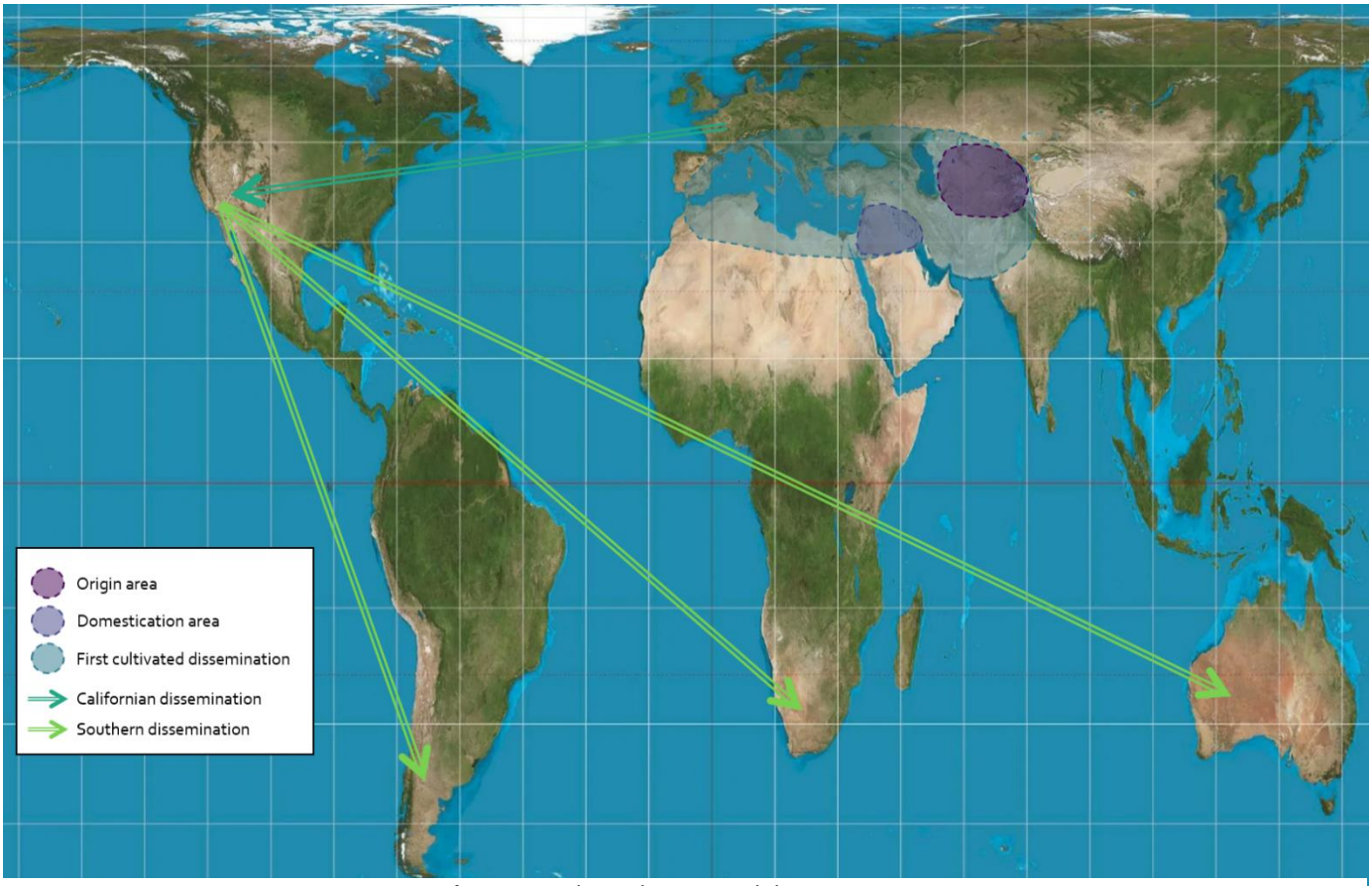


Figure I.1. Almond origin and dissemination.

During its dissemination, almond cultivation was established along the Mediterranean. By the time of the Roman Empire, its cultivation was generalized all over the Empire. The adaptation to different environments, accelerated by propagation by seeds and the outcrossing character of the species, formed different genetic populations. Several studies have analyzed the genetic structure of the cultivated almond, finding a strong geographical component in those populations of Mediterranean origin (Shiran et al., 2007; Gouta et al., 2010; Elhamzaoui et al., 2012; Cabrita et al., 2014; Fernández i Martí et al., 2015; Halász et al., 2019; Pavan et al., 2021).

After this first expansion, almond cultivation remained relatively static for hundreds of years. Only a few exchanges of genetic material were produced due to the movements of the borders of that time. For instance, the introduction of North African material during the Arab occupation of the Iberian Peninsula (711 – 1492). The next significant dissemination took place during the 16th and 17th centuries, when Spanish missions and explorers introduced almonds to America (Figure I.1). Franciscan monks are credited with planting the first almond trees in California. However, an important introduction of soft-shelled French cultivars to California was produced during the 1850-1900. Today's Californian commercial cultivars were bred from that material.

The last stage in the almond dissemination process is the introduction of its cultivation to the Southern Hemisphere (Figure I.1). Californian almond cultivars and cultural managements methods were introduced to Argentina, Australia, Chile and South Africa during the early to mid-19th century.

Post-domestication hybridization with wild species have been also reported during the almond dissemination. Denisov, 1988 detected recombinant and intermediate phenotypes of *P. fenzliana* and *P. dulcis* in the Caucasus, in some areas where both species co-exist. Additionally, *P. webbi*, a wild almond species native to Southern Italy, is presumed to be the source of self-compatibility in Italian almond cultivars (Godini, 2000). Finally, Delplancke et al., 2012 detected wild-to-crop and crop-to-wild gene flow between almond and its wild counterpart *P. orientalis*.

I.3. Production

Almond world production (1.76 kernel million tons in 2021) is led by the state of California, in the USA (Australian Almond Board, 2022). California represents a 79% of the almonds produced worldwide. It is followed by Australia and Spain, with an 8% and 6% of the worldwide production. The rest of the production is shared by several countries like Turkey (1%), Tunisia (1%), Portugal (1%), Morocco (1%), Chile (1%), Greece (0.5%), Italy (0.2%), Iran (0.2%) and others (2%) (Figure I.2).

In worldwide terms, we can differentiate two almond production models: The traditional Asian-Mediterranean and the Californian-Australian. In the Mediterranean Region and Western and Central Asia, traditional orchards are characterized by the use of locally-adapted hard-shelled cultivars grafted onto almond rootstocks. This culture system minimizes the inputs of water, labour and fertilizers, resulting in a very low productivity. However, due to the increasing demand and attractive prices, very productive orchards using intensive cultural practices, irrigation and improved cultivars and rootstocks have been developed in Spain, Portugal, Italy, Morocco, Tunisia, Israel, Syria and other countries. Almond production in California and Australia is characterized for the reduced germplasm in which the industry is based. Most of the cultivars currently used in these countries were developed from soft-shelled almonds introduced from France to California in the early 1900s. Among these cultivars, 'Nonpareil' is the dominant one, since it represents the 39% and 47% of the total almond production of California and Australia, respectively (Australian Almond Board, 2022; Californian Almond Board, 2022). In these two countries, almonds are usually grafted onto peach, almond x peach or almond x plum hybrid rootstocks, more tolerant to irrigated and heavy soils. These orchards are designed to optimize productivity, so the use of intensive management systems and high water and fertilizers inputs are common. Additionally, countries like Chile, Argentina or South Africa are adapting this production model to their new orchards.

I.4. Classical breeding

I.4.1. History of breeding

The establishment of almond cultivation in the Mediterranean Region and Western and Central Asia started a process of selection and adaptation sustained over thousands of years. This process continued until the 20th century, where almond production relied on locally adapted cultivars and landraces. During this century, several almond-producing countries established the first breeding programs with the idea of almond production efficiency, based on controlled crosses and offspring selection.

The first almond breeding programs started in California (EEUU) in the 1920s. This effort was followed by the former Soviet Union in the 1930s, France and Israel in the 1960s and Spain, Italy,

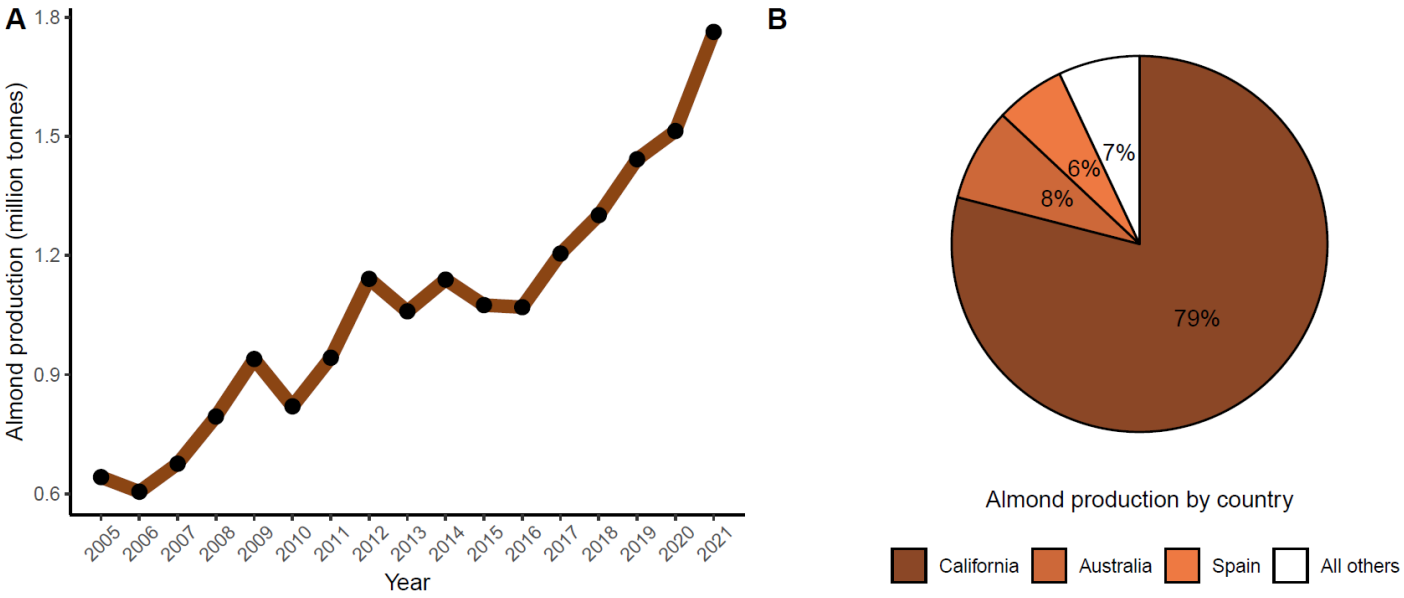


Figure I.2. A) Almond production from 2005 to 2021. B) Almond production by country in 2021.

Greece, Tunisia and Argentina in the 1970s. More recently, breeding programs were started in Australia in the early 2000s and in Turkey in the 2010s. Although private breeding programs exist in California, most of them are public or public-private initiatives.

Although California was the first almond-growing region to start a breeding program, the success of their breeding efforts has been limited. 'Nonpareil' (a seedling selection of unknown origin) remains the main cultivar planted, while modern cultivars coming from breeding initiatives have a low impact in the market. Australia has a situation similar to California, even if the University of Adelaide breeding program has released several cultivars like 'Mira', 'Capella', or 'Vela', their production is residual. Only 'Independence', a self-compatible cultivar released by Zaiger Genetics in 2012, is among the top ten cultivars more planted in both California and Australia (Australian Almond Board, 2022; Californian Almond Board, 2022).

The INRA almond breeding program was the most successful in Europe for many years. Cultivars like 'Ferragnès', 'Ferraduel' or the self-compatible 'Lauranne' had a great impact in the market, being widely planted in France and also in Spain. In fact, 'Lauranne' is still one of the cultivars most planted in Spain. Additionally, these cultivars have been used as parents in the three Spanish breeding programs for many years, being ancestors of most of the Spanish modern cultivars.

In Spain, 'Guara' (syn. 'Tuono'), a traditional Italian self-compatible cultivar released by the CITA breeding program in 1987 by error, was the first cultivar supposedly coming from a Spanish breeding program widely planted. Currently, it represents a 27% of the almond orchards in Spain (Nuts production in Spain, 2021). Other cultivars like 'Masbovera' or 'Antoñeta' were released in the same period, but less planted. More recently, modern cultivars such as 'Vairo', 'Marinada', 'Penta', 'Makako', 'Mardía' or 'Vialfas' are replacing traditional cultivars. Only a few traditional cultivars such as 'Marcona' or 'Desmayo Largueta' and their pollinizers are still planted (Nuts production in Spain, 2021).

I.4.2. Breeding cycle

Almond breeding is aimed at developing new almond varieties with improved traits. The new almond varieties have to meet the demands of three different players in the almond industry: farmers, traders/processors and final costumers. Farmers demands are focused on agronomic traits such as yield, disease resistance, adaptability to different growing conditions, etc. On the other hand, traders, processors and final costumers focused on kernel quality traits such as flavor, physical traits, chemical composition, etc. Develop almonds trees meeting those demands involves carefully selecting and crossing almond trees to create offspring combining desirable characteristics of both parents.

The first step in almond breeding is to define the specific goals and objectives of the breeding program. For instance, major objectives in Californian breeding efforts include improved pollinizers for 'Nonpareil', high production, soft shell, resistance to the noninfectious bud failure disorder and kernel characteristics similar to 'Nonpareil'. In the Mediterranean Region, all the breeding programs have similar objectives: self-compatibility, late blooming, high production, hard shell, disease resistance and kernel quality.

Controlled crosses and selection are the main strategy used by almond breeders worldwide. Breeders choose parents that possess desirable traits and cross them during the blooming season. This involves transferring pollen from the flowers of one selected parent (commonly referred as the father) to the stigma of another parent (the mother). After maturing, fruits are collected, stratified during the winter and after germination, the seedlings are planted in early spring.

The seedlings are grown on their own roots, evaluated and selected in field conditions during years. Typically, breeders establish selection criteria based on the desired traits, and seedlings that meet these criteria are chosen for further evaluation. The selection criteria may vary depending on the breeding objectives and the specific traits being targeted.

The selected seedlings or selections are then propagated onto commercial rootstocks and their performance is evaluated in different growing conditions, including multiple locations and seasons. This helps to determine the adaptability and stability of the selections and ensures they can perform well across different environments.

After multiple rounds of evaluation and field trials, the best-performing selections are chosen as advanced selections. These advanced selections possess the desired traits and show consistent performance across various environments. Advanced selections are thoroughly tested, and if they demonstrate superior performance and market potential, they are registered and released for commercial production.

Almond breeding is a long-term process that requires patience, expertise, and careful observation. It often takes decades from the initial cross-pollination to the release of a new almond variety. However, through diligent selection and breeding efforts, almond breeders contribute to the continuous improvement and innovation in the almond industry.

I.5. Molecular breeding

I.5.1. Marker-assisted breeding

Marker-assisted breeding includes all the breeding activities based on molecular markers. These activities mainly include germplasm characterization, marker-assisted selection (MAS) and marker-assisted introgression (MAI).

Germplasm characterization

When a breeding program starts using molecular markers, the genetic characterization of its germplasm is one of the first must-do. It refers to the analysis and evaluation of the genetic composition of a collection of plant genetic resources. Germplasms typically represent a diverse range of plant varieties and wild relatives, which serve as a valuable source of genetic diversity for plant breeding. Its genetic characterization provides valuable information to plant breeders, enabling them to make informed decisions, optimize breeding strategies, avoid inbreeding, and preserve the genetic variability within the germplasm.

In almond, many efforts have focused on characterize genetically the germplasms developed in producing countries such as the USA (California), Australia, Spain, Italy, Tunisia, Portugal, etc (Shiran et al., 2007; Gouta et al., 2010; Elhamzaoui et al., 2012; Cabrita et al., 2014; Fernández i Martí et al., 2015; Halász et al., 2019; Pavan et al., 2021). These studies have detected a high genetic diversity and a strong population structure mainly based on the geographic origin of the accessions under study. Additionally, other studies have focused on determining the genetic relatedness of almond with other wild relatives (as mention in section I.2.1.).

Marker-assisted selection

MAS involves the use of molecular markers (SNPs, SSRs, etc) associated with a particular trait of interest, to identify and select plants with that desired trait before it is expressed. Markers are used to screen the seedlings produced by the breeding program every year. Each seedling is genotyped and those possessing the desired alleles associated with the trait of interest are selected for further evaluation (Figure I.3). MAS accelerates the breeding process by allowing breeders to focus their efforts on plants with the highest probability of carrying the desired traits. It helps reduce the time and resources required for field-based evaluations and increases the precision of selection. Currently, three almond traits are suitable for MAS. Self-compatibility has been one of the major objectives in many breeding programs, and the first molecular markers to detect it were developed more than 20 years ago (Tamura et al., 2000; Sutherland et al., 2004; Ortega et al., 2005; Gómez et al., 2019b). More recently, a SNP was detected to be the causal mutation conferring the sweet kernel trait to almond, open the door to perform MAS for this trait (Sánchez-Pérez et al., 2019a). Additionally, the IRTA almond breeding program designed a SNP to detect the *Lb* allele (delaying the bloom around 10-15 days, details can be found in section 1.5.2.) and it is routinely applied to select for late blooming (Ignasi Batlle, personal communication).

Marker-assisted introgression

MAI consists in the use of molecular markers to efficiently transfer a specific trait or a set of traits from a donor parent to a recipient parent. It allows breeders to introduce desirable traits from wild relatives or exotic germplasm (usually the donor parent) into elite breeding lines or cultivars (usually the recipient parent) minimizing the loss of desirable genetic background.

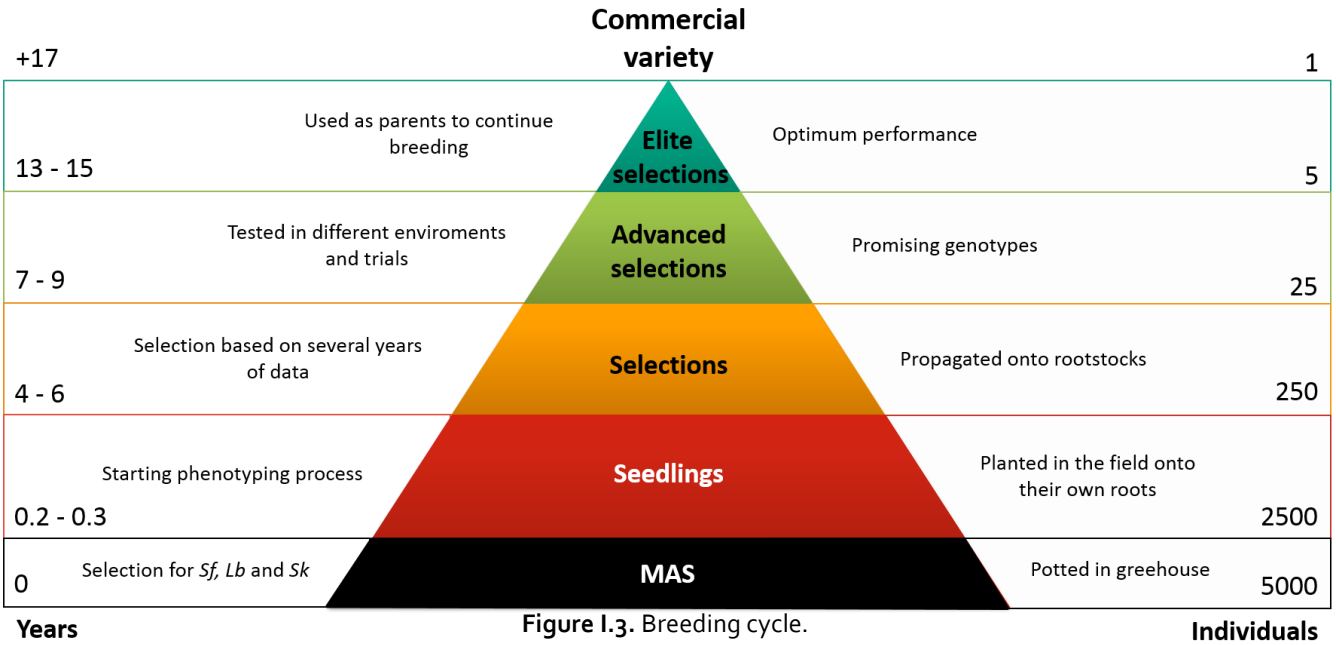


Figure I.3. Breeding cycle.

Unfortunately, the examples of MAI in almond breeding are scarce. Even if the UCD breeding program used several *Prunus* species (from peach to wild almonds) to transfer self-compatibility to almond, they did not use any molecular marker during the process (Gradziel et al., 2001; Gradziel, 2022). The only effort focusing on MAI and using almond as recipient parent is being done by the Israelite breeding program. In this project, they are transferring photosynthetic stem capability from *Prunus arabica* to the cultivated almond (Brukental et al., 2021; Trainin et al., 2022). Another effort of MAI using almond donor parent, is the development of the first collection of ILs in a tree species where each line present a unique almond chromosomic fragment in the peach genetic background (Kalluri et al., 2022).

I.5.2. Development of new molecular markers

For the development of new molecular markers applied in breeding, trait loci analyses are the approaches mainly used. As the name indicates, trait loci analyses are methods that links or associates two different types of information: phenotypic data (trait measurements) and genotypic data (usually from molecular markers). The aim of these analyses is to identify genomic regions associated with a trait of interest. They may vary depending on the complexity of the character under study and the type of population used. Basically, there are two types of traits: mendelian (or qualitative) traits and quantitative traits. Mendelian traits are controlled by a single gene, and their inheritance follows the principles of Mendelian genetics (Mendel, 1865). They exhibit a discrete or discontinuous variation, meaning they can be classified into distinct categories (e.g. self-compatibility or self-incompatibility, sweet kernel or bitter kernel, etc). Quantitative traits are controlled by multiple genes, often with a complex interaction between genetic and environmental factors (Doerge, 2002). They exhibit continuous variation, meaning they fall on a spectrum and can be measured on a quantitative scale (e.g. blooming time, yield, etc).

Quantitative trait loci mapping

Quantitative trait loci mapping, or QTL mapping, is primarily used in the study of complex traits, which are influenced by multiple genes and environmental factors. It aims to identify specific genomic regions, known as QTLs, that contribute to the observed variation in the trait

of interest. QTL mapping typically involves analyzing the genetic variation in a controlled population, such as an experimental cross between different individuals, and correlating that variation with the trait measurements. The main goal is to determine which genetic markers (often SSRs or SNPs) are linked to the trait, providing insights into the genetic architecture underlying the trait.

In almond, QTL mapping is usually performed in F1 populations (the cross between two genotypes). In this kind of populations, the variability is reduced to a maximum of four alleles per locus (two coming from the female parent and other two from the male parent) making it less representative of real-world scenarios.

Genome-wide association analysis

Genome-wide association analysis, or GWAS, is a screening of the entire genome of a species looking for associations between genetic markers and traits of interest. It uses thousands or even millions of molecular markers (usually SNPs), covering the whole genome of the species. GWAS is typically conducted in populations including individuals from different backgrounds, making it more representative of real-world scenarios.

This technique has been applied in almond with relative success. Three different GWAS have been performed so far, detecting different QTLs for traits of interest to breeding, as detailed in section 1.5.3 (Di Guardo et al., 2021; Pavan et al., 2021; Sideli et al., 2023). However, the germplasms used in these studies was reduced, limiting the effectiveness of this technique.

Genomic prediction

Genomic prediction (or genomic selection) aims to predict the performance of individuals based on their genome. As GWAS, it uses thousands or even millions of molecular markers covering the whole genome of the species. Then, that genotypic information is combined with phenotypic data to build a statistical model that link molecular markers to the trait of interest. It is especially useful for traits that are influenced by multiple genes or difficult to phenotype. The main difference with GWAS is that in GWAS, a mathematical model is created to test if a genetic variant is associated to a trait of interest. So, for every molecular marker, that model is run. In genomic prediction, a unique mathematical model using all the molecular markers is created to predict the trait of interest.

This technique requires a large number of individuals and molecular markers to have enough statistical power, limiting its applicability in almond or other tree crops species. Due to these drawbacks, there are no studies focusing on genomic prediction in almond and examples in other *Prunus* species are scarce (Fu et al., 2022; Hardner et al., 2022; Li et al., 2023).

I.5. Almond genetics

I.5.1. Almond genome

Almond has a diploid genome with eight chromosomes, like most species in the *Prunus* genus (2n = 2x = 16). It has a small genome, with approximately 300 Mb (Sánchez-Pérez et al., 2019a; Alioto et al., 2020a; D'amico-Willman et al., 2022). According to its out-crossing nature, it is one of the most polymorphic species among the fruit trees. Indeed, almond diversity levels have been reported as high as those from wild almond species (Delplancke et al., 2012). The almond genome presents a high synteny and collinearity with other *Prunus* species. This allow the

production of fertile hybrids between species of different sections within the *Prunus* genus, such as almonds and peaches (Brukental et al., 2021; Gradziel, 2022; Kalluri et al., 2022).

I.5.2. Qualitative characters and major genes

Self-incompatibility

Almond, as other *Prunus* species, shows a gametophytic self-incompatibility system. The viability of a cross depends on the *S*-locus, since the haplotype pollen genome and diploid pistil genome have to carry different *S* alleles in order to allow the pollen tube to grow. Although the specificity of the GSI reaction can be explained by assuming a single locus with multiple co-dominant S alleles, two separate genes at the *S* locus control pollen and pistil self-incompatibility system. Since these two genes are tightly linked to each other and behave as if they are a single locus, the term *S* haplotype is used to refer to this situation. The pistil *S* allele is controlled by a Ribonuclease (RNase) and the pollen allele by an F-box gene (Kao and Tsukamoto, 2004).

Kernel bitterness

Kernels of wild almond species are usually bitter and highly toxic to humans and predators because cyanogenic diglucoside amygdalin accumulates in the cotyledons. Genetic studies showed that sweet almond kernels in the cultivated almond originated from a dominant mutation within the almond linkage group 5, at a locus referred to as Sweet kernel (*Sk*) (Sánchez-Pérez et al., 2007). A study that de novo assembled the homozygous sweet cultivar 'Lauranne', showed that a mutation in the *bHLH2* transcription factor, involved in the synthesis of amygdalin, resulted in the almond sweet kernel trait (Sánchez-Pérez et al., 2019a).

Late blooming

Blooming time is considered to be quantitatively inherited, and most results on the transmission of flowering time in almond show this quantitative behavior (Kester, 1965; Grasselly and Gall, 1967; Grasselly, 1978; Vargas and Romero, 2001). However, Kester (1965) (Kester, 1965) suggested the existence of a single dominant gene determining blooming time in the late-blooming budsport 'Tardy Nonpareil'. Progenies of 'Tardy Nonpareil' showed a bimodal distribution in blooming time, indicating that a single gene was responsible of most of the phenotypic variation of the trait. Today, many studies have confirmed the existence of a major gene, named *Lb*, which delays bloom 10-15 days on average (Ballester et al., 2001; Silva et al., 2005; Sánchez-Pérez et al., 2007). The *Lb* allele has been mapped in the linkage group four in almond, although the responsible gene has not been identified yet.

I.5.3. Quantitative traits

To date, several studies have focused on mapping QTLs in almond. The first efforts were based on SSRs using traditional QTL mapping in F1 progenies (Sánchez-Pérez et al., 2007; Fernández i Martí et al., 2011, 2013a; Font i Forcada et al., 2012) or association mapping in germplasm populations (Font i Forcada et al., 2015b, 2015a). These studies found several QTLs associated to different agronomic and kernel quality traits. More recently, the development of high-throughput sequencing technologies and the almond 60K SNP array (Duval et al., 2023a) allowed to perform genome-wide association analysis (GWAS) in almond. Up to three studies have used this technique in almond so far (Di Guardo et al., 2021; Pavan et al., 2021; Sideli et al., 2023), focusing on kernel quality traits. A major QTL associated to crack-out percentage in chromosome two has been reported in several of the mentioned studies. QTLs reported in almond are further discussed in Section 2.1 and Section 2.4.3.

I.6. Genomic resources and tools

I.6.1 The *Prunus* reference map and other almond linkage maps

One of the first genomic resources in almond and other *Prunus* species was the *Prunus* reference map, a linkage map of the F2 population from 'Texas' (almond) x 'Early Gold' (peach), also known as the TxE map. The map was initially constructed with 226 RFLPs and 11 isozyme markers (Joobeur et al., 1998) and has been improved over the years with more RFLPs and the addition of SSRs markers (Joobeur et al., 2000) and later reconstructed using only SNPs and SSRs (Donoso et al., 2015). As a result of the high transferability of the T×E markers developed and the high synteny between *Prunus* species, several maps constructed with markers in common were interconnected and the position of 28 major genes was integrated into an interspecific consensus reference map (Dirlewanger et al., 2004). Among these 28 mapped major genes, 19 were mapped in peach progenies, 6 in almond or almond × peach, 2 in apricot and 1 in a myrobalan plum.

Apart from the *Prunus* reference map and other interspecific peach x almond maps, several genetic maps have been developed specifically in almond. The first linkage analysis in almond was performed by Arús et al. 1994 (Arus et al., 1994), but it only included 10 isozyme genes. The first complete map for almond was constructed by Viruel et al. 1995 (Viruel et al., 1995) using 120 restriction fragments polymorphisms (RFLPs) and seven isoenzymes in a F1 progeny between 'Ferragnes' and 'Tuono'. This map comprised the eight expected linkage groups and spanned approximately 400 cM. The next map, coming from the cross 'Felisia' x 'Bertina', used 81 RFLPs and five random amplified polymorphic DNAs (RAPDs) was published in 1998 (Ballester et al., 2001). In 2007, a map using 79 SSRs was stablished in a 'R1000' x 'Desmayo Largueta' progreny (Sánchez-Pérez et al., 2007). Tavassolian et al. 2010 (Tavassolian et al., 2010) created a linkage map coming from the cross 'Nonpareil' x 'Lauranne' using 157 markers (93 SSRs, 35 ISSRs, 14 SNPs, 4 S-alleles, and 11 RAPDs). This map was later saturated using genotyping-by-sequencing and improved including other progenies with 'Nonpareil' and 'Lauranne' as recurrent parents (Goonetilleke et al., 2018). Finally, a map was developed for the F1 population 'Vivot' x 'Blanquerna' using 56 SSRs (Font i Forcada et al., 2012; Fernández i Martí et al., 2013a).

I.6.2 The almond reference genomes

In 2019, the almond cv. 'Lauranne' reference genome was published (Sánchez-Pérez et al., 2019a). It was sequenced using a combination of Illumina and PacBio technologies and the final assembly had 246 Mb. In that study, it was reported that a SNP was the causal mutation conferring the sweet kernel trait to almond, allowing its domestication. That SNP was reported as a non-synonymous point mutation (Leu to Phe) in the dimerization domain of the bHLH2 transcription factor, preventing transcription of the two cytochrome P450 genes and resulting in the sweet kernel trait. Recently, another point mutation in the bHLH2 transcription factor associated to the sweet kernel trait has been reported (Lotti et al., 2023).

Soon after, an independent study published the almond cv. 'Texas' reference genome (Alioto et al., 2020a). In this case, it was sequenced using Illumina and Oxford Nanopore technologies and the genome size was 238 Mb. In this study, they performed a comprehensive comparison of the peach and almond genomes, focusing on the role of transposable as a source of diversification in *Prunus*.

Last but not least, in 2022, the cv. 'Nonpareil' reference genome was released (D'amico-Willman et al., 2022). It was sequenced using a combination of Illumina and PacBio technologies, with a genome size of 256 Mb. Additionally, they reported the first draft plastid and mitochondrion assemblies for almond and whole-genome methylome data of different tissues of 'Nonpareil' was provided as a supplementary resource to the almond research community.

I.6.3 The almond 60k SNP array

SNPs arrays are a high-throughput and cost-effective genotyping technology used to detect and analyze genetic variations (SNPs) in an organism's DNA. In Duval et al. 2023 (Duval et al., 2023a), the first high-density almond SNP array was presented to the almond scientific community. In this study, they used 81 almond resequences to filter and pre-select a set of 71,846 SNPs. After that, they genotyped 210 almond accessions to finally chose 60,581 SNPs, including those linked to *RMja* (nematode resistance) (Van Ghelder et al., 2010) and *Sk* (sweet kernel) genes (Sánchez-Pérez et al., 2019a). The rate of missing data was between 0.4% and 2.7% for almond accessions and less than 15.5% for peach and wild almond accessions, suggesting that this array can be used also for peach, interspecific peach × almond and wild almond genetic studies.

I.7. Bioinformatics in almond breeding

Bioinformatics is an interdisciplinary field that combines biology, computer science, and statistics to gather, store, analyze, and interpret biological data. It involves the development and application of computational tools, algorithms, and databases to study and understand biological systems at the molecular level. The field of bioinformatics emerged as a response to the explosion of biological data generated by various high-throughput technologies, such as DNA or RNA sequencing. It provides the means to handle and extract meaningful information from these large and complex datasets.

In the context of almond breeding, bioinformatics is applied mostly in genomics, transcriptomics and their association with phenotypic data. One of the most significant contributions of bioinformatics to almond breeding have been the de novo assemblies and annotation of three almond genomes, 'Lauranne', 'Texas' and 'Nonpareil'. Another successful application of bioinformatics in almond genomics was the development of the almond 60K SNP array. The SNPs in this array were filtered and selected by aligning 81 almond resequences against 'Texas' reference genome. And finally, one of the most widely used tools in breeding integrating genomic and phenotypic data is GWAS. Its uses in almond breeding are discussed in section I.5.2.

In the context of transcriptomics, RNA-Sequencing (RNA-Seq) is the method most widely used. It is a high-resolution, sensitive and high-throughput approach used to study the regulation of complex traits at the transcriptome level. The usual pipeline in a RNA-Seq is to quantify gene expression levels and identify differentially expressed genes between samples or conditions. After that, the biological pathways, gene ontology terms or functional categories that are overrepresented among the differentially expressed genes are identified through a functional enrichment analysis. In almond, several studies have focused on RNA-Seq to study diverse biological processes such as response to biotic or abiotic stresses (Mousavi et al., 2014; Moll et al., 2022), flowering time (Prudencio et al., 2021), self-incompatibility (Gómez et al., 2019a) or fruit drop (Guo et al., 2021).

Another approach used in transcriptomics to study complex traits or biological processes is the creation of gene coexpression networks (GCNs). No GCN has ever been developed in almond, but some of them have been reported in other *Prunus* species such as peach or apricot (García-Gómez et al., 2020; Jiang et al., 2023; Wang et al., 2023). GCNs are discussed in more detail in Section 4.1.

Although almond breeding has focused on genomics and transcriptomics, other bioinformatics approaches such as epigenomics have been used to study non-infectious bud failure (D'Amico-Willman et al., 2022b). Even if bioinformatics plays a crucial role in modern plant breeding, its implementation in almond breeding is at its first steps. Integrating these computational tools into almond scion breeding programs is needed to create more efficient, precise, and targeted breeding programs and accelerating the development of improved almond varieties.

I.8. References

Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., et al. (2020). Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *Plant J.* 101, 455–472. doi: 10.1111/tpj.14538.

Arus, P., Olarte, C., Romero, M., and Vargas, F. (1994). Linkage analysis of ten isozyme genes in F1 segregating almond progenies. *J. Am. Soc. Hortic. Sci.* 119, 339–344. doi: 10.21273/jashs.119.2.339.

Australian Almond Board (2022). Almond insights 2021/22.

Ballester, J., Socias I Company, R., Arus, P., and De Vicente, M. C. (2001). Genetic mapping of a major gene delaying blooming time in almond. *Plant Breed.* 120, 268–270. doi: 10.1046/j.1439-0523.2001.00604.x.

Browicz, K., and Zohary, D. (1996). The genus Amygdalus L.(Rosaceae): species relationships, distribution and evolution under domestication. *Genet. Resour. Crop Evol.* 43, 229–247.

Brukental, H., Doron-Faigenboim, A., Bar-Ya'akov, I., Harel-Beja, R., Attia, Z., Azoulay-Shemer, T., et al. (2021). Revealing the Genetic Components Responsible for the Unique Photosynthetic Stem Capability of the Wild Almond Prunus arabica (Olivier) Meikle. *Front. Plant Sci.* 12, 1–14. doi: 10.3389/fpls.2021.779970.

Cabrita, L., Apostolova, E., Neves, A., Marreiros, A., and Leitão, J. (2014). Genetic diversity assessment of the almond (*Prunus dulcis* (Mill.) D.A. Webb) traditional germplasm of Algarve, Portugal, using molecular markers. *Plant Genet. Resour. Characterisation Util.* 12, S164–S167. doi: 10.1017/S1479262114000471.

Californian Almond Board (2022). Almond Almanac 2022.

Chin, S. W., Shaw, J., Haberle, R., Wen, J., and Potter, D. (2014). Diversification of almonds, peaches, plums and cherries - Molecular systematics and biogeographic history of Prunus (Rosaceae). *Mol. Phylogenet. Evol.* 76, 34–48. doi: 10.1016/j.ympev.2014.02.024.

D'amico-Willman, K. M., Ouma, W. Z., Meulia, T., Sideli, G. M., Gradziel, T. M., and Fresnedo-Ram Irez, J. (2022). Whole-genome sequence and methylome profiling of the almond [Prunus dulcis (Mill.) D.A. Webb] cultivar 'Nonpareil.' *G3 Genes|Genomes|Genetics*. doi: 10.1093/G3JOURNAL/JKAC065.

D'Amico-Willman, K. M., Sideli, G. M., Allen, B. J., Anderson, E. S., Gradziel, T. M., and Fresnedo-Ramírez, J. (2022). Identification of Putative Markers of Non-infectious Bud Failure in Almond [Prunus dulcis (Mill.) D.A. Webb] Through Genome Wide DNA Methylation Profiling and Gene Expression Analysis in an Almond × Peach Hybrid Population. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.804145.

Delplancke, M., Alvarez, N., Benoit, L., Espíndola, A., Joly, H. I., Neuenschwander, S., et al. (2013). Evolutionary history of almond tree domestication in the Mediterranean basin. *Mol. Ecol.* 22, 1092. doi: 10.1111/mec.12129.

Delplancke, M., Alvarez, N., Espíndola, A., Joly, H., Benoit, L., Brouck, E., et al. (2012). Gene flow among wild and domesticated almond species: Insights from chloroplast and nuclear markers. *Evol. Appl.* 5, 317–329. doi: 10.1111/j.1752-4571.2011.00223.x.

Delplancke, M., Yazbek, M., Arrigo, N., Espíndola, A., Joly, H., and Alvarez, N. (2016). Combining conservative and variable markers to infer the evolutionary history of Prunus subgen. Amygdalus s.l. under domestication. *Genet. Resour. Crop Evol.* 63, 221–234. doi: 10.1007/s10722-015-0242-6.

Denisov, V. P. (1988). Almond genetic resources in the USRR and their use in production and breeding. *Acta Hortic.* 224, 299–306.

Di Guardo, M., Farneti, B., Khomenko, I., Modica, G., Mosca, A., Distefano, G., et al. (2021). Genetic characterization of an almond germplasm collection and volatilome profiling of raw and roasted kernels. *Hortic. Res.* 8, 27. doi: 10.1038/s41438-021-00465-7.

Dirlewanger, E., Graziano, E., Joobeur, T., Garriga-Calderé, F., Cosson, P., Howad, W., et al. (2004). Comparative mapping and marker-assisted selection in Rosaceae fruit crops. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9891–9896. doi: 10.1073/pnas.0307937101.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.* 3, 43–52. doi: 10.1038/nrg703.

Donoso, J. M., Eduardo, I., Picañol, R., Batlle, I., Howad, W., Aranzana, M. J., et al. (2015). High-density mapping suggests cytoplasmic male sterility with two restorer genes in almond × peach progenies. *Hortic. Res.* 2. doi: 10.1038/hortres.2015.16.

Duval, H., Coindre, E., Ramos-Onsins, S. E., Alexiou, K. G., Rubio-Cabetas, M. J., Martínez-García, P. J., et al. (2023). Development and Evaluation of an AxiomTM 60K SNP Array for Almond (Prunus dulcis). *Plants* 12. doi: 10.3390/plants12020242.

Elhamzaoui, A., Oukabli, A., Charafi, J., and Moumni, M. (2012). Assessment of genetic diversity of Moroccan cultivated almond (*Prunus dulcis* Mill. DA Webb) in its area of extreme diffusion, using nuclear microsatellites. *Am. J. Plant Sci.* 3, 1294–1303.

Evreinoff, V. A. (1958). Contribution a letude de l'amandier. *Fruits et primeurs Afrique* 28, 99–104.

Fernández i Martí, A., Font i Forcada, C., Kamali, K., Rubio-Cabetas, M. J., Wirthensohn, M., and Socias i Company, R. (2015). Molecular analyses of evolution and population structure in a worldwide almond [*Prunus dulcis* (Mill.) D.A. Webb syn. *P. amygdalus* Batsch] pool assessed by microsatellite markers. *Genet. Resour. Crop Evol.* 62, 205–219. doi: 10.1007/s10722-014-0146-x.

Fernández i Martí, A., Font i Forcada, C., and Socias i Company, R. (2013). Genetic analysis for physical nut traits in almond. *Tree Genet. Genomes* 9, 455–465. doi: 10.1007/s11295-012-0566-8.

Fernández i Martí, À., Howad, W., Tao, R., Segura, J. M. A., Arús, P., and Socias i Company, R. (2011). Identification of quantitative trait loci associated with self-compatibility in a Prunus species. *Tree Genet. Genomes* 7, 629–639. doi: 10.1007/s11295-010-0362-2.

Font i Forcada, C., i Martí, À. F., and I Company, R. S. (2012). Mapping quantitative trait loci for kernel composition in almond. *BMC Genet.* 13, 1–9. doi: 10.1186/1471-2156-13-47/TABLES/3.

Font i Forcada, C., Oraguzie, N., Reyes-Chin-Wo, S., Espiau, M. T., Company, R. S. I., and Fernández I Martí, A. (2015a). Identification of genetic loci associated with quality traits in almond via association mapping. *PLoS One* 10. doi: 10.1371/journal.pone.0127656.

Font i Forcada, C., Velasco, L., Socias i Company, R., and Fernández i Martí, Á. (2015b). Association mapping for kernel phytosterol content in almond. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00530.

Fu, W., da Silva Linge, C., Lawton, J. M., and Gasic, K. (2022). Feasibility of genomic prediction for brown rot (*Monilinia* spp.) resistance in peach. *Fruit Res.* 2, 1–12. doi: 10.48130/frures-2022-0002.

García-Gómez, B. E., Ruiz, D., Salazar, J. A., Rubio, M., Martínez-García, P. J., and Martínez-Gómez, P. (2020). Analysis of Metabolites and Gene Expression Changes Relative to Apricot (Prunus

armeniaca L.) Fruit Quality During Development and Ripening. *Front. Plant Sci.* 0, 1269. doi: 10.3389/FPLS.2020.01269.

Godini, A. (2000). About the possible relationships between Amygdalus webbi Spach and Amygdalus communis L. *NUCIS Newsl.* 9, 17–19.

Gómez, E. M., Buti, M., Sargent, D. J., Dicenta, F., and Ortega, E. (2019a). Transcriptomic analysis of pollen-pistil interactions in almond (Prunus dulcis) identifies candidate genes for components of gametophytic self-incompatibility. *Tree Genet. Genomes* 15. doi: 10.1007/s11295-019-1360-7.

Gómez, E. M., Dicenta, F., Batlle, I., Romero, A., and Ortega, E. (2019b). Cross-incompatibility in the cultivated almond (Prunus dulcis): Updating, revision and correction. *Sci. Hortic. (Amsterdam).* 245, 218–223. doi: 10.1016/j.scienta.2018.09.054.

Goonetilleke, S. N., March, T. J., Wirthensohn, M. G., Arús, P., Walker, A. R., and Mather, D. E. (2018). Genotyping by sequencing in almond: SNP discovery, linkage mapping, and marker design. *G3 Genes, Genomes, Genet.* 8, 161–172. doi: 10.1534/g3.117.300376.

Gouta, H., Ksia, E., Buhner, T., Moreno, M. Á., Zarrouk, M., Mliki, A., et al. (2010). Assessment of genetic diversity and relatedness among Tunisian almond germplasm using SSR markers. *Hereditas* 147, 283–292. doi: 10.1111/j.1601-5223.2009.02147.x.

Gradziel, T. ., Martínez-Gómez, P., Dicenta, F., and Kester, D. . (2001). The utilization of related prunus species for almond variety improvement. *J. Am. Pomol. Soc.* 55. Available at: https://search.proquest.com/docview/209767315?pq-origsite=gscholar [Accessed May 17, 2020].

Gradziel, T. M. (2022). Transfer of Self-Fruitfulness to Cultivated Almond from Peach and Wild Almond. *Horticulturae* 8. doi: 10.3390/horticulturae8100965.

Grasselly, C. (1976). Les espèces sauvages d'amandier. *Options méditerranéennes* 32, 28–43.

Grasselly, C. (1978). Observations sur l'utilization d'un mutant d'amandier a floraison tardize dans un programme d'hybridizaiton. *Ann. Amelior. Plant.* 28, 685–695.

Grasselly, C., and Gall, H. (1967). Étude sur la possibilité de combinaison de quelques caractères agronomiques chez l'amandier Cristomorto hybridé par trois autres variétés. *Ann. Amélior. Plant.* 17, 83–91.

Guo, C., Wei, Y., Yang, B., Ayup, M., Li, N., Liu, J., et al. (2021). Developmental transcriptome profiling uncovered carbon signaling genes associated with almond fruit drop. *Sci. Rep.* 11, 1–12. doi: 10.1038/s41598-020-69395-z.

Halász, J., Kodad, O., Galiba, G. M., Skola, I., Ercisli, S., Ledbetter, C. A., et al. (2019). Genetic variability is preserved among strongly differentiated and geographically diverse almond germplasm: an assessment by simple sequence repeat markers. *Tree Genet. Genomes* 15, 1–13. doi: 10.1007/s11295-019-1319-8.

Hansen, J. M. (1991). The Palaeoethnobotany of Franchthi Cave. Excavations at Franchthi Cave, Greece. *Paléorient* 18–1, 135–137.

Hardner, C. M., Fikere, M., Gasic, K., da Silva Linge, C., Worthington, M., Byrne, D., et al. (2022). Multi-environment genomic prediction for soluble solids content in peach (Prunus persica). *Front. Plant Sci.* 13, 1–18. doi: 10.3389/fpls.2022.960449.

Jiang, X., Liu, K., Peng, H., Fang, J., Zhang, A., Han, Y., et al. (2023). Comparative network analysis reveals the dynamics of organic acid diversity during fruit ripening in peach (Prunus persica L. Batsch). *BMC Plant Biol.* 23, 1–14. doi: 10.1186/S12870-023-04037-W/TABLES/1.

Joobeur, T., Periam, N., De Vicente, M. C., King, G. J., and Arus, P. (2000). Development of a second generation linkage map for almond using RAPD and SSR markers. *Genome* 43, 649–655. doi: 10.1139/g00-040.

Joobeur, T., Viruel, M. A., De Vicente, M. C., Jáuregui, B., Ballester, J., Dettori, M. T., et al. (1998). Construction of a saturated linkage map for Prunus using an almond x peach F2 progeny. *Theor. Appl. Genet.* 97, 1034–1041. doi: 10.1007/s001220050988.

Kalluri, N., Serra, O., Donoso, J. M., Picañol, R., Howad, W., Eduardo, I., et al. (2022). Construction of a collection of introgression lines of "Texas" almond DNA fragments in the "Earlygold" peach genetic background. *Hortic. Res.* 9, 1–11. doi: 10.1093/hr/uhac070.

Kao, T., and Tsukamoto, T. (2004). The Molecular and Genetic Bases of S-RNase-Based Self-Incompatibility. *Plant Cell* 16, 72–83. doi: 10.1105/tpc.016154.S-RNase-Based.

Kester, D. E. (1965). Inheritance of time of bloom in certain progenies of almond. *Proc. Am. Soc. Hortic. Sci.* 87, 214–221.

Kislev, M. E., Nadel, D., and Carmi, I. (1992). Epipalaeolithic (19,000 BP) cereal and fruit diet at Ohalo II, Sea of Galilee, Israel. *Rev. Palaeobot. Palynol.* 73, 161–166. doi: 10.1016/0034-6667(92)90054-K.

Kislev, M., Melamed, Y., Simchoni, O., and Marmorstein, M. (1997). Computerized key of grass grains of the Mediterranean basin. *Lagascalia* 19 (1–2), 289–294.

Ladizinsky, G. (1999). On the origin of almond. *Genet. Resour. Crop Evol.* 46, 143–147. doi: 10.1023/A:1008690409554.

Li, X., Wang, J., Su, M., Zhang, M., Hu, Y., Du, J., et al. (2023). Multiple-statistical genome-wide association analysis and genomic prediction of fruit aroma and agronomic traits in peaches. *Hortic. Res.* doi: 10.1093/hr/uhad117.

Lotti, C., Minervini, A. P., Delvento, C., Losciale, P., Gaeta, L., Sánchez-Pérez, R., et al. (2023). Detection and distribution of two dominant alleles associated with the sweet kernel phenotype in almond cultivated germplasm. *Front. Plant Sci.* 14, 1–7. doi: 10.3389/fpls.2023.1171195.

McCreery, D. W. (1979). Flotation of the Bab edh-Dhra and Numeira plant remains. *Annu. Am. Sch. Orient. Res.* 46, 165.

Mendel, G. (1865). Experiments in plant hybridization. Available at: http://www.netspace.org./MendelWeb/ [Accessed October 11, 2019].

Moll, L., Baró, A., Montesinos, L., Badosa, E., Bonaterra, A., and Montesinos, E. (2022). Induction of Defense Responses and Protection of Almond Plants Against Xylella fastidiosa by Endotherapy with a Bifunctional Peptide. *Phytopathology* 112, 1907–1916. doi: 10.1094/PHYTO-12-21-0525-R.

Mousavi, S., Alisoltani, A., Shiran, B., Fallahi, H., Ebrahimie, E., Imani, A., et al. (2014). De novo transcriptome assembly and comparative analysis of differentially expressed genes in Prunus dulcis Mill. in response to freezing stress. *PLoS One* 9, 1–13. doi: 10.1371/journal.pone.0104541.

Nuts production in Spain (2021). Madrid Available at: https://www.mapa.gob.es/es/agricultura/temas/producciones-agricolas/analisisdelarealidadproductivafrutossecos2020_tcm30-584009.pdf.

Ortega, E., Sutherland, B. G., Dicenta, F., Boskovic, R., and Tobutt, K. R. (2005). Determination of incompatibility genotypes in almond using first and second intron consensus primers: detection of new S alleles and correction of reported S genotypes. *Plant Breed.* 124, 188–196. doi: 10.1111/j.1439-0523.2004.01058.x.

Pavan, S., Delvento, C., Mazzeo, R., Ricciardi, F., Losciale, P., Gaeta, L., et al. (2021). Almond diversity and homozygosity define structure, kinship, inbreeding, and linkage disequilibrium in cultivated germplasm, and reveal genomic associations with nut and seed weight. *Hortic. Res.* 8, 1–12. doi: 10.1038/s41438-020-00447-1.

Pérez-Jordà, G., Alonso, N., Rovira, N., Figueiral, I., López-Reyes, D., Marinval, P., et al. (2021). The Emergence of Arboriculture in the 1st Millennium BC along the Mediterranean's "Far West." *Agronomy.* 11, 902. doi: 10.3390/agronomy11050902.

Prudencio, Á. S., Hoeberichts, F. A., Dicenta, F., Martínez-Gómez, P., and Sánchez-Pérez, R. (2021). Identification of early and late flowering time candidate genes in endodormant and ecodormant almond flower buds. *Tree Physiol.* 41, 589–605. doi: 10.1093/treephys/tpaa151.

Rosenberg, M., Nesbitt, M., Redding, R. W., and Strasser, T. F. (1995). Some preliminary observations concerning early Neolithic subsistence behaviors in eastern Anatolia. *Anatolica* 21, 1–12.

Sánchez-Pérez, R., Howad, W., Dicenta, F., Arús, P., and Martínez-Gómez, P. (2007). Mapping major genes and quantitative trait loci controlling agronomic traits in almond. *Plant Breed.* 126, 310–318. doi: 10.1111/j.1439-0523.2007.01329.x.

Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Aiese Cigliano, R., Del Cueto, J., et al. (2019). Mutation of a bHLH transcription factor allowed almond domestication. *Science (80-.).* 364, 1095–1098. doi: 10.1126/science.aav8197.

Shiran, B., Amirbakhtiar, N., Kiani, S., Mohammadi, S., Sayed-Tabatabaei, B. E., and Moradi, H. (2007). Molecular characterization and genetic relationship among almond cultivars assessed by RAPD and SSR markers. *Sci. Hortic. (Amsterdam).* 111, 280–292. doi: 10.1016/j.scienta.2006.10.024.

Sideli, G. M., Mather, D., Wirthensohn, M., Dicenta, F., Goonetilleke, S. N., Martínez-García, P. J., et al. (2023). Genome-wide association analysis and validation with KASP markers for nut and shell traits in almond (Prunus dulcis [Mill.] D.A.Webb). *Tree Genet. genomes.* 19, 13. doi: 10.1007/s11295-023-01588-9.

Silva, C., Garcia-Mas, J., Sánchez, A. M., Arús, P., and Oliveira, M. M. (2005). Looking into flowering time in almond (Prunus dulcis (Mill) D. A. Webb): The candidate gene approach. *Theor. Appl. Genet.* 110, 959–968. doi: 10.1007/s00122-004-1918-z.

Sutherland, B. G., Robbins, T. P., and Tobutt, K. R. (2004). Primers amplifying a range of Prunus S-alleles. *Plant Breed.* 123, 582–584. doi: 10.1111/j.1439-0523.2004.01016.x.

Tamura, M., Ushijima, K., Sassa, H., Hirano, H., Tao, R., Gradziel, T. M., et al. (2000). Identification of self-incompatibility genotypes of almond by allele-specific PCR analysis. *Theor. Appl. Genet.* 101, 344–349.

Tavassolian, I., Rabiei, G., Gregory, D., Mnejja, M., Wirthensohn, M. G., Hunt, P. W., et al. (2010). Construction of an almond linkage map in an Australian population Nonpareil × Lauranne. *BMC Genomics* 11. doi: 10.1186/1471-2164-11-551.

Trainin, T., Brukental, H., Shapira, O., Attia, Z., Tiwari, V., Hatib, K., et al. (2022). Physiological characterization of the wild almond Prunus arabica stem photosynthetic capability. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.941504.

Van Ghelder, C., Lafargue, B., Dirlewanger, E., Ouassa, A., Voisin, R., Polidori, J., et al. (2010). Characterization of the RMja gene for resistance to root-knot nematodes in almond: Spectrum, location, and interest for Prunus breeding. *Tree Genet. Genomes* 6, 503–511. doi: 10.1007/s11295-010-0268-z.

Van Zeist, W., and de Roller, G. J. (1992). The plant husbandry of aceramic Çayönü, SE Turkey. *Palaeohistoria* 33–34, 65–96.

Vargas, F. J., and Romero, M. A. (2001). Blooming time in almond progenies. *Options Méditerranéennes* 56, 29–34.

Viruel, M. A., Messeguer, R., de Vicente J Garcia-Mas, M. C., Puigdom, P., Vargas P Arfis, nech F., Garcia-Mas, J., et al. (1995). A linkage map with RFLP and isozyme markers for almond. *Theor. Appl. Genet.* 91, 964–971. doi: https://doi.org/10.1007/BF00223907.

Wang, Q., Cao, K., Li, Y., Wu, J., Fan, J., Ding, T., et al. (2023). Identification of co-expressed networks and key genes associated with organic acid in peach fruit. *Sci. Hortic. (Amsterdam).* 307, 111496. doi: 10.1016/j.scienta.2022.111496.

Wen, J., Berggren, S. T., Lee, C.-H., Ickert-Bond, S., and Yi, T.-S. (2008). Phylogenetic inferences in Prunus (Rosaceae) using chloroplast ndhF and nuclear ribosomal ITS sequences. *J. Syst. Evol.* 46, 322. doi: 10.3724/SP.J.1002.2008.08065.

Willcox, G., Fornite, S., and Herveux, L. (2008). Early Holocene cultivation before domestication in northern Syria. *Veg. Hist. Archaeobot.* 17, 313–325. doi: 10.1007/s00334-007-0121-y.

Zeinalabedini, M., Khayam-Nekoui, M., Grigorian, V., Gradziel, T. M., and Martínez-Gómez, P. (2010). The origin and dissemination of the cultivated almond as determined by nuclear and chloroplast SSR marker analysis. *Sci. Hortic. (Amsterdam).* 125, 593–601. doi: 10.1016/j.scienta.2010.05.007.

Objectives

Objectives

The main objective of this PhD thesis was the designing and application of genomic and bioinformatic tools in almond breeding. To achieve this goal, four specific objectives were proposed:

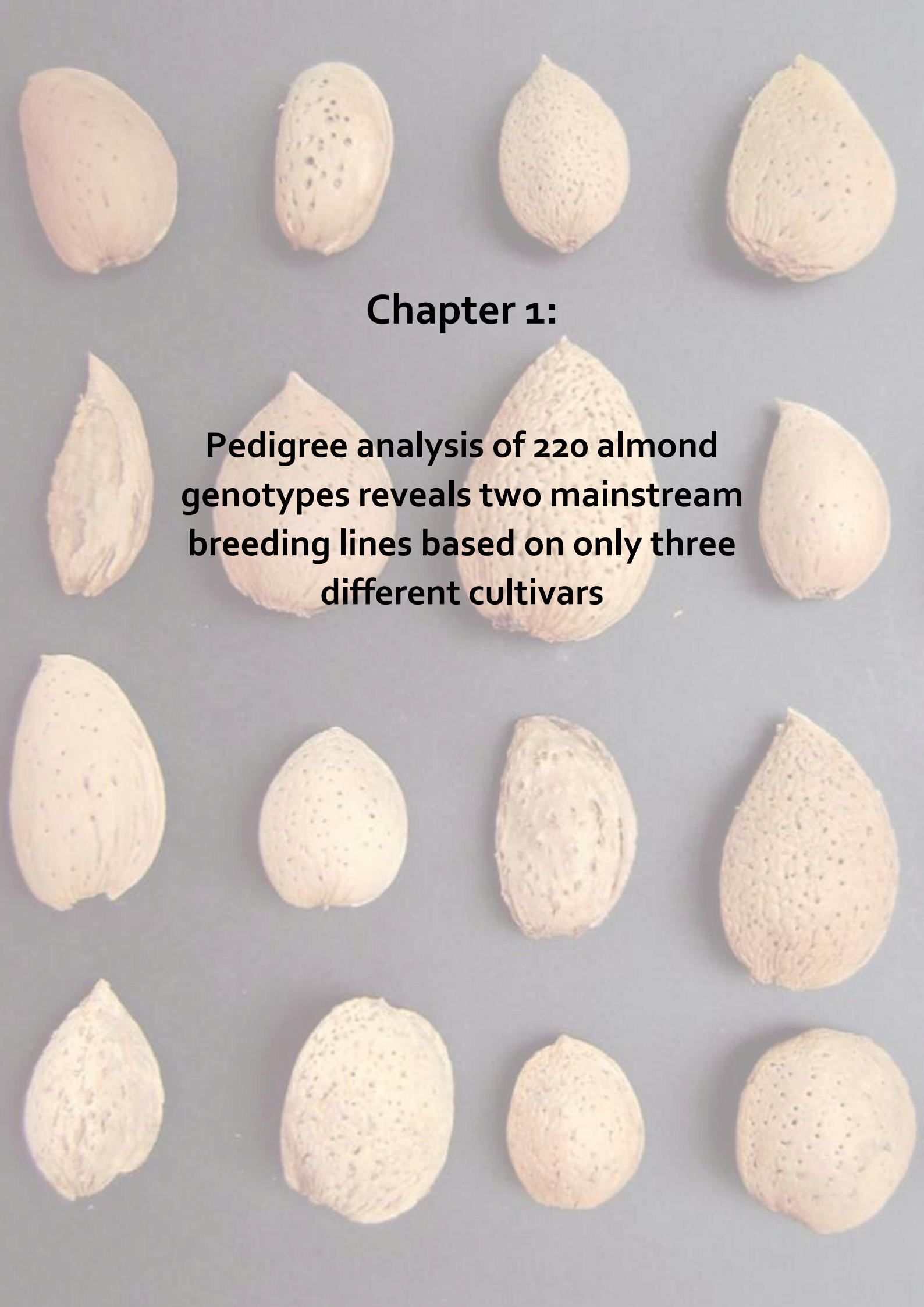
1.- Study the breeding tendencies followed by almond breeders worldwide through a pedigree analysis.

2.- Mapping QTLs related to kernel quality traits in a F1 population coming from the cross 'Marcona' x 'Marinada'.

3.- Study the genetic structure and look for additive and non-additive genotype-phenotype associations in a population formed by 211 almond accessions from 20 different countries.

4.- Develop a new tool for predicting gene function in almond and other *Prunus* species.

Objectives



Chapter 1:

Pedigree analysis of 220 almond genotypes reveals two mainstream breeding lines based on only three different cultivars

Pedigree analysis of 220 almond genotypes reveals two world mainstream breeding lines based on only three different cultivars

Felipe Pérez de los Cobos[1,4], Pedro J. Martínez-García[2], Agustí Romero[1], Xavier Miarnau[3], Iban Eduardo[4], Werner Howad[4], Mourad Mnejja[4], Federico Dicenta[2], Rafel Socias i Company[5], Maria J. Rubio[5], Thomas M. Gradziel[6], Michelle Wirthensohn[7], Henri Duval[8], Doron Holland[9], Pere Arús[4], Francisco J. Vargas[1], Ignasi Batlle[1]

[1]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Mas Bové, Ctra. Reus-El Morell Km 3,8, 43120 Constantí, Tarragona, Spain

[2]Centro de Edafología y Biología Aplicada del Segura, Consejo Superior de Investigaciones Científicas (CEBAS-CSIC), P.O. Box 164, 30100 Espinardo, Murcia, Spain

[3]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Fruitcentre, PCiTAL, Gardeny Park, Fruitcentre Building, 25003 Lleida, Spain

[4]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Centre de Recerca en Agrigenòmica (CRAG), CSIC-IRTA-UAB-UB. Cerdanyola del Vallès (Bellaterra), 08193 Barcelona, Spain

[5]Centro de Investigación y Tecnología Agroalimentaria de Aragón (CITA), Avda. Montañana 930, 50059, Zaragoza, Instituto Agroalimentario de Aragón IA2 (CITA-Universidad de Zaragoza), Zaragoza, Spain

[6]University of California, 1 Shields Avenue, Davis, California 95616, USA

[7]University of Adelaide, Waite Research, School of Agriculture, Food and Wine, PMB 1, Glen Osmond, SA, 5064, Adelaide, Australia

[8]Institut National de la Recherche Agronomique (INRA), Domain St. Maurice CS 60094, 84143, Montfavet Cedex, France

[9]Agricultural Research Organization, Newe-Ya'ar Research Center, P.O. Box 1021, Ramat Yishad 30095, Israel

Horticulture Research, Volume 8, 2021, 11, https://doi.org/10.1038/s41438-020-00444-4

Abstract

Loss of genetic variability is an increasing challenge in tree breeding programs due to the repeated use of a reduced number of founder genotypes. However, in almond, little is known about the genetic variability in current breeding stocks, although several cases of inbreeding depression have been reported. To gain insights into the genetic structure in modern breeding programs worldwide, marker-verified pedigree data of 220 almond cultivars and breeding selections were analyzed. Inbreeding coefficients, pairwise relatedness and genetic contribution were calculated for these genotypes. The results reveal two mainstream breeding lines based on three cultivars: 'Tuono'-'Cristomorto' and 'Nonpareil'. Descendants from 'Tuono' or 'Cristomorto' number 76 (sharing 34 descendants), while 'Nonpareil' has 71 descendants. The mean inbreeding coefficient of the analyzed genotypes was 0.041, with 14 genotypes presenting a high inbreeding coefficient, over 0.250. Breeding programs from France, the USA and Spain showed inbreeding coefficients of 0.075, 0.070 and 0.037, respectively. According to their genetic contribution, modern cultivars from Israel, France, the USA, Spain and Australia trace back to a maximum of six main founding genotypes. Among the group of 65 genotypes carrying the S_f allele for self-compatibility, the mean relatedness coefficient was 0.125, with 'Tuono' as the main founding genotype (24.7% of total genetic contribution). The results broaden our understanding about the tendencies followed in almond breeding over the last 50 years and will have a large impact into breeding decision-making process worldwide. Increasing current genetic variability is required in almond breeding programs to assure genetic gain and continuing breeding progress.

1. Chapter 1

1.1. Introduction

Almond [*Prunus dulcis* (Miller) D.A. Webb, syn. *P. amygdalus* (L) Batsch] is the most economically important temperate tree nut crop worldwide. Due to increasing demand, production areas are expanding into warm and cold climatic regions of both hemispheres. Almond world production (1,258,324 kernel tonnes) is led by the USA (80%), Australia (6%) and Spain (5%) (International Nut & Dried Fruits Council, 2019).

The origin of almond within the Amygdalus subgenus, including cultivated almond and its wild relatives such as *P. fenzliana* Fritsh, *P. bucharica* (Korsh.) Fedtsch, *P. kuramica* (Korsh.) Kitam. and *P. triloba* Lindl. (Grasselly and Crossa-Raynaud, 1980; Zeinalabedini et al., 2010a) took place approximately 5.88 million years ago(Alioto et al., 2020a). Almond originated in the arid mountainous regions of Central Asia, where it was first cultivated around 5000 years ago (Velasco et al., 2016) and then moved to the Mediterranean region and later to California and the southern hemisphere (South America, Australia and South Africa) (Kester et al., 1991). Wide cultivation of almond, often under the more severe environments of Central Asia and the Mediterranean region, was possible because of the availability of a highly diverse gene pool, genetic recombination promoted by its self-incompatibility and possibly, by interspecific hybridization and gene introgression involving other members of the Amygdalus subgenus. As a result, almond is an extremely variable species, with a high morphological and physiological diversity. This variability, measured with biochemical and molecular markers (Arulsekar et al., 1986; Arús et al., 2009; Fernández i Martí et al., 2015), has revealed that almond is the most genetically variable of the diploid *Prunus* cultivated species (Byrne, 1990; Mnejja et al., 2010) .

In the Mediterranean Region, two thousand years of almond culture concentrated production to specific areas, where well-defined seedling ecotypes and local cultivars evolved (Grasselly and Crossa-Raynaud, 1980). By the turn of the 20 th century, most of these almond producing countries had identified locally desirable cultivars that were often seedling selections of unknown origin (Gradziel et al., 2017). Thus, growers selected cultivars and landraces, which represented a rich genetic diversity. Most of these Mediterranean local cultivars have largely disappeared from cultivation in the last 50 years (Batlle et al., 2017). Modern almond cultivation is based on a reduced number of cultivars (preferably self-compatible) grafted onto soil adapted clonal rootstocks and cultivated under irrigated conditions when possible.

Modern almond breeding started in the 1920's with the making of controlled crosses and seedling selections to meet changing agronomic and market demands. Currently, there are six active public breeding programs worldwide: the USA (UCD-USDA), Spain (CITA, IRTA and CEBAS-CSIC), Australia (University of Adelaide) and Israel (ARO). Some private breeding programs exist also in the USA. In addition, there were various breeding initiatives in Russia, France, Greece, Italy and Argentina (Batlle et al., 2017). Different breeding objectives were developed according to regional agronomic, commercial and market requirements. One of the main differences in the objectives is nut shell hardness. Two types of almonds are bred: soft-shelled (in the USA and Australia mainly) and hard-shelled (in most Mediterranean countries). Common aims of Mediterranean breeding programs are self-compatibility and late-blooming, as most traditional almond cultivars are self-incompatible and early-blooming. Self-compatibility is controlled by a single self-compatibility S_f dominant allele (López et al., 2006). During the last 50 years, almond breeding for self-compatibility has mainly used two sources of S_f, local landraces originated in Italy ('Tuono' and 'Genco') and related species as *P. persica* and *P. webbii* (Socias i Company, 2017).

Almond breeders have relied mainly on outcrossing and, occasionally, on introgression from other *Prunus* species, for the development of new cultivars. Initially, in the USA (with limited accessible genetic resources) and later in Russia and Mediterranean region (with more diverse germplasm available) rapid genetic advances were achieved. In California, 'Carmel' (introduced in 1966), as 'Nonpareil' pollinizer, was the first cultivar release with extensive commercial impact. In Russia and the former Soviet Union, several late flowering and frost hardy cultivars were obtained in the 1950's with Primorskyi (date unknown) later used extensively for breeding in Europe. In the Mediterranean region, late flowering, productive, well-adapted and resilient cultivars like Ferragnès (1973) or Masbovera (1992) were released with great success. The French self-compatible cultivar Lauranne (1991) showed a broad environmental adaptation, high production and regular cropping.

Although improved cultivars continued to be released, the amount of progress per generation diminishes since parents were continually drawn from the same genepool (Batlle et al., 2017). This situation has resulted in a potential loss of genetic variability in new breeding stocks and cultivars. Inbreeding depression in almond, expressed as low vigor, reduced flower number and fruit set, increased fruit abortion, lower seed germination and seedling survival, increased leaf and wood abnormalities and loss of disease resistance have been reported (Grasselly, 1976b; Grasselly and Olivier, 1981; Socias i Company, 2011; Martínez-García et al., 2012). In addition, low self-fruitfulness in self-compatible almond genotypes was suspected to be due to inbreeding (Alonso and Socias i Company, 2005).

Regarding breeding for self-compatibility, male parents carrying the S_f allele and sharing the other *S*-allele with the female parent are commonly used. In addition, crossing heterozygous self-compatible parents in breeding programs has been suggested to obtain homozygous self-compatible genotypes to be used in further breeding (Ortega and Dicenta, 2003). Such breeding strategies can narrow the genetic variability of crops when they lead to a reduced number of genotypes utilized as parents.

Summarizing, modern almond breeding and production are dominated by a small number of widely distributed and related cultivars. This situation can lead to a potential increase of inbreeding depression and genetic vulnerability, i.e. susceptibility of most of the grown cultivars to biotic and abiotic stresses due to similarities in their genotypes (Van De Wouw et al., 2010; Keneni et al., 2012). Therefore, it is needed to have up-to-date information of the relationships among genotypes used at breeding and production levels.

Several almond populations have been analyzed with molecular markers in order to determine genetic variability and relatedness (Gouta et al., 2010; Cabrita et al., 2014; Fernández i Martí et al., 2015; Halász et al., 2019). However, these studies were performed with material from limited geographic areas and do not represent the current worldwide status of almond breeding stocks. Although genomic measures of inbreeding are more accurate than those obtained from pedigree data (Kardos et al., 2015; Wang, 2016), pedigree-based analysis is a cost-effective technique to estimate these parameters in breeding populations and an alternative when genomic-measures are unviable. Several reports have evaluated inbreeding based on pedigree data in breeding populations of fruit and nut tree crops (Choi and Kappel, 2004; Debuse et al., 2005; Son et al., 2012; Marrano et al., 2019). In almond, a pedigree analysis of 123 different genotypes from the USA, France, Spain, Israel and Russia was reported (Lansari et al., 1994). However, their work was mainly focused on North American genotypes and did not include many cultivars that have subsequently been released worldwide. This study aimed to determine the genetic structure of current breeding stocks and breeding tendencies over the last 50 years using marker-verified pedigree data.

1.2.Materials and methods

1.2.1Marker-verified pedigree data

Pedigree data of 220 almond genotypes (169 of known origin and 51 of unknown origin) were compiled from available bibliography and breeding records. From the 220 almond genotypes, 37 genotypes were no longer available (17% of the studied genotypes) as they were eliminated some time ago or were from discontinued breeding programs. To verify parental relationships of the rest of genotypes (183), we used SSRs, SNPs and self-incompatibility S alleles data from previous studies performed by the breeding programs taking part in this study (Supplementary Material 1.1). Marker data confirmed both parents of 71 genotypes and one parent of four genotypes (146 confirmed relationships) and found three erroneous parentages. Two wrong parentages were found on the male parent of 'Capella' and 'Davey', changing their pedigree to open-pollinated and a third incorrect parentage on 'Yosemite' female parent, eliminating this genotype from the analysis.

After the corrections made, pedigrees of 169 genotypes of known origin (77 of them marker-verified, approximately 54% of the available genotypes) were analyzed (Supplementary Material 1.1). The origin of the genotypes were: 59 from Spain, 56 from the USA, 16 from Russia, 11 from Israel, 10 from France, 7 from Australia, 7 from Greece, 2 from Argentina and 2 from Italy.

A pedigree data file was created. Each record in the file contained one cultivar or selection name, the female parent and the male parent, in that order. Once entered, these data were available for inbreeding analyses such as determining the number of times a cultivar appeared in a pedigree as a male or female genitor. Genotypes of known origin were classified into two groups according to self-compatibility: 104 self-incompatible and 65 self-compatible.

1.2.2. Inbreeding coefficient, pairwise relatedness and genetic contribution

The inbreeding coefficient (*F*) is defined as the probability that a pair of alleles at any locus in an individual are identical by descent and it is calculated by the following formula (Wright, 1922):

$$F_x = \sum \left[\left(\frac{1}{2}\right)^{n_1+n_2+1} (1 + F_A) \right]$$

Where n_1 = number of generations from one parent back to the common ancestor, n_2 = number of generations from the other parent back to the common ancestor and F_A = inbreeding coefficient of the common ancestor.

Pairwise relatedness (*r*) or coancestry coefficient, the degree of relationship by descent of two parents, equals the inbreeding coefficient of their prospective progeny.

The genetic contribution (*GC*) of a founder to a cultivar is calculated by the following formula (Sjulin and Dale, 1987):

$$GC = \sum_{1}^{x} \left(\frac{1}{2}\right)^{n}$$

Where *n* = number of generations in a pedigree pathway between the founding clone and the cultivar and *x* = number of pathways between the founding clone and the cultivar. The three

parameters were calculated using the SAS INBRED procedure (SAS 9.4 SAS Institute, Cary NC USA).

In summary, the inbreeding coefficient measures the probability that two alleles in a locus are identical by descent and so copies of the same allele from a previous generation. The pairwise relatedness measures the probability that two alleles at any locus are identical by descent (copies of the same allele in a previous generation) between two different individuals. *F* and *r* range from 0 to 1, with values close to 0 indicating a low degree of inbreeding or relatedness and values close to 1 indicating a high degree of inbreeding or relatedness. The genetic contribution estimates the proportion of genome that comes from the same individual. Thus, a child will have 0.5 genome of either parent and a grandchild will have 0.25 genomes of his grandparents.

1.2.3. Analysis description

To calculate *F*, *r*, and *GC*, parents of unknown origin were assumed to be unrelated and noninbred. The seed parent involved in all open-pollinations was also assumed to be unrelated to the pollen parent. These assumptions, based on the fact that most almond cultivars are obligate outcrossers because of their self-incompatibility, may lead to an underestimation of inbreeding. In the cases of genotypes of open-pollinated origin (OP), numbers OP1, OP2, OP3, etc. were given to the pollen parent in order to be distinguishable for genetic studies. Also, all mutants were considered to have no genetic differences from the original cultivar, thus *GC* = 1. Since the differences between such mutants and the original cultivar are expected to be caused by a few mutations in the DNA, this simplification avoids the overestimation of inbreeding coefficients. Cultivars like Supernova and Guara were considered as 'Tuono' clones (Marchese et al., 2008; Dicenta et al., 2015). Regarding the different clones of the French paper-shell cultivar Princesse, used in both the USA and Russian breeding programs, we adopted the approach of Lansari et al. (1994) by analyzing both clones as the same cultivar. Historical reports suggest that the Hatch series 'Nonpareil', 'I.X.L.' and 'Ne Plus Ultra' were seedling selections from an open pollination progeny of the early-introduced cultivar Princesse. This cultivar probably originated from the Languedoc region in France (Wood, 1925a; Kester et al., 1991; Kester and Gradziel, 1996; Bartolozzi et al., 1998). Also, 'Nikitskij' was selected in France in 1902 (Rikhter, 1972). Because their specific origins remain uncertain, we analyzed these genotypes as non-related, which, however, could lead to an underestimation of inbreeding.

Pedigree data were analyzed at four levels: worldwide, by country (Australia, France, Israel, Spain and the USA), by breeding program (when different programs exist within a country: CITA, IRTA, CEBAS-CSIC and, UCD-USDA) and by genotypes carrying the S_f allele for self-compatibility.

1.3. Results

1.3.1. Founding clones

The entire almond pedigree traced back to 51 founding clones (Supplementary Figure 1.1). 'Nonpareil', 'Cristomorto', 'Mission' and 'Tuono' were the founders with the largest number of descendants in the pedigree: 140 of the 169 genotypes of known parentage traced back to one or more of these founding clones (Figure 1.1). No genotype was derived from all four cultivars, i.e. did not trace back to the four founding clones. There were only five genotypes that came from a 3-way shared progeny, all of them tracing back to 'Tuono'-'Cristomorto'-'Nonpareil'. The largest 2-way shared genotype sub in set were 'Tuono'-'Cristomorto' and 'Nonpareil'-'Mission'

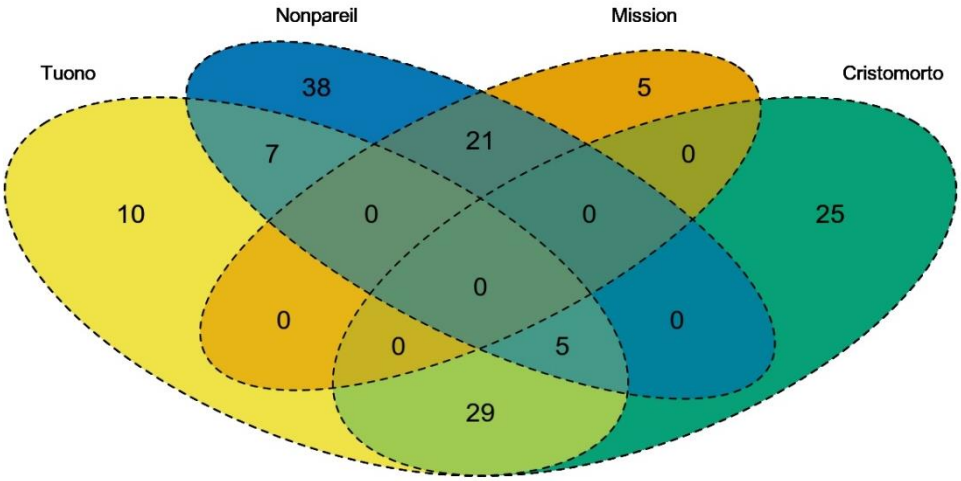


Figure 1.1. Venn diagram showing the number of descendants shared by 'Tuono', 'Nonpareil', 'Mission' and 'Cristomorto'.

with 29 and 21 descendants respectively. 'Mission' only shared progeny with 'Nonpareil' (Figure 1.1).

Analyzing the results by country, breeding programs from the USA had two main founding clones, 'Nonpareil' and 'Mission', with 46 and 24 descendants respectively out of 56. These two founders were followed by 'Eureka' and 'Harriott', with 14 and 11 descendants each. Breeding programs from Spain had three main founding clones, 'Tuono', 'Cristomorto' and 'Primorskyi', with 32, 31 and 24 descendants respectively. Cultivars from the discontinued French program had three main founding clones from two geographical origins, 'Cristomorto' and 'Tuono' (from Italy) with nine and five descendants, respectively and 'Aï' (from France), with eight descendants. The Australian program had only two main founding clones, 'Nonpareil' and 'Lauranne', with six and five derived genotypes, respectively. The Israeli breeding program showed the most balanced pedigree with six main founding clones, 'Marcona', 'Greek', 'Um ElFahem', 'Tuono', 'Nonpareil' and 'Ferragnès'.

Table 4.1. Genotypes with the highest inbreeding coefficient.

Line Name	Female Parent	Male Parent	Origin	Country	Inbreeding
A2-198	C1328	C1328	CEBAS-CSIC	SPAIN	0.5
Solano	21-19W	22-20	UCD	USA	0.375
Sonora	21-19W	22-20	UCD	USA	0.375
Vesta	Nonpareil	Solano	UCD	USA	0.375
Ferralise	Ferraduel	Ferragnès	INRA	FRANCE	0.25
FGFD2	Ferragnès	Ferraduel	INRA	FRANCE	0.25
21-19W	Nonpareil	A1-30	UCD	USA	0.25
22-20	Nonpareil	A1-30	UCD	USA	0.25
6-27	Nonpareil	Jordanolo	UCD	USA	0.25
Calif. 24-6	Eureka	A5-25	UCD	USA	0.25
Emerald	Mission	S2	PRIVATE	USA	0.25
Profuse	Nonpareil	Jordanolo	PRIVATE	USA	0.25
Supareil	Nonpareil	Carmel	PRIVATE	USA	0.25
D01-462	A2-198	S5133	CEBAS-CSIC	SPAIN	0.25

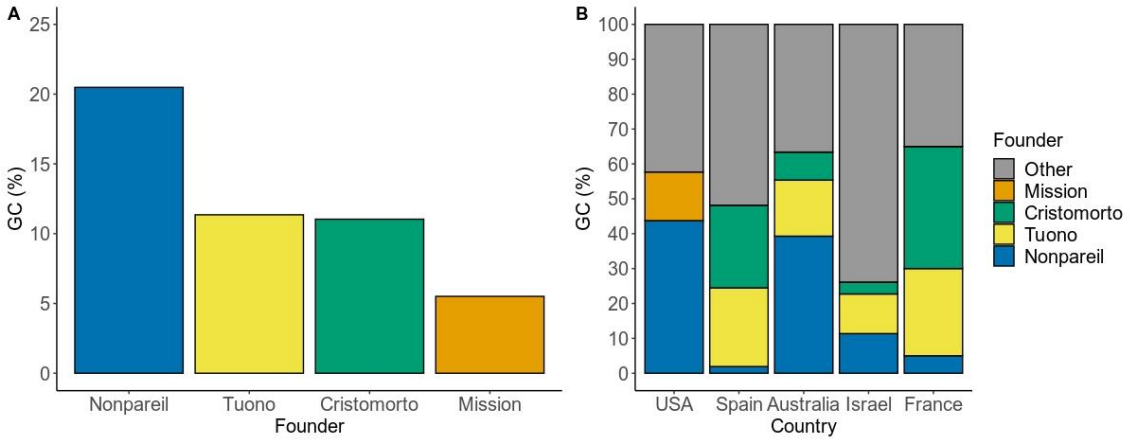


Figure 1.2. Mean genetic contribution (GC) of founders 'Nonpareil', 'Tuono', 'Cristomorto' and 'Mission' worldwide (A) and by country (B).

The UCD breeding program had 'Nonpareil' as main founding clone with 29 descendants. Cultivars Eureka, Mission and Harriott had a slight influence in the pedigree with 14, 12 and 10 descendants respectively. Within Spain, CITA breeding program had Italian 'Tuono' as the main founding clone with seven descendants. The IRTA breeding program showed three main founding clones, 'Cristomorto', 'Primorskyi' and 'Tuono' with 30, 19 and 16 descendants respectively. The CEBAS-CSIC breeding program had three main founding clones, 'Tuono', 'Ferragnès' and 'Primorskyi' with 15, nine and eight descendants respectively. The French local cultivar Aï was also present in the three Spanish programs through the largely used French 'Ferraduel' and 'Ferragnès'. These two cultivars were the ancestors of 25 genotypes.

Analyzing the 65 genotypes carrying the S_f allele for self-compatibility, the founding clones that traced back to the origin of this allele are 'Tuono', 'Genco' and genotypes originated from introgression crosses with *P. persica* and *P. webbii*.

1.3.2. Inbreeding coefficients

The mean inbreeding coefficient (*F*) of the 169 genotypes of known parentage analyzed was 0.041 (Supplementary Material 1.2). Some 43 genotypes presented an $F > 0$, with 14 over 0.250 (Table 1.1).

Considering results within each country, programs showing more inbreeding were France, the USA and Spain with 0.075, 0.070 and 0.037 mean *F*, respectively (Supplementary Material 1.2). The programs from Australia and Israel had $F = 0$. The USA accessions ranged from $F = 0$ to 0.375 with 21 of the 56 genotypes having $F > 0$. The French cultivar Ferralise and selection FGFD2, derived from the same reciprocal cross, had $F = 0.250$. The Spanish selection A2-198 from CEBAS-CSIC, showed the highest inbreeding coefficient ($F = 0.500$) as it is a selfing from selection C1328 and was raised to obtain homozygous S_fS_f individuals.

The UCD-USDA breeding program had a mean *F* of 0.096. Within Spain, the CITA program had $F = 0$. The CEBAS-CSIC program had only three genotypes with $F > 0$, but presented an average *F* of 0.048. The IRTA program holds 15 genotypes with $F > 0$ and a mean *F* of 0.043 (Supplementary Material 1.2). Considering only the 65 self-compatible genotypes, they had a mean *F* of 0.042, ranging from 0 to 0.500 (Supplementary Material 1.2).

1.3.3. Genetic contribution

'Nonpareil', 'Tuono', 'Cristomorto' and 'Mission' were the founding clones with the highest mean genetic contribution (*GC*; Figure 1.2). These four cultivars accounted for 48.4% of the total *GC* worldwide. 'Nonpareil' represented 20.5% of *GC* worldwide, 'Tuono' and 'Cristomorto' were around 11% and 'Mission' slightly exceeded 5%. Nevertheless, the mean *GC* of these founding clones within each country was variable. The breeding programs most dependent on these founders were Australia and France, where 'Nonpareil', 'Tuono' and 'Cristomorto' represented more than 60% of the total *GC*. Israel was the least dependent country as these founders represented approximately 25% of the total *GC*. Cultivar Nonpareil was the founder with the highest mean *GC* in the USA and Australia, while in Spain and France were 'Tuono' and 'Cristomorto'. The cultivar Mission was used only in the American programs.

Table 1.5. Genetic contribution (*GC*) of mean founding clones by country.

Founding clone	Country of origin	GC (%)	GC Total (%)
Australia			
Nonpareil	USA	39.3	**71.4**
Lauranne	France	32.1	
France			
Cristomorto	Italy	35.0	**100.0**
Aï	France	30.0	
Tuono	Italy	25.0	
Ardechoise	France	5.0	
Tardy Nonpareil	USA	5.0	
Israel			
Greek	Israel	20.5	**81.9**
Marcona	Spain	18.2	
Um ElFahem	Israel	13.6	
Tuono	Italy	11.4	
Nonpareil	USA	11.4	
Ferragnès	France	6.8	
Spain			
Cristomorto	Italy	23.7	**69.4**
Tuono	Italy	22.6	
Primorksyi	Russia	15.6	
Aï	France	7.5	
USA			
Nonpareil	USA	43.7	**71.8**
Mission	USA	13.9	
Eureka	USA	8.7	
Harriott	USA	5.5	

Table 1.2 shows the *GC* of the mean founders by country. In the Australian breeding program, only two founders, 'Nonpareil' and 'Lauranne', represented the 71.4% of the total *GC*. The French breeding program was characterized by the extensive use of three founders 'Cristomorto' (*GC* = 35.0%), 'Aï' (*GC* = 30.0%) and 'Tuono' (*GC* = 25.0%). These cultivars together with 'Ardèchoise' and 'Tardy Nonpareil' (both *GC* = 5.0%) accounted for 100% of the total *GC*. The Israeli breeding program presented six main founders, 'Greek' (*GC* = 20.5%), 'Marcona' (*GC* = 18.2%), 'Um ElFahem' (*GC* = 13.6%), 'Tuono' (*GC* = 11.4%), 'Nonpareil' (*GC* = 11.4%) and 'Ferragnès (*GC* = 6.8%) which together accounted for 81.9% of the total *GC*. The USA breeding

programs was largely dependent on 'Nonpareil' (*GC* = 43.7%) followed by 'Mission' (*GC* = 13.9%), 'Eureka' (*GC* = 8.7%), and 'Harriott' (*GC* = 5.5%) which all accounted for 71.8% of the total *GC*. The cultivars released by the three Spanish breeding programs were based mainly on four founders: 'Cristomorto' (*GC* = 23.7%), 'Tuono' (*GC* = 22.6%), 'Primorskyi' (*GC* = 15.6%) and 'Aï' (*GC* = 7.5%), accounting for 69.4% of the total *GC*.

The UCD-USDA breeding program had the same founders as the overall American programs, 'Nonpareil' (*GC* = 43.2%), 'Eureka' (*GC* = 14.8%), 'Harriott' (*GC* = 8.5%) and 'Mission' (*GC* = 5.5%). Differences were observed in the use of founding cultivars between Spanish breeding programs. The CITA program was mainly based on four cultivars 'Tuono' (*GC* = 35.0%), and 'Belle d'Aurons', 'Bertina' and 'Genco' (*GC* = 10.0% each). These cultivars were accounting for 65.0% of the total *GC*. The CEBAS-CSIC program was based also on four founders, 'Tuono' (*GC* = 28.9%), 'Ferragnès' (*GC* = 18.4%), 'Genco' (*GC* = 12.5%) and 'Primorskyi' (*GC* = 11.8%). The IRTA program was based on four founding clones too: 'Cristomorto' (*GC* = 39.9%), 'Primorskyi' (*GC* = 21.5%), 'Tuono' (*GC* = 14.4%) and 'Aï' (*GC* = 8.0%). The self-compatible Italian cultivar Tuono was the S_f donor most commonly used by the three Spanish programs. Within the 65 genotypes bred carrying the S_f allele, the 24.7% of the total *GC* came from 'Tuono' (Supplementary Material 1.3).

Table 1.6. Genotypes with the highest mean relatedness (*r*).

Genotype	Mean *r*
Nonpareil	0.153
Tardy Nonpareil	0.153
Jeffries	0.153
Kern Royal	0.153
Vesta	0.143
A97001-1bT4	0.137
Carina	0.136
Mira	0.133
Maxima	0.133

1.3.4. Pairwise relatedness

Pairwise relatedness (*r*) between all cultivars and breeding selections is showed in Supplementary Material 1.4. Cultivars with the highest mean *r* worldwide are present in Table 1.3. The genotype with the highest mean *r* was 'Nonpareil' followed by its mutants ('Tardy Nonpareil', 'Jeffries' and 'Kern Royal'). 'Vesta', from the cross 'Nonpareil' x 'Solano', was next. Carina, Mira and Maxima (Australian genotypes originated from the cross 'Nonpareil' x 'Lauranne'), followed. These three genotypes were first generation of 'Nonpareil', second generation of 'Tuono' and third generation of 'Cristomorto'.

Table 1.4 shows the mean *r* among breeding programs by country. Programs from Australia and France had the highest mean *r* (0.256 and 0.357 respectively). In contrast, Israel showed the lowest mean *r*. Comparing relatedness results between countries, Spain and the USA breeding programs were the least related. The most related breeding programs were those of France and Spain and also, Australia and France.

In the Australian breeding program, the selection A97001-1BT47 had the highest mean *r* with a value of 0.375. 'Rhea' was not related with the rest of the genotypes so its mean *r* was zero. The rest of the genotypes have a mean *r* between 0.188 and 0.333 showing a high degree of relationship.

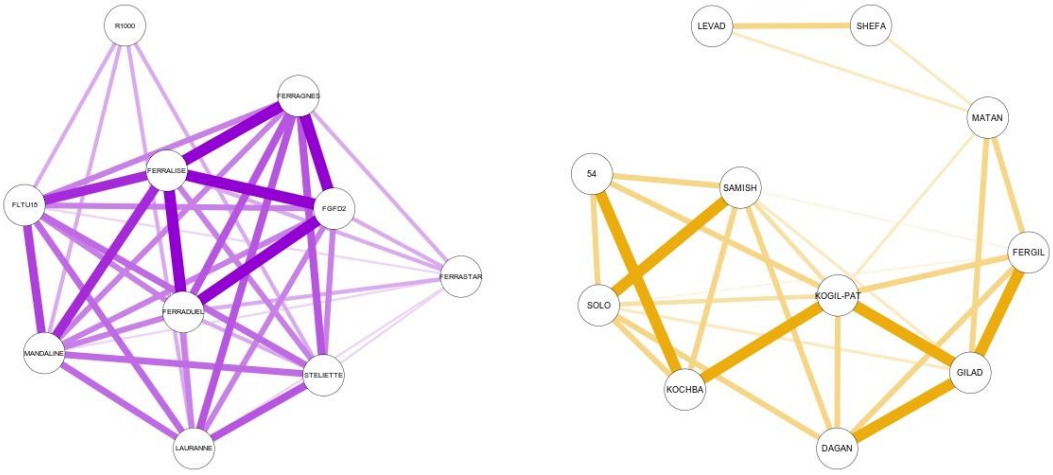


Figure 1.3. Relationship matrix of genotypes from France (left) and Israel (right). Line thickness shows degree of relationship, being the thicker lines the more related genotypes.

In the French breeding program, 'Ferralise' had the highest mean *r* (0.500). 'Ferrastar' and 'R1000' had the lowest mean *r*, 0.167 and 0.111 respectively. The rest of French genotypes had a mean *r* over 0.300, being the breeding program with the most related genotypes.

Table 1.7. Mean of pairwise relatedness (*r*) among breeding programs from five different countries.

	Australia	France	Israel	Spain	USA
Australia	**0.256**	0.156	0.081	0.094	0.172
France	-	**0.357**	0.070	0.195	0.022
Israel	-	-	**0.134**	0.047	0.050
Spain	-	-	-	**0.162**	0.009
USA	-	-	-	-	**0.232**

Genotypes from the Israeli program had a mean *r* under 0.225. The highest *r* observed between the ten cultivars released was 0.500 between two pairs: 'Dagan'-'Gilad' and 'Fergil'-'Gilad'. Selection 54 showed *r* of 0.500 with 'Kochba' and 0.250 with 'Kogil-Pat', 'Samish' and 'Solo'. Figure 1.3 compares the breeding program with the most related genotypes (France) with the breeding program with the least related genotypes (Israel).

Within the Spanish breeding programs, the highest *r* among released cultivars was 0.500 ('Antoñeta'-'Marta' and 'Makako'-'Penta'). 'Makako'-'Tardona' and 'Penta'-'Tardona' had an *r* = 0.313. The CEBAS-CSIC's selections A2-192 and C1328 had the highest *r* with a value of 1. In the CEBAS-CSIC program, 'D01-462' had the highest mean *r* (0.273). The genotypes with a higher mean *r* in the CITA breeding program were 'Guara' and 'Felisia' with values of 0.278 and 0.250 respectively. The remaining CITA genotypes had a mean *r* under 0.200. Within the IRTA breeding program, the highest *r* among released cultivars was 0.563 ('Glorieta'-'Marinada'). Among IRTA's selections, '29-47' and '35-164', showed the highest relationship with an *r* of 0.719. The selection '29-47' had the highest mean *r* (0.350). The rest of IRTA's genotypes had mean *r* over 0.130 (Supplementary Material 1.4). In the USA breeding programs, 'Nonpareil' and its mutations ('Tardy Nonpareil', 'Jeffries' and 'Kern Royal') and 'Vesta' had a mean *r* over 0.400. 'Independence' and 'Bell' had a mean *r* equal to 0. The rest of North American genotypes showed a high degree of relatedness between them. Two combinations, 'Solano'-'Vesta' and 'Sonora'-'Vesta', had *r* = 1, with 'Sonora'-'Vesta' *r* = 0.875. Analyzing the highest *r* values among

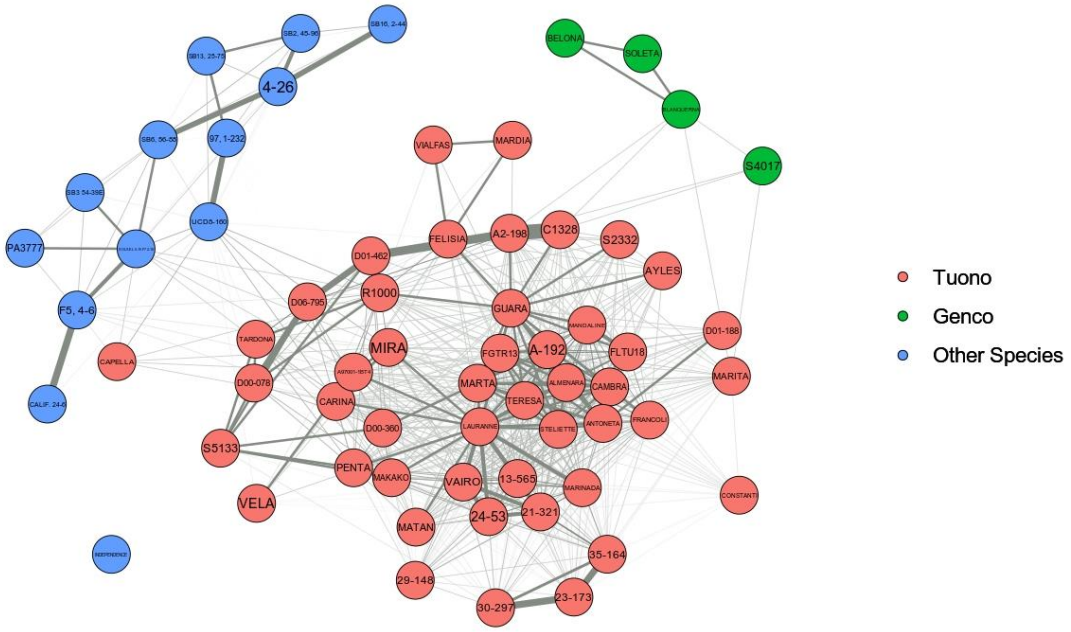


Figure 1.4. Relationship matrix of the 65 self-compatible genotypes carrying the S_f allele and its origin. Line thickness shows degree of relationship, being the thicker lines the more related genotypes.

selections and cultivars, four combinations had an $r = 1$ ('21-19W'-'Solano', '22-20'-'Solano', '21-19W'-'Sonora' and '22-30'-'Sonora'). In addition, two other pairs: '21-19W'-'Vesta' and '22-20'-'Vesta' had an r of 0.875 (Supplementary Material 1.4). Within the UCD breeding program, 'Vesta', 'Sonora' and 'Solano' had a mean r over 0.400.

Among the group of 65 genotypes carrying the S_f allele, the mean r was 0.125. Grouping the genotypes by origin of the S_f allele source ('Tuono', 'Genco' and other *Prunus spp*) the mean r were 0.210, 0.333 and 0.173 respectively (Supplementary Material 1.4). Figure 1.4 shows the main self-compatibility sources used when breeding for this character with `Tuono', 'Genco' and other *Prunus* species involved in 48, 4 and 13 genotypes respectively.

1.4. Discussion

1.4.1. Two mainstream breeding lines based on three different cultivars

Our genetic study of almond breeding programs worldwide demonstrated that the most widely used cultivars were Nonpareil, Tuono, Cristomorto and Mission. 'Nonpareil' had a large influence in USA and Australian programs, where soft-shelled nuts are bred. This reference cultivar was present in all the breeding programs studied (in some cases through its late blooming mutant Tardy Nonpareil). The self-compatible 'Tuono' and the late blooming 'Cristomorto' were extensively used in the Mediterranean programs, where hard-shelled nuts are bred. 'Mission' initially showed a considerable importance worldwide, but deeper analysis demonstrated that it was mainly influential in private American programs. Taking into account these results, we can establish two main breeding lines based on the use of three different founders: the European programs based mainly on 'Tuono' and 'Cristomorto' (hard-shell), and the North American-Australian programs based on 'Nonpareil' (soft shell). The French and Spanish breeding programs were based directly on 'Tuono' and 'Cristomorto'. In the French INRA program, the Italian cultivars Tuono and Cristomorto account for 60.0% of total *GC* and were present in the pedigree of all ten cultivars and selections evaluated. Also, the local French late-flowering and *Monilinia* resistant cultivar Aï was a parent to both 'Ferragnès' and 'Ferraduel'. In the three Spanish breeding programs, the importance of 'Tuono' and 'Cristomorto' cultivars was very high, accounting to 46.2% of total *GC*. These two cultivars were

present in the pedigree of 53 out of 59 cultivars and breeding selections from Spain. These results can be explained by the large influence of the French germplasm on the Spanish breeding programs, causing a high relationship between the programs of both countries (mean *r* = 0.195). In the North American breeding programs, 'Nonpareil' accounts for 43.7% of the total *GC* and was present in the pedigree of 48 out of 56 cultivars and breeding selections from the USA. In Australia, 'Nonpareil' accounts for 39.3% of the total *GC* and is present in the pedigree of 6 out of 7 cultivars and breeding selections. Also, 'Lauranne' (32.1% of the total *GC*) reaches an importance similar to 'Nonpareil', explaining the close relationship between the Australian and French programs (mean *r* = 0.156). Even in other countries with non-continuous breeding initiatives, such as Russia, Greece or Argentina, the use of 'Nonpareil' as a founder was common. Israel was the only country where these cultivars had a relatively low influence. This may be due to the extreme Israeli climatic conditions, forcing breeders to use locally-adapted selections as parents. In Spain, the use of locally-adapted cultivars such as Bertina at CITA as a donor for *Polystigma ochraceum (Wahlenb.) Sacc.* resistance was successful but used only to a limited extent. Other examples of secondary founders include 'Primorskyi', used regularly as late-blooming and *Fusicoccum* resistance donor in two of the Spanish breeding programs (IRTA and CEBAS-CSIC) and 'Eureka' and 'Harriott' in the North American breeding programs.

1.4.2. Loss of genetic variability and increasing of inbreeding at breeding and production level

Comparing our results on almond inbreeding with other *Prunus* species, the mean inbreeding coefficient worldwide of all genotypes (*F*=0.036) was lower than that of Japanese plum (Byrne, 1989) and apple (Noiton and Alspach, 1996) and several orders of magnitude lower than those calculated for peach (Scorza et al., 1985; Gradziel et al., 1993) and cherry (Choi and Kappel, 2004). Within almond, inbreeding and relatedness coefficients obtained in this study were higher than those reported by Lansari et al., 1994. While they documented only 10 genotypes with $F > 0$ (four of them with $F \geq 0.250$), we found 43 genotypes meeting this condition (14 of them with $F \geq 0.250$). Analyzing mean *r* by country, in the case of France and the USA (with a number of cultivars comparable in both studies), this coefficient increased. This loss of variability and an associated increase of inbreeding is due to the repeated use of a limited number of parents ('Nonpareil', 'Tuono' and 'Cristomorto') and their related genotypes, as we have shown for almond breeding.

Among the group of the 65 genotypes carrying the S_f allele for self-compatibility, the mean *r* was 0.125. In cherry self-compatible selections, coefficients of coancestry ranged from 0.102 to 0.256 (Choi and Kappel, 2004) and thus were of similar magnitude. In Western Europe, the Italian cultivar Tuono was used extensively as a source of self-compatibility, late blooming and spur type cropping. More recently, it has become important in Israel and Australia (in Australia through 'Lauranne' ('Ferragnès' x 'Tuono')). This 'Ferragnès' x 'Tuono' cross also originated the cultivar Steliette and was later successfully used in two of the Spanish breeding programs, resulting in three self-compatible cultivars: Cambra at CITA, and Antoñeta and Marta at CEBAS-CSIC. Thus, these five cultivars are full-siblings. In addition, in the USA, breeders are using 'Guara' (syn 'Tuono') as S_f donor. A similar case occurred in sweet cherry with the cultivar Stella as it was the most frequently utilized parent for self-compatible selections in North America (Choi and Kappel, 2004).

A lack of diverse germplasm may limit continued progress in almond breeding programs. This genetic limitation is of particular concern in the main producing countries. Thus, Californian and Australian production rely mainly on 'Nonpareil' and closely related cultivars (Australian Almond Board, 2019; Californian Almond Board, 2019), while in Spain, some new Spanish cultivars like Vairo and Penta, derived from second generation of 'Tuono' and 'Cristomorto', as well as 'Belona' and 'Soleta', derived from second generation of 'Genco', are replacing

traditional cultivars in new orchards. This trend is also favored by the almond industry needs. Only in some regions of Central Asia, Middle East and North Africa, local and well adapted traditional selections still play an important role in commercial production (Gouta et al., 2010; Elhamzaoui et al., 2012; Zaurov et al., 2015; Hamadeh et al., 2018).

1.4.3. Usefulness of pedigree data analyzing breeding tendencies

Pedigree analysis is a cost-effective and well-established way to monitoring inbreeding and relatedness among controlled breeding populations. However, the veracity of any analysis based on this kind of data relies on the accuracy of records collected across multiple institutions and by many breeders. In order to verify parental relationships of the genotypes under study, we used SSRs, SNPs and self-incompatibility S alleles data from previous analysis carried out by the breeding programs taking part in this study. Our molecular marker analysis confirmed 146 parentage relationships and found three errors (2% error rate), which were corrected accordingly. Thus, the marker-based pedigree analysis performed showed only small parental changes and corroborate the consistency of the results reached by this study.

However, several reports have demonstrated that large-scale genomic analysis may provide more accurate results than pedigree analysis (Kardos et al., 2015; Wang, 2016). This kind of genome based pedigree analysis has already been performed in apple (Muranty et al., 2020). The recent publication of two almond reference genomes (Sánchez-Pérez et al., 2019a; Alioto et al., 2020a) and the increasing availability of quality genomic data opens opportunities to complement our study and obtain more complete and accurate pedigrees based on genomic variability. This kind of studies can be useful even when some genotypes were discarded due to breeding process, as is the case in our almond pedigree work.

Although almond showed a higher genetic variability than other *Prunus* species, the historical expansion of almond from the Mediterranean region to California and from California to Australia could have caused a bottleneck effect in the breeding population under study. Different studies have reported a high genetic relatedness between Australian and Californian cultivars (Martínez-Gómez et al., 2003; Fernández i Martí et al., 2015), possibly caused by the introduction of a limited number of cultivars from Europe to these countries. In addition, breeding programs worldwide have used cultivars from French origin as main founders as Aï, Princesse, Ardechoise, Nonpareil, IXL, Ne Plus Ultra or Nikitskij. This situation could have led to an underestimation of relatedness and inbreeding. The use of large-scale genomic data would provide most valuable information in this respect, expanding the almond pedigree beyond breeding records.

1.5. Conclusions

This almond pedigree study reviews the progress made in breeding over the last 50 years. Results showed that two main breeding lineages, based on only three cultivars (Nonpareil, Tuono and Cristomorto) have dominated modern breeding worldwide. This limitation has led to the high level of inbreeding found in modern cultivars. The inbreeding observed in our study could explain the phenotypic depression early reported in breeding populations (Grasselly, 1976b; Grasselly and Olivier, 1981; Alonso and Socias i Company, 2005; Socias i Company, 2011; Martínez-García et al., 2012). Thus, future almond breeding should avoid inbreeding and favor genetic gain. Diversify the sources of self-compatibility, which are presently dominated by 'Tuono', and broaden the germplasm used when breeding are urgent needs. Additional analyses based on genomic data are needed to more accurately determine the levels of inbreeding and the loss of genetic variability among almond breeding programs worldwide.

1.6. References

Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., et al. (2020). Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *Plant J.* 101, 455–472. doi: 10.1111/tpj.14538.

Alonso, J., and Socias i Company, R. (2005). Self-incompatibility expression in self-compatible almond genotypes may be due to inbreeding. *J. Am. Soc. Hortic. Sci.* 130, 865–869. Available at: https://journals.ashs.org/jashs/view/journals/jashs/130/6/article-p865.xml

Arulsekar, S., Parfitt, D. E., and Kester, D. E. (1986). Comparison of isozyme variability in peach and almond cultivars. *J. Hered.* 77, 272–274. doi: 10.1093/oxfordjournals.jhered.a110235.

Arús, P., Gradziel, T., Oliveira, M. M., and Tao, R. (2009). "Genomics of Almond," in *Genetics and Genomics of Rosaceae* (New York, NY: Springer New York), 187–219. doi: 10.1007/978-0-387-77491-6_9.

Australian Almond Board (2019). Almond insights 2018-19. Available at: https://www.australianalmonds.com.au/ [Accessed April 20, 2020].

Bartolozzi, F., Warburton, M. L., Arulsekar, S., and Gradziel, T. M. (1998). Genetic characterization and relatedness among california almond cultivars and breeding lines detected by randomly amplified polymorphic DNA (RAPD) analysis. *J. Am. Soc. Hortic. Sci.* 123, 381–387.

Batlle, I., Dicenta, F., Gradziel, T. M., Wirthensohn, M., Duval, H., and Vargas, F. J. (2017). "Classical genetics and breeding," in *Almonds: Botany, production and uses* (Boston: CABI), 111–148.

Byrne, D. (1989). Inbreeding, coancestry, and founding clones of Japanese-type plums of California and the southeastern United States. *J. Am. Soc. Hortic. Sci.*

Byrne, D. H. (1990). Isozyme Variability in Four Diploid Stone Fruits Compared with Other Woody Perennial Plants. *J. Hered.* 81, 68–71. doi: 10.1093/oxfordjournals.jhered.a110927.

Cabrita, L., Apostolova, E., Neves, A., Marreiros, A., and Leitão, J. (2014). Genetic diversity assessment of the almond (*Prunus dulcis* (Mill.) D.A. Webb) traditional germplasm of Algarve, Portugal, using molecular markers. *Plant Genet. Resour. Characterisation Util.* 12, S164–S167. doi: 10.1017/S1479262114000471.

Californian Almond Board (2019). Californian Almond Board. Available at: https://www.almonds.com/ [Accessed April 21, 2020].

Choi, C., and Kappel, F. (2004). Inbreeding, coancestry, and founding clones of sweet cherries from North America. *J. Am. Soc. Hortic. Sci.* 129, 535–543. doi: 10.21273/JASHS.129.4.0535.

International Nut & Dried Fuits Council. (2019). Nuts & dried fruits statistical yearbook 2018/2019. Reus Available at: https://www.nutfruit.org/files/tech/1553521370_INC_Statistical_Yearbook_2018.pdf [Accessed April 12, 2020].

Debuse, C. J., Shaw, D. V., and Dejong, T. M. (2005). Response to inbreeding of seedling traits in a *Prunus domestica* L. breeding population. *J. Am. Soc. Hortic. Sci.* 130, 904–911. doi: 10.21273/JASHS.130.6.904.

Dicenta, F., Sánchez-Pérez, R., Rubio, M., Egea, J., Batlle, I., Miarnau, X., et al. (2015). The origin of the self-compatible almond "Guara." *Sci. Hortic. (Amsterdam).* 197, 1–4. doi: 10.1016/j.scienta.2015.11.005.

Elhamzaoui, A., Oukabli, A., Charafi, J., and Moumni, M. (2012). Assessment of genetic diversity of Moroccan cultivated almond (*Prunus dulcis* Mill. DA Webb) in its area of extreme diffusion, using nuclear microsatellites. *Am. J. Plant Sci.* 3, 1294–1303.

Fernández i Martí, A., Font i Forcada, C., Kamali, K., Rubio-Cabetas, M. J., Wirthensohn, M., and Socias i Company, R. (2015). Molecular analyses of evolution and population structure in a worldwide almond [*Prunus dulcis* (Mill.) D.A. Webb syn. *P. amygdalus* Batsch] pool assessed by microsatellite markers. *Genet. Resour. Crop Evol.* 62, 205–219. doi: 10.1007/s10722-014-0146-x.

Gouta, H., Ksia, E., Buhner, T., Moreno, M. Á., Zarrouk, M., Mliki, A., et al. (2010). Assessment of genetic diversity and relatedness among Tunisian almond germplasm using SSR markers. *Hereditas* 147, 283–292. doi: 10.1111/j.1601-5223.2009.02147.x.

Gradziel, T., Beres, W., and Pelletreau, K. (1993). Inbreeding in California canning clingstone peach cultivars. *Fruit Var. J.*

Gradziel, T. M., Curtis, R., and Socias i Company, R. (2017). "Production and growing regions," in *Almonds: Botany, production and uses* (Boston: CABI), 70–86.

Grasselly, C. (1976). Mise en évidence de quelques types autocompatibles parmi les cultivars d'amandier (P. amygdalus Batsch) de la population des Pouilles. Available at: https://pascal-

francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=PASCAL7638005468 [Accessed April 22, 2020].

Grasselly, C., and Crossa-Raynaud, P. (1980). *The almond tree*. Paris: Maisonneuve et Larose.

Grasselly, C., and Olivier, G. (1981). Difficulté de survie de jeunes semis d'amandiers dans certaines descendances. *Options Mediterr.* Available at: http://ressources.ciheam.org/om/pdf/s01/CI010763.pdf [Accessed April 22, 2020].

Halász, J., Kodad, O., Galiba, G. M., Skola, I., Ercisli, S., Ledbetter, C. A., et al. (2019). Genetic variability is preserved among strongly differentiated and geographically diverse almond germplasm: an assessment by simple sequence repeat markers. *Tree Genet. Genomes* 15, 1–13. doi: 10.1007/s11295-019-1319-8.

Hamadeh, B., Chalak, L., Coppens d'Eeckenbrugge, G., Benoit, L., and Joly, H. I. (2018). Evolution of almond genetic diversity and farmer practices in Lebanon: Impacts of the diffusion of a graft-propagated cultivar in a traditional system based on seed-propagation. *BMC Plant Biol.* 18, 1–18. doi: 10.1186/s12870-018-1372-8.

Kardos, M., Luikart, G., and Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity (Edinb).* 115, 63–72. doi: 10.1038/hdy.2015.17.

Keneni, G., Bekele, E., Imtiaz, M., and Dagne, K. (2012). Genetic vulnerability of modern crop cultivars: causes, mechanism and remedies. *Int. J. Plant Res.* 2, 69–79. doi: 10.5923/j.plant.20120203.05.

Kester, D. E., and Gradziel, T. M. (1996). "Fruit breeding," in, eds. J. Janick and J. N. Moore (Wiley), 1–97.

Kester, D. E., Gradziel, T. M., and Grasselly, C. (1991). Almonds (Prunus). *Acta Hortic.*, 701–760. doi: 10.17660/actahortic.1991.290.16.

Lansari, A., Kester, D. E., and Iezzoni, A. F. (1994). Inbreeding, coancestry, and founding clones of almonds of California, Mediterranean shores, and Russia. *J. Am. Soc. Hortic. Sci.* 119, 1279–1285. doi: 10.21273/JASHS.119.6.1279.

López, M., Vargas, F. J., and Batlle, I. (2006). Self-(in)compatibility almond genotypes: A review. *Euphytica* 150, 1–16. doi: 10.1007/s10681-005-9009-z.

Marchese, A., Bošković, R. I., Martínez-García, P. J., and Tobutt, K. R. (2008). The origin of the self-compatible almond 'Supernova.' *Plant Breed.* 127, 105–107. doi: 10.1111/j.1439-0523.2008.01421.x.

Marrano, A., Martínez-García, P. J., Bianco, L., Sideli, G. M., Di Pierro, E. A., Leslie, C. A., et al. (2019). A new genomic tool for walnut (Juglans regia L.): development and validation of the high-density Axiom™ J. regia 700K SNP genotyping array. *Plant Biotechnol. J.* 17, 1027–1036.

Martínez-García, P. J., Dicenta, F., and Ortega, E. (2012). Anomalous embryo sac development and fruit abortion caused by inbreeding depression in almond (*Prunus dulcis*). *Sci. Hortic. (Amsterdam).* 133, 23–30. doi: 10.1016/j.scienta.2011.10.001.

Martínez-Gómez, P., Arulsekar, S., Potter, D., and Gradziel, T. M. (2003). An extended interspecific gene pool available to peach and almond breeding as characterized using simple sequence repeat (SSR) markers. *Euphytica* 131, 313–322. doi: 10.1023/A:1024028518263.

Mnejja, M., Garcia-Mas, J., Audergon, J. M., and Arús, P. (2010). Prunus microsatellite marker transferability across rosaceous crops. *Tree Genet. Genomes* 6, 689–700. doi: 10.1007/s11295-010-0284-z.

Muranty, H., Denancé, C., Feugey, L., Crépin, J. L., Barbier, Y., Tartarini, S., et al. (2020). Using whole-genome SNP data to reconstruct a large multi-generation pedigree in apple germplasm. *BMC Plant Biol.* 20, 1–18. doi: 10.1186/s12870-019-2171-6.

Noiton, D., and Alspach, P. (1996). Founding Clones, Inbreeding, Coancestry, and Status Number of Modern Apple Cultivars. *J. Am. Soc. Hortic. Sci.* 121, 773–782. Available at: https://journals.ashs.org/jashs/view/journals/jashs/121/5/article-p773.xml [Accessed October 18, 2019].

Ortega, E., and Dicenta, F. (2003). Inheritance of self-compatibility in almond: Breeding strategies to assure self-compatibility in the progeny. *Theor. Appl. Genet.* 106, 904–911. doi: 10.1007/s00122-002-1159-y.

Rikhter, A. (1972). Biological basis for the creation of almond cultivars and commercial orchards. *Akad. Nauk SSSR*.

Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Aiese Cigliano, R., Del Cueto, J., et al. (2019). Mutation of a bHLH transcription factor allowed almond domestication. *Science (80-.).* 364, 1095–1098. doi: 10.1126/science.aav8197.

Scorza, R., Mehlenbacher, S. ., and Lightner, G. . (1985). Inbreeding and coancestry of freestone peach cultivars of the eastern United States and implications for peach germplasm improvement. *J. Am. Soc. Hortic. Sci.*

Sjulin, T., and Dale, A. (1987). Genetic diversity of North American strawberry cultivars. *J. Am. Soc. Hortic. Sci.*, 375–385. Available at: http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=ACERVO.xis&method=post&formato=2&cantidad=1&expresion=mfn=032363 [Accessed April 20, 2020].

Socias i Company, R. (2011). Breeding self-compatible almonds. *Plant Breed. Rev.* 8, 313–338. doi: 10.1002/9781118061053.ch9.

Socias i Company, R. (2017). "Pollen-style (in)compatibility: development of autogamous cultivars," in *Almonds: Botany, production and uses* (Boston: CABI), 188–208.

Son, K. M., Kwon, S. Il, and Choi, C. (2012). Inbreeding, coancestry, and founding clones of apple cultivars released from Korea. *Hortic. Environ. Biotechnol.* 53, 404–409. doi: 10.1007/s13580-012-0012-8.

Van De Wouw, M., Kik, C., Van Hintum, T., Van Treuren, R., and Visser, B. (2010). Genetic erosion in crops: concept, research results and challenges. *Plant Genet. Resour. Characterisation Util.* 8, 1–15. doi: 10.1017/S1479262109990062.

Velasco, D., Hough, J., Aradhya, M., and Ross-Ibarra, J. (2016). Evolutionary Genomics of Peach and Almond Domestication. *G3 (Bethesda).* 6, 3985–3993. doi: 10.1534/g3.116.032672.

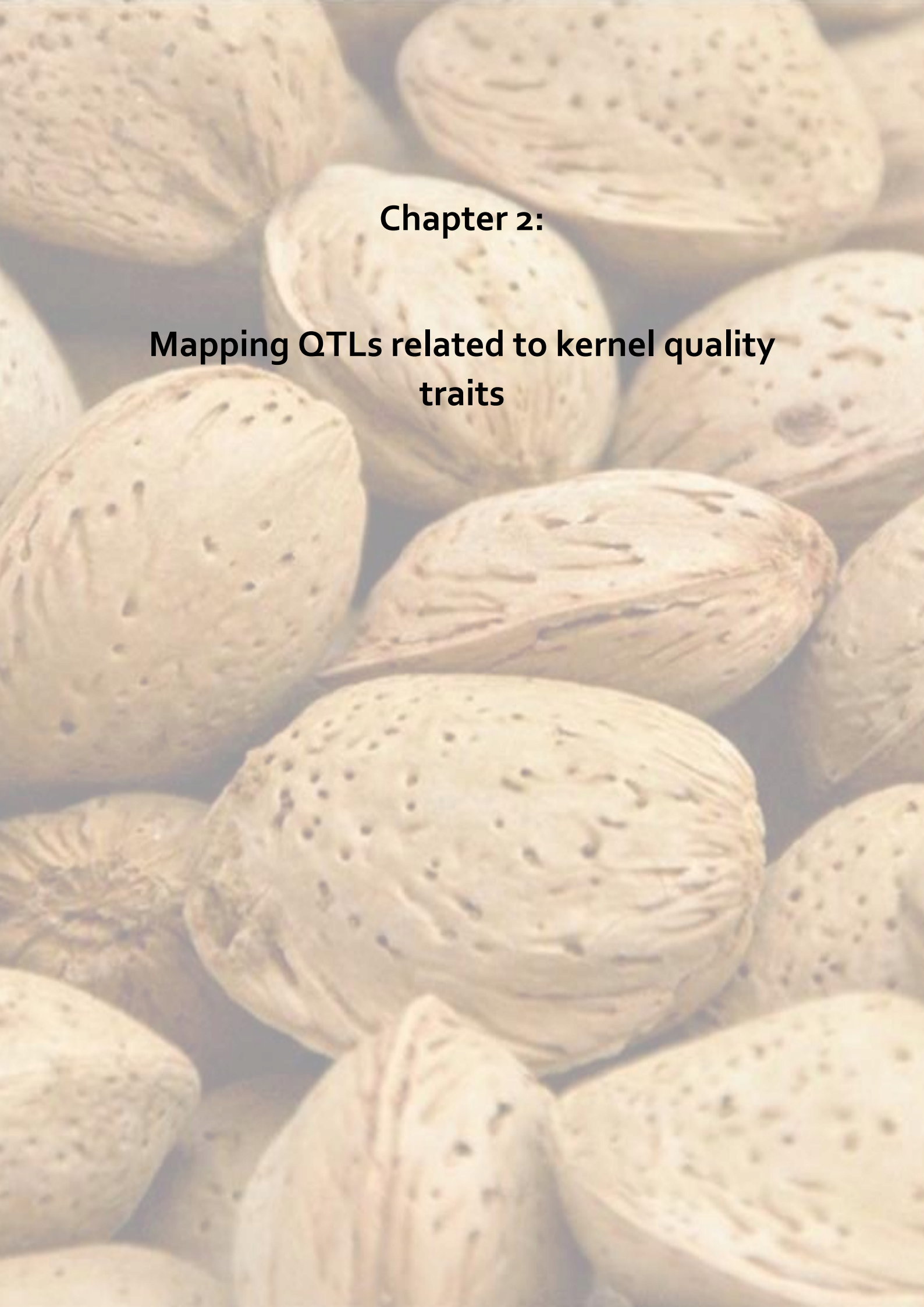
Wang, J. (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theor. Popul. Biol.* 107, 4–13. doi: 10.1016/j.tpb.2015.08.006.

Wood, M. N. (1925). *Almond varieties in the United States*. Washington D.C.: US Department of Agriculture.

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat.* 56, 330–338. doi: 10.1086/279872.

Zaurov, D., Eisenman, S., Ford, T., Khokhlov, S., Kenjebaev, S., Shalpykov, K., et al. (2015). Genetic resources of almond species in the former USSR. *J. Am. Soc. Hortic. Sci.* 50, 18–29.

Zeinalabedini, M., Khayam-Nekoui, M., Grigorian, V., Gradziel, T. M., and Martínez-Gómez, P. (2010). The origin and dissemination of the cultivated almond as determined by nuclear and chloroplast SSR marker analysis. *Sci. Hortic. (Amsterdam).* 125, 593–601. doi: 10.1016/j.scienta.2010.05.007.



Chapter 2:

Mapping QTLs related to kernel quality traits

QTL mapping of kernel quality traits in almond

Felipe Pérez de los Cobos[1,2,3], Agustí Romero[1], Pere Arús[2,3], Iban Eduardo[2,3], Ignasi Batlle[1]

[1]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Mas Bové, Ctra. Reus-El Morell Km 3,8 43120 Constantí Tarragona, Spain

[2]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Centre de Recerca en Agrigenòmica (CRAG), CSIC-IRTA-UAB-UB. Cerdanyola del Vallès (Bellaterra), 08193 Barcelona, Spain

[3]Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Cerdanyola del Vallès (Bellaterra), 08193 Barcelona, Spain

Manuscript in preparation

Abstract

Almond [*Prunus dulcis* (Miller) D.A. Webb] stands out for its adaptability to different conditions and its kernel's high nutritional value. Additionally, its antioxidant properties reduce the risk of diseases such as arthritis, vasculitis, high blood pressure, cancer or Alzheimer's. That makes the almond a crop with a high potential to adapt to an agriculture threatened by the climate change and the needs of feeding an increasing world population. However, our knowledge about the inheritance of kernel quality traits is still limited. In this study, we performed a QTL mapping of kernel physical and chemical quality traits. A F1 population, coming from the cross 'Marcona' x 'Marinada', was phenotyped during four years using conventional and image analysis methods. The use of the almond 60K SNP array allowed us to build high quality linkage maps. In total, 12 major and minor QTLs were mapped for the traits under study. The QTLs found for symmetry and kernel shoulder are of particular interest in almond, since it is the first time any QTL has been mapped for these traits. Another QTL was mapped to margaric acid content, a fatty acid closely related to oleic acid content. These QTLs will allow the development of molecular markers for kernel quality traits in almond and therefore the implementation of marker assisted selection in breeding programs.

2. Chapter 2

2.1. Introduction

Humanity is facing one of its greatest challenges. According to FAO (https://www.fao.org/home/en/), by 2050 the world population will have increased to 10 billion people, rising by 50% the need for food production to meet global demand. At the same time, land degradation, water scarcity and climate change put at risk food production. In this scenario, the sustainability of agricultural systems, their adaption to climate change and the optimization in the use of limited resources such as water become key to achieving food security.

In this context, almond [*Prunus dulcis* (Miller) D.A. Webb, syn. *P. amygdalus* (L) Batsch] stands out as one of the crops that can best meet these needs. Almond shows one of the widest range of blooming dates among crops, adapting it to different climates (Alonso Segura et al., 2017; Martínez-Gómez et al., 2017). In addition, it is a crop adapted to different irrigation regimes, from the total irrigation practiced in California, to the traditional dryland farming established on the Mediterranean coast (Gradziel et al., 2017). Apart from its extreme adaptability to different environments and irrigation regimes, almonds have a high nutritive value (Becerra-Tomás et al., 2019; Barreca et al., 2020). It arises partly from its high proportion of oleic acid, that constitutes an important source of calories, but does not contribute to cholesterol formation in humans. Additionally, the high protein fraction of almond kernels is of high quality and readily assimilable. Its antioxidant properties reduce the risk of diseases such as arthritis, vasculitis, high blood pressure, cancer or Alzheimer's (Becerra-Tomás et al., 2019; Barreca et al., 2020). These characteristics make the almond a crop with a high potential to adapt to an agriculture stressed by the effects of climate change and the needs of healthily feeding an increasing world population.

Apart from its importance for the final consumer, kernel quality traits must also be taken in account from the point of view of the kernel industrial aptitude. Almond uses are largely diverse, including raw consumption, snacks, chocolates, marzipans, cookies, ice creams, etc. Each use has its own specifications for quality and different varieties can adapt better to specific uses. Each almond use has its own preferred kernel type in terms of size, shape, physical properties and chemical composition. Kernels of more than 1,5 g are considered large, while those weighing less than 1,0 g are considered small. Almond breeding programs should include industrial aptitude as breeding objectives due to its relevance for the future acceptance of new cultivars (Batlle et al., 2017). Almond quality requirements for industrial aptitude was revised by Romero, 2014.

Kernel quality has been a major breeding objective for many almond breeding programs worldwide. Therefore, several studies have focused on mapping QTLs related to physical and chemical almond kernel quality traits using different approaches. Fernández i Martí et al., 2013 and Font i Forcada et al., 2012 used the same F1 population to study kernel physical and chemical traits, respectively, founding several QTLs related to those traits. In Font i Forcada et al., 2015, they identified several SSRs associated to tocopherol, protein and fatty acids content via association mapping. Finally, Di Guardo et al., 2021, found several QTLs related to volatiles composition in raw and roasted almonds. However, due to the polygenic character of these traits, the complexity of the phenotyping process or the limited number of molecular markers used in the mentioned studies, these results have not been translated into molecular markers useful for breeding. Only one study has been able to propose a candidate gene for a trait related to kernel quality, in this case the accumulation of amygdalin (Sánchez-Pérez et al., 2019a). As

a result of these drawbacks, there are currently no marker assisted selection (MAS) strategies for kernel quality traits in almond breeding.

In this study, a QTL mapping of kernel quality traits such as kernel weight, shape-related traits, color and chemical traits, has been carried out. A F1 population, coming from a 'Marcona' x 'Marinada' cross, has been phenotyped during 4 years. High-quality, highly saturated linkage maps were created using the 60K SNPs array (Duval et al., 2023a) available for almond. Finally, a QTL mapping was carried out. We identified 12 QTLs associated with the traits under study. The results of this study will allow the development of molecular markers for kernel quality traits in almond and therefore the implementation of MAS into breeding programs.

2.2. Materials and methods

2.2.1. Plant material and Genotyping

Plant material consisted in an F1 population coming from the cross 'Marcona' x 'Marinada' (Marc x Mari). 'Marcona' is a Spanish traditional cultivar, with a rounded shape kernel and high level of fatty acids, above the rest of the varieties. 'Marinada' is a modern cultivar released by IRTA in 2008, self-compatible and with a very sweet kernel. The number of individuals forming the population was 91. Seedlings were grafted onto 'Garnem' roostock, planted at 4m x 1,8m in 2015. Marc x Mari population was kept in Mas Bove IRTA experimental station (41.170723 N, 1.172942 E) under standard agricultural practices.

Total genomic DNA from the 91 individuals and two parents was isolated using the protocol followed in Sonneveld et al. 2001. After that, samples were genotyped using the almond 60K SNP array (Duval et al., 2023a).

2.2.2. Phenotypic data collection and processing

Marc x Mari population and both parents were evaluated for kernel weight, kernel shape-related traits, crack-out, color traits and chemical traits. As kernel shape-related traits, we phenotyped kernel length, width, thickness, roundness, globosity, shoulder and symmetry. As kernel chemical traits, we phenotyped kernel protein, fiber and fat content and fatty acids profile, including myristic, palmitic, palmitoleic, margaric, cis-10-heptadecenoic, stearic, oleic, vaccenic, linoelaidic, γ-linoleic, arachidic and cis-11-eicosenoic acid.

Fifty mature fruits were randomly collected from each individual of the Marc x Mari population, including the parents. The fruits were considered mature when the mesocarp was fully dry and split along the fruit suture and the peduncle was near to complete abscission. After measuring nut weight, shells were cracked to obtain the kernels. Weights were obtained using an electronic balance. Crack-out was calculated according to: crack-out = (nut weight – kernel weight) / nut weight. Crack-out was measured from 2019 to 2021.

After that, we measured kernel length, width and thickness with a digital Vernier caliper. Roundness and globosity were estimated using the ratios width/length and width/thickness. These traits were measured during four consecutive years, from 2018 to 2021.

For image analysis, we took a standard photo of six kernels from each individual and we analyze them using the shape analyzer software (Jurado-Ruiz et al., 2023). This image analysis tool is based on deep learning and it is able to detect almond kernels and measure them automatically within an image. From the different parameters this tool is able to measure, we only included in this study kernel length IA, width IA, roundness IA, symmetry SSIM and symmetry jaccard. These traits were phenotyped during two consecutive years, 2020 and 2021. Additionally, in the year 2020, we measured the angle of the kernel shoulder using tomato analyzer (Gonzalo

et al., 2009) and we performed a visual estimation of the kernel shoulder using values from one to five, one meaning no shoulder and five a large shoulder (Supplementary Material 2.1).

Tegument (the brown kernel skin) and kernel color were determined with a Minolta Chroma Meter (CR-300; Minolta, Ramsey, NJ, USA) tri-stimulus color analyser calibrated to a white porcelain reference plate using a CIELAB scale with color space coordinates L, a and b. Tegument and kernel color were measured from 2018 to 2021.

Fat content was analyzed by Soxhlet method, using 5 – 6 g of crushed kernels (without skin) and petroleum ether (boiling point 40 to 60 ºC) for 7 h in Soxhlet apparatus. This trait was measured only in 2021.

Crude protein was analyzed by Dumas' combustion procedure using Leco FP-528 analyzer. Briefly, 0.2 g of grounded sample was weighed in a porcelain sample holder (boat) for introduction into the combustion chamber (850±1ºC) utilizing an automated sample loader. The combustion process converts covalently bound 130 nitrogen into nitrogen gas (N2) that is quantified by passing the gas through a conductivity cell. Protein content was computed using a 6.25 factor.

Crude fiber was measured using 1 g of ground sample by adding boiling 0.26 N sulfuric acid (30 min) followed by boiling 0.23 N potassium hydroxide (30 min). The extracted residue was dried at 103 ±1ºC (3 h) and the dried sample weighed, put in a furnace (550±1ºC for 3 h), and finally the ashes were weighed.

Fatty acids were analyzed by gas-chromatography with flame ionization detector (GC-FID) using a capillary column. The fatty acid methyl esters (FAMEs) were prepared by trans-esterification with 0.5 M potassium hydroxide, following the official method UNE-EN ISO 5509:2000. FAMEs (1 mL) were separated using a gas-chromatograph (HP 6890; Agilent Technologies, Barcelona, Spain) equipped with an FID detector and a capillary column [30 m · 0.25 mm i.d. (HP-Innowax, Agilent Technologies)]. The carrier gas was helium, and the flow rate was 1 mL·min–1. The injector and detector temperatures were 220 and 275ºC, respectively. The FAME identification was based on retention time relative to those of a standard FAME mixture (Sigma-Aldrich, Madrid, Spain). Fatty acids were measured only in 2021.

For traits with more than one year of data, least square means were calculated accordingly to the following equations:

$$P_i = Genotype_i + e_i \quad (1)$$

Where P_i is the phenotypic value of the ith genotype, $Genotype_i$ is the genotypic effect of the ith genotype, and e_i is the residual error of the model.

$$P_{ij} = Genotype_i + Year_j + e_{ij} \quad (2)$$

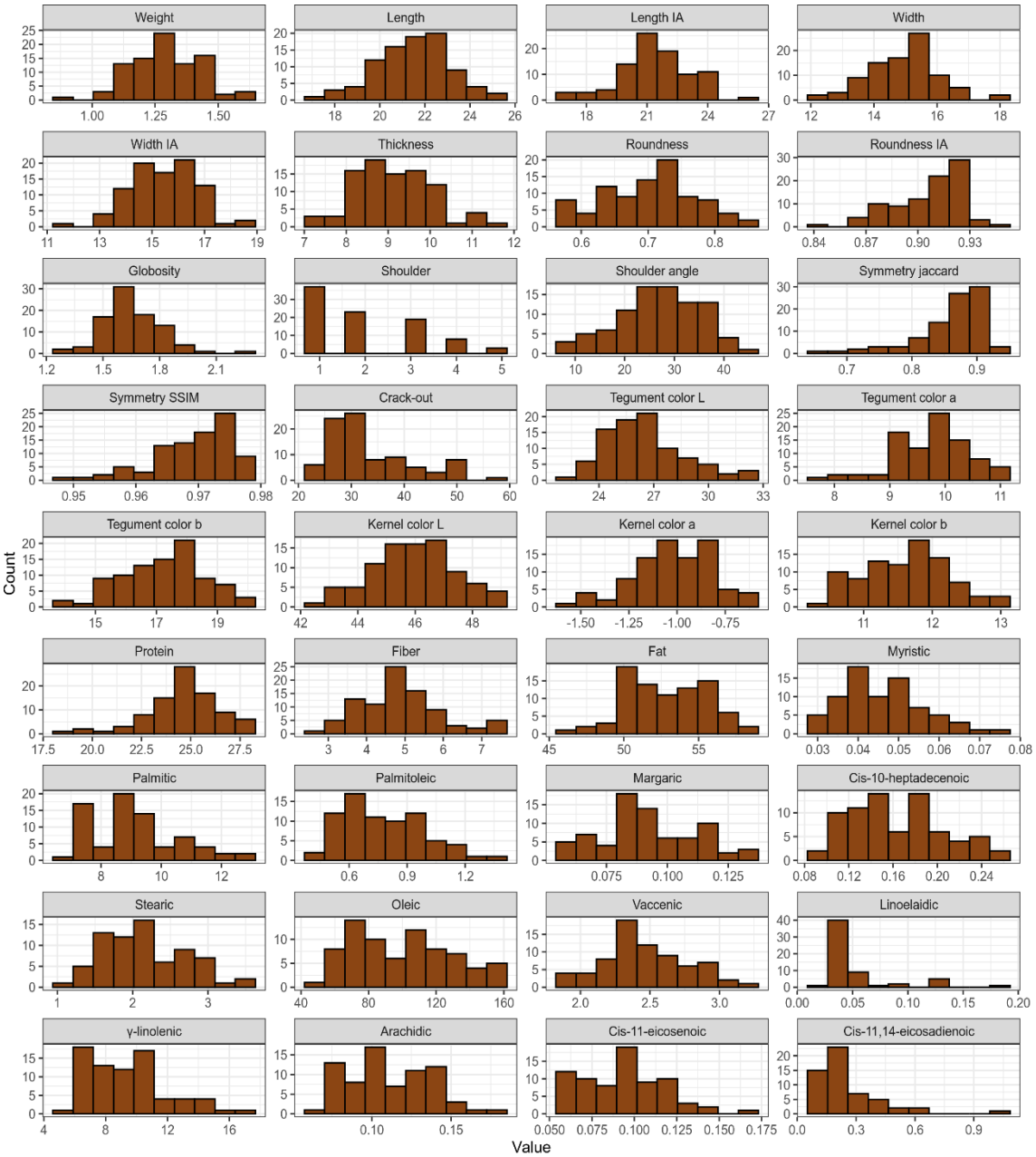


Figure 2.1. Histograms of frequency for the traits under study. For traits with more than one year of data, lsmean values are used. X axis represents phenotypic values, Y axis represents frequency.

Where P_{ij} is the phenotypic value of the ith genotype in the jth year, $Genotype_i$ is the genotypic effect of the ith genotype, $Year_j$ is the effect of the jth year, and e_{ij} is the residual error of the model.

Results from lsmean regression 1 and 2 were compared and the one with the highest R^2 was selected as phenotypic data. After that, Pearson's correlation coefficient was calculated for all traits.

2.2.3. SNP filtering, linkage map construction and QTL mapping

Genotypic data was retrieved using the Axiom Analysis Suite. Samples were filtered following the Axiom best practices workflow, but setting average call rate > 95. Then, we filtered out SNPs with the following characteristics: (i) monomorfic SNPs in the progeny; (ii) heterozygous SNPs in 'Marcona' and 'Marinada', but with only two genotypic classes in the progeny; (iii)

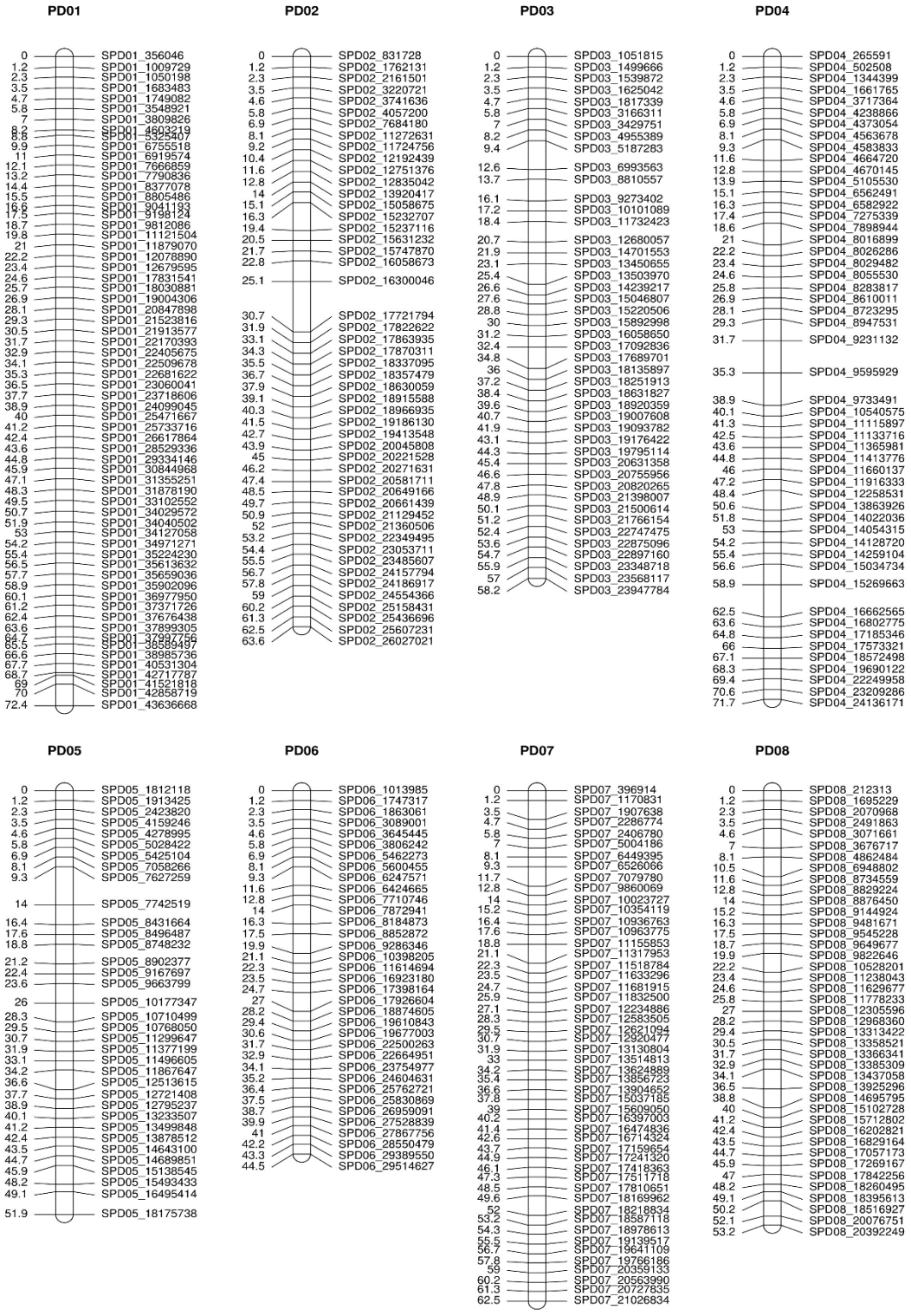


Figure 2.2. Marc x Mari CP linkage map.

homozygous SNPs in 'Marcona' and heterozygous in 'Marinada', but with three genotypic classes in the progeny; (iv) homozygous SNPs in 'Marinada' and heterozygous in 'Marcona', but with three genotypic classes in the progeny.

After filtering based on segregation, AlphaFamImputed (Whalen et al., 2020) with default settings was used to impute missing data and detect genotyping errors. Then, using a homemade R script, SNPs were phased. We ordered the SNPs based on their physical position

and established a set of bins (i.e. groups of SNPs with identical genotype for all the individuals), where each bin was separated from the adjacent bin by a single or a few recombination events.

Finally, three linkage maps were built using JoinMap 5 (https://www.kyazma.nl/index.php/JoinMap/): Marc x Mari CP map was built using all the bins previously selected, Marcona map was built using only bins segregating in 'Marcona', and Marinada map was built using only bins segregating in 'Marinada'. For the CP map construction process, maximum likehood's algorithm was used, for Marcona map and Marinada map, ksambi's algorithm was used. QTLs for all the traits were analyzed in the three maps using MapQTL 6.0 (https://www.kyazma.nl/index.php/MapQTL/). Lsmean and raw data for all traits were used for QTL mapping. For QTL mapping process, interval mapping and Kruskal-Wallis algorithms were used. All QTLs significant in the K-W analysis and with LOD ≥ 4.0 in the lsmean data and consistent between years in the raw data were considered as significant. These QTLs were named according to the recommendations for standard QTL nomenclature and reporting of the Genome Database for Rosaceae. Those QTLs explaining more than 25% of the variance were considered as major QTLs.

2.3. Results

2.3.1. Including the variable year improved lsmean regressions

In all cases, lsmean regression 1 had a lower R^2 than lsmean regression 2 (Table 2.1). However, these differences were higher in traits such as weight, length, width, thickness and all color and chemical traits, where lsmean regression 1 had a R^2 lower than 0.5. In the case of traits measured by image analysis such as length IA, width IA, roundness IA, symmetry jaccard and symmetry SSIM, lsmean regression 1 had R^2 values really close to those of lsmean regression 2.

Table 2.1. R^2 values of the lsmean regressions 1 and 2 for every trait.

Trait	R^2 lsmean 1 (Trait = Genotype + e)	R^2 lsmean 2 (Trait = Genotype + year + e)
Weight	0.382	0.712
Length	0.487	0.641
Length IA	0.736	0.850
Width	0.325	0.744
Width IA	0.751	0.803
Thickness	0.325	0.598
Roundness	0.629	0.771
Roundness IA	0.894	0.911
Globosity	0.506	0.628
Symmetry jaccard	0.912	0.913
Symmetry SSIM	0.873	0.878
Crack-out	0.787	0.800
Tegument color L	0.310	0.472
Tegument color a	0.180	0.743
Tegument color b	0.139	0.780
Kernel color L	0.153	0.661
Kernel color a	0.344	0.650
Kernel color b	0.301	0.744
Protein	0.416	0.654
Fiber	0.294	0.453

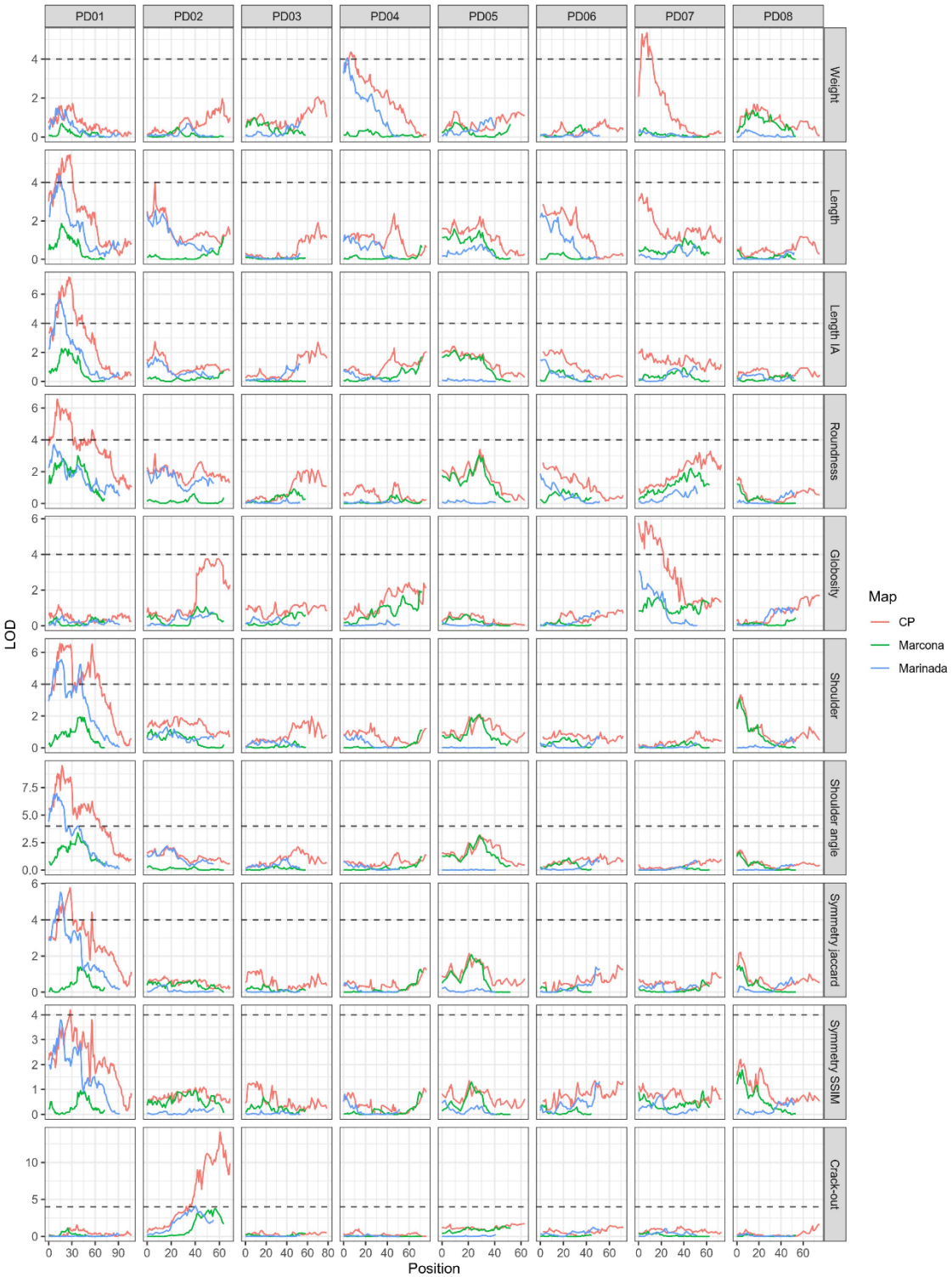


Figure 2.3. QTL mapping of kernel weight, shape-related traits and crack-out. In columns, the eight almond chromosomes, in rows the different traits. X axis represents the position measured in cM, Y axis represents the LOD value. The horizontal dashed line indicates LOD = 4.

In general, all traits showed a normal distribution, with some exceptions (Figure 2.1). In the case of fat content and some fatty acids such as myristic, palmitic, palmitoleic, cis-10-heptadecenoid, oleic and γ-linoleic, the frequency distribution fitted with a bimodal distribution. Distributions of roundness IA, symmetry jaccard and symmetry SSIM did not fit with a normal distribution.

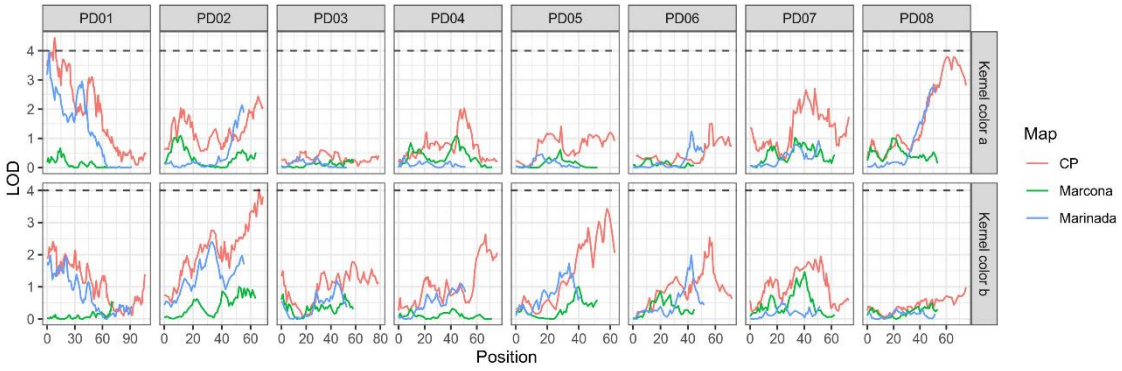


Figure 2.4. QTL mapping of color traits. In columns, the eight almond chromosomes, in rows the different traits. X axis represents the position measured in cM, Y axis represents the LOD value. The horizontal dashed line indicates LOD = 4.

2.3.2. Linkage maps

After the filtering based on segregation errors, imputation and phasing, 9237 SNPs were left. From these SNPs, 3810 were segregating in 'Marcona' (they were heterozygous in 'Marcona' and homozygous in 'Marinada'), 4060 were segregating in 'Marinada' (they were heterozygous in 'Marinada' and homozygous in 'Marcona') and 1367 SNPs were segregating in both 'Marcona' and 'Marinada' (they were heterozygous in both 'Marcona' and 'Marinada'). These SNPs were distributed in 1044 bins (SNPs with identical genotype for all the individuals), 11.3 % of the SNPs used to build the map.

All linkage maps were distributed in eight linkage groups. Marc x Mari CP map had a length of 536 cM (Figure 2.2). The average distance between loci was 0.51 cM/locus. The biggest linkage group, LG1, had a length of 87.3 cM, while the smallest group, LG5, had 59.2 cM. Marcona's map had a length of 478 cM. The average distance between loci was 0.12 cM/locus. The biggest linkage goup, Lg1, had a length of 72.37 cM, while the smallest group, LG6, had 44.47 cM. Marinada's map had a length of 447 cM. The average distance between loci was 0.11 cM/locus. The biggest linkage goup, LG1, had a length of 91.82 cM, while the smallest group, LG5, had 40.82 cM.

2.3.3. QTL mapping

In total, 12 QTLs were mapped for the traits under study (Table 2.2). For kernel weight, two QTLs were detected: qP-KWe4 and qPKWe7, situated in chromosomes four and seven and explaining a 20.6 and 24.7 % of the variance, respectively. For length and length IA a major QTL, qP-KLe1, was mapped. It was situated in chromosome one and explained a 25 % of the variance. For roundness, a major QTL, qP-KRo1, was mapped in chromosome 1, explaining a 29.3 % of the variance. For globosity, a major QTL was mapped, qP-KGlo7, explaining a 26.6 % of the variance and situated in chromosome seven. Crack-out percentage had the biggest QTL mapped in this study, with a LOD of 14.09 and an explained variance of 52.6 % (Figure 2.3). For shoulder and shoulder angle, a major QTL was detected, qP-Sho1, explaining a 29.3 % of the variance. For symmetry jaccard and symmetry SSIM, a major QTL was detected, qP-KSy1, explaining approximately a 26 % of the variance (Figure 2.3).

Regarding color, two QTLs were detected, one for kernel color a, qP-KCoa1, situated in chromosome one and qP-KCob2, for kernel color b, situated in chromosome two (Figure 2.4). For chemical traits, two QTLs were detected, one for protein content and another for margaric acid content. The QTL for protein content, qP-KPro3, was situated in chromosome three and explained a 24.8 % of the variance. The QTL for margaric acid content, qP-MarA1, was situated

in chromosome one and explained a 23.6 % of the variance. However, even if the margaric acid content was the only one with a LOD score higher than 4, other fatty acids such as palmitic acid, stearic acid, oleic acid, vaccenic acid and linolenic acid had a LOD peak in the same region as margaric acid (Figure 2.5).

Table 2.2. QTLs found, indicating the name, trait, Map, chromosome, Top SNP, LOD and variance explained.

QTL name	Trait	Map	Chr	Top SNP	QTL range (cM)	LOD	VE (%)
qP-KWe4	Weight	CP, Marcona	4	SPD04_2563039	0 - 12.66	4.36	20.6
qP-KWe7	Weight	CP	7	SPD07_6526066	0.852-12.05	5.35	24.7
qP-KLe1	Length Length IA	CP, Marcona	1	SPD01_9127114	14.866 - 30.90	5.41	25.0
qP-KRo1	Roundness	CP	1	SPD01_5637925	9.73 - 14.72	6.55	29.3
qP-KGlo7	Globosity	CP	7	SPD07_6449395	5.13 - 9.55	5.85	26.6
qP-KSho1	Shoulder Shoulder angle	CP	1	SPD01_6915000	9.72 - 30.15	6.55	29.3
qP-KSy1	Symmetry jaccard Symmetry SSIM	CP	1	SPD01_9127114	20.46 - 30.15	5.77	26.1
qP-Cro2	Crack-out	CP, Marcona, Marinada	2	SPD02_23393185	59.53 - 61.68	14.09	52.6
qP-KCoa1	Kernel color a	CP	1	SPD01_3861313	0.00 - 11.44	4.44	21.0
qP-KCob2	Kernel color b	CP	2	SPD02_25278518	58.96 - 68.66	4.04	19.2
qP-KPro3	Protein	CP	3	SPD03_21636173	64.00 - 70.17	5.39	24.8
qP-MarA1	Margaric acid	CP	1	SPD01_40336063	96.781 - 107.04	4.22	23.6

2.4. Discussion

2.4.1. Recent advances in almond genomics and bioinformatics tools allowed the construction of a high quality linkage map

The development of the almond 60k SNP array (Duval et al., 2023a) has open the door to perform genome-wide analyses in almond. This cost-effective genotyping tool, combined with a bioinformatics pipeline aimed to detect and correct genotyping errors, have allowed us to build a high quality linkage map. Marc x Mari map was formed by 9,237 SNPs, but only 1044 bins, that indicates a level of saturation never reached in almond linkage maps before (Sánchez-Pérez et al., 2007, 2012; Font i Forcada et al., 2012; Fernández i Martí et al., 2013b). However, the fact that from the 9,237 SNPs forming the linkage map there were only 1044 bins, indicates that the main limiting factor on this linkage map was the number of recombinations and thus, the number of individuals forming Marc x Mari population.

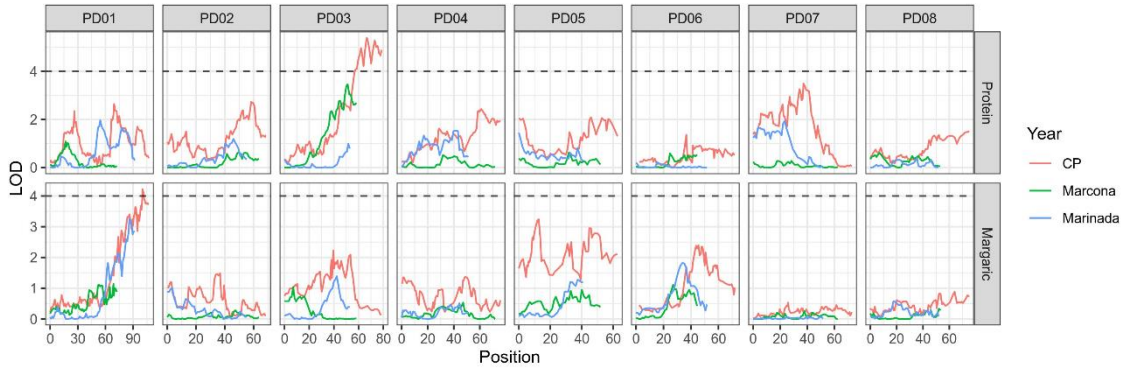


Figure 2.5. QTL mapping of chemical traits. In columns, the eight almond chromosomes, in rows the different traits. X axis represents the position measured in cM, Y axis represents the LOD value. The horizontal dashed line indicates LOD = 4.

2.4.2. Image analysis and lsmean data transformation stand out as an efficient alternatives to conventional phenotyping methods

The highly correlation between traits measured with conventional methods and image analysis methods (kernel length, width and roundness) and the fact that the same QTLs were found indicates the accuracy of image data methods. This should encourage breeders to incorporate these phenotyping protocols into the breeding cycle as they are less time-consuming.
Another important application of image analysis in this study was its use to phenotype kernel symmetry. Its high correlation with the shoulder visual phenotyping and manual analysis of the shoulder angle will allow to phenotype this trait in an efficient way. This will facilitate the development of molecular markers to apply MAS for symmetry, a trait looked for costumers and the almond industry.

Another important innovation carried out in this study was the use of lsmean data instead of raw phenotypic data. This is an approach already mainstream in other trait-loci analyses such as GWAS, but, as far as we know, never applied in QTL mapping before. By comparing models including only the variability in the phenotypic data caused by each individual with models including the year as well, researcher can obtain an indirect measured of the heritability of the trait and also the effect caused by the environment. This information has a high value in this kind of analyses, since the effectiveness of QTL mapping rely on the heritability of the trait under study.

2.4.3. The novel QTLs mapped in this study will allow the implementation of MAS strategies applied to kernel quality traits

Kernel quality has been a major breeding objective for many almond breeding programs worldwide. Several studies have focused on study the inheritance of physical and chemical kernel quality traits. Indeed, three QTLs reported in this study were mapped in previous studies. QP-KWe4, associated to kernel weight and situated in chromosome four, was already mapped in a F1 population coming from the cross 'R1000' x 'Desmayo Largueta' (Sánchez-Pérez et al., 2007). QP-KLe1, associated to kernel length and situated in chromosome one, has been already mapped in a panel of 98 almond cultivars (Font i Forcada et al., 2015a) and in a F1 population coming from the cross 'Vivot' x 'Blanquerna' (Fernández i Martí et al., 2013b). Finally, qP-CRO2, associated to crack-out percentage and situated in chromosome two, have been reported by several studies (Sánchez-Pérez et al., 2007; Goonetilleke et al., 2018; Pavan et al., 2021; Sideli et al., 2023).

The high quality linkage map build, along the QTLs reported in this study, will allow the implementation of efficient MAS strategies applied to kernel quality traits. In total, two QTLs

were mapped for kernel weight, five were mapped for shape-related traits, one for crack-out, two for kernel color and two for kernel chemical traits. Additionally, QTLs for two traits, kernel shoulder and symmetry, were mapped for the first time in almond. Any molecular marker capable of predict these two traits will be extremely useful for breeders, since only symmetrical almonds and with no shoulder are selected.

Only a single QTL was mapped related to fatty acids, qP-MarA1. However, according to our results, the content of this particular fatty acid was closely related with other fatty acids such as myristic acid, palmitic acid, stearic acid, arachidic acid, palmitoleic acid, cis-10-heptadecenoid acid, oleic acid and γ-linoleic acid. This was also confirmed by the similar trend observed in the LOD score of all these fatty acids. Taken together, this indicates that a single molecular marker could predict the content of all these fatty acids. This is of particular interest in the case of the oleic acid, since it is the predominant fatty acid in almond and the responsible of the health benefits associated to it (Becerra-Tomás et al., 2019; Barreca et al., 2020).

2.5. Conclusions

In this study, we carried out a QTL mapping focusing on physical and chemical kernel quality traits in almond. We built a high quality linkage map using the almond 60k SNP array. Additionally, we compared conventional and image analysis phenotyping methods, highlighting the usefulness of image analysis in almond breeding. We encourage breeders to incorporate these phenotyping protocols into the breeding cycle, since they are more efficient. Finally, 12 QTL were mapped, five of them associated to shape-related traits. The use of these QTLs will allow the implementation of MAS strategies applied to kernel quality traits into almond breeding.

2.6. References

Alonso Segura, J. M., Socias i Company, R., & Kodad, O. (2017, October 20). Late-blooming in almond: A controversial objective. *Scientia Horticulturae*. Elsevier B.V. https://doi.org/10.1016/j.scienta.2017.05.036

Barreca, D., Nabavi, S. M., Sureda, A., Rasekhian, M., Raciti, R., Silva, A. S., … Mandalari, G. (2020). Almonds (Prunus Dulcis Mill. D. A. Webb): A Source of Nutrients and Health-Promoting Compounds. *Nutrients 2020, Vol. 12, Page 672*, *12*(3), 672. https://doi.org/10.3390/NU12030672

Batlle, I., Dicenta, F., Gradziel, T. M., Wirthensohn, M., Duval, H., & Vargas, F. J. (2017). Classical genetics and breeding. In *Almonds: Botany, production and uses* (pp. 111–148). Boston: CABI.

Becerra-Tomás, N., Paz-Graniel, I., Kendall, C., Kahleova, H., Rahelić, D., Sievenpiper, J. L., & Salas-Salvadó, J. (2019). Nut consumption and incidence of cardiovascular diseases and cardiovascular disease mortality: A meta-analysis of prospective cohort studies. *Nutrition Reviews*, *77*(10), 691–709. https://doi.org/10.1093/nutrit/nuz042

Di Guardo, M., Farneti, B., Khomenko, I., Modica, G., Mosca, A., Distefano, G., … Gentile, A. (2021). Genetic characterization of an almond germplasm collection and volatilome profiling of raw and roasted kernels. *Horticulture Research*, *8*(1), 27. https://doi.org/10.1038/s41438-021-00465-7

Duval, H., Coindre, E., Ramos-Onsins, S. E., Alexiou, K. G., Rubio-Cabetas, M. J., Martínez-García, P. J., … Arús, P. (2023). Development and Evaluation of an AxiomTM 60K SNP Array for Almond (Prunus dulcis). *Plants*, *12*(2). https://doi.org/10.3390/plants12020242

Fernández i Martí, A., Font i Forcada, C., & Socias i Company, R. (2013a). Genetic analysis for physical nut traits in almond. *Tree Genetics and Genomes*, *9*(2), 455–465. https://doi.org/10.1007/s11295-012-0566-8

Fernández i Martí, A., Font i Forcada, C., & Socias i Company, R. (2013b). Genetic analysis for physical nut traits in almond. *Tree Genetics and Genomes*, *9*(2), 455–465. https://doi.org/10.1007/S11295-012-0566-8/TABLES/5

Font i Forcada, C., i Martí, À. F., & I Company, R. S. (2012). Mapping quantitative trait loci for kernel composition in almond. *BMC Genetics*, *13*(1), 1–9. https://doi.org/10.1186/1471-2156-13-47/TABLES/3

Font i Forcada, C., Oraguzie, N., Reyes-Chin-Wo, S., Espiau, M. T., Company, R. S. I., & Fernández I Martí, A. (2015). Identification of genetic loci associated with quality traits in almond via association mapping. *PLoS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0127656

Gonzalo, M. J., Brewer, M. T., Anderson, C., Sullivan, D., Gray, S., & Van Der Knaap, E. (2009). Tomato fruit shape analysis using morphometric and morphology attributes implemented in tomato analyzer software program. *Journal of the American Society for Horticultural Science*, *134*(1), 77–87. https://doi.org/10.21273/jashs.134.1.77

Goonetilleke, S. N., March, T. J., Wirthensohn, M. G., Arús, P., Walker, A. R., & Mather, D. E. (2018). Genotyping by sequencing in almond: SNP discovery, linkage mapping, and marker design. *G3: Genes, Genomes, Genetics*, *8*(1), 161–172. https://doi.org/10.1534/g3.117.300376

Gradziel, T. M., Curtis, R., & Socias i Company, R. (2017). Production and growing regions. In *Almonds: Botany, production and uses* (pp. 70–86). Boston: CABI.

Jurado-Ruiz, F., Onielfa, C., Dujak, C., Pradas, N., Pérez de los Cobos, F., & Aranzana, M. J. (2023). Shape Analyzer: An application for fruit morphology phenotyping. *Manuscript in Preparation*.

Martínez-Gómez, P., Prudencio, A. S., Gradziel, T. M., & Dicenta, F. (2017, August 1). The delay of flowering time in almond: a review of the combined effect of adaptation, mutation and breeding. *Euphytica*. Springer Netherlands. https://doi.org/10.1007/s10681-017-1974-5

Pavan, S., Delvento, C., Mazzeo, R., Ricciardi, F., Losciale, P., Gaeta, L., … Lotti, C. (2021). Almond diversity and homozygosity define structure, kinship, inbreeding, and linkage disequilibrium in cultivated germplasm, and reveal genomic associations with nut and seed weight. *Horticulture Research*, *8*(1), 1–12. https://doi.org/10.1038/s41438-020-00447-1

Romero, A. (2014). Almond quality requirements for industrial purposes - Its relevance for the future acceptance of new cultivars from breeding programs. In *Acta Horticulturae* (Vol. 1028, pp. 213–220). International Society for Horticultural Science. https://doi.org/10.17660/ActaHortic.2014.1028.34

Sánchez-Pérez, R., Dicenta, F., & Martínez-Gómez, P. (2012). Inheritance of chilling and heat requirements for flowering in almond and QTL analysis. *Tree Genetics and Genomes*, *8*(2), 379–389. https://doi.org/10.1007/s11295-011-0448-5

Sánchez-Pérez, R., Howad, W., Dicenta, F., Arús, P., & Martínez-Gómez, P. (2007). Mapping major genes and quantitative trait loci controlling agronomic traits in almond. *Plant Breeding*, *126*(3), 310–318. https://doi.org/10.1111/j.1439-0523.2007.01329.x

Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Aiese Cigliano, R., Del Cueto, J., … Lindberg Møller, B. (2019). Mutation of a bHLH transcription factor allowed almond domestication. *Science*, *364*(6445), 1095–1098. https://doi.org/10.1126/science.aav8197

Sideli, G. M., Mather, D., Wirthensohn, M., Dicenta, F., Goonetilleke, S. N., Martínez-García, P. J., & Gradziel, T. M. (2023). Genome-wide association analysis and validation with KASP markers for nut and shell traits in almond (Prunus dulcis [Mill.] D.A.Webb). *Tree Genetics & Genomes.*, *19*(2), 13. article. https://doi.org/10.1007/s11295-023-01588-9

Sonneveld, T., Robbins, T. P., Bošković, R., & Tobutt, K. R. (2001). Cloning of six cherry self-incompatibility alleles and development of allele-specific PCR detection. *Theoretical and Applied Genetics*, *102*(6–7), 1046–1055. https://doi.org/10.1007/s001220000525

Whalen, A., Gorjanc, G., & Hickey, J. M. (2020). AlphaFamImpute: High-accuracy imputation in full-sib families from genotype-by-sequencing data. *Bioinformatics*, *36*(15), 4369–4371. https://doi.org/10.1093/bioinformatics/btaa499

Chapter 3:

Almond genetic structure and non-additive GWAS reveal new insights into almond dissemination history and candidate genes for nut traits and blooming time

Almond population genomics and non-additive GWAS reveal new insights into almond dissemination history and candidate genes for nut traits and blooming time

Felipe Pérez de los Cobos[1,2,3], Eva Coindre[4], Naima Dlalah[4], Bénédicte Quilot-Turion[4], Ignasi Batlle[1], Pere Arús[2,3], Iban Eduardo[2,3], Henri Duval[4]

[1]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Mas Bové, Ctra. Reus-El Morell Km 3,8 43120 Constantí Tarragona, Spain

[2]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Centre de Recerca en Agrigenòmica (CRAG), CSIC-IRTA-UAB-UB. Cerdanyola del Vallès (Bellaterra), 08193 Barcelona, Spain

[3]Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Cerdanyola del Vallès (Bellaterra), 08193 Barcelona, Spain

[4]INRAE, GAFL, F-84143, Montfavet, France

Under revision in Horticulture Research

Abstract

Domestication drastically changed crop genomes, fixing alleles of interest and creating different genetic populations. Genome-wide association studies (GWASs) are a powerful tool to detect these alleles of interest (and so QTLs). In this study, we explored the genetic structure as well as additive and non-additive genotype-phenotype associations in a collection of 243 almond accessions. Our genetic structure analysis strongly supported the subdivision of the accessions into five ancestral groups, all formed by accessions with a common origin. One of these groups was formed exclusively by Spanish accessions, while the rest were mainly formed by accessions from China, Italy, France and the USA. These results agree with archaeological and historical evidence that separate modern almond dissemination into four phases: Asiatic, Mediterranean, Californian and southern hemisphere. In total, we found 13 independent QTLs for nut weight, crack-out percentage, double kernels percentage and blooming time. Of the 13 QTLs found, only one had an additive effect. Through candidate gene analysis, we proposed *Prudul26A013473* as a candidate gene responsible for the main QTL found in crack-out percentage, *Prudul26A012082* and *Prudul26A017782* as candidate genes for the QTLs found in double kernels percentage, and *Prudul26A000954* as a candidate gene for the QTL found in blooming time. Our study enhances our knowledge of almond dissemination history and will have a great impact on almond breeding.

3. Chapter 3

3.1. Introduction

One of the landmarks of human history was the transition from nomadic hunter-gatherer societies to settled agriculture-based societies. This transition, known as the Neolithic Revolution, marked the beginning of the domestication of wild plant species as cultivated crops (Weisdorf, 2005). Domestication, and later dispersal and diversification of crops, introduced substantial changes into their genomes, fixing alleles of interest, creating different genetic populations and adapting these groups to different environmental conditions (Doebley et al., 2006; Gross and Olsen, 2010). These changes, accumulated over thousands of years, led to the crops we consume today.

Nowadays, the main actor changing crop genomes is modern breeding. The efficient implementation of breeding strategies requires the correct management of germplasms, optimized genotyping and phenotyping methods, concise knowledge of crop genetic structure and the study of genetic determinism behind traits of interest (Swarup et al., 2021; Thudi et al., 2021). Genome-wide association study (GWAS) is a powerful tool to study quantitative traits in plant breeding. It aims to find polymorphic genetic markers (typically SNPs) significantly associated with phenotypic variation (Korte and Farlow, 2013). However, one of the weaknesses of this technique is that most GWAS models assume that the genotypic variation has an additive effect on the phenotype. This means that non-additive effects, such as dominant-recessive or overdominant interactions, are not included in the models even when they may be relevant for most traits (Tsepilov et al., 2015).

Almond [*Prunus dulcis* Miller (D.A. Webb)] is the most economically important temperate nut tree worldwide. In the period 2011-2021, its production increased 54%, reaching 1,684,395 metric tons of kernel(Council, 2021). It belongs to the *Rosaceae* family and the *Prunus* genus with other important crops including peach, plum, apricot and cherry.

While research on the almond domestication process is in the early stages, some important insights have been made. The most accepted theory to date is that almond originated from hybridizations with several wild relatives somewhere between the Eastern Mediterranean and Southwest Asia, expanding rapidly to Central Asia and the Western Mediterranean. Many studies using different approaches support this theory, from analyses based on morphology, habitat and/or coexistence in cultivated areas (Grassely, 1976; Denisov, 1988a; Ladizinsky, 1999), through genomic analyses (Zeinalabedini et al., 2010b; Delplancke, 2013; Delplancke et al., 2016), to archaeobotanic evidence (Zohary and Hopf, 1993; Willcox et al., 2008b; Pérez-Jordà et al., 2021). In this sense, many efforts have focused on analyzing the population structure of different almond germplasms (Font i Forcada et al., 2015a; Di Guardo et al., 2021; Pavan et al., 2021). Nevertheless, these studies have been limited by the geographical origin of the accessions (most accessions came from the same region) or by the low number of markers used. As a result of these drawbacks, our knowledge of the genetic structure of the cultivated almond is still limited.

The recent publication of three almond reference genomes (Sánchez-Pérez et al., 2019b; Alioto et al., 2020b; D'Amico-Willman et al., 2022a) and the development of a 60K SNP array (Duval et al., 2023b) have opened the door to performing genome-wide analyses on almond. So far, there have been three GWASs using genome-wide marker data (Di Guardo et al., 2021; Pavan et al., 2021; Sideli et al., 2023), and several QTLs linked to shell and kernel quality traits were identified. Nevertheless, these studies only focused on additive GWAS models and the variability of the plant material was reduced. Studying a broader germplasm and analyzing non-additive genotype-phenotype associations would allow the exploration of the origin and

historic dissemination of the cultivated almond at the same time that would help to find alleles of interest and QTLs fixed over thousands of years of domestication.

In this study, we explored the genetic structure and genotype-phenotype associations in a collection of 243 almond accessions from different origins. For this purpose, we first characterized the genetic diversity of the collection using the almond 60K SNP array. Then we carried out a GWAS using additive and non-additive models for different traits, including kernel and nut weight, crack-out percentage, double kernels percentage and blooming time. As far as we know, this is the first non-additive GWAS in Rosaceae species. Using candidate gene analysis, we also proposed candidate genes responsible for the main QTLs found in this analysis.

3.2. Materials and methods

3.2.1. Plant material and genotyping

We used a diversity panel of 243 accessions from 21 countries and five continents (Supplementary Material 3.1). Of the 243 accessions, 161 were maintained in the INRAE collection (43.948611 N, 4.808333 E) and 97 in the IRTA collection (41.170723 N, 1.172942 E), with 78 accessions in common at the two locations. DNA of the 180 accessions from the INRAE and IRTA collections was extracted from leaves according to Antanaviciute et al. 2015. After DNA extraction, samples were genotyped using the 60K SNP array (Duval et al., 2023).

For the remaining 63 accessions, genotype information was obtained from two different sources: 45 resequences were from a previous study (Duval et al., 2023) and resequences from 18 accessions were downloaded from NCBI (Supplementary Material 3.1).

3.2.2. Genotypic data filtering and datasets

SNP calling of samples from the DNA libraries was according to Duval et al. 2023. Only SNPs present in the 60K almond SNP array were selected (60581 SNPs). All the samples were merged, giving a dataset with 243 accessions and 60581 SNPs. These SNPs were filtered following these criteria: i) Call rate per sample higher than 82% ii) Call rate per SNP higher than 90% and iii) Minimum allele frequency (MAF) higher than 5%.

Using this initial dataset (Table 3.1), we calculated the identity-by-state, i.e. the number of SNPs with the same allelic state shared between accessions. Accessions with an identity-by-state higher than 98% were declared clonal groups. In total, 22 clonal groups with two or more accessions were detected. Within each clonal group, the accession with the highest number of SNPs was selected (Supplementary Material 3.1). The remaining accessions were classified as landraces and breeding cultivars based on pedigree information (Supplementary Material 3.1).

From the initial dataset, three more datasets were created for each analysis in this study. For the genetic structure analysis, only the 152 accessions classified as landraces were selected. We also created two more datasets for GWAS, including only phenotyped individuals for nut traits and blooming time, with 79 and 167 accessions respectively. After selecting the accessions, the datasets were filtered again following the same criteria described above. For the datasets used in GWAS, we included two more criteria: iv) SNPs with three genotypic classes v) Minimum genotypic class frequency higher than 5%. The four datasets used in this study are presented in Table 3.1.

3.2.3. Genetic structure analysis

A population structure analysis, an additive kinship, a phylogenetic tree and a principal component analysis (PCA) were used to determine the genetic structure of the 152 accessions classified as landraces. The population structure analysis was performed using the LEA R

package (Frichot et al., 2014). The number of ancestral groups tested were from 1 to 15 with ten repetitions. An accession was considered to belong to an ancestral group when the coefficient of belonging to that specific group was higher than 60%. If an accession did not belong to any ancestral group, it was considered admixed. Additive kinship was estimated with the rrBLUP R package (Endelman, 2011). The phylogenetic tree was built using the unweighted pair group method with arithmetic mean algorithm included in the ape R package (Paradis et al., 2004), and PCA using the factoMineR R package (Lê et al., 2008).

Table 3.1. Description of the four datasets used.

Dataset	Nº accessions	Accessions included	Nº SNPs
Initial	243	All	54,112
Structure	152	Classified as landraces	53,985
Nut traits	79	Phenotyped for nut traits	22,928
Blooming time	167	Phenotyped for blooming time	16,804

3.2.4. Homozygosity analysis

Runs of homozygosity (*ROHs*) were analyzed in all 243 accessions using the detectRUNS R package (https://cran.r-project.org/package=detectRUNS). Two lengths of *ROHs* were analyzed: ROH_2 higher than 4,163,686 bp and $ROH_{0.25}$ higher than 520,461 bp (2% and 0,25% of the almond genome size according to "Texas" reference genome v2.0 (Alioto et al., 2020a), respectively). ROH_2 and $ROH_{0.25}$ were detected using a window size equal to 20 SNPs. The maximum gap between SNPs was equal to 1,000,000 bp for ROH_2 and 100,000 bp for $ROH_{0.25}$. For every accession, the overall inbreeding values F_2 and $F_{0.25}$ were calculated using ROH_2 and $ROH_{0.25}$, respectively. Finally, we calculated the frequencies $Freq_2$ and $Freq_{0.25}$ with which every SNP was located in a ROH_2 and $ROH_{0.25}$, respectively.

3.2.5. Linkage Disequilibrium decay

The squared correlated coefficient, r^2, was estimated in the 152 individuals classified as landraces using VCFTools v0.1.16 (Danecek et al., 2011). As it was calculated individually for every chromosome using a 250,000 bp window, the r^2 was calculated for every combination of SNPs within that window. We used a threshold of 0.2 to set the LD decay which was then represented graphically using a loess regression function with a span of 0.1.

3.2.6. Phenotypic data collection and analysis

As nut traits, we phenotyped nut weight (NW), kernel weight (KW), crack-out percentage (CRO) and double kernels percentage (DK). Each accession was phenotyped between nine to twelve years in the IRTA collection. From each accession, at least 100 mature fruits were randomly collected. The fruit was considered mature when the mesocarp was fully dry and split along the fruit suture and the peduncle was near to complete abscission. Samples were stored at room temperature for at least two weeks. After measuring NW, the shells were cracked to measure the weight of the kernels. All weights were measured using an electronic balance. DK was measuring by counting the number of shells containing double kernels. CRO was calculated according to Equation 1:

$$CRO = (NW - KW)/NW \qquad (1)$$

Blooming time (BLO) was phenotyped for three consecutive years (2020-2022) as Julian days when about 5% of flower buds were fully open for each tree. This trait was measured in the INRAE collection.

Best Linear Unbiased Prediction (BLUP) for NW, KW, CRO and BLO was estimated for each genotype using a linear mixed model according to Equation 2:

$$P_{ijk} = \mu + Y_{ik} + G_j + e_{ijk} \qquad (2)$$

Where P_{ijk} is the phenotypic value (=BLUP) of the kth repetition of the jth genotype in the ith year, μ is the mean value of the phenotypic trait, Y_{ik} is the fixed effect of the kth repetition of the ith year, G_j is the random genotypic effect of genotype j, and e_{ijk} is the residual error of the model.

BLUP for DK was estimated for each genotype using a linear mixed model according to Equation 3:

$$P_{jk} = \mu + G_j + e_{jk} \qquad (3)$$

Where for equation 2, P_{jk} is the phenotypic value of the kth repetition of the jth genotype, μ and G_j have the same meanings as in equation 1, e_{jk} is the residual error.

For every trait, broad-sense heritability (h^2) was estimated as:

$$h^2 = \frac{{\sigma^2}_G}{{\sigma^2}_G + \frac{{\sigma^2}_\varepsilon}{n}}$$

Where ${\sigma^2}_G$ is the genotype variance, ${\sigma^2}_\varepsilon$ is the residual variance and n is the mean number of measures.

3.2.7. Genome-Wide Association Study (GWAS)

We explored additive and non-additive genotype-phenotype associations in two different datasets: nut traits and blooming time datasets, with 79 and 167 accessions respectively. For this purpose, we transformed these genotypic datasets as follow. For additive effects, the three possible genotypes of a biallelic marker with a reference allele (*a1*) and an alternative allele (*a2*), were written in numeric representation as 1 (*a1a1*, homozygous for the reference allele), 0 (*a1a2*, heterozygous) and -1 (*a2a2*, homozygous for the alternative allele). For dominant effects, genotypes *a1a1* and *a1a2* have the same effect in the phenotype, so *a1a1* and *a1a2* were codified as 1 and *a2a2* as -1. For recessive effects, genotypes *a1a2* and *a2a2* have the same effect in the phenotype, so *a1a2* and *a2a2* were codified as -1 and *a1a1* as 1. Note that the dominant and recessive transformations correspond to a dominant-recessive genotype-phenotype interaction, but we had to differentiate the effects of a dominant reference allele or a dominant alternative allele. For overdominant effects, genotypes *a1a1* and *a2a2* have the same effect in the phenotype, so genotypes *a1a1* and *a2a2* were codified as 1 and *a1a2* as 0 (Supplementary Material 3.2) (Tsepilov et al., 2015).

The mixed model from rrBLUP R package(Endelman, 2011) was used in this study. BLUPs were used as phenotypic data for each trait. For every model, we used three different corrections: including the additive kinship (K), the population structure (Q) or both (K+Q):

$$Y = \mu + X\beta + Qv + Zu + \varepsilon \qquad (4)$$

Where Y is the vector of phenotypic values, μ the overall mean, X the allelic state matrix, β the allelic effect of each SNP 4, Q is the structural matrix estimated by the LEA R package, v is an

effect vector estimated by the model and used as a fixed effect, Z is an incidence matrix linking observations to the vector u that is a polygenic random effect with a covariance structure defined by the kinship (K as previously estimated) u~N (0;2KVg), and ε the residual effect.

The choice of each correction was based on the adjustment of the p-values obtained to a uniform distribution as expected under the null hypothesis. The corrected Bonferroni threshold at 5% was used to identify significant association between phenotypic data and genotypic markers.

Before considering any significant genotype-phenotype association found as a QTL, we used visual analysis to confirm that the phenotypic data distribution matched the genotype-phenotype interaction searched (e.g. if an association found with the additive transformation matched an additive phenotypic distribution). We considered any significant genotype-phenotype association matching its phenotypic data distribution as a true positive QTL.

For the QTLs considered as true positives, we assumed the Simple model's R^2 as the variance explained for those QTL. We also calculated the combined variance of the QTLs detected for every trait. In this case, we assumed as the combined variance explained the R^2 of a linear regression using all the top SNPs detected for a trait.

3.2.8. Candidate gene analysis

For every trait, we selected the QTL with the highest -log10(p-value). If the QTL selected had a -log10(p-value) higher than 6.5, we performed a candidate gene analysis. In the case of DK, we used candidate gene analysis on the two QTLs we found, as both had a -log10(p-value) higher than 6.5.

Every QTL region was defined using the position of the top SNP and the estimated LD decay for every chromosome. The beginning of the QTL was defined as the top SNP position minus the estimated LD decay and the end of the QTL was defined as the top SNP position plus the estimated LD decay. We determined the number of genes located in the QTL region using the "Texas" reference genome v2.0 (Alioto et al., 2020a). Any gene located less than 2,000 bp from the QTL region was included in it, as we considered that the regulatory region of that gene was situated within the QTL region. Then we searched the homologous genes from peach (*Prunus persica*) and Arabidopsis (*Arabidopsis thaliana*).

To obtain more information on the function of the most suitable candidate genes, *Prudul26A013473*, *Prudul26A012082*, *Prudul26A017782* and *Prudul26A000954*, we used the PeachGCN v1 (Pérez de los Cobos et al., 2023) for gene coexpression network analysis. We first determined the homologous of our candidate genes in peach, then extracted coexpressing genes. Finally, we performed an enrichment analysis of the coexpressing subnetworks using Gene Ontology (GO) and Mapman ontologies (Ashburner et al., 2000; Thimm et al., 2004). The significance threshold was held at q-value < 0.05. Enriched terms annotating at least 2% of the genes in the coexpressing subnetworks were classified as top enriched terms (TET).

3.3. Results

3.3.1 Almond genetic structure defined five ancestral groups

The results from the population structure analysis, additive kinship, phylogenetic tree and PCA, indicated that the most acceptable genetic structure of the 152 landraces included in this analysis is a model with five ancestral populations (Figure 3.1). According to the population structure analysis (Figure 3.1D), 80 accessions were part of one of the five ancestral groups, while 72 were considered admixed. The number of accessions was homogeneous between

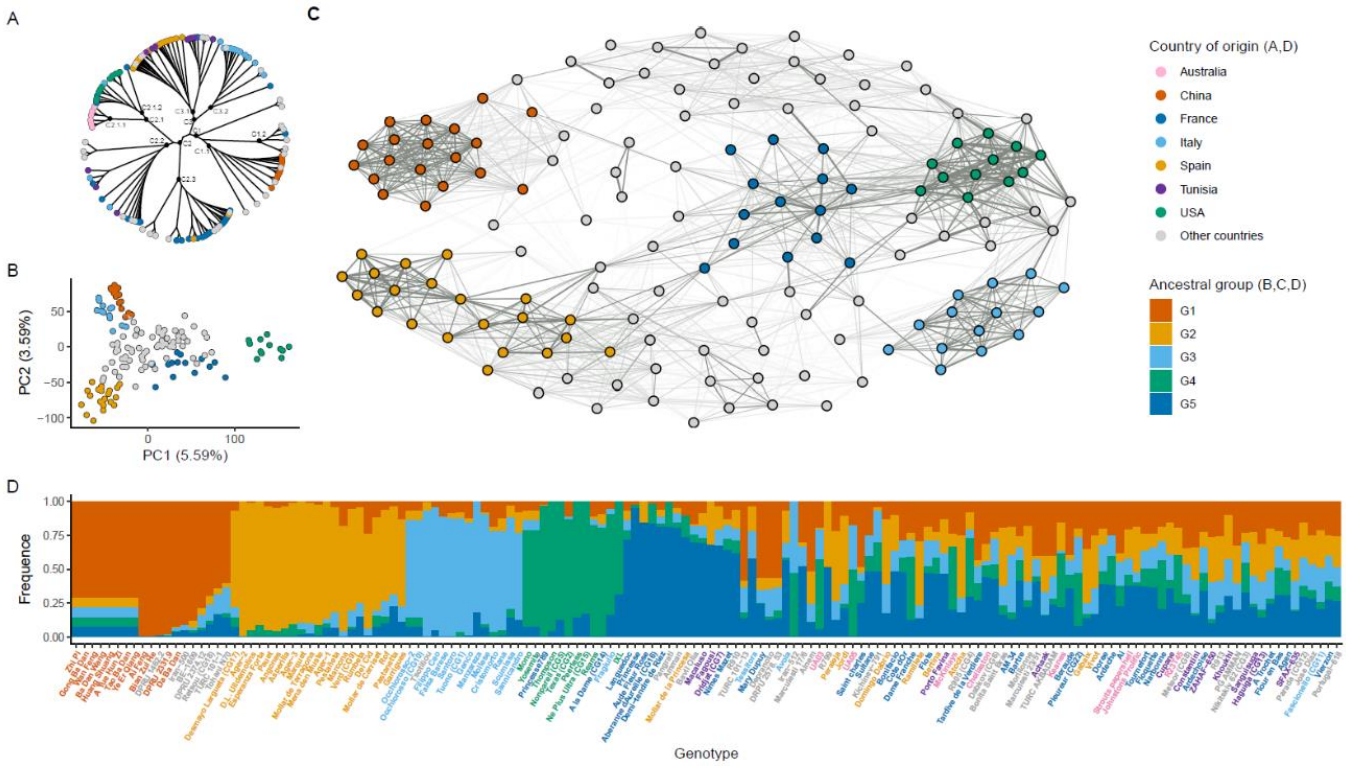


Figure 3.1. Genetic structure analysis. A) Phylogenetic tree. B) Principal components analysis C) Additive kinship. Edges with absolute weight less than 0.05 are not represented. D) Population structure analysis.

groups, ranging from 12 to 21 accessions (G4 and G2 respectively). Ancestral group G1 was mainly composed by Asian accessions: 12 Chinese, two Iranian, two Turkish and one Pakistani. This group also included one Greek and one Romanian accession. Ancestral group G2 was entirely formed by Spanish accessions. Ancestral group G3 was formed by 13 Italian accessions and one Greek. In ancestral group G4, nine out of 12 accessions were from the USA, along with two French and one Italian accession. Ancestral group G5 had six French accessions, three Tunisian, one Greek, one Jordanian, one Iranian, one Moroccan and one Spanish accession.

Additive kinship results showed five dense clusters (Figure 3.1C). These clusters included all the accessions forming the five ancestral groups from the population structure analysis. Clusters formed by G1 and G3 had less connections with other accessions. On the other hand, clusters formed by G2, G4 and G5 accessions had several connections between them and accessions from other countries. G2 accessions, separated in two sub-clusters, were also connected with accessions from Australia and North Africa, among others. G4 and G5 clusters were strongly connected, and the G5 cluster was connected to several accessions from Australia.

The phylogenetic tree had three main clades: C1, C2 and C3 (Figure 3.1A). C1, mainly formed by Asian accessions, was subdivided in two secondary clades, C1.1 and C1.2. All the Chinese accessions were situated in C1.1. C2 was mainly formed by French, American and Australian accessions. American and Australian accessions were found in the same secondary clade, C2.1, but separated in two tertiary clades, C2.1.1 and C2.1.2. French accessions were in two different secondary clades, C2.2 and C2.3, along with some Tunisian, Italian and Spanish accessions. The Spanish, Italian and Tunisian accessions were in C3. And while the Spanish and Tunisian accessions were found in the same secondary clade, C3.1, most of the Italian accessions formed another secondary clade, C3.2.

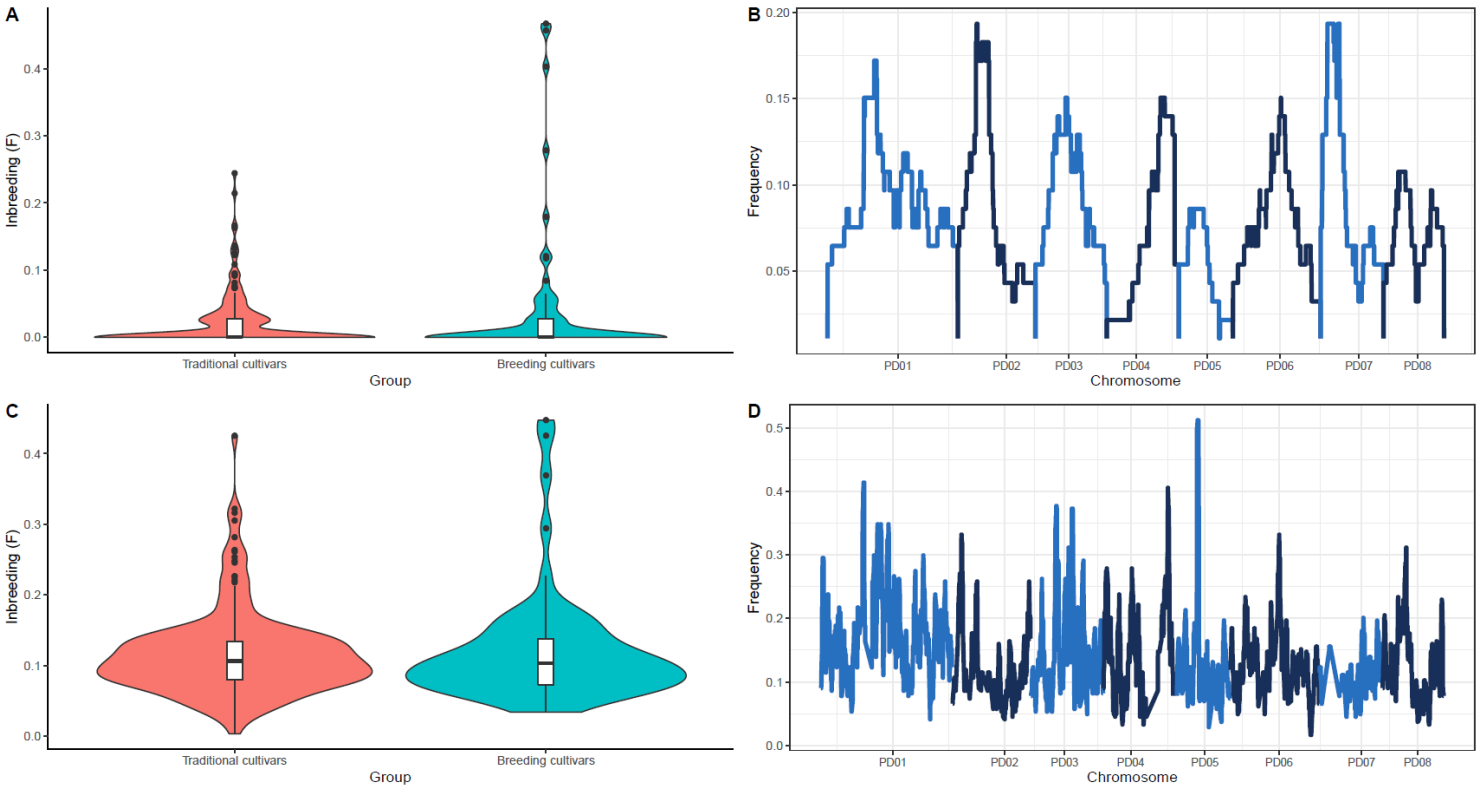


Figure 3.2. Homozygosity analysis. A) Inbreeding using F_2. B) $Freq_2$ results. C) Inbreeding using $F_{0.25}$. D) $Freq_{0.25}$ results.

Based on the Principal Compound Analysis, the ancestral groups were well separated according to the first two principal components, explaining 9.18% of the overall variation (Figure 3.1B). Admixture accessions were mostly in the center of the scatter plot with values close to 0 for the two principal components. Genotypes from G4 group were located to the right of the X axis while G1, G2 and G3 groups were to the left of the X axis and separated along the Y axis.

3.3.2. Only three breeding cultivars had a F2 value over 0.4

Landraces and breeding cultivars had low levels of inbreeding using F_2 or $F_{0.25}$ values (Figure 3.2A and 3.2C). In the case of F_2, used to detect recent inbreeding events, inbreeding was low for both landraces and breeding cultivars. Most of the cultivars had an F_2 value equal to zero, with only nine landraces and seven breeding cultivars exceeding $F_2 = 0.1$.

In the case of breeding cultivars, there were some exceptions with high levels of inbreeding, such as "Ayles", "Amandier rose" and "Garfi", with F_2 values over 0.4. Using $F_{0.25}$ as the inbreeding measure, the inbreeding values of both landraces and breeding cultivars were slightly higher, with a mean $F_{0.25}$ around 0.1 in both cases. Only accessions "Ayles", "Amandier rose" and "DPRU 487-A" had a $F_{0.25}$ value over 0.4.

Comparing $Freq_2$ and $Freq_{0.25}$ there were large differences (Figure 3.2B and 3.2D). No SNP had a $Freq_2$ higher than 0.2, while there were three genomic regions with $Freq_{0.25}$ higher than 0.4, in chromosomes 1, 4 and 5. In the case of PD01 and PD05, the genomic regions with high $Freq_{0.25}$ contained 24 and 27 genes, respectively according to the "Texas" reference genome v2.0(Alioto et al., 2020a). In contrast, the genomic region with high $Freq_{0.25}$ in PD04 contained only one gene, *Prudul26A014015*.

3.3.3. Linkage disequilibrium decay was between 4,259 and 6,904 bp

The mean size of the LD block was 6,061 bp and ranged from 4,259 bp (PD08) to 6,904 bp (PD02) (Supplementary Material 3.3).

3.3.4. All the traits had heritability higher than 0.90

A significant negative correlation was found between NW and CRO (-0.77) (Supplementary Material 3.4). A weaker correlation was also found between NW and KW (0.49). Genetic variance and the variance due to the environment (residual error) were extracted from the linear mixed models for each trait (Table 3.2). Heritability for all traits was higher than 0.90.

Table 3.2. Partition of variance of the traits under study.

	KW	NW	CRO	DK	BLO
Genotype	0.024	2.05	136.62	0.0098	59.91
Error	0.023	0.43	9.27	0.0038	19.59
Number of repetitions	12.24	12.33	12.24	9.45	4.05
Heritability	0.93	0.98	0.99	0.96	0.93

NW: nut weight, CRO: crack-out percentage, DK: double kernels percentage, BLO: blooming time, VE: Variance explained, CVE: Combined variance explained.

Table 3.3. Summary of the QTLs identified, indicating the trait and the QTL, the genetic effect, the correction model used, the closest SNP and its chromosome location, the p-value, the variance explained and the combined variance explained.

Trait	Name	Effect	Correction	Top SNP	Chr	-log10(p-value)	VE	CVE
NW	**qP-NW2**	Recessive	Q	SPD02_19131932	2	6.07	33.71 %	45.74%
	qP-NW3	Overdominant	K+Q	SPD03_1356626	3	6.13	18.77 %	
	qP-NW6	Overdominant	K+Q	SPD06_24556800	6	6.1	9.04 %	
CRO	**qP-CRO6**	Additive	Q	SPD06_28386604	6	7.19	22.66%	71.78%
	qP-CRO5.1	Dominant	Q	SPD05_2529870	5	6	23.22%	
	qP-CRO1.1	Recessive	Q	SPD01_31737488	1	5.75	47.5%	
	qP-CRO2	Recessive	Q	SPD02_19131932	2	8.28	49.13%	
	qP-CRO1.2	Overdominant	K+Q	SPD01_39980467	1	7.76	33.44%	
	qP-CRO3	Overdominant	K+Q	SPD03_1356626	3	7.37	27.9%	
	qP-CRO5.2	Overdominant	K+Q	SPD05_14094224	5	6.23	23.57%	
DK	**qP-DK7.1**	Dominant	K+Q	SPD07_7765332	7	7.46	30.7%	65.87%
	qP-DK7.2	Dominant	K+Q	SPD07_11278816	7	7.54	47.57%	
BLO	**qP-BLO2**	Recessive	Q	SPD02_17409544	2	6.99	16.44%	-

K: kinship; Q: population structure, NW: nut weight, CRO: crack-out percentage, DK: double kernels percentage, BLO: blooming time, VE: Variance explained, CVE: Combined variance explained.

3.3.5. 13 true positive QTLs were found for the traits under study

In total, 19 associations were detected for the traits under study (Figure 3.3). Of these associations, six were considered false positives, since the phenotypic data distribution did not match the genotype-phenotype interaction searched. Therefore, 13 QTLs were considered true

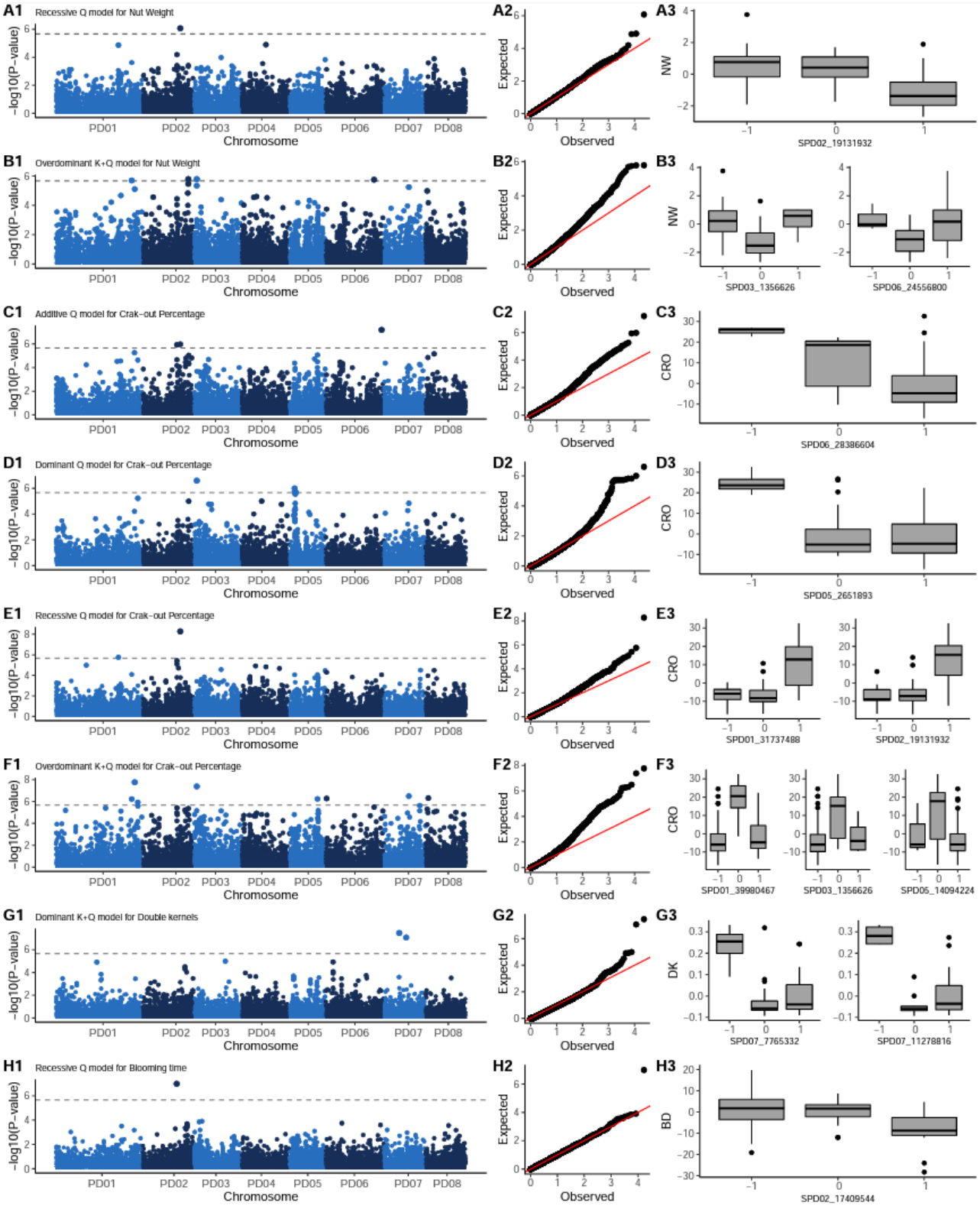


Figure 3.3. Genome-wide association analysis. Row A) Recessive Q model for Nut weight. Row B) Over-dominant K+Q model for Nut weight. Row C) Additive Q model for Crack-out percentage. Row D) Dominant Q model for Crack-out percentage. Row E) Recessive Q model for Crack-out percentage. Row F) Over-dominant K+Q model for Crack-out percentage. Row G) Dominant K+Q model for Double kernels percentage. Row H) Recessive Q model for blooming time. Column 1) Manhattan plots. Column 2) Q-Q plots. Column 3) Boxplots of the true positive QTLs.

positives (Figure 3.3, column 3). These QTLs were named according to the recommendations

for standard QTL nomenclature and reporting of the Genome Database for Rosaceae (Table 3.3).

Within these 13 QTLs, only qP-CRO6 had an additive effect, the rest had non-additive effects on the phenotype. By trait, we found three, seven, two and one QTL for NW, CRO, DK and BLO, respectively. No QTL was found for KW. NW and CRO had QTLs located at the same position: qP-NW2 and qP-CRO2 had SPD02_19131932 as top SNP with a recessive effect, while qP-NW3 and qP-CRO3 had SPD03_1356626 as top SNP with an overdominant effect. The variance explained by the QTLs ranged from 9 % (qP-NW6) to 49 % (qP-CRO2). The combined variance explained for all the QTLs was 45.74% for NW, 71.78% for CRO, 65.87% for DK and the only QTL found for BLO explained 16.44% of the variance.

3.3.6. Candidate genes controlling crack-out percentage, double kernels percentage and blooming time

The length of the qP-CRO2 region was 13.8kb, containing just one gene (Table 3.4). This gene, *Prudul26A013473*, was annotated as a NAC domain-containing protein.

The length of the qP-DK7.1 region was 11.5kb, containing two genes (Table 3.4) encoding a eukaryotic translation initiation factor 3 subunit E and an alpha-carbonic anhydrase domain-containing protein.

The length of qP-DK7.2 was 11.5kb, containing four genes (Table 3.4) encoding an ML domain-containing protein, a ribosomal L6 domain-containing protein and two uncharacterized proteins.

The length of qP-BLO2 was 13.8kb, containing three genes (Table 3.4) encoding a cyclic nucleotide-binding domain-containing protein, a beta-amylase and an uncharacterized protein.

Table 3.4. List of candidate genes.

QTL	Candidate genes	*Prunus Persica* homolog	*Arabidopsis thaliana* homolog	Uniprot function prediction
qP-CRO2	*Prudul26A013473*	*Prupe.2G196600*	*AT3G61910* *AT2G46770*	NAC domain-containing protein
qP-DK7.1	*Prudul26A012082*	*Prupe.7G052700*	*AT3G57290*	Eukaryotic translation initiation factor 3 subunit E
	Prudul26A002885	*Prupe.7G052800*	*AT4G21000* *AT4G20990*	Alpha-carbonic anhydrase domain-containing protein
qP-DK7.2	*Prudul26A005959*	*Prupe.7G092000*	*AT5G06470*	Uncharacterized protein
	Prudul26A029836	*Prupe.7G092100*	*AT5G06480* *AT3G11780*	ML domain-containing protein
	Prudul26A017782	*Prupe.7G092200*	*AT2G18400*	Ribosomal_L6 domain-containing protein
	Prudul26A008330	*Prupe.7G092300*	-	Uncharacterized protein
qP-BLO2	*Prudul26A000954*	*Prupe.2G169700*	*AT4G00520* *AT1G01710*	Cyclic nucleotide-binding domain-containing protein
	Prudul26A028547	*Prupe.2G169800*	*AT3G49050*	Uncharacterized protein
	Prudul26A019171	*Prupe.2G169900*	*AT2G45880*	Beta-amylase

The gene coexpression network (GCN) analysis of *Prupe.2G196600*, the homologous gene of *Prudul26A013473*, indicated 25 enriched terms, with nine classified as top enriched terms (TET) (Supplementary Material 3.5; Figure 3.4). Within GObp, we found two TETs: '*plant-type secondary cell wall biogenesis'* and *'cell wall organization'*. Within GOmf, we found two TETs:

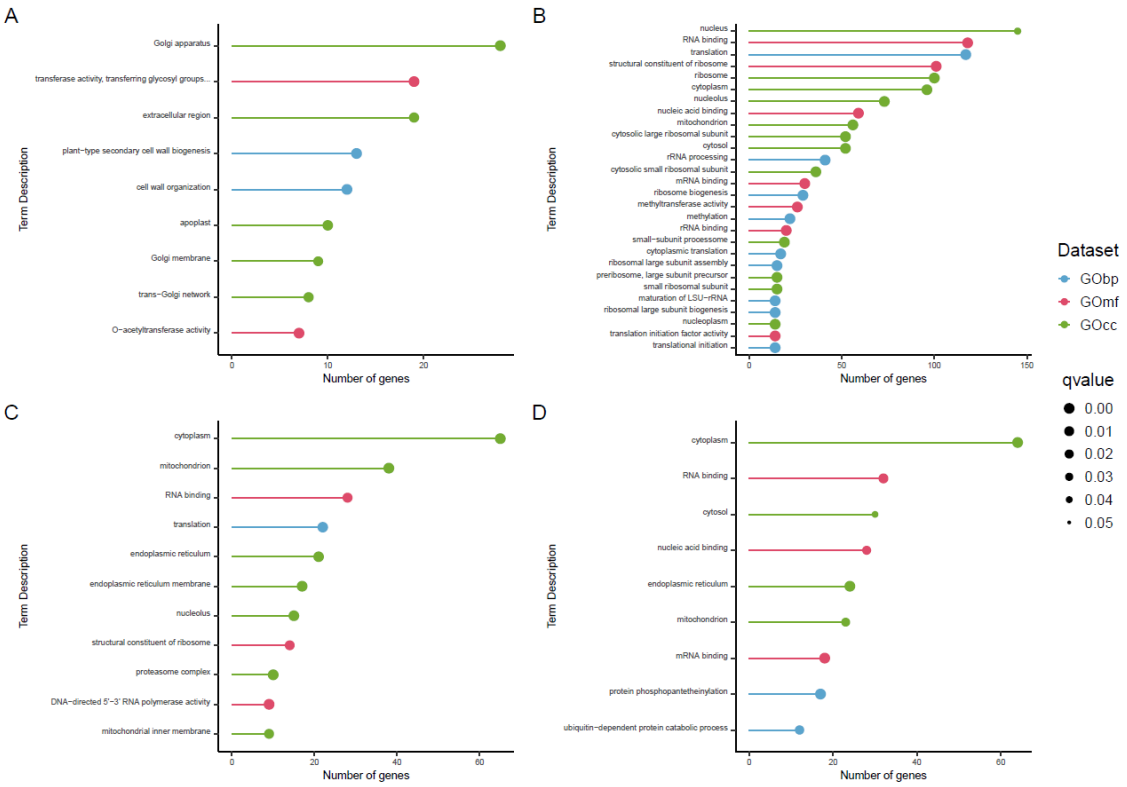


Figure 3.4. Top enriched terms of the most suitable candidate genes. A) *Prupe.2G196600* B) *Prupe.2G169700* C) *Prupe.7G052700* D) *Prupe.7G092200*.

'*transferase activity, transferring glycosyl groups*' and '*O-acetyltransferase activity*'. Within GOcc, we found five TETs: '*Golgi apparatus*', '*extracellular region*', '*apoplast*', '*Golgi membrane*' and '*trans-Golgi network*'.

The GCN analysis of *Prupe.7G052700*, the homologous gene of *Prudul26A012082*, indicated 192 enriched terms, and 28 were classified as TETs (Supplementary Material 3.5; Figure 3.4B). Within GObp, we found nine TETs: '*translation*', '*rRNA processing*', '*ribosome biogenesis*', '*methylation*', '*cytoplasmic translation*', '*ribosomal large subunit assembly*', '*maturation of LSU-rRNA*', '*ribosomal large subunit biogenesis*' and '*translational initiation*'. Within GOmf, we found seven TETs: '*RNA binding*', '*structural constituent of ribosome*', '*nucleic acid binding*', '*mRNA binding*', '*methyltransferase activity*', '*rRNA binding*' and '*translation initiation factor activity*'. Within GOcc, we found 12 TETs: '*nucleus*', '*ribosome*', '*cytoplasm*', '*nucleolus*', '*mitochondrion*', '*cytosolic large ribosomal subunit*', '*cytosol*', '*cytosolic small ribosomal subunit*', 'small-*subunit processome*', '*preribosome large subunit precursor*', '*small ribosomal subunit*' and '*nucleoplasm*'.

The GCN analysis of *Prupe.7G092200*, the homologous gene of *Prudul26A017782*, indicated 161 enriched terms, and 11 were classified as TETs (Supplementary Material 3.5; Figure 3.4C). Within GObp, we found one TET: '*translation*'. Within GOmf, we found three TETs: '*RNA binding*', '*structural constituent of ribosome*' and '*DNA-directed 5'-3' RNA polymerase activity*'. Within GOcc, we found seven TETs: '*cytoplasm*', '*mitochondrion*', '*endoplasmic reticulum*', '*endoplasmic reticulum membrane*', '*nucleolus*', '*proteasome complex*' and '*mitochondrial inner membrane*'.

The GCN analysis of *Prupe.2G169700*, the homologous gene of *Prudul26A000954*, indicated 32 enriched terms, with nine classified as TETs (Supplementary Material 3.5; Figure 3.4D). Within GObp, we found two TETs: '*protein phosphopantetheinylation*' and '*ubiquitin-dependent protein*

catabolic process'. Within GOmf, we found three TETs: '*RNA binding', 'nucleic acid binding'* and *'mRNA binding'*. Within GOcc, we found four TETs: '*cytoplasm', 'cytosol', 'endoplasmic reticulum'* and *'mitochondrion'*.

3.4.Discussion

3.4.1. Almond genetic structure explains its historical worldwide dissemination

Our results strongly supported the subdivision of these accessions into five ancestral groups. Each group was formed by accessions with a common geographical origin: G2 exclusively by Spanish accessions, while G1, G3, G4 and G5 were mainly formed by Chinese, Italian, American and French accessions, respectively. These results are in agreement with Pavan et al. 2021, who, with a more limited germplasm, found four ancestral groups, each formed mainly by accessions from Spain, France, Italy and the USA.

Apart from Chinese accessions, G1 included accessions from Pakistan, Iran, Turkey, Greece, and Romania. Due to the diverse origin of the accessions forming this group, they may be considered as part of a more primitive almond pool. This agrees with the first dissemination of almond from its center of origin, spreading throughout south-western Asia, eventually reaching modern Turkey, Greece and other regions of the Eastern Mediterranean (Gradziel and Socias i Company, 2017).

After reaching the Eastern Mediterranean, Greeks and Phoenicians introduced the almond to other adapted areas of the Mediterranean. By the time of the Roman Empire, almond cultivation had spread all along the Mediterranean coast. Ancestral groups G2, G3 and G5 may have been established during this period.

G2 was entirely formed by Spanish accessions. The fact that the Spanish accessions were related to Tunisian and other North African accessions could be explained by the introduction of new North African genetic material during the Arab occupation of the Iberian Peninsula.

Even though almond cultivation was introduced in California by the Spanish missions, our results showed a close genetic relationship between Californian and French accessions. This was noticeable in both phylogenetic tree, kinship and population structure analysis. In the phylogenetic tree, all Californian accessions were located in only one branch, clustering together with the French accessions. In the kinship analysis, both G4 and G5 clusters were connected by several edges, being the clusters with most connections between each other, indicating a relatively recent common ancestor between these two groups. Finally, the population structure analysis included two French accessions, "Princesse789" and "A la Dame (CG14)" within group G4. This genetic relatedness between French and Californian accessions is explained by the introduction of French commercial cultivars in California from 1850 to 1900 (Wood, 1925b). In this sense, "Princesse789" and "A la Dame (CG14)" may be among the cultivars introduced in California during that period, or at least, close relatives.

G5 formation also included three Tunisian accessions. This close genetic relatedness may be explained by the exchange of material between the two countries during the French occupation of Tunisia from 1881 to 1956.

Finally, Californian cultivars were introduced to adapted regions of the southern hemisphere, including Chile, Argentina, South Africa and Australia. This explains the genetic relatedness between Californian and Australian accessions.

3.4.2. The cultivated almond shows signs of inbreeding and domestication

Homozygosity analysis was used to study the inbreeding levels in the studied germplasm. Different *ROH* lengths allowed us to detect modern (ROH_2) and ancient ($ROH_{0.25}$) inbreeding events. In general, all accessions had a low $F_{0.25}$ and F_2 value, regardless of whether they were classified as landraces or breeding cultivars. Nevertheless, there were some breeding cultivars with an F_2 value over 0.4, such as "Ayles", "Amandier rose" and "Garfi". These results are in agreement with Pérez de los Cobos et al. 2021, who concluded that breeding practices could be increasing inbreeding levels.

Looking at $Freq_{0.25}$ and $Freq_2$, we found no region with high $Freq_2$. On the other hand, there were several regions with a high $Freq_{0.25}$, in chromosomes 1, 4 and 5. This may respond to selection pressure during the almond domestication process, fixing in the genome alleles of interest. In the case of chromosome 4, the peak of high $Freq_{0.25}$ was formed only by one gene, *Prudul26A014015*. According to phylomeDB, *Prudul26A014015* had two homologous genes in Arabidopsis, *AT1G21390* and *AT1G76980*. *AT1G21390* was annotated as *EMBRYO DEFECTIVE* (*EMB*) *2170*, while *AT1G76980* was annotated as patatin-like phospholipase domain protein. *EMB* genes are related with embryo development in Arabidopsis (Meinke, 2020). Since the edible commercial product of the almond is the seed, selective pressure may be exerted on a gene related with embryo and seed development.

3.4.3. Non-additive GWAS, short LD decay and GCN analysis allowed the discovery of several candidate genes for breeding traits

Among the 13 QTLs detected in this study, only one had an additive effect. This indicates that non-additive effects could be the main source of genotype-phenotype interactions in almond. Furthermore, due to the similarity between *Prunus* genus genomes, this phenomenon could be repeated in other cross-pollinating *Prunus* species. In peach, a self-compatible *Prunus* species that still maintains an important level of heterozygosity, this may have predominantly led to the fixation of dominant/recessive QTLs and to the selection for heterozygosis in those that are overdominant.

Another remarkable aspect was that most of the QTLs detected in this study were defined by only one SNP (the top SNP). This is caused by the short LD decay found in the population, affecting this study in two different ways: first, it has facilitated the search for candidate genes, since the genomic regions associated with the traits of interest were small and only a few genes were found in these regions (in the case of qP-CRO2, only one gene was found in the QTL region). Second, this could have limited the detection of some regions of interest. In genomic regions with a lower concentration of SNPs, some regions of interest might have been lost because the distance between SNPs was greater than the LD decay. This could be the case for KW, were no QTL was found.

For the rest of the traits under study, we found several QTLs that could be used in marker assisted selection. However, the mathematical models used to calculate the combined variance explained by these QTLs did not take into account possible epistatic interactions between the QTLs. More research is needed before these QTLs can be used in marker assisted selection.

Among the 13 QTLs detected in this study, only qP-CRO2 has been reported in other studies (Sánchez-Pérez et al., 2007; Goonetilleke et al., 2018; Pavan et al., 2021; Sideli et al., 2023). Only one gene was found in the qP-CRO2 region, *Prudul26A013473*. This gene, annotated as a NAC TF, was homologous to *NST1* in Arabidopsis. *NST1* has been reported as a key regulator of the formation of secondary cell walls in woody tissues (Zhong et al., 2008) and specifically associated with secondary cell wall formation within the enb layer in Arabidopsis seeds (Mitsuda et al., 2007). Furthermore, the GCN analysis of *Prupe.2G196600*, the homologous

gene of *Prudul26A013473*, gave several enriched terms related to secondary cell wall biogenesis and organization. All this evidence indicates that *Prudul26A013473* is the gene responsible for qP-CRO2, having a major role in the transcriptional regulation of the almond endocarp lignification.

For double kernels percentage, we found two different QTLs, qP-DK7.1 and qP-DK7.2. In almond, the presence of double kernels in a shell is due to the development and fertilization of two ovules in the ovary when the secondary ovule does not degenerate (Gradziel and Martínez-Gómez, 2002). In the case of qP-DK7.1, only two candidate genes were found in the QTL region, *Prudul26A012082* and *Prudul26A002885*. These genes were annotated as a eukaryotic translation initiation factor 3 subunit E (*eIF3e*) and alpha-carbonic anhydrase domain-containing protein, respectively. *eIF3e* has been reported as essential for embryo development and normal plant cell growth in Arabidopsis (Yahalom et al., 2008; Xia et al., 2010). The GCN analysis of *Prupe.7G052700*, the homologous gene of *Prudul26A012082*, gave several terms related to embryo development (Supplementary Material 3.5). This indicates that *Prudul26A012082* is the gene responsible for qP-DK7.1, having a major role in almond ovule and embryo development.

There were four candidate genes in qP-DK7.2 region, *Prudul26A029836*, *Prudul26A017782*, *Prudul26A005959* and *Prudul26A008330*. These genes were annotated as a ML domain-containing protein, a ribosomal large subunit 6 (*RL6*) and two uncharacterized proteins, respectively. It has been suggested that *RL6*, among other ribosomal subunits, is essential for embryogenesis in Arabidopsis (Romani et al., 2012). The GCN analysis of *Prupe.7G052700*, homologous gene of *Prudul26A012082*, showed several terms related to cell cycle regulation and cell development (Supplementary Material 3.5): *Prudul26A017782* is most likely the gene responsible for qP-DK7.2, having a major role in almond ovule and embryo development.

Within qP-BLO2, three genes were found, *Prudul26A000954*, *Prudul26A028547* and *Prudul26A019171*. These genes were annotated as a cyclic nucleotide-binding domain-containing protein, an uncharacterized protein and a beta-amylase. Both *AT1G01710* and *AT4G00520*, the homologous genes of *Prudul26A000954* in Arabidopsis, were annotated as Acyl-CoA thioesterases. These proteins catalyze the hydrolysis of acyl-CoAs to free fatty acids and coenzyme A. During dormancy breaking in perennial fruit trees, reactive oxygen species (ROS) are produced. One of the pathways that produces these ROS starts from fatty acids, beta-oxidated to monosaccharides and these monosaccharides produce ROS via mitochondrial respiration or are oxidized via the Pentose Phosphate Pathway (Beauvieux et al., 2018). The GCN analysis of *Prupe.2G169700*, the homologous gene of *Prudul26A000954*, gave several terms related with SWI/SNF complex and BAF60 (Supplementary Material 3.5). SWI/SNF complexes have been shown to participate in the control of flower development and blooming time (Reyes, 2014). BAF60 is a SWI/SNF subunit, and induces a change at the high-order chromatin level, repressing the photoperiod flowering pathway in Arabidopsis (Jégu et al., 2014). *Prudul26A000954* therefore appears to be the gene responsible for qP-BLO2, having a major role in blooming time through fatty acids metabolism.

3.5. Conclusions

In this study, we carried out genetic structure analysis and non-additive GWAS in a set of 243 almond accessions. Our genetic results agreed with the archaeological and historical evidence that separate modern almond dissemination into four phases: Asiatic, Mediterranean, Californian and southern hemisphere. Of the 13 QTLs found for the traits of interest, only one had an additive effect, suggesting that non-additive effects could be the major source of genotype-phenotype interactions in almond and other *Prunus* species. Based on the fast LD

decay and the use of the peach GCN we propose four candidate genes for the main QTLs found in this study.

3.6. References

Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., et al. (2020a). Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *Plant J.* 101, 455–472. doi: 10.1111/tpj.14538.

Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., et al. (2020b). Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *Plant J.* 101, 455–472. doi: 10.1111/tpj.14538.

Antanaviciute, L., Harrison, N., Battey, N. H., and Harrison, R. J. (2015). An inexpensive and rapid genomic DNA extraction protocol for rosaceous species. *J. Hortic. Sci. Biotechnol.* 90, 427–432. doi: 10.1080/14620316.2015.11513205.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet. 2000 251* 25, 25–29. doi: 10.1038/75556.

Beauvieux, R., Wenden, B., and Dirlewanger, E. (2018). Bud Dormancy in Perennial Fruit Tree Species: A Pivotal Role for Oxidative Cues. *Front. plant Sci.* 9, 657. doi: 10.3389/fpls.2018.00657.

Council, I. N. and D. F. (2021). Nuts & dried fruit statistical yearbook 2020/2021.

D'Amico-Willman, K. M., Ouma, W. Z., Meulia, T., Sideli, G. M., Gradziel, T. M., and Fresnedo-Ramírez, J. (2022). Whole-genome sequence and methylome profiling of the almond [Prunus dulcis (Mill.) D.A. Webb] cultivar "Nonpareil." *G3 Genes, Genomes, Genet.* 12. doi: 10.1093/g3journal/jkac065.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/BIOINFORMATICS/BTR330.

Delplancke, M. (2013). Evolutionary history of almond tree domestication in the Mediterranean basin. *Mol. Ecol.* 22, 1092. doi: 10.1111/mec.12129.

Delplancke, M., Yazbek, M., Arrigo, N., Espíndola, A., Joly, H., and Alvarez, N. (2016). Combining conservative and variable markers to infer the evolutionary history of Prunus subgen. Amygdalus s.l. under domestication. *Genet. Resour. Crop Evol.* 63, 221–234. doi: 10.1007/s10722-015-0242-6.

Denisov, V. (1988). Almond genetic resources in the USSR and their use in production and breeding. *Acta Hortic.* 224, 299–306.

Di Guardo, M., Farneti, B., Khomenko, I., Modica, G., Mosca, A., Distefano, G., et al. (2021). Genetic characterization of an almond germplasm collection and volatilome profiling of raw and roasted kernels. *Hortic. Res.* 8, 27. doi: 10.1038/s41438-021-00465-7.

Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The Molecular Genetics of Crop Domestication. *Cell* 127, 1309–1321. doi: 10.1016/J.CELL.2006.12.006.

Duval, H., Coindre, E., Ramos-Onsins, S. E., Alexiou, K. G., Rubio-Cabetas, M. J., Martínez-García, P. J., et al. (2023). Development and Evaluation of an AxiomTM 60K SNP Array for Almond (Prunus dulcis). *Plants 2023, Vol. 12, Page 242* 12, 242. doi: 10.3390/PLANTS12020242.

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024.

Font i Forcada, C., Oraguzie, N., Reyes-Chin-Wo, S., Espiau, M. T., Company, R. S. I., and Fernández I Martí, A. (2015). Identification of genetic loci associated with quality traits in almond via association mapping. *PLoS One* 10. doi: 10.1371/journal.pone.0127656.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196, 973–983. doi: 10.1534/genetics.113.160572.

Goonetilleke, S. N., March, T. J., Wirthensohn, M. G., Arús, P., Walker, A. R., and Mather, D. E. (2018). Genotyping by sequencing in almond: SNP discovery, linkage mapping, and marker design. *G3 Genes, Genomes, Genet.* 8, 161–172. doi: 10.1534/g3.117.300376.

Gradziel, T. M., and Martínez-Gómez, P. (2002). Shell seal breakdown in almond is associated with the site of secondary ovule abortion. *J. Am. Soc. Hortic. Sci.* 127, 69–74. doi: 10.21273/jashs.127.1.69.

Gradziel, T. M., and Socias i Company, R. (2017). *Almonds: Botany, production and uses*.

Grassely, C. (1976). Origine et évolution de l'amandier cultivé. *Options Méditerr* 32, 45–49.

Gross, B. L., and Olsen, K. M. (2010). Genetic perspectives on crop domestication. *Trends Plant Sci.* 15, 529–537. doi: 10.1016/J.TPLANTS.2010.05.008.

Jégu, T., Latrasse, D., Delarue, M., Hirt, H., Domenichini, S., Ariel, F., et al. (2014). The BAF60 Subunit of

the SWI/SNF Chromatin-Remodeling Complex Directly Controls the Formation of a Gene Loop at FLOWERING LOCUS C in Arabidopsis. *Plant Cell* 26, 538–551. doi: 10.1105/tpc.113.114454.

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* 9, 1–9. doi: 10.1186/1746-4811-9-29/FIGURES/4.

Ladizinsky, G. (1999). On the origin of almond. *Genet. Resour. Crop Evol.* 46, 143–147. doi: 10.1023/A:1008690409554.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* 25, 1–18. doi: 10.18637/JSS.V025.I01.

Meinke, D. (2020). Genome-wide identification of EMBRYO. *New Phytol.* 226, 306. doi: 10.1111/nph.16071.

Mitsuda, N., Iwase, A., Yamamoto, H., Yoshida, M., Seki, M., Shinozaki, K., et al. (2007). NAC Transcription Factors, NST1 and NST3, Are Key Regulators of the Formation of Secondary Walls in Woody Tissues of Arabidopsis. *Plant Cell* 19, 270–280. doi: 10.1105/tpc.106.047043.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/BIOINFORMATICS/BTG412.

Pavan, S., Delvento, C., Mazzeo, R., Ricciardi, F., Losciale, P., Gaeta, L., et al. (2021). Almond diversity and homozygosity define structure, kinship, inbreeding, and linkage disequilibrium in cultivated germplasm, and reveal genomic associations with nut and seed weight. *Hortic. Res.* 8, 1–12. doi: 10.1038/s41438-020-00447-1.

Pérez-Jordà, G., Alonso, N., Rovira, N., Figueiral, I., López-Reyes, D., Marinval, P., et al. (2021). The Emergence of Arboriculture in the 1st Millennium BC along the Mediterranean's "Far West." *Agronomy.* 11, 902. doi: 10.3390/agronomy11050902.

Pérez de los Cobos, F., García-Gómez, B. E., Orduña-Rubio, L., Batlle, I., Arús, P., Matus, J. T., et al. (2023). First large-scale peach gene coexpression network: A new tool for predicting gene function. *biorxiv.org*. doi: 10.1101/2023.06.22.546058.

Pérez de los Cobos, F., Martínez-García, P. J., Romero, A., Miarnau, X., Eduardo, I., Howad, W., et al. (2021). Pedigree analysis of 220 almond genotypes reveals two world mainstream breeding lines based on only three different cultivars. *Hortic. Res.* 8. doi: 10.1038/S41438-020-00444-4/42041731/41438_2020_ARTICLE_444.PDF.

Reyes, J. C. (2014). The many faces of plant SWI/SNF complex. *Mol. plant.* 7, 454–458. doi: 10.1093/mp/sst147.

Romani, I., Tadini, L., Rossi, F., Masiero, S., Pribil, M., Jahns, P., et al. (2012). Versatile roles of Arabidopsis plastid ribosomal proteins in plant growth and development. *The Plant journal.* 72, 922–934. doi: 10.1111/tpj.12000.

Sánchez-Pérez, R., Howad, W., Dicenta, F., Arús, P., and Martínez-Gómez, P. (2007). Mapping major genes and quantitative trait loci controlling agronomic traits in almond. *Plant Breed.* 126, 310–318. doi: 10.1111/j.1439-0523.2007.01329.x.

Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Cigliano, R. A., Cueto, J. Del, et al. (2019). Mutation of a bHLH transcription factor allowed almond domestication. *Plant Sci.*, 1095–1998. doi: 10.1126/science.aav8197.

Sideli, G. M., Mather, D., Wirthensohn, M., Dicenta, F., Goonetilleke, S. N., Martínez-García, P. J., et al. (2023). Genome-wide association analysis and validation with KASP markers for nut and shell traits in almond (Prunus dulcis [Mill.] D.A.Webb). *Tree Genet. genomes.* 19, 13. doi: 10.1007/s11295-023-01588-9.

Swarup, S., Cargill, E. J., Crosby, K., Flagel, L., Kniskern, J., and Glenn, K. C. (2021). Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci.* 61, 839–852. doi: 10.1002/CSC2.20377.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939. doi: 10.1111/J.1365-313X.2004.02016.X.

Thudi, M., Palakurthi, R., Schnable, J. C., Chitikineni, A., Dreisigacker, S., Mace, E., et al. (2021). Genomic resources in plant breeding for sustainable agriculture. *J. Plant Physiol.* 257, 153351. doi: 10.1016/J.JPLPH.2020.153351.

Tsepilov, Y. A., Shin, S. Y., Soranzo, N., Spector, T. D., Prehn, C., Adamski, J., et al. (2015). Nonadditive effects of genes in human metabolomics. *Genetics* 200, 707–718. doi: 10.1534/genetics.115.175760.

Weisdorf, J. L. (2005). From Foraging To Farming: Explaining The Neolithic Revolution. *J. Econ. Surv.* 19, 561–586. doi: 10.1111/J.0950-0804.2005.00259.X.

Willcox, G., Fornite, S., and Herveux, L. (2008). Early Holocene cultivation before domestication in

northern Syria. *Veg. Hist. archaeobotany* 17, 313–325. doi: 10.1007/s00334-007-0121-y.

Wood, M. N. (1925). *Almond varieties in the United States*.

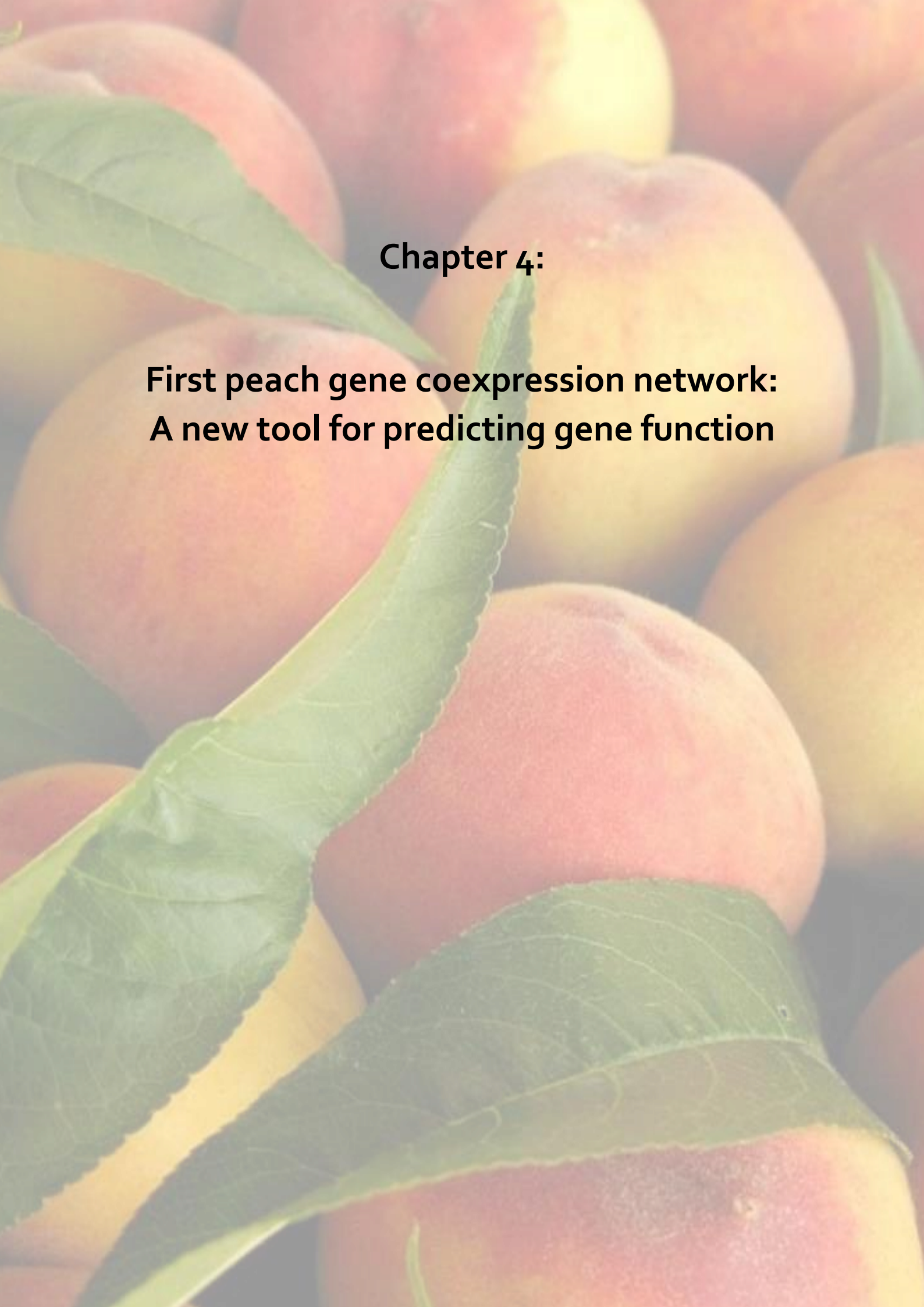
Xia, C., Wang, Y. J., Li, W. Q., Chen, Y. R., Deng, Y., Zhang, X. Q., et al. (2010). The Arabidopsis eukaryotic translation initiation factor 3, subunit F (AteIF3f), is required for pollen germination and embryogenesis. *Plant J.* 63, 189–202. doi: 10.1111/J.1365-313X.2010.04237.X.

Yahalom, A., Kim, T.-H., Roy, B., Singer, R., Von Arnim, A. G., and Chamovitz, D. A. (2008). Arabidopsis eIF3e is regulated by the COP9 signalosome and has an impact on development and protein translation. *The Plant journal.* 53, 300–311. doi: 10.1111/j.1365-313X.2007.03347.x.

Zeinalabedini, M., Khayam-Nekoui, M., Grigorian, V., Gradziel, T., and Martínez-Gómez, P. (2010). The origin and dissemination of the cultivated almond as determined by nuclear and chloroplast SSR marker analysis. *Sci. Hortic. (Amsterdam).* 125, 593–601. doi: 10.1016/j.scienta.2010.05.007.

Zhong, R., Lee, C., Zhou, J., McCarthy, R. L., and Ye, Z. H. (2008). A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. *Plant Cell* 20, 2763–2782. doi: 10.1105/tpc.108.061325.

Zohary, D., and Hopf, M. (1993). *Domestication of plants in the old world*. 4th Editio. Oxford, UK: Clarendon Press.



Chapter 4:

First peach gene coexpression network: A new tool for predicting gene function

First peach large-scale gene coexpression network: a new tool for predicting gene function

Felipe Pérez de los Cobos[1,2,3], Beatriz E. García-Gómez[2,3], Luis Orduña-Rubio[4], Ignasi Batlle[1], Pere Arús[2,3], José Tomás Matus[4], Iban Eduardo[2,3]

[1]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Mas Bové, Ctra. Reus-El Morell Km 3,8 43120 Constantí Tarragona, Spain

[2]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Centre de Recerca en Agrigenòmica (CRAG), CSIC-IRTA-UAB-UB. Cerdanyola del Vallès (Bellaterra), 08193 Barcelona, Spain

[3]Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Cerdanyola del Vallès (Bellaterra), 08193 Barcelona, Spain

[4]Institute for Integrative Systems Biology (I2SysBio), Universitat de Valencia-CSIC, Paterna, 46908, Valencia, Spain

Under revision in Horticulture Research

Abstract

Transcriptomics studies generate enormous amounts of biological information. Nowadays, representing this complex data as gene coexpression networks (GCNs) is becoming commonplace. Peach is a model for *Prunus* genetics and genomics, but identifying and validating genes associated to peach breeding traits is a complex task. A GCN capable of capturing stable gene-gene relationships would help researchers overcome the intrinsic limitations of peach genetics and genomics approaches and outline future research opportunities. In this study, we created four GCNs from 604 Illumina RNA-Seq libraries. We evaluated the performance of every GCN in predicting functional annotations using an algorithm based on the 'guilty-by-association' principle. The GCN with the best performance was COO300, encompassing 21,956 genes. To validate its performance predicting gene function, we used two well-characterized genes involved in fruit flesh softening: the endopolygalacturonases *PpPG21* and *PpPG22*. Genes coexpressing with *PpPG21* and *PpPG22* were extracted and named as melting flesh (MF) subnetwork. Finally, we performed an enrichment analysis of MF subnetwork and compared the results with the current knowledge regarding peach fruit softening. The MF subnetwork mainly included genes involved in cell wall expansion and remodeling, with expression triggered by ripening-related phytohormones such as ethylene, auxin and methyl jasmonates. All these processes are closely related with peach fruit softening and therefore related to the function of *PpPG21* and *PpPG22*. These results validate COO300 as a powerful tool for peach and *Prunus* research. COO300, renamed as PeachGCN v1.0, and the scripts necessary to perform a function prediction analysis using it, are available at https://github.com/felipecobos/PeachGCN.

4. Chapter 4

4.1. Introduction

The advent of omics technologies has allowed the scientific community to generate enormous amounts of biological information. In parallel, increasingly efficient bioinformatic tools help us transform this information into structured biological knowledge. To date, more than seven million RNA-Seq libraries are available at the National Center of Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/), representing a great opportunity for large-scale bioinformatics analysis and biological data integration. Therefore, taking advantage of this valuable resource is essential in the age of big data analytics.

In transcriptomics, representing this complex data as gene coexpression networks (GCNs) is becoming a widespread practice. GCNs are usually represented as undirected graphs, where nodes correspond to genes and edges correspond to correlations in expression patterns of genes. GCNs can be built across multiple experimental conditions (condition-independent GCNs) or in specific experimental conditions (condition-dependent GCNs, e.g., tissue specific GCNs). They are based on the 'guilt-by-association' (GBA) principle (Oliver, 2000), which states that genes with related functions share similar expression patterns. Following this principle, and using the functional annotation of the genes forming the network, GCNs can be a very powerful tool to infer potential gene functions to specific genes or gene families and to understand the regulation of specific metabolic pathways. For this reason, GCNs are extremely useful in crop species, where most of the bioinformatic and genetic tools are modest and our understanding of gene function is still limited (Schaefer et al., 2017). Several studies have already created GCNs in the plant model *Arabidopsis thaliana* (Amrine, Blanco-Ulate, & Cantu, 2015; Furuya et al., 2021; Liu et al., 2019; Mao, Van Hemert, Dash, & Dickerson, 2009), maize (Huang et al., 2017; Ma et al., 2017), rice (Childs et al., 2011; Ficklin et al., 2010), wheat (Lv et al., 2020) and grapevine (Orduña et al., 2022; Orduña-Rubio et al., 2023; Wong, 2020; Wong et al., 2016).

Peach [*Prunus persica* L. (Batsch)] has been used as a model organism for genetics and genomics in the *Rosaceae*, and more specifically in the *Prunus* genus, which also encompasses other crops such as sweet and tart cherry, European and Japanese plum, apricot and almond. However, in peach, the validation of genes responsible for breeding traits is a complex task. Long intergeneration times and phenological cycles and space constraints due to the large size of the individuals under study are some of the hindrances for the work of peach geneticists (Aranzana et al., 2019). Moreover, there is a lack of efficient genetic transformation systems (Limera et al., 2017; Ricci et al., 2020). As a result of these limitations, only two genes to date, *DRO1* and *TAC1*, have been biologically validated based on mutant analysis (Dardick et al., 2013; Guseman et al., 2017).

Although small-scale condition-dependent GCNs have been reported in peach and other *Prunus* species (García-Gómez et al., 2020; Jiang et al., 2023; Wang et al., 2023; Wu et al., 2021; Xi, Feng, Liu, Zhang, & Zhao, 2019; Zhang et al., 2019), these were created *ad-hoc* to study specific biological processes and so cannot be used in other experimental contexts. Therefore, a GCN capable of capturing robust gene-gene relationships under different experimental conditions, developmental stages and tissues is needed. A GCN with these characteristics will help researchers overcome the intrinsic limitations of peach genetics and genomics approaches and outline future research opportunities.

In this study, we present the first large-scale GCN in peach. We constructed four GCNs from publicly available RNA-Seq data and evaluated the performance of every GCN using a machine-learning algorithm based on the GBA principle. The GCN with the best performance was validated by predicting gene functions of well-characterized genes. Finally, we provide the scripts and data needed for function prediction analyses using the GCN presented in this study. These resources can be found at https://github.com/felipecobos/PeachGCN.

4.2. Materials and methods

4.2.1. Data compilation

Forty-nine independent Sequence Read Archive (SRA) Bioprojects, encompassing 608 RNA-Seq libraries (Supplementary Material 4.1) were downloaded from the SRA database (Leinonen et al., 2011) in the NCBI (Sayers et al., 2022a). These RNA-Seq libraries represented all the libraries available in the NCBI to date 09/04/2020. The peach reference genome 'Lovell' version 2.1 (Verde et al., 2013, 2017) and its functional annotation were downloaded from Genome Database for Rosaceae (GDR) (Jung et al., 2019). Finally, seven functional gene annotation datasets were retrieved using the methods described below. Gene Ontology peach functional terms for biological process (GObp), molecular function (GOmf) and cellular component (GOcc) (Ashburner et al., 2000; Carbon et al., 2021) and Pfam database peach classification (Mistry et al., 2021) were retrieved using the biomaRt R package (Durinck et al., 2009). Kyoto Encyclopedia of Genes and Genomes (KEGG) peach pathway annotations (Kanehisa & Goto, 2000) were retrieved using the KEGG API (https://www.kegg.jp/kegg/rest/keggapi.html). PANTHER HMM peach classifications version 16 (Mi et al., 2021) and MapMan Pathways version 4.2 (Thimm et al., 2004) were downloaded from the public repositories.

4.2.2. Mapping and quality filtering

We performed a sequencing-quality filtering and adapter removal using Trim Galore! version 0.6.1 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Reads with terminal Ns were trimmed, then reads with a Phred score lower than 28 or smaller than 35 nucleotides were filtered. Filtered libraries were quality checked using FastQC version 0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). HISAT2 version 2.1 (Kim et al., 2015) was used to map RNA sequencing libraries to the reference peach genome 'Lovell' version 2.1 (Verde et al., 2013, 2017) with default parameters. Mapped Binary Alignment Map (BAM) files were filtered by alignment quality using SAMtools version 1.9 (Danecek et al., 2021; Li et al., 2009). Reads with mapping quality lower than 40 were filtered out. After this filtering, BAM files with less than 5,000,000 reads were discarded, leaving a total of 498 RNA-Seq libraries from 43 independent Bioprojects for further analyses.

4.2.3. Aggregated and non-aggregated GCNs inference

A raw count matrix was calculated using featureCounts (Liao et al., 2014), from Subread R package version 2.0.0 (http://subread.sourceforge.net/). For the raw count matrix construction, we excluded chimeric fragments and we used the coding DNA sequences as feature type and gene IDs as attribute type. The raw count matrix was then normalized to fragments per Kilobase million (FPKM) mapped fragments (Z. Wang, Gerstein, & Snyder, 2009), obtaining a FPKM matrix. We then applied two different methodologies: aggregated and non-aggregated network inference with two sparsity thresholds set at top 100 (stringent threshold) and 300 (relaxed threshold) ranked genes (Supplementary Figure 4.1).

For non-aggregated analysis, genes with less than 0.5 FPKM in 50% of the RNA-Seq libraries were removed. Pearson's correlation coefficient (PCC) was calculated for the remaining genes and ranked according to descending PCC, giving a PCC matrix. High reciprocal rank networks for the top 100 (HRR100) and top 300 (HRR300) were constructed according to the formula:

$$HRR\ (x, y)\ =\ [max(rank(x, y), rank(y, x))]$$

Whereby *rank(x, y)* is the descending sorted rank of gene *y* according to the coexpression list of gene *x* and vice versa for *rank(y, x)*.

For aggregated analysis, we clustered the samples into 43 different groups according to the Bioproject study ID. We filtered Bioprojects with less than six RNA-Seq libraries, leaving 26 different groups with a total of 450 RNA-Seq libraries. Genes with less than 0.5 FPKM in 50% of the libraries within each group were removed and from each filtered FPKM matrix, a high reciprocal rank network for the top 100 and top 300 was constructed. Frequency of gene coexpression interactions in all groups was calculated and ranked in a co-occurrence matrix. Finally, co-occurrence networks for top 100 (COO100) and top 300 (COO300) interactions were obtained.

4.2.4. Networks performance assay

Networks were evaluated for their ability to connect peach genes sharing functional annotations. For this purpose, GBA neighbor voting, a machine learning algorithm based on the GBA principle (Ballouz et al., 2017), was assessed over the GObp, GOmf, GOcc, Pfam, KEGG, PANTHER and MapMan datasets. Each network was scored by the area under the receiver operator characteristic curve (AUROC) across all functional categories annotated for the seven datasets. Annotations were limited to groups containing 20-1,000 genes to ensure robustness and stable performance when using neighbor voting. The AUROC value threshold for an acceptable network functional annotation was set at 0.7.

We also evaluated the impact of adding individual Bioprojects to the different networks created, HRR300, HRR100, COO300 and COO100. For this purpose, we selected five subsets each of two Bioprojects computing the top 100 and top 300 HRR and COO GCNs, evaluating their AUROC using GObp, GOmf, GOcc and MapMan datasets. We repeated this process adding one Bioproject to the initial subset to reach five subsets each with 26 Bioprojects, the maximum number of Bioprojects used in this study. The final subsets corresponded to the full HRR300, HRR100, COO300 and COO100.

4.2.5. Network validation

To validate the performance of COO300 in predicting gene functional annotations, we selected two well-characterized genes responsible for fruit flesh softening in peach, the endopolygalacturonases *PpPG21* and *PpPG22*, located on chromosome 4 (Table 4.1) (Cheng et al., 2022; Gu et al., 2016; Jiang et al., 2020; Nakano et al., 2020; Qian et al., 2021; Zhu et al., 2017). Based on the evidence available to date, the variability of flesh softening and stone adhesion during fruit ripening is due to the allelic combination of these two homologous genes. Both genes, *PpPG21* or *PpPG22*, are associated with the development of melting, non-melting or non-softening fruits, while *PpPG22* is associated with the development of freestone or clingstone fruits.

Genes coexpressed with *PpPG21* and *PpPG22* were extracted. Since both genes are involved in the peach fruit flesh softening process, we selected genes present in both subnetworks. The selected subnetwork, named melting flesh (MF) subnetwork, had 238 genes. With an enrichment analysis of the MF subnetwork using GObp, GOmf, GOcc and Mapman datasets we were able to identify the functional annotations statistically over-represented in each of the subnetworks studied. The significance threshold was held at *q*-value < 0.05. Finally, we compared the enriched terms (the functional annotations statistically over-represented) of the MF subnetwork with the current knowledge on the peach fruit softening process. In addition, as a negative control, we created 20 subnetworks with 238 randomly selected genes from COO300 and carried out an enrichment analysis of all negative control subnetworks following the steps described above.

Table 4.1. Candidate genes selected for network validation. The gene IDs were referred to the peach reference genome version 1 and 2.0 (Verde et al., 2013, 2017) and NCBI (Sayers et al., 2022b) while genomic coordinates and annotation were referred to the peach reference genome version 2.0 (Verde et al., 2017).

Gene ID	Gene name	Genomic coordinates	Annotation
Prupe.4G261900 **ppa006839m** **LOC18781156**	*PpPG21* *PpPG2* *PpPGM*	Chr04:19046344-19049605 +	Involved in fruit ripening. Promotes flesh softening.
Prupe.4G262200 **ppa006857m** **LOC18779267**	*PpPG22* *PpPG1* *PpPGF*	Chr04:19081325-19083984 +	Involved in fruit ripening. Promotes flesh softening and stone detaching from mesocarp.

4.3. Results

4.3.1. Aggregated GCNs had 21,956 genes, 81.7% of the protein-coding genes annotated in the peach reference genome

To understand the differences between the GCNs, we analyzed the general topological characteristics of the four GCNs inferred in this study (Table 4.2). The two aggregated GCNs (COO100 and COO300) had 21,956 genes, while the two built by non-aggregated methods (HRR100 and HRR300) had 17,505 genes. Of the total number of 26,873 protein-coding genes annotated in the peach reference genome, this represented 81.7 % for aggregated and 65.1 % for non-aggregated networks. The number of genes conforming the aggregated GCNs represented 16.6 % more genes (4,451) from the peach whole-genome annotation than non-aggregated GCNs.

Table 4.2. General topological characteristics of non-aggregated and aggregated GCNs with 100 and 300 top coexpressed genes (HRR100, HRR300, COO100 and COO300).

GCN	Number of genes	*P. persica* genes included in the GCN (%)	Range of node degree connectivity (min-max)	Average node degree connectivity
HRR100	17,505	65.1	649 (100-749)	161
HRR300	17,505	65.1	1490 (300-1790)	470
COO100	21,956	81.7	315 (100-415)	149
COO300	21,956	81.7	785 (300-1085)	442

The different methods used not only affected the number of genes included in the network, but also the node degree connectivity (number of coexpressed genes by gene) across all nodes of the GCN. Average node degree connectivity was higher in networks with relaxed sparsity (442 in COO300 and 470 in HRR300) in comparison to stringent sparsity (149 in COO100 and 161 in HRR100). The range between minimum and maximum node degree connectivity is wider in

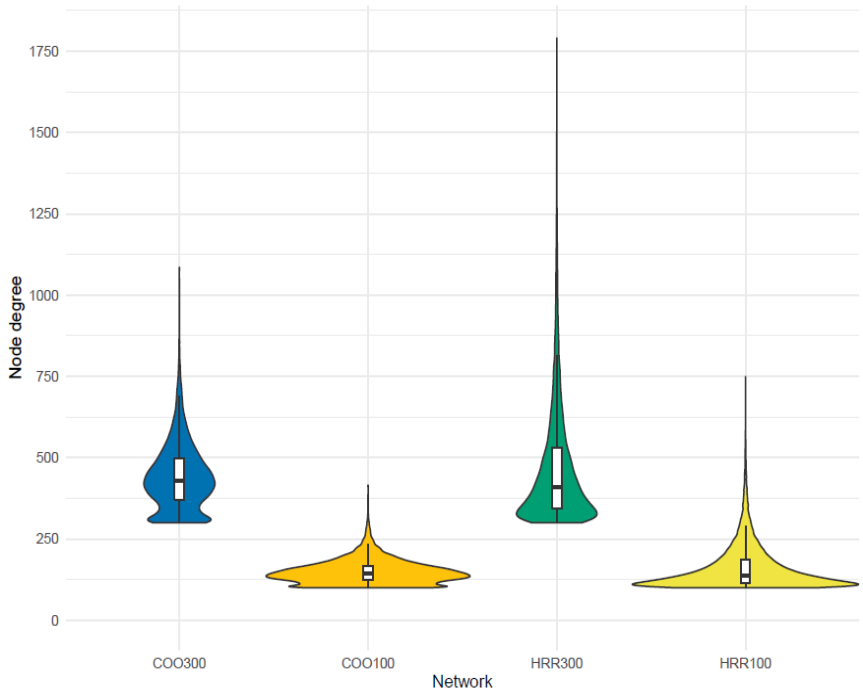


Figure 4.1. Violin plot of node degree connectivity in each of the aggregated and non-aggregated networks with relaxed or stringent sparsity (COO300, COO100, HRR300 and HRR100). Boxplots of node degree connectivity were added for each violin plot.

non-aggregated GCNs compared to aggregated GCNs with the same sparsity threshold (comparing HRR300 with COO300 and HRR100 with COO100). The minimum node degree connectivity was set by the sparsity threshold in all the networks: 100 for stringent sparsity (HRR100 and COO100) and 300 for relaxed sparsity (HRR300 and COO300). The highest node degree connectivity was found in HRR300, with a maximum of 1790 coexpressed genes with one single gene. In addition, aggregated GCNs showed a bimodal node degree connectivity distribution while non-aggregated GCNs had a unimodal distribution (Figure 4.1).

4.3.2. COO300 was the GCN with the highest AUROC value

When considering sparsity threshold, both GCNs with relaxed sparsity (HRR300 and COO300) had AUROC values over 0.7 for all the databases annotated (Table 4.3). COO300 was the GCN with the highest average AUROC value (0.746), outperforming the other GCNs. COO300 had the highest mean AUROC in almost all the datasets, except for Pfam and PANTHER, where the performance of HRR300 was better than that of COO300. HRR100 and COO100 had AUROC values under 0.7 in almost all the functional annotation databases, except for the GOcc and PANTHER datasets. The best functional annotation performance in all the networks was for the functional annotation GOcc with an average AUROC value of 0.761, followed by PANTHER (0.724), KEGG (0.718), Pfam (0.714), GObp (0.709), Mapman (0.706) and GOmf (0.693).

The method used for network building also affected its performance, but the effect was not consistent. Considering the effect of the sparsity threshold, average AUROC values for relaxed sparsity threshold were always higher (HRR300 and COO300 = 0.741) than for the stringent threshold (HRR100 and COO100 = 0.694). When comparing GCNs by aggregation method, at relaxed sparsity (HRR300 and COO300), the average AUROC value for the aggregated method was higher but comparing at the stringent threshold (HRR100 and COO100), the average AUROC value was better for the non-aggregated method.

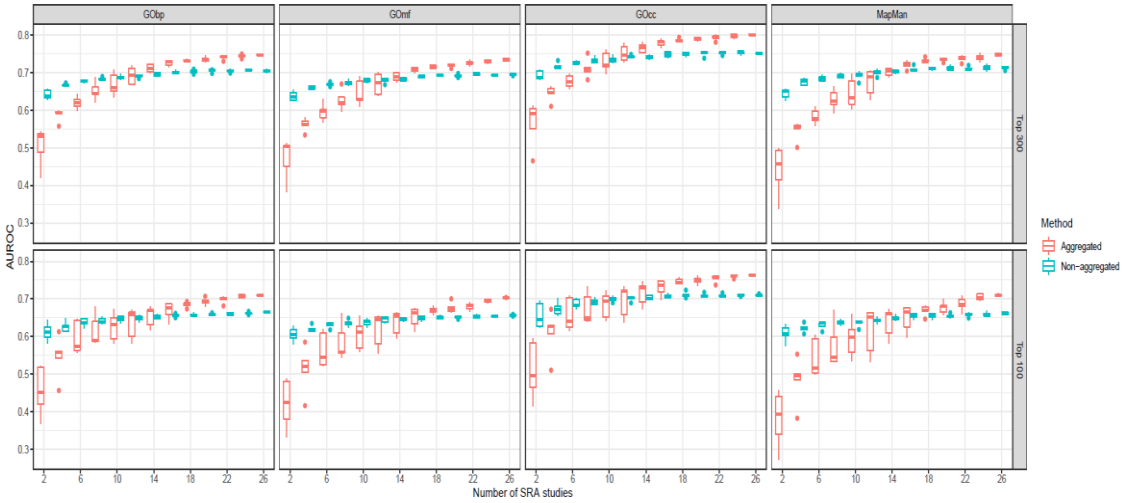


Figure 4.2. Boxplots of the AUROC value for every subset of Bioprojects (from 2 to 26) and method used.

Table 4.3. AUROC values for each GCN (COO300, HRR300, COO100, HRR100) performance in the different datasets. The best performance by dataset was highlighted with an asterisk.

GCN	GObp	GOmf	GOcc	Pfam	KEGG	PANTHER	MapMan	Average
COO300	0.738*	0.723*	0.788*	0.736	0.750*	0.746	0.741*	0.746*
HRR300	0.724	0.705	0.773	0.745*	0.728	0.749*	0.732	0.736
COO100	0.681	0.670	0.733	0.680	0.697	0.688	0.664	0.687
HRR100	0.692	0.673	0.748	0.695	0.695	0.712	0.686	0.700
Average	0.709	0.693	0.761	0.714	0.718	0.724	0.706	0.717

Finally, we evaluated the effects of adding Bioprojects on the AUROC value in every GCN built in this study. Figure 4.2 shows the correlation between the network AUROC value and the number of Bioprojects used. For every combination of GCN building method (aggregated or non-aggregated), threshold (top 300 or top 100) and dataset used (GObp, GOmf, GOcc and MapMan) we observed similar trends, where the AUROC value increased with the number of Bioprojects. This trend was more pronounced for aggregated GCNs than non-aggregated GCNs, reaching a plateau after adding 10-12 Bioprojects. In all cases, the standard deviation of aggregated GCNs decreased as the number of Bioprojects increased.

4.3.3. Aggregated GCNs showed a positive trend between average node degree connectivity and AUROC score of individual functional annotations for GObp, GOmf, GOcc, KEGG and Mapman

To assess the relationship between the AUROC score of individual functional annotations and the average node degree connectivity of the genes sharing that annotation we used a Loess regression (Figure 4.3). For example, an individual functional annotation could be *GOcc: cell wall*, we studied if the individual AUROC score of *GOcc: cell wall* was related to the average number of connections of the genes sharing *GOcc: cell wall* annotation. We then repeated the analysis for all the functional annotations within a dataset (*GOcc: apoplast*, *GOcc: extracellular region*, etc). In the case of aggregated GCNs, there was a positive trend between average node degree connectivity and AUROC score of individual functional annotations for GObp, GOmf, GOcc, KEGG and Mapman. In the case of non-aggregated GCNs the only dataset with a positive trend between average node degree connectivity and AUROC score of individual functional annotations was KEGG. The average node degree connectivity had no effect on the AUROC score of individual functional annotations in the Pfam dataset in any of the GCNs studied.

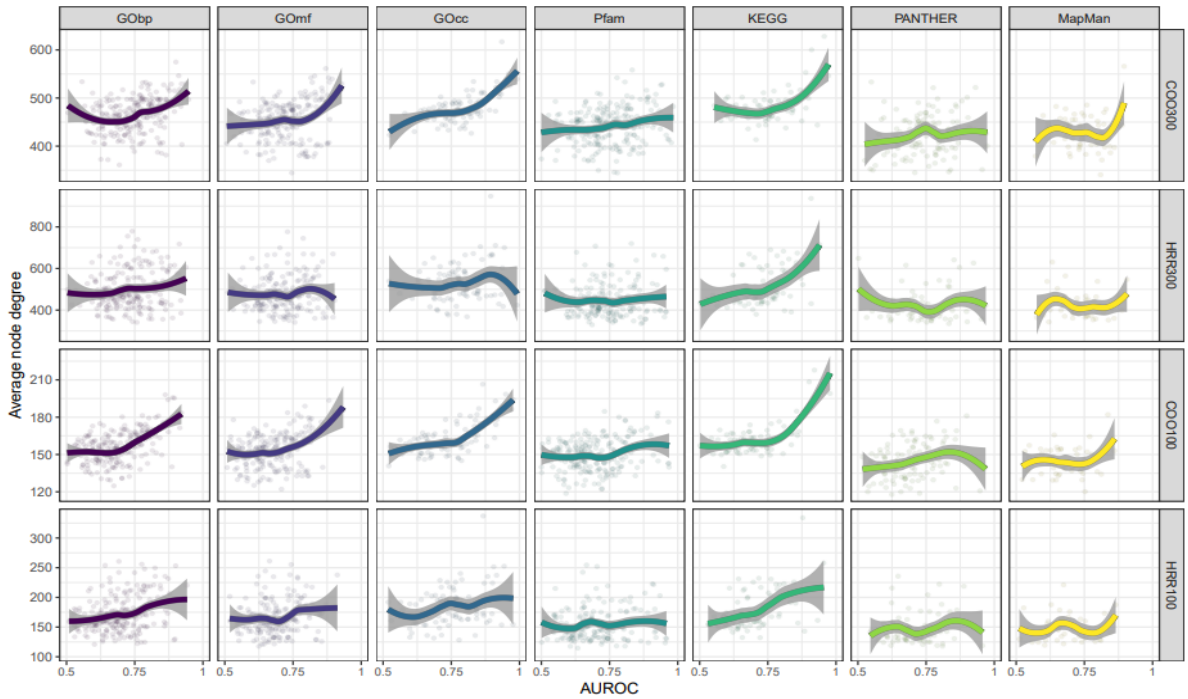


Figure 4.3. Scatter plot and Loess regression representation of average node degree connectivity by AUROC value for each of the GCNs (COO300, HRR300, COO100 and HRR100) in all the datasets used for network annotation (CObp, GOmf, GOcc, Pfam, KEEG, PANTHER and Mapman).

4.3.4. MF subnetwork was enriched in 33 terms

PpPG21 and *PpPG22* were annotated using GObp, GOmf, GOcc, Pfam, KEGG, PANTHER and MapMan. Both genes shared several annotations: '*GOcc: extracellular region'*, '*GOcc: cell wall'*, '*GObp: metabolic process'*, '*GObp: cell wall organization'*, '*GOmf: hydrolase activity, acting on glycosyl bounds'*, '*GOmf: polygalacturonase activity'*, '*Mapman: enzyme classification. hydrolases. Glycoxylases'* and '*Pfam: glycosyl hydrolases family 28'*. *PpPG22* only had two terms not shared with *PpPG21*, '*GObp: fruit ripening'* and '*GObp: carbohydrate metabolic process'*.

The *PpPG21* and *PpPG22* subnetworks were constituted by 485 and 354 genes, respectively. Even if *PpPG21* and *PpPG22* were not coexpressed, both subnetworks shared 238 genes. These genes were selected and named the melting flesh (MF) subnetwork (Figure 4.4; Supplementary Material 4.2). This MF subnetwork was annotated in GObp, GOcc, GOmf and Mapman datasets. Of the 238 genes in the MF subnetwork, 136 were annotated in GObp, 123 in GOcc, 156 in GOmf and 116 in Mapman (Supplementary Material 4.2).

After MF subnetwork annotation, we performed an enrichment analysis. The MF subnetwork was enriched in 33 different terms, so 33 terms were significantly over-represented in this subnetwork. Of these 33 terms, 12 belonged to the GOmf dataset, nine to Mapman, eight to GObp and four to GOcc (Figure 4.5; Supplementary Material 4.2).

Within GOmf, up to 26 genes were annotated as *hydrolase activity* or as its child term (direct descendant), *hydrolase activity, acting on glycosyl bonds*. The next term was '*xyloglucan:xyloglucosyl transferase activity'*, with four genes annotated. With three genes annotated, we found the terms '*methyl indole-3-acetate esterase activity'*, '*methyl salicylate esterase activity'*, '*methyl jasmonate esterase activity'*, '*oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two*

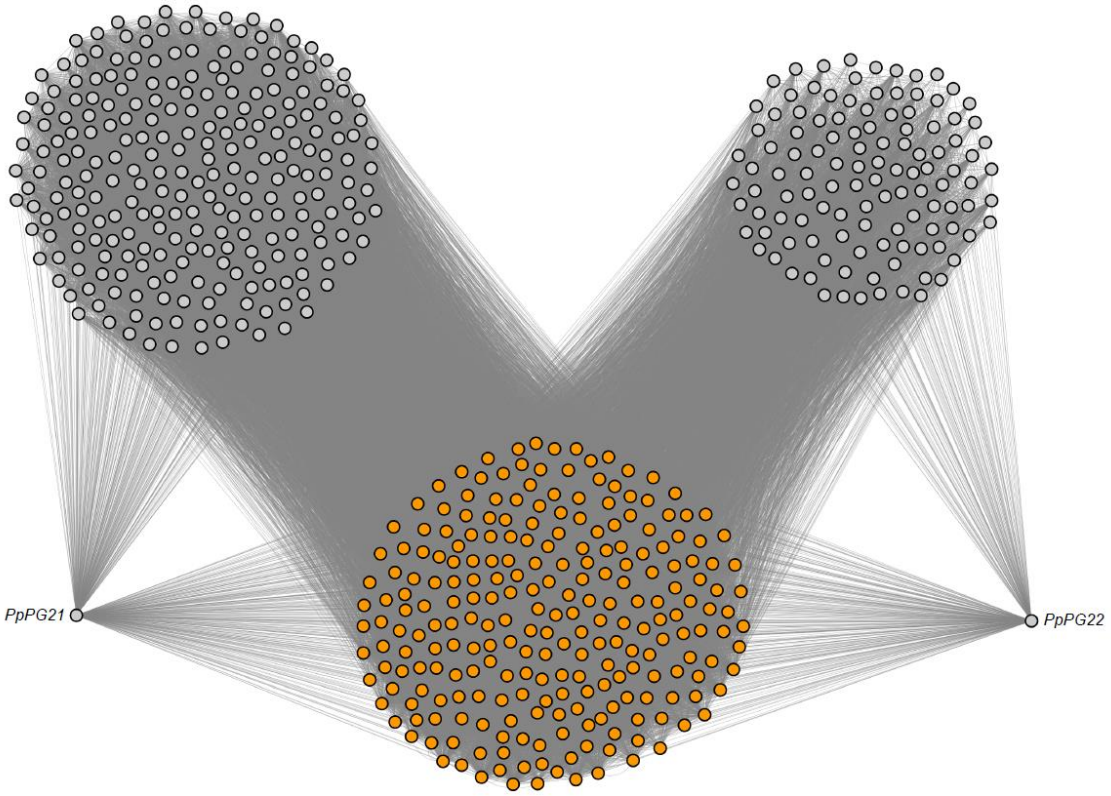


Figure 4.4. *PpPG21, PpPG22* and MF subnetworks. MF subnetwork is highlighted in orange.

molecules of water and *metal ion transmembrane transporter activity'*. Finally, with two genes annotated, we found the terms *'inositol hexakisphosphate binding'*, *'phosphate ion transmembrane transporter activity'*, *'protein-disulfide reductase activity'* and *'acid-amino acid ligase activity'*.

Using Mapman as the annotation dataset, 27 genes were annotated as *'enzyme classification'*. There were eight genes annotated as *'glycosyltransferase'*, a child term of *'enzyme classification'*. The next term, with 17 genes annotated, was *'phytohormone action'*. There were four genes annotated as *'auxin'* or *'auxin.conjugation and degradation'* and three as *'ethylene'*, child terms of *'phytohormone action'*. With two genes annotated, we found the terms *'Solute transport.carrier-mediated transport.IT superfamily.phosphate transporter (PHO)'*, *'Nutrient uptake.phosphorus assimilation.phosphate uptake.phosphate transporter (PHO1)'* and *'Lipid metabolism.fatty acid biosynthesis.fatty acid desaturation.omega-3/omega-6 fatty acid desaturase (FAD2/3/6-8)'*.

Within GObp, there were 26 genes annotated as *'oxidation-reduction process'*. Up to 11 genes were annotated as *'metabolic process'*. There were nine genes annotated as *'cell wall organization'* and four as *'cell wall biogenesis'*, child terms of *'cell wall organization or biogenesis'*. There were four genes annotated as *'cellular glucan metabolic process'* and its child term, *'xyloglucan metabolic process'*, four as *'jasmonic acid metabolic process'* and three as *'salicylic acid metabolic process'*.

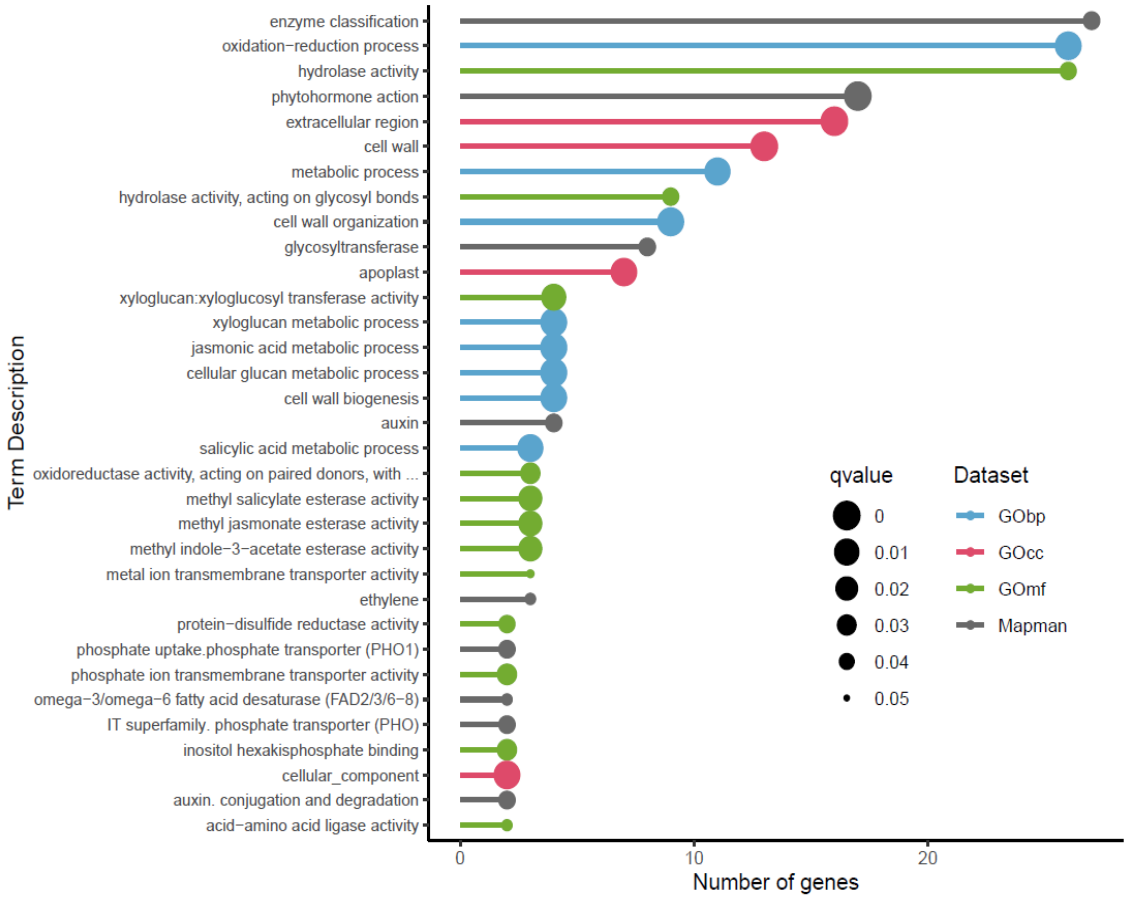


Figure 4.5. Lollipop plot of enriched terms found in the MF subnetwork. Enriched terms were sorted by the number of genes annotated by each term.

Using GOcc as the annotation dataset, 16 genes were annotated as '*extracellular region*', seven genes as '*apoplast*', child term of '*extracellular region*', and up to 13 genes were annotated as '*cell wall*'.

Among the 20 negative control subnetworks created, the mean number of enriched terms was 1.35, while the melting subnetwork had 33 enriched terms (Supplementary Material 4.3).

4.4.Discussion

4.4.1. The GCN topological characteristics are affected by the different algorithms used

To achieve the best results during gene coexpression networks (GCNs) building, two variables were tuned, aggregation method and sparsity threshold. The four GCNs obtained were evaluated, with substantial differences in the general topological characteristics of the GCNs inferred.

When considering GCN building methods, a major difference between aggregated and non-aggregated GCNs was the number of genes forming the network. Aggregated GCNs had 21,956 genes (81.7 % of *P. persica* genes), while non-aggregated GCNs only had 17,505 (65.1 % of *P. persica* genes). This difference comes from the low-expression gene filtering. In non-aggregated GCNs all the genes with less than 0.5 FPKM in 50% of the 498 RNA-Seq libraries were filtered, while in aggregated GCNs this filtering is independently performed for each of the 26 Bioproject groups. That allowed the inclusion in the GCN of genes expressed in more specific conditions and therefore involved in more specific processes. This indicates that both

aggregated and non-aggregated networks were able to capture stable gene-gene relationships expressed in most of the RNA-Seq libraries used in the analysis, but only aggregated GCNs were able to detect gene-gene interactions produced in specific conditions. Condition-independent gene-gene connections could be related to basal metabolic pathways, while condition-dependent gene-gene interactions could be associated to specific metabolic pathways.

This could explain the difference in the distribution of node degree connectivity between aggregated and non-aggregated GCNs. As shown in Figure 4.1, aggregated GCNs had a bimodal distribution of node degree connectivity, while non-aggregated GCNs had a unimodal distribution. As mentioned previously, aggregated GCNs may be able to detect genes involved in specific and basal metabolic processes. The two modes detected in aggregated GCNs node degree connectivity distribution could be associated with these two groups of genes. The group with the lower node degree distribution could be associated with genes involved in more specific metabolic pathways, coexpressed with a lower number of genes. The group with the higher node degree distribution could be associated with genes involved in basal metabolic pathways and coexpressed with a higher number of genes. On the other hand, non-aggregated GCNs may only detect genes involved in basal metabolic pathways, having only one mode in their node degree distribution.

Another factor affecting the topology of the networks was the sparsity threshold selected. HRR300 and COO300 had a node degree connectivity higher than HRR100 and COO100. This was an expected result, since a higher number of ranked genes allows a higher number of connections between genes.

4.4.2. Sparsity threshold and the number of Bioprojects determine network performance

According to the results, sparsity was a key factor affecting network performance. The average AUROC of relaxed sparsity threshold networks (HRR300 and COO300) was 0.741, while that of stringent sparsity threshold networks (HRR100 and COO100) was 0.694. Applying relaxed sparsity threshold during network building represented an increment of 6.3% in the AUROC score in comparison to stringent sparsity threshold.

The number of Bioprojects used to build the GCN was a key factor in the case of aggregated methods, indicating the minimum number of Bioprojects necessary to reach a sufficiently high AUROC score (Figure 4.2). In every case, aggregated methods had a lower AUROC value than non-aggregated methods using a low number of Bioprojects. By increasing this number, aggregated methods overtook non-aggregated methods, as found in other studies (Orduña et al., 2022; Orduña-Rubio et al., 2023). For future GCNs construction, increasing the number of Bioprojects could improve the performance of the GCNs.

Studying the effect of functional annotations average node degree on the AUROC value, we found major differences depending on the type of dataset used. There was a positive correlation between functional annotations average node degree and functional annotations individual AUROC in datasets based on evidence such as GObp, GOmf, GOcc, KEGG and Mapman. On the other hand, this correlation was lost with datasets based on domain identification by sequence similarity, such as PANTHER and Pfam. These results are in agreement with the GBA principle, which states that coexpressed genes share function, and not necessarily similar sequences.

4.4.3. COO300 validated as a powerful tool for peach and *Prunus* research

In peach, fruit flesh softening has been extensively studied at fruit ripening and postharvest due to its implication in fruit shelf life. Fruit softening involves several cellular processes, such as the disassembly of the cell wall and the dissolution of the middle lamella. These modifications are

the result of hydrolytic changes in the polysaccharides forming the cell wall, including celluloses, hemicelluloses (mainly xyloglucan) and pectins. Several terms found in the MF subnetwork were associated to this process, such as '*GOcc: cell wall*', '*GObp: cell wall organization*' and '*GObp: cell wall biogenesis*', '*GOmf: hydrolase activity*', '*GOmf: hydrolase activity, actin on glycosyl bonds*', '*Mapman: enzyme classification.EC_2 transferases.EC_2.4 glycosyltransferase*' and '*GOmf: xyloglucan:xyloglucosyl transferase activity*'.

Peach flesh softening is a synergistic process triggered by an extensive phytohormone signaling network. As a climacteric fruit, cross talk between ethylene and auxin occurs during peach ripening (Trainotti et al., 2007). Moreover, methyl jasmonates (MeJAs) play an important role in slowing down fruit ripening by inhibiting ethylene production and fruit flesh softening (Soto et al., 2012; Wei et al., 2017). Up to seven enriched terms were related to these phytohormones in the MF subnetwork, such as '*Mapman: phytohormone action*', '*Mapman: phytohormone action. Auxin*', '*GObp: jasmonic acid metabolic process*', '*Mapman: phytohormone action. ethylene*', '*GOmf: methyl indole-3-esterase activity*', '*GOmf: methyl jasmonate esterase activity*' and '*Mapman: phytohormone action. auxin. auxin conjugation and degradation*'.

We found 25 genes in the MF subnetwork that have previously been reported as associated to ripening and softening (Supplementary Material 4.2). Among them, we identified several genes involved in the enzymatic machinery responsible for cell wall disassembly, such as a pectin methylesterase (*Prupe.7G192800*), a pectin methylesterase inhibitor (*Prupe.1G114500*), a pectate lyase (*Prupe.4G116600*), a β-galactosidase (*Prupe.3G050200*) and a xyloglucan endotransglycosylase hydrolase (*Prupe.1G255100*). Additionally, we found an expansin, a cell wall structural protein (*Prupe.6G075100*). Related to ethylene, we identified a 1-amino-cyclopropane-1-carboxylate synthase (*PpACS1*, *Prupe.2G176900*) and 1-amino-cyclopropane-1-carboxylate oxidase (*PpACO1*, *Prupe.3G209900*), both genes codifying the key enzymes catalyzing the final steps of the ethylene biosynthetic pathway (Tonutti et al., 1997). In fact, *PpACS1* has been previously reported as a regulator of *PpPG21* (Tatsuki et al., 2013). Another gene related to ethylene production was an ethylene receptor 2 (*PpETR2*, *Prupe.1G034300*). The implication of this gene in the ethylene transduction signal has been verified at the transcriptional level in the final stages of fruit ripening in melting flesh peaches (Wang et al., 2017). Regarding genes related to auxin biosynthesis, we found a YUCCA-like auxin-biosynthesis gene (*PpYUC11*, *Prupe.6G157500*) and an IAA amino acid synthase (*PpGH3*, *Prupe.6G226100*). Both genes have been reported to have the same expression pattern as *PpACS1* at late ripening stages in response to high auxins levels in melting flesh fruits (Pan et al., 2015).

Based on these results, we can affirm that the MF subnetwork is mainly formed by genes involved in cell wall organization and biogenesis, with expression regulated by ripening-related phytohormones such as ethylene, auxin and MeJA. Moreover, we found 25 genes previously reported as involved in softening, some taking part in key steps of these processes. These results demonstrate that the MF subnetwork is closely related to peach fruit softening and therefore to the function of *PpPG21* and *PpPG22*. Taken together, this validates COO300 as an accurate and powerful tool for peach and *Prunus* research.

4.4.4. Gene coexpression networks as catalysts for *Prunus* research

While large-scale GCNs have been unexplored as tools in *Prunus* research until now, they are widely used in the model organism *Arabidopsis thaliana* and other crop species. Depending on the needs of the researcher, GCNs have been exploited in different ways. One of the most common is to identify different modules (also known as clusters) within the GCN through a clusterization analysis. These gene modules, which represent groups of genes highly connected between them and relatively isolated from the rest of the GCN, are particularly useful to study

uncharacterized biological processes. For example, Childs et al., 2011 used this approach in rice to annotate 13,537 genes, 2,980 of which had no previous annotation.

Another approach that uses group of genes to study specific biological processes is the guide gene analysis. In this case, a list of well-characterized genes involved in a specific biological process are selected and genes coexpressing with the list of genes of interest are extracted from the network. In this way, the selected genes are used as a guide to study the transcriptional regulation of the biological process of interest. Huang et al., 2017 successfully applied this approach to study the cell wall biosynthesis in maize. Pathway-centered network analysis has also been helpful in the identification of members or regulators of secondary metabolic pathways (Orduña-Rubio et al., 2023).

GCNs can also be used to study specific gene families, being particularly useful for studying transcription factor families. For instance, Wong et al., 2016 developed a MYB-centered GCN to study the potential processes being regulated by this family in grapevine.

Finally, GCNs can be used to infer the function of a gene of interest. This is a situation of special interest in peach and *Prunus* research, where most trait-loci analyses lead to a list of candidate genes associated with the trait under study. With poor or no functional information, identifying the responsible gene from this list of candidates can be almost impossible. Even when a high-confidence candidate gene is identified, the lack of an efficient genetic transformation system is still one of the main limitations for functional, mutant, or transgenic based validation. Having a tool such as the GCN presented in this study, with which obtaining useful information about the biological processes in which a gene is involved, may be of critical importance.

4.5. Conclusions

In this study, we performed the widest overview of transcriptomic analysis carried out to date in peach or other *Prunus* species. The GCN inference methods used, aggregated or non-aggregated, affected the topological characteristics and performance of the GCNs created. Using two well-characterized genes in peach, *PpPG21* and *PpPG22*, we were able to validate the network with the best performance, COO300. The GCN tool presented in this study will help *Prunus* researchers overcome the intrinsic limitations of working with crop tree species, prioritize research lines and outline new ones. COO300, named as PeachGCN v1.0, and the scripts necessary to run a function prediction analysis using it, are available at https://github.com/felipecobos/PeachGCN.

4.6. References

Amrine, K. C. H., Blanco-Ulate, B., & Cantu, D. (2015). Discovery of Core Biotic Stress Responsive Genes in Arabidopsis by Weighted Gene Co-Expression Network Analysis. *PLOS ONE*, *10*(3), e0118731. https://doi.org/10.1371/JOURNAL.PONE.0118731

Aranzana, M. J., Decroocq, V., Dirlewanger, E., Eduardo, I., Gao, Z. S., Gasic, K., Iezzoni, A., Jung, S., Peace, C., Prieto, H., Tao, R., Verde, I., Abbott, A. G., & Arús, P. (2019). Prunus genetics and applications after de novo genome sequencing: achievements and prospects. In *Horticulture Research* (Vol. 6, Issue 1, p. 58). Nature Publishing Group. https://doi.org/10.1038/s41438-019-0140-8

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics 2000 25:1*, *25*(1), 25–29. https://doi.org/10.1038/75556

Ballouz, S., Weber, M., Pavlidis, P., & Gillis, J. (2017). EGAD: Ultra-fast functional analysis of gene networks. *Bioinformatics*, *33*(4), 612–614. https://doi.org/10.1093/bioinformatics/btw695

Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L.-P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., Mushayahama, T., & LaBonte, S. A. (2021). The Gene Ontology resource: enriching a GOld mine [Article]. *Nucleic Acids Research.*, *49*(D1), D325–D334. https://doi.org/10.1093/nar/gkaa1113

Cheng, C., Liu, J., Wang, X., Wang, Y., Yuan, Y., & Yang, S. (2022). PpERF/ABR1 functions as an activator to regulate PpPG expression resulting in fruit softening during storage in peach (Prunus persica). *Postharvest Biology and Technology*, *189*, 111919. https://doi.org/10.1016/J.POSTHARVBIO.2022.111919

Childs, K. L., Davidson, R. M., & Buell, C. R. (2011). Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. *PLoS ONE*, *6*(7), e22196. https://doi.org/10.1371/journal.pone.0022196

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Dardick, C., Callahan, A., Horn, R., Ruiz, K. B., Zhebentyayeva, T., Hollender, C., Whitaker, M., Abbott, A., & Scorza, R. (2013). PpeTAC1 promotes the horizontal growth of branches in peach trees and is a member of a functionally conserved gene family found in diverse plants species. *The Plant Journal*, *75*(4), 618–630. https://doi.org/10.1111/TPJ.12234

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols 2009 4:8*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Ficklin, S. P., Luo, F., & Feltus, F. A. (2010). The Association of Multiple Interacting Genes with Specific Phenotypes in Rice Using Gene Coexpression Networks. *Plant Physiology*, *154*(1), 13–24. https://doi.org/10.1104/PP.110.159459

Furuya, T., Saito, M., Uchimura, H., Satake, A., Nosaki, S., Miyakawa, T., Shimadzu, S., Yamori, W., Tanokura, M., Fukuda, H., & Kondo, Y. (2021). Gene co-expression network analysis identifies BEH3 as a stabilizer of secondary vascular development in Arabidopsis. *The Plant Cell*. https://doi.org/10.1093/PLCELL/KOAB151

García-Gómez, B. E., Ruiz, D., Salazar, J. A., Rubio, M., Martínez-García, P. J., & Martínez-Gómez, P. (2020). Analysis of Metabolites and Gene Expression Changes Relative to Apricot (Prunus armeniaca L.) Fruit Quality During Development and Ripening. *Frontiers in Plant Science*, *0*, 1269. https://doi.org/10.3389/FPLS.2020.01269

Gu, C., Wang, L., Wang, W., Zhou, H., Ma, B., Zheng, H., Fang, T., Ogutu, C., Vimolmangkang, S., & Han, Y. (2016). Copy number variation of a gene cluster encoding endopolygalacturonase mediates flesh texture and stone adhesion in peach. *Journal of Experimental Botany*, *67*(6), 1993–2005. https://doi.org/10.1093/JXB/ERW021

Guseman, J. M., Webb, K., Srinivasan, C., & Dardick, C. (2017). DRO1 influences root system architecture in Arabidopsis and Prunus species. *The Plant Journal*, *89*(6), 1093–1105. https://doi.org/10.1111/TPJ.13470

Huang, J., Vendramin, S., Shi, L., & McGinnis, K. M. (2017). Construction and optimization of a large gene coexpression network in maize using RNA-seq data. *Plant Physiology*, *175*(1), 568–583. https://doi.org/10.1104/pp.17.00825

Jiang, L., Kang, R., Feng, L., Yu, Z., & Luo, H. (2020). iTRAQ-based quantitative proteomic analysis of peach fruit (Prunus persica L.) at different ripening and postharvest storage stages. *Postharvest Biology and Technology*, *164*, 111137. https://doi.org/10.1016/J.POSTHARVBIO.2020.111137

Jiang, X., Liu, K., Peng, H., Fang, J., Zhang, A., Han, Y., & Zhang, X. (2023). Comparative network analysis reveals the dynamics of organic acid diversity during fruit ripening in peach (Prunus persica L. Batsch). *BMC Plant Biology*, *23*(1), 1–14. https://doi.org/10.1186/S12870-023-04037-W/TABLES/1

Jung, S., Lee, T., Cheng, C.-H., Buble, K., Zheng, P., Yu, J., Humann, J., Ficklin, S. P., Gasic, K., Scott, K., Frank, M., Ru, S., Hough, H., Evans, K., Peace, C., Olmstead, M., DeVetter, L. W., McFerson, J., Coe, M., … Main, D. (2019). 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Research*, *47*(D1), D1137–D1145. https://doi.org/10.1093/nar/gky1000

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/NAR/28.1.27

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods 2015 12:4*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(SUPPL. 1), D19–D21. https://doi.org/10.1093/nar/gkq1019

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/BIOINFORMATICS/BTT656

Limera, C., Sabbadini, S., Sweet, J. B., & Mezzetti, B. (2017). New Biotechnological Tools for the Genetic Improvement of Major Woody Fruit Species. *Frontiers in Plant Science*, *0*, 1418. https://doi.org/10.3389/FPLS.2017.01418

Liu, W., Lin, L., Zhang, Z., Liu, S., Gao, K., Lv, Y., Tao, H., & He, H. (2019). Gene co-expression network analysis identifies trait-related modules in Arabidopsis thaliana. *Planta 2019 249:5*, *249*(5), 1487–1501. https://doi.org/10.1007/S00425-019-03102-9

Lv, L., Zhang, W., Sun, L., Zhao, A., Zhang, Y., Wang, L., Liu, Y., Li, Z., Li, H., & Chen, X. (2020). Gene co-expression network analysis to identify critical modules and candidate genes of drought-resistance in wheat. *PLOS ONE*, *15*(8), e0236186. https://doi.org/10.1371/JOURNAL.PONE.0236186

Ma, S., Ding, Z., & Li, P. (2017). Maize network analysis revealed gene modules involved in development, nutrients utilization, metabolism, and stress response. *BMC Plant Biology*, *17*(1), 1–17. https://doi.org/10.1186/s12870-017-1077-4

Mao, L., Van Hemert, J. L., Dash, S., & Dickerson, J. A. (2009). Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics 2009 10:1*, *10*(1), 1–24. https://doi.org/10.1186/1471-2105-10-346

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, *49*(D1), D394–D403. https://doi.org/10.1093/NAR/GKAA1106

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/NAR/GKAA913

Nakano, R., Kawai, T., Fukamatsu, Y., Akita, K., Watanabe, S., Asano, T., Takata, D., Sato, M., Fukuda, F., & Ushijima, K. (2020). Postharvest Properties of Ultra-Late Maturing Peach Cultivars and Their Attributions to Melting Flesh (M) Locus: Re-evaluation of M Locus in Association With Flesh Texture. *Frontiers in Plant Science*, *11*, 1817. https://doi.org/10.3389/FPLS.2020.554158/BIBTEX

Oliver, S. (2000). Guilt-by-association goes global. *Nature 2000 403:6770*, *403*(6770), 601–602. https://doi.org/10.1038/35001165

Orduña, L., Li, M., Navarro-Payá, D., Zhang, C., Santiago, A., Romero, P., Ramšak, Ž., Magon, G., Höll, J., Merz, P., Gruden, K., Vannozzi, A., Cantu, D., Bogs, J., Wong, D. C. J., Huang, S. shan C., & Matus, J. T. (2022). Direct regulation of shikimate, early phenylpropanoid, and stilbenoid pathways by Subgroup 2 R2R3-MYBs in grapevine. *Plant Journal*, *110*(2), 529–547. https://doi.org/10.1111/TPJ.15686

Orduña-Rubio, L., Santiago, A., Navarro-Payá, D., Zhang, C., Wong, D. C. J., & Matus, J. T. (2023). Aggregated gene co-expression networks for predicting transcription factor regulatory landscapes in a non-model plant species. *BioRxiv*. https://doi.org/https://doi.org/10.1101/2023.04.24.538042

Pan, L., Zeng, W., Niu, L., Lu, Z., Liu, H., Cui, G., Zhu, Y., Chu, J., Li, W., Fang, W., Cai, Z., Li, G., & Wang, Z. (2015). PpYUC11, a strong candidate gene for the stony hard phenotype in peach (Prunus persica L. Batsch), participates in IAA biosynthesis during fruit ripening. *Journal of Experimental Botany*, *66*(22), 7031–7044. https://doi.org/10.1093/JXB/ERV400

Qian, M., Xu, Z., Zhang, Z., Li, Q., Yan, X., Liu, H., Han, M., Li, F., Zheng, J., Zhang, D., & Zhao, C. (2021). The downregulation of PpPG21 and PpPG22 influences peach fruit texture and softening. *Planta 2021 254:2*, *254*(2), 1–12. https://doi.org/10.1007/S00425-021-03673-6

Ricci, A., Sabbadini, S., Prieto, H., Padilla, I. M., Dardick, C., Li, Z., Scorza, R., Limera, C., Mezzetti, B., Perez-Jimenez, M., Burgos, L., & Petri, C. (2020). Genetic Transformation in Peach (Prunus persica L.): Challenges and Ways Forward. *Plants 2020, Vol. 9, Page 971*, *9*(8), 971. https://doi.org/10.3390/PLANTS9080971

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022a). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/NAR/GKAB1112

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022b). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/NAR/GKAB1112

Schaefer, R. J., Michno, J. M., & Myers, C. L. (2017). Unraveling gene function in agricultural species using gene co-expression networks. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1860*(1), 53–63. https://doi.org/10.1016/J.BBAGRM.2016.07.016

Soto, A., Ruiz, K. B., Ziosi, V., Costa, G., & Torrigiani, P. (2012). Ethylene and auxin biosynthesis and signaling are impaired by methyl jasmonate leading to a transient slowing down of ripening in peach fruit. *Journal of Plant Physiology*, *169*(18), 1858–1865. https://doi.org/10.1016/J.JPLPH.2012.07.007

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y., & Stitt, M. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, *37*(6), 914–939. https://doi.org/10.1111/J.1365-313X.2004.02016.X

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M. T., Grimwood, J., Cattonaro, F., Zuccolo, A., Rossini, L., Jenkins, J., Vendramin, E., Meisel, L. A., Decroocq, V., Sosinski, B., Prochnik, S., Mitros, T., … Rokhsar, D. S. (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, *45*(5), 487–494. https://doi.org/10.1038/ng.2586

Verde, I., Jenkins, J., Dondini, L., Micali, S., Pagliarani, G., Vendramin, E., Paris, R., Aramini, V., Gazza, L., Rossini, L., Bassi, D., Troggio, M., Shu, S., Grimwood, J., Tartarini, S., Dettori, M. T., & Schmutz, J. (2017). The Peach v2.0 release: High-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3606-9

Wang, Q., Cao, K., Li, Y., Wu, J., Fan, J., Ding, T., Khan, I. A., & Wang, L. (2023). Identification of co-expressed networks and key genes associated with organic acid in peach fruit. *Scientia Horticulturae*, *307*, 111496. https://doi.org/10.1016/j.scienta.2022.111496

Wang, X., Ding, Y., Wang, Y., Pan, L., Niu, L., Lu, Z., Cui, G., Zeng, W., & Wang, Z. (2017). Genes involved in ethylene signal transduction in peach (Prunus persica) and their expression profiles during fruit maturation. *Scientia Horticulturae*, *224*, 306–316. https://doi.org/10.1016/J.SCIENTA.2017.06.035

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 2008 10:1*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wong, D. C. J. (2020). Network aggregation improves gene function prediction of grapevine gene co-expression networks. *Plant Molecular Biology*, *103*(4–5), 425–441. https://doi.org/10.1007/s11103-020-01001-2

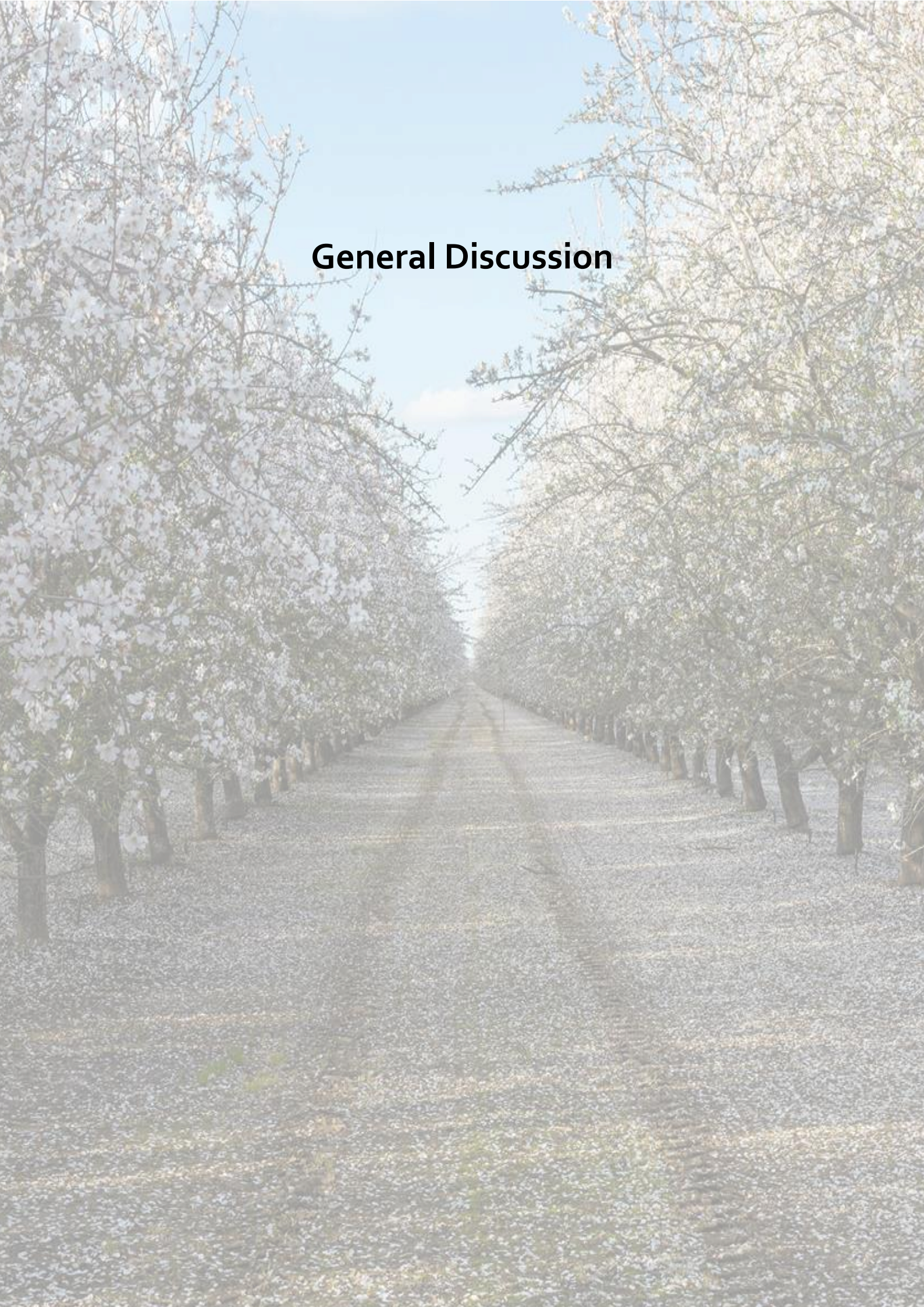
Wong, D. C. J., Schlechter, R., Vannozzi, A., Höll, J., Hmmam, I., Bogs, J., Tornielli, G. B., Castellarin, S. D., & Matus, J. T. (2016). A systems-oriented analysis of the grapevine R2R3-MYB transcription factor family uncovers new insights into the regulation of stilbene accumulation. *DNA Research*, *23*(5), 451–466. https://doi.org/10.1093/dnares/dsw028

Wu, X., Du, A., Zhang, S., Wang, W., Liang, J., Peng, F., & Xiao, Y. (2021). Regulation of growth in peach roots by exogenous hydrogen sulfide based on RNA-Seq. *Plant Physiology and Biochemistry*, *159*, 179–192. https://doi.org/10.1016/J.PLAPHY.2020.12.018

Xi, W., Feng, J., Liu, Y., Zhang, S., & Zhao, G. (2019). The R2R3-MYB transcription factor PaMYB10 is involved in anthocyanin biosynthesis in apricots and determines red blushed skin. *BMC Plant Biology*, *19*(1). https://doi.org/10.1186/s12870-019-1898-4

Zhang, Q., Feng, C., Li, W., Qu, Z., Zeng, M., & Xi, W. (2019). Transcriptional regulatory networks controlling taste and aroma quality of apricot (Prunus armeniaca L.) fruit during ripening. *BMC Genomics*, *20*(1), 45. https://doi.org/10.1186/s12864-019-5424-8

Zhu, Y., Zeng, W., Wang, X., Pan, L., Niu, L., Lu, Z., Cui, G., & Wang, Z. (2017). Characterization and Transcript Profiling of PME and PMEI Gene Families during Peach Fruit Maturation. *Journal of the American Society for Horticultural Science*, *142*(4), 246–259. https://doi.org/10.21273/JASHS04039-17



General Discussion

5. General Discussion

Plant breeding has long been a crucial aspect of agricultural research, aimed at improving crops production, quality, and overall resilience. In the last decades, significant advancements in genomics and bioinformatics have revolutionized the field of plant breeding, and almond breeding is no exception. These cutting-edge tools have opened up new possibilities, enabling researchers and breeders to explore the almond genome in unprecedented detail and accelerate the process of developing improved almond varieties. This thesis aimed to design, develop and implement several genetic tools in almond breeding.

With the objective of study breeding tendencies in the last 50 years of almond breeding, in Chapter 1, we performed a pedigree analysis using 220 pedigrees from breeding records and molecular marker information. Our results detected two worldwide mainstream breeding lines: one European, based mainly in 'Tuono' and 'Cristomorto' cultivars as founders, and the Californian-Australian line, based primarily in 'Nonpareil'. Indeed, the repeated use of these three founders and their related genotypes by almond breeders resulted in a loss of genetic variability and an increase of inbreeding. This situation was even more evident when we studied the breeding strategies to introduce self-compatibility in new almond varieties. The use of the cultivar 'Tuono' as a source of self-compatibility has been a common practice in most breeding programs, creating a bottleneck effect. This lack of genetic variability could limit almond breeding progress. Now that this important trait is already present in most breeding programs and can be selected with molecular markers, breeders can look for new variability in the high genetic variability available in the genetic pool of local cultivars from different origins. In addition, new strategies based on whole genome selection with markers, as the marker assisted introgression strategy (MAI) developed in peach (Serra et al., 2016; Kalluri et al., 2022), could help to introduce the high genetic variability available in the wild almond (Gradziel, 2022) in a more efficient way. Finally, other breeding strategies based on whole genome selection as the resynthesis (Eduardo et al., 2020) can contribute to introgress key traits as self-compatibility, in local cultivars with a high value, as 'Marcona', keeping a very similar genomic background. All these strategies can help to maximizing the genetic gain while minimize the loss of genetic diversity.

Currently, three almond traits are suitable for marker-assisted selection (MAS), self-compatibility, sweet kernel and blooming time (discussed in Section I.5.1). To increase the number of genes that could be implement in the MAS pipeline, in Chapter 2, the genetic inheritance of several kernel quality traits has been studied using a gene mapping approach in the interspecific segregating population between the cultivars 'Marcona' and 'Marinada'. The use of the almond 60K SNP array combined with novel bioinformatic pipelines allowed us to build a high quality and highly saturated linkage map. The QTLs reported here, will allow the implementation of efficient MAS strategies applied to kernel quality traits. Of particular importance were the QTLs found for symmetry, as they were reported for the first time in almond. Additionally, the high correlation found between traits measured with conventional methods and the more efficient image analysis methods and the fact that the same QTLs were identified open the door to implement these methods in almond breeding programs as a new phenotyping tool. Another important innovation carried out in chapter 2 was the use of lsmean data instead of raw phenotypic data. This phenotypic data modelization allow researchers to minimize the bias produced in the data by uncontrolled variables (like the changing environment), having a more accurate phenotypic data. This is an approach already mainstream in other trait-loci analyses such as GWAS, but never applied in QTL mapping before.

To further understand the genetic variability present in a wide collection of almond germplasm and to identify molecular markers associated to other important traits, in Chapter 3, we carried out a genetic structure analysis and non-additive GWAS in a set of 243 almond accessions from from 21 countries and five continents. Our results strongly supported the subdivision of these accessions into five ancestral groups. Each group was formed by accessions with a common geographical origin, agreeing with the archaeological and historical evidence that separate almond dissemination into four phases: Asiatic, Mediterranean, Californian and southern hemisphere. Through a homozygosity analysis, we detected low levels of inbreeding in most of the accessions under study. However, high levels of inbreeding were detected in some modern cultivars, agreeing with the results found in Chapter 1, where we concluded that breeding practices could be increasing inbreeding in almond. Also, signals of domestication were detected in chromosomes one, four and five. Among the 13 QTLs detected in Chapter 3, only one had an additive effect. This indicated that non-additive effects could be the main source of genotype-phenotype interactions in almond and other *Prunus* species. Finally, the use of the peachGCN, developed in this thesis, allowed us to propose four candidate genes for the main QTLs mapped.

Finally, in Chapter 4, a new tool to further increase the accuracy of *Prunus* gene function prediction was developed. For that, we constructed four GCNs from publicly available RNA-Seq data, we evaluated the performance of every GCN and finally, we validated the GCN with the best performance. To validate the performance of the GCN, we selected two well-characterized genes responsible for fruit flesh softening in peach, the endopolygalacturonases *PpPG21* and *PpPG22*. MF subnetwork, constituted by the genes coexpressing with *PpPG21* and *PpPG22*, was mainly formed by genes involved in cell wall organization and biogenesis, with expression regulated by ripening-related phytohormones such as ethylene, auxin and MeJA. Additionaly, we found in MF subnetwork 25 genes previously reported as involved in softening, some taking part in key steps of that process. These results demonstrated that the MF subnetwork was closely related to peach fruit softening and therefore to the function of *PpPG21* and *PpPG22*. The next step on this research line will be making the tool more accessible to all researchers, no matter their background in bioinformatics. For that, we are developing a user-friendly and online interface for this tool, where researchers will interact with the PeachGCN in an easy way.

Taken together, in this thesis we have studied the natural and human-driven dissemination of the modern almond throughout the world and followed the breeding tendencies applied by almond breeding worldwide. We have applied novel methods of phenotyping and data modelization. We have mapped numerous QTLs using different trait-loci analysis such as QTL mapping or GWAS and proposed responsible genes for key breeding traits. And finally, we have created a new and powerful tool for predicting gene function in almond and other *Prunus* species.

This is just an example of how genomics and bioinformatics can change almond breeding. Traditional breeding methods rely heavily on phenotypic evaluations, which can be time-consuming, labor-intensive, and prone to environmental influences. With the advent of genomics and bioinformatics, breeders can now identify genetic markers associated with desirable traits, allowing for MAS. As these tools continue to evolve and become more accessible, their application in almond breeding will undoubtedly play a pivotal role in shaping the future of almond cultivation. Additionally, the integration of genomic data with other types of omics data, such as transcriptomics, proteomics, or metabolomics, will offer a more comprehensive understanding of the complex interactions within the almond genome and its

response to environmental factors. This systems-level understanding will enhance the capacity to predict the performance of specific genotypes under varying conditions and will assist breeders in making well-informed decisions regarding almond variety development.

General Conclusions

6. General Conclusions

1. The pedigree analysis of 220 almond accessions of different origins detected that two main breeding lineages based on only three cultivars, 'Nonpareil' and 'Cristomorto'-'Tuono' have dominated modern breeding worldwide. This situation can increase the risk of almond inbreeding depression in future generations.

2. Additional analyses based on genome-wide data are needed to more accurately determine the levels of inbreeding and the loss of genetic variability among almond breeding programs worldwide.

3. The use of the recently developed almond 60K SNP array combined with novel bioinformatic protocols allowed the development of a high quality and highly saturated linkage map of a F1 population coming from the cross 'Marcona' x 'Marinada'.

4. The use of image analysis tools for kernel analysis presented a high correlation with convention time-consuming methods, suggesting that these methods should be implemented in almond breeding programs to increase phenotyping efficiency.

5. Using the high quality map developed in the F1 population derived from the cross 'Marcona' x 'Marinada', 12 QTLs related to kernel quality traits were mapped. From those QTLs, two were associated to kernel weight, five to shape-related traits, one to crack-out, two to color traits and two to chemical traits.

6. A QTL for kernel symmetry has been reported for the first time in almond, allowing the implementation of MAS applied to this important trait in breeding.

7. A genetic structure analysis of 243 almond accessions strongly supported the subdivision of the population in five ancestral populations, agreeing with the archaeological and historical evidence that separate modern almond dissemination into four phases: Asiatic, Mediterranean, Californian and southern hemisphere.

8. A genome-wide association study (GWAS) looking for additive and non-additive phenotype-genotype interactions allowed us to map 13 QTLs associated to the traits under study. From those QTLs, two were associated to nut weight, seven to crack-out, two to double kernels and one to blooming time.

9. From the 13 QTLs found for the traits of interest, only one had an additive effect. These results suggest that non-additive effects could be the major source of genotype-phenotype interactions in almond and other Prunus species.

10. The fast linkage desequilibrium decay observed and the use of the PeachGCN developed in this thesis allowed the identification of four positional candidate genes for the main QTLs found in the GWAS. *Prudul26A013473* was proposed as the gene responsible for qP-CRO2, a QTL related to crack-out percentage. *Prudul26A012082* and *Prudul26A017782* were proposed as the responsible genes for qP-DK7.1 and qP-DK7.2 respectively, both QTLs associated to double kernel percentage. *Prudul26A000954* was proposed as the responsible gene for qP-BLO2, a QTL associated to blooming time.

11. Four gene coexpression networks (GCNs) were created with aggregated and non-aggregated methods. We measured the performance of every GCN using the GBA neighbor voting, a machine learning algorithm based on the 'Guilty-by-association' principle. Over the four GCNs, COO300 was the one with the best performance.

12. Using two well-characterized genes in peach, endopolygalacturonases *PpPG21* and *PpPG22*, we were able to validate the usefulness of COO300 as a tool for predicting gene function. This GCN, named as the PeachGCN v1, will help *Prunus* researchers overcome the intrinsic limitations of working with crop tree species, prioritize research lines and outline new ones.

7. General References

Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., et al. (2020). Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *Plant J.* 101, 455–472. doi: 10.1111/tpj.14538.

Alonso, J., and Socias i Company, R. (2005). Self-incompatibility expression in self-compatible almond genotypes may be due to inbreeding. *J. Am. Soc. Hortic. Sci.* 130, 865–869.

Alonso Segura, J. M., Socias i Company, R., and Kodad, O. (2017). Late-blooming in almond: A controversial objective. *Sci. Hortic. (Amsterdam).* 224, 61–67. doi: 10.1016/j.scienta.2017.05.036.

Antanaviciute, L., Harrison, N., Battey, N. H., and Harrison, R. J. (2015). An inexpensive and rapid genomic DNA extraction protocol for rosaceous species. *J. Hortic. Sci. Biotechnol.* 90, 427–432. doi: 10.1080/14620316.2015.11513205.

Arulsekar, S., Parfitt, D. E., and Kester, D. E. (1986). Comparison of isozyme variability in peach and almond cultivars. *J. Hered.* 77, 272–274. doi: 10.1093/oxfordjournals.jhered.a110235.

Arús, P., Gradziel, T., Oliveira, M. M., and Tao, R. (2009). "Genomics of Almond," in *Genetics and Genomics of Rosaceae* (New York, NY: Springer New York), 187–219. doi: 10.1007/978-0-387-77491-6_9.

Arus, P., Olarte, C., Romero, M., and Vargas, F. (1994). Linkage analysis of ten isozyme genes in F1 segregating almond progenies. *J. Am. Soc. Hortic. Sci.* 119, 339–344. doi: 10.21273/jashs.119.2.339.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet. 2000 251* 25, 25–29. doi: 10.1038/75556.

Australian Almond Board (2019). Almond insights 2018-19.

Australian Almond Board (2022). Almond insights 2021/22.

Ballester, J., Socias I Company, R., Arus, P., and De Vicente, M. C. (2001). Genetic mapping of a major gene delaying blooming time in almond. *Plant Breed.* 120, 268–270. doi: 10.1046/j.1439-0523.2001.00604.x.

Barreca, D., Nabavi, S. M., Sureda, A., Rasekhian, M., Raciti, R., Silva, A. S., et al. (2020). Almonds (Prunus Dulcis Mill. D. A. Webb): A Source of Nutrients and Health-Promoting Compounds. *Nutr. 2020, Vol. 12, Page 672* 12, 672. doi: 10.3390/NU12030672.

Bartolozzi, F., Warburton, M. L., Arulsekar, S., and Gradziel, T. M. (1998). Genetic characterization and relatedness among california almond cultivars and breeding lines detected by randomly amplified polymorphic DNA (RAPD) analysis. *J. Am. Soc. Hortic. Sci.* 123, 381–387.

Batlle, I., Dicenta, F., Gradziel, T. M., Wirthensohn, M., Duval, H., and Vargas, F. J. (2017). "Classical genetics and breeding," in *Almonds: Botany, production and uses* (Boston: CABI), 111–148.

Beauvieux, R., Wenden, B., and Dirlewanger, E. (2018). Bud Dormancy in Perennial Fruit Tree Species: A Pivotal Role for Oxidative Cues. *Front. plant Sci.* 9, 657. doi: 10.3389/fpls.2018.00657.

Becerra-Tomás, N., Paz-Graniel, I., Kendall, C., Kahleova, H., Rahelić, D., Sievenpiper, J. L., et al. (2019). Nut consumption and incidence of cardiovascular diseases and cardiovascular disease mortality: A meta-analysis of prospective cohort studies. *Nutr. Rev.* 77, 691–709. doi: 10.1093/nutrit/nuz042.

Browicz, K., and Zohary, D. (1996). The genus Amygdalus L.(Rosaceae): species relationships, distribution and evolution under domestication. *Genet. Resour. Crop Evol.* 43, 229–247.

Brukental, H., Doron-Faigenboim, A., Bar-Ya'akov, I., Harel-Beja, R., Attia, Z., Azoulay-Shemer, T., et al. (2021). Revealing the Genetic Components Responsible for the Unique Photosynthetic Stem Capability of the Wild Almond Prunus arabica (Olivier) Meikle. *Front. Plant Sci.* 12, 1–14. doi: 10.3389/fpls.2021.779970.

Byrne, D. (1989). Inbreeding, coancestry, and founding clones of Japanese-type plums of California and the southeastern United States. *J. Am. Soc. Hortic. Sci.*

Byrne, D. H. (1990). Isozyme Variability in Four Diploid Stone Fruits Compared with Other Woody Perennial Plants. *J. Hered.* 81, 68–71. doi: 10.1093/oxfordjournals.jhered.a110927.

Cabrita, L., Apostolova, E., Neves, A., Marreiros, A., and Leitão, J. (2014). Genetic diversity assessment of the almond (*Prunus dulcis* (Mill.) D.A. Webb) traditional germplasm of Algarve, Portugal, using molecular markers. *Plant Genet. Resour. Characterisation Util.* 12, S164–S167. doi: 10.1017/S1479262114000471.

Californian Almond Board (2019). Almond Almanac 2019.

Californian Almond Board (2022). Almond Almanac 2022.

Chin, S. W., Shaw, J., Haberle, R., Wen, J., and Potter, D. (2014). Diversification of almonds, peaches, plums and cherries - Molecular systematics and biogeographic history of Prunus (Rosaceae). *Mol.*

Phylogenet. Evol. 76, 34–48. doi: 10.1016/j.ympev.2014.02.024.
Choi, C., and Kappel, F. (2004). Inbreeding, coancestry, and founding clones of sweet cherries from North America. *J. Am. Soc. Hortic. Sci.* 129, 535–543. doi: 10.21273/JASHS.129.4.0535.
International Nut and Dried Fruits Council (2019). Nuts & dried fruits statistical yearbook 2018/2019.
International Nut and Dried Fruits Council (2021). Nuts & dried fruit statistical yearbook 2020/2021.
D'Amico-Willman, K. M., Ouma, W. Z., Meulia, T., Sideli, G. M., Gradziel, T. M., and Fresnedo-Ramírez, J. (2022). Whole-genome sequence and methylome profiling of the almond [Prunus dulcis (Mill.) D.A. Webb] cultivar "Nonpareil." *G3 Genes, Genomes, Genet.* 12. doi: 10.1093/g3journal/jkac065.
D'Amico-Willman, K. M., Sideli, G. M., Allen, B. J., Anderson, E. S., Gradziel, T. M., and Fresnedo-Ramírez, J. (2022). Identification of Putative Markers of Non-infectious Bud Failure in Almond [Prunus dulcis (Mill.) D.A. Webb] Through Genome Wide DNA Methylation Profiling and Gene Expression Analysis in an Almond × Peach Hybrid Population. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.804145.
Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/BIOINFORMATICS/BTR330.
Debuse, C. J., Shaw, D. V., and Dejong, T. M. (2005). Response to inbreeding of seedling traits in a *Prunus domestica* L. breeding population. *J. Am. Soc. Hortic. Sci.* 130, 904–911. doi: 10.21273/JASHS.130.6.904.
Delplancke, M., Alvarez, N., Benoit, L., Espíndola, A., Joly, H. I., Neuenschwander, S., et al. (2013). Evolutionary history of almond tree domestication in the Mediterranean basin. *Mol. Ecol.* 22, 1092. doi: 10.1111/mec.12129.
Delplancke, M., Alvarez, N., Espíndola, A., Joly, H., Benoit, L., Brouck, E., et al. (2012). Gene flow among wild and domesticated almond species: Insights from chloroplast and nuclear markers. *Evol. Appl.* 5, 317–329. doi: 10.1111/j.1752-4571.2011.00223.x.
Delplancke, M., Yazbek, M., Arrigo, N., Espíndola, A., Joly, H., and Alvarez, N. (2016). Combining conservative and variable markers to infer the evolutionary history of Prunus subgen. Amygdalus s.l. under domestication. *Genet. Resour. Crop Evol.* 63, 221–234. doi: 10.1007/s10722-015-0242-6.
Denisov, V. (1988). Almond genetic resources in the USSR and their use in production and breeding. *Acta Hortic.* 224, 299–306.
Di Guardo, M., Farneti, B., Khomenko, I., Modica, G., Mosca, A., Distefano, G., et al. (2021). Genetic characterization of an almond germplasm collection and volatilome profiling of raw and roasted kernels. *Hortic. Res.* 8, 27. doi: 10.1038/s41438-021-00465-7.
Dicenta, F., Sánchez-Pérez, R., Rubio, M., Egea, J., Batlle, I., Miarnau, X., et al. (2015). The origin of the self-compatible almond "Guara." *Sci. Hortic. (Amsterdam).* 197, 1–4. doi: 10.1016/j.scienta.2015.11.005.
Dirlewanger, E., Graziano, E., Joobeur, T., Garriga-Calderé, F., Cosson, P., Howad, W., et al. (2004). Comparative mapping and marker-assisted selection in Rosaceae fruit crops. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9891–9896. doi: 10.1073/pnas.0307937101.
Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The Molecular Genetics of Crop Domestication. *Cell* 127, 1309–1321. doi: 10.1016/J.CELL.2006.12.006.
Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.* 3, 43–52. doi: 10.1038/nrg703.
Donoso, J. M., Eduardo, I., Picañol, R., Batlle, I., Howad, W., Aranzana, M. J., et al. (2015). High-density mapping suggests cytoplasmic male sterility with two restorer genes in almond × peach progenies. *Hortic. Res.* 2. doi: 10.1038/hortres.2015.16.
Duval, H., Coindre, E., Ramos-Onsins, S. E., Alexiou, K. G., Rubio-Cabetas, M. J., Martínez-García, P. J., et al. (2023). Development and Evaluation of an AxiomTM 60K SNP Array for Almond (Prunus dulcis). *Plants* 12. doi: 10.3390/plants12020242.
Eduardo, I., Alegre, S., Alexiou, K. G., and Arús, P. (2020). Resynthesis: Marker-Based Partial Reconstruction of Elite Genotypes in Clonally-Reproducing Plant Species. *Front. Plant Sci.* 11, 1–8. doi: 10.3389/fpls.2020.01205.
Elhamzaoui, A., Oukabli, A., Charafi, J., and Moumni, M. (2012). Assessment of genetic diversity of Moroccan cultivated almond (*Prunus dulcis* Mill. DA Webb) in its area of extreme diffusion, using nuclear microsatellites. *Am. J. Plant Sci.* 3, 1294–1303.
Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024.
Evreinoff, V. A. (1958). Contribution a letude de l'amandier. *Fruits et primeurs Afrique* 28, 99–104.

Fernández i Martí, A., Font i Forcada, C., Kamali, K., Rubio-Cabetas, M. J., Wirthensohn, M., and Socias i Company, R. (2015). Molecular analyses of evolution and population structure in a worldwide almond [*Prunus dulcis* (Mill.) D.A. Webb syn. *P. amygdalus* Batsch] pool assessed by microsatellite markers. *Genet. Resour. Crop Evol.* 62, 205–219. doi: 10.1007/s10722-014-0146-x.

Fernández i Martí, A., Font i Forcada, C., and Socias i Company, R. (2013). Genetic analysis for physical nut traits in almond. *Tree Genet. Genomes* 9, 455–465. doi: 10.1007/s11295-012-0566-8.

Fernández i Martí, À., Howad, W., Tao, R., Segura, J. M. A., Arús, P., and Socias i Company, R. (2011). Identification of quantitative trait loci associated with self-compatibility in a Prunus species. *Tree Genet. Genomes* 7, 629–639. doi: 10.1007/s11295-010-0362-2.

Font i Forcada, C., i Martí, À. F., and I Company, R. S. (2012). Mapping quantitative trait loci for kernel composition in almond. *BMC Genet.* 13, 1–9. doi: 10.1186/1471-2156-13-47/TABLES/3.

Font i Forcada, C., Oraguzie, N., Reyes-Chin-Wo, S., Espiau, M. T., Company, R. S. I., and Fernández I Martí, A. (2015a). Identification of genetic loci associated with quality traits in almond via association mapping. *PLoS One* 10. doi: 10.1371/journal.pone.0127656.

Font i Forcada, C., Velasco, L., Socias i Company, R., and Fernández i Martí, Á. (2015b). Association mapping for kernel phytosterol content in almond. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00530.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196, 973–983. doi: 10.1534/genetics.113.160572.

Fu, W., da Silva Linge, C., Lawton, J. M., and Gasic, K. (2022). Feasibility of genomic prediction for brown rot (*Monilinia* spp.) resistance in peach. *Fruit Res.* 2, 1–12. doi: 10.48130/frures-2022-0002.

García-Gómez, B. E., Ruiz, D., Salazar, J. A., Rubio, M., Martínez-García, P. J., and Martínez-Gómez, P. (2020). Analysis of Metabolites and Gene Expression Changes Relative to Apricot (Prunus armeniaca L.) Fruit Quality During Development and Ripening. *Front. Plant Sci.* 0, 1269. doi: 10.3389/FPLS.2020.01269.

Godini, A. (2000). About the possible relationships between Amygdalus webbi Spach and Amygdalus communis L. *NUCIS Newsl.* 9, 17–19.

Gómez, E. M., Buti, M., Sargent, D. J., Dicenta, F., and Ortega, E. (2019a). Transcriptomic analysis of pollen-pistil interactions in almond (Prunus dulcis) identifies candidate genes for components of gametophytic self-incompatibility. *Tree Genet. Genomes* 15. doi: 10.1007/s11295-019-1360-7.

Gómez, E. M., Dicenta, F., Batlle, I., Romero, A., and Ortega, E. (2019b). Cross-incompatibility in the cultivated almond (Prunus dulcis): Updating, revision and correction. *Sci. Hortic. (Amsterdam).* 245, 218–223. doi: 10.1016/j.scienta.2018.09.054.

Gonzalo, M. J., Brewer, M. T., Anderson, C., Sullivan, D., Gray, S., and Van Der Knaap, E. (2009). Tomato fruit shape analysis using morphometric and morphology attributes implemented in tomato analyzer software program. *J. Am. Soc. Hortic. Sci.* 134, 77–87. doi: 10.21273/jashs.134.1.77.

Goonetilleke, S. N., March, T. J., Wirthensohn, M. G., Arús, P., Walker, A. R., and Mather, D. E. (2018). Genotyping by sequencing in almond: SNP discovery, linkage mapping, and marker design. *G3 Genes, Genomes, Genet.* 8, 161–172. doi: 10.1534/g3.117.300376.

Gouta, H., Ksia, E., Buhner, T., Moreno, M. Á., Zarrouk, M., Mliki, A., et al. (2010). Assessment of genetic diversity and relatedness among Tunisian almond germplasm using SSR markers. *Hereditas* 147, 283–292. doi: 10.1111/j.1601-5223.2009.02147.x.

Gradziel, T. ., Martínez-Gómez, P., Dicenta, F., and Kester, D. . (2001). The utilization of related prunus species for almond variety improvement. *J. Am. Pomol. Soc.* 55. Available at: https://search.proquest.com/docview/209767315?pq-origsite=gscholar [Accessed May 17, 2020].

Gradziel, T., Beres, W., and Pelletreau, K. (1993). Inbreeding in California canning clingstone peach cultivars. *Fruit Var. J.*

Gradziel, T. M. (2022). Transfer of Self-Fruitfulness to Cultivated Almond from Peach and Wild Almond. *Horticulturae* 8. doi: 10.3390/horticulturae8100965.

Gradziel, T. M., Curtis, R., and Socias i Company, R. (2017). "Production and growing regions," in *Almonds: Botany, production and uses* (Boston: CABI), 70–86.

Gradziel, T. M., and Martínez-Gómez, P. (2002). Shell seal breakdown in almond is associated with the site of secondary ovule abortion. *J. Am. Soc. Hortic. Sci.* 127, 69–74. doi: 10.21273/jashs.127.1.69.

Gradziel, T. M., and Socias i Company, R. (2017). *Almonds: Botany, production and uses*.

Grasselly, C. (1976a). Les espèces sauvages d'amandier. *Options méditerranéennes* 32, 28–43.

Grasselly, C. (1976b). Mise en évidence de quelques types autocompatibles parmi les cultivars d'amandier (P. amygdalus Batsch) de la population des Pouilles. Available at: https://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=PASCAL7638005468 [Accessed April

22, 2020].
Grasselly, C. (1978). Observations sur l'utilization d'un mutant d'amandier a floraison tardize dans un programme d'hybridizaiton. *Ann. Amelior. Plant.* 28, 685–695.
Grasselly, C., and Crossa-Raynaud, P. (1980). *The almond tree*. Paris: Maisonneuve et Larose.
Grasselly, C., and Gall, H. (1967). Étude sur la possibilité de combinaison de quelques caractères agronomiques chez l'amandier Cristomorto hybridé par trois autres variétés. *Ann. Amélior. Plant.* 17, 83–91.
Grasselly, C., and Olivier, G. (1981). Difficulté de survie de jeunes semis d'amandiers dans certaines descendances. *Options Mediterr.* Available at: http://ressources.ciheam.org/om/pdf/s01/CI010763.pdf [Accessed April 22, 2020].
Grassely, C. (1976). Origine et évolution de l'amandier cultivé. *Options Méditerr* 32, 45–49.
Gross, B. L., and Olsen, K. M. (2010). Genetic perspectives on crop domestication. *Trends Plant Sci.* 15, 529–537. doi: 10.1016/J.TPLANTS.2010.05.008.
Guo, C., Wei, Y., Yang, B., Ayup, M., Li, N., Liu, J., et al. (2021). Developmental transcriptome profiling uncovered carbon signaling genes associated with almond fruit drop. *Sci. Rep.* 11, 1–12. doi: 10.1038/s41598-020-69395-z.
Halász, J., Kodad, O., Galiba, G. M., Skola, I., Ercisli, S., Ledbetter, C. A., et al. (2019). Genetic variability is preserved among strongly differentiated and geographically diverse almond germplasm: an assessment by simple sequence repeat markers. *Tree Genet. Genomes* 15, 1–13. doi: 10.1007/s11295-019-1319-8.
Hamadeh, B., Chalak, L., Coppens d'Eeckenbrugge, G., Benoit, L., and Joly, H. I. (2018). Evolution of almond genetic diversity and farmer practices in Lebanon: Impacts of the diffusion of a graft-propagated cultivar in a traditional system based on seed-propagation. *BMC Plant Biol.* 18, 1–18. doi: 10.1186/s12870-018-1372-8.
Hansen, J. M. (1991). The Palaeoethnobotany of Franchthi Cave. Excavations at Franchthi Cave, Greece. *Paléorient* 18–1, 135–137.
Hardner, C. M., Fikere, M., Gasic, K., da Silva Linge, C., Worthington, M., Byrne, D., et al. (2022). Multi-environment genomic prediction for soluble solids content in peach (Prunus persica). *Front. Plant Sci.* 13, 1–18. doi: 10.3389/fpls.2022.960449.
Jégu, T., Latrasse, D., Delarue, M., Hirt, H., Domenichini, S., Ariel, F., et al. (2014). The BAF60 Subunit of the SWI/SNF Chromatin-Remodeling Complex Directly Controls the Formation of a Gene Loop at FLOWERING LOCUS C in Arabidopsis. *Plant Cell* 26, 538–551. doi: 10.1105/tpc.113.114454.
Jiang, X., Liu, K., Peng, H., Fang, J., Zhang, A., Han, Y., et al. (2023). Comparative network analysis reveals the dynamics of organic acid diversity during fruit ripening in peach (Prunus persica L. Batsch). *BMC Plant Biol.* 23, 1–14. doi: 10.1186/S12870-023-04037-W/TABLES/1.
Joobeur, T., Periam, N., De Vicente, M. C., King, G. J., and Arus, P. (2000). Development of a second generation linkage map for almond using RAPD and SSR markers. *Genome* 43, 649–655. doi: 10.1139/g00-040.
Joobeur, T., Viruel, M. A., De Vicente, M. C., Jáuregui, B., Ballester, J., Dettori, M. T., et al. (1998). Construction of a saturated linkage map for Prunus using an almond x peach F2 progeny. *Theor. Appl. Genet.* 97, 1034–1041. doi: 10.1007/s001220050988.
Jurado-Ruiz, F., Onielfa, C., Dujak, C., Pradas, N., Pérez de los Cobos, F., and Aranzana, M. J. (2023). Shape Analyzer: An application for fruit morphology phenotyping. *Manuscript in preparation.*
Kalluri, N., Serra, O., Donoso, J. M., Picañol, R., Howad, W., Eduardo, I., et al. (2022). Construction of a collection of introgression lines of "Texas" almond DNA fragments in the "Earlygold" peach genetic background. *Hortic. Res.* 9, 1–11. doi: 10.1093/hr/uhac070.
Kao, T., and Tsukamoto, T. (2004). The Molecular and Genetic Bases of S-RNase-Based Self-Incompatibility. *Plant Cell* 16, 72–83. doi: 10.1105/tpc.016154.S-RNase-Based.
Kardos, M., Luikart, G., and Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity (Edinb).* 115, 63–72. doi: 10.1038/hdy.2015.17.
Keneni, G., Bekele, E., Imtiaz, M., and Dagne, K. (2012). Genetic vulnerability of modern crop cultivars: causes, mechanism and remedies. *Int. J. Plant Res.* 2, 69–79. doi: 10.5923/j.plant.20120203.05.
Kester, D. E. (1965). Inheritance of time of bloom in certain progenies of almond. *Proc. Am. Soc. Hortic. Sci.* 87, 214–221.
Kester, D. E., and Gradziel, T. M. (1996). "Fruit breeding," in, eds. J. Janick and J. N. Moore (Wiley), 1–97.
Kester, D. E., Gradziel, T. M., and Grasselly, C. (1991). Almonds (Prunus). *Acta Hortic.*, 701–760. doi:

10.17660/actahortic.1991.290.16.
Kislev, M. E., Nadel, D., and Carmi, I. (1992). Epipalaeolithic (19,000 BP) cereal and fruit diet at Ohalo II, Sea of Galilee, Israel. *Rev. Palaeobot. Palynol.* 73, 161–166. doi: 10.1016/0034-6667(92)90054-K.
Kislev, M., Melamed, Y., Simchoni, O., and Marmorstein, M. (1997). Computerized key of grass grains of the Mediterranean basin. *Lagascalia* 19 (1–2), 289–294.
Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* 9, 1–9. doi: 10.1186/1746-4811-9-29/FIGURES/4.
Ladizinsky, G. (1999). On the origin of almond. *Genet. Resour. Crop Evol.* 46, 143–147. doi: 10.1023/A:1008690409554.
Lansari, A., Kester, D. E., and Iezzoni, A. F. (1994). Inbreeding, coancestry, and founding clones of almonds of California, Mediterranean shores, and Russia. *J. Am. Soc. Hortic. Sci.* 119, 1279–1285. doi: 10.21273/JASHS.119.6.1279.
Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* 25, 1–18. doi: 10.18637/JSS.V025.I01.
Li, X., Wang, J., Su, M., Zhang, M., Hu, Y., Du, J., et al. (2023). Multiple-statistical genome-wide association analysis and genomic prediction of fruit aroma and agronomic traits in peaches. *Hortic. Res.* doi: 10.1093/hr/uhad117.
López, M., Vargas, F. J., and Batlle, I. (2006). Self-(in)compatibility almond genotypes: A review. *Euphytica* 150, 1–16. doi: 10.1007/s10681-005-9009-z.
Lotti, C., Minervini, A. P., Delvento, C., Losciale, P., Gaeta, L., Sánchez-Pérez, R., et al. (2023). Detection and distribution of two dominant alleles associated with the sweet kernel phenotype in almond cultivated germplasm. *Front. Plant Sci.* 14, 1–7. doi: 10.3389/fpls.2023.1171195.
Marchese, A., Bošković, R. I., Martínez-García, P. J., and Tobutt, K. R. (2008). The origin of the self-compatible almond 'Supernova.' *Plant Breed.* 127, 105–107. doi: 10.1111/j.1439-0523.2008.01421.x.
Marrano, A., Martínez-García, P. J., Bianco, L., Sideli, G. M., Di Pierro, E. A., Leslie, C. A., et al. (2019). A new genomic tool for walnut (Juglans regia L.): development and validation of the high-density Axiom™ J. regia 700K SNP genotyping array. *Plant Biotechnol. J.* 17, 1027–1036.
Martínez-García, P. J., Dicenta, F., and Ortega, E. (2012). Anomalous embryo sac development and fruit abortion caused by inbreeding depression in almond (*Prunus dulcis*). *Sci. Hortic. (Amsterdam).* 133, 23–30. doi: 10.1016/j.scienta.2011.10.001.
Martínez-Gómez, P., Arulsekar, S., Potter, D., and Gradziel, T. M. (2003). An extended interspecific gene pool available to peach and almond breeding as characterized using simple sequence repeat (SSR) markers. *Euphytica* 131, 313–322. doi: 10.1023/A:1024028518263.
Martínez-Gómez, P., Prudencio, A. S., Gradziel, T. M., and Dicenta, F. (2017). The delay of flowering time in almond: a review of the combined effect of adaptation, mutation and breeding. *Euphytica* 213. doi: 10.1007/s10681-017-1974-5.
McCreery, D. W. (1979). Flotation of the Bab edh-Dhra and Numeira plant remains. *Annu. Am. Sch. Orient. Res.* 46, 165.
Meinke, D. (2020). Genome-wide identification of EMBRYO. *New Phytol.* 226, 306. doi: 10.1111/nph.16071.
Mendel, G. (1865). Experiments in plant hybridization.
Mitsuda, N., Iwase, A., Yamamoto, H., Yoshida, M., Seki, M., Shinozaki, K., et al. (2007). NAC Transcription Factors, NST1 and NST3, Are Key Regulators of the Formation of Secondary Walls in Woody Tissues of Arabidopsis. *Plant Cell* 19, 270–280. doi: 10.1105/tpc.106.047043.
Mnejja, M., Garcia-Mas, J., Audergon, J. M., and Arús, P. (2010). Prunus microsatellite marker transferability across rosaceous crops. *Tree Genet. Genomes* 6, 689–700. doi: 10.1007/s11295-010-0284-z.
Moll, L., Baró, A., Montesinos, L., Badosa, E., Bonaterra, A., and Montesinos, E. (2022). Induction of Defense Responses and Protection of Almond Plants Against Xylella fastidiosa by Endotherapy with a Bifunctional Peptide. *Phytopathology* 112, 1907–1916. doi: 10.1094/PHYTO-12-21-0525-R.
Mousavi, S., Alisoltani, A., Shiran, B., Fallahi, H., Ebrahimie, E., Imani, A., et al. (2014). De novo transcriptome assembly and comparative analysis of differentially expressed genes in Prunus dulcis Mill. in response to freezing stress. *PLoS One* 9, 1–13. doi: 10.1371/journal.pone.0104541.
Muranty, H., Denancé, C., Feugey, L., Crépin, J. L., Barbier, Y., Tartarini, S., et al. (2020). Using whole-genome SNP data to reconstruct a large multi-generation pedigree in apple germplasm. *BMC Plant Biol.* 20, 1–18. doi: 10.1186/s12870-019-2171-6.
Noiton, D., and Alspach, P. (1996). Founding Clones, Inbreeding, Coancestry, and Status Number of

Modern Apple Cultivars. *J. Am. Soc. Hortic. Sci.* 121, 773–782. Available at: https://journals.ashs.org/jashs/view/journals/jashs/121/5/article-p773.xml [Accessed October 18, 2019].

Nuts production in Spain (2021). Madrid Available at: https://www.mapa.gob.es/es/agricultura/temas/producciones-agricolas/analisisdelarealidadproductivafrutossecos2020_tcm30-584009.pdf.

Orduña-Rubio, L., Santiago, A., Navarro-Payá, D., Zhang, C., Wong, D. C. J., and Matus, J. T. (2023). Aggregated gene co-expression networks for predicting transcription factor regulatory landscapes in a non-model plant species. *bioRxiv*. doi: https://doi.org/10.1101/2023.04.24.538042.

Ortega, E., and Dicenta, F. (2003). Inheritance of self-compatibility in almond: Breeding strategies to assure self-compatibility in the progeny. *Theor. Appl. Genet.* 106, 904–911. doi: 10.1007/s00122-002-1159-y.

Ortega, E., Sutherland, B. G., Dicenta, F., Boskovic, R., and Tobutt, K. R. (2005). Determination of incompatibility genotypes in almond using first and second intron consensus primers: detection of new S alleles and correction of reported S genotypes. *Plant Breed.* 124, 188–196. doi: 10.1111/j.1439-0523.2004.01058.x.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/BIOINFORMATICS/BTG412.

Pavan, S., Delvento, C., Mazzeo, R., Ricciardi, F., Losciale, P., Gaeta, L., et al. (2021). Almond diversity and homozygosity define structure, kinship, inbreeding, and linkage disequilibrium in cultivated germplasm, and reveal genomic associations with nut and seed weight. *Hortic. Res.* 8, 1–12. doi: 10.1038/s41438-020-00447-1.

Pérez-Jordà, G., Alonso, N., Rovira, N., Figueiral, I., López-Reyes, D., Marinval, P., et al. (2021). The Emergence of Arboriculture in the 1st Millennium BC along the Mediterranean's "Far West." *Agronomy.* 11, 902. doi: 10.3390/agronomy11050902.

Pérez de los Cobos, F., García-Gómez, B. E., Orduña-Rubio, L., Batlle, I., Arús, P., Matus, J. T., et al. (2023). First large-scale peach gene coexpression network: A new tool for predicting gene function. *biorxiv.org*. doi: 10.1101/2023.06.22.546058.

Pérez de los Cobos, F., Martínez-García, P. J., Romero, A., Miarnau, X., Eduardo, I., Howad, W., et al. (2021). Pedigree analysis of 220 almond genotypes reveals two world mainstream breeding lines based on only three different cultivars. *Hortic. Res.* 8. doi: 10.1038/S41438-020-00444-4.

Prudencio, Á. S., Hoeberichts, F. A., Dicenta, F., Martínez-Gómez, P., and Sánchez-Pérez, R. (2021). Identification of early and late flowering time candidate genes in endodormant and ecodormant almond flower buds. *Tree Physiol.* 41, 589–605. doi: 10.1093/treephys/tpaa151.

Reyes, J. C. (2014). The many faces of plant SWI/SNF complex. *Mol. plant.* 7, 454–458. doi: 10.1093/mp/sst147.

Rikhter, A. (1972). Biological basis for the creation of almond cultivars and commercial orchards. *Akad. Nauk SSSR*.

Romani, I., Tadini, L., Rossi, F., Masiero, S., Pribil, M., Jahns, P., et al. (2012). Versatile roles of Arabidopsis plastid ribosomal proteins in plant growth and development. *The Plant journal.* 72, 922–934. doi: 10.1111/tpj.12000.

Romero, A. (2014). Almond quality requirements for industrial purposes - Its relevance for the future acceptance of new cultivars from breeding programs. in *Acta Horticulturae* (International Society for Horticultural Science), 213–220. doi: 10.17660/ActaHortic.2014.1028.34.

Rosenberg, M., Nesbitt, M., Redding, R. W., and Strasser, T. F. (1995). Some preliminary observations concerning early Neolithic subsistence behaviors in eastern Anatolia. *Anatolica* 21, 1–12.

Sánchez-Pérez, R., Dicenta, F., and Martínez-Gómez, P. (2012). Inheritance of chilling and heat requirements for flowering in almond and QTL analysis. *Tree Genet. Genomes* 8, 379–389. doi: 10.1007/s11295-011-0448-5.

Sánchez-Pérez, R., Howad, W., Dicenta, F., Arús, P., and Martínez-Gómez, P. (2007). Mapping major genes and quantitative trait loci controlling agronomic traits in almond. *Plant Breed.* 126, 310–318. doi: 10.1111/j.1439-0523.2007.01329.x.

Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Aiese Cigliano, R., Del Cueto, J., et al. (2019). Mutation of a bHLH transcription factor allowed almond domestication. *Science (80-.).* 364, 1095–1098. doi: 10.1126/science.aav8197.

Scorza, R., Mehlenbacher, S. ., and Lightner, G. . (1985). Inbreeding and coancestry of freestone peach cultivars of the eastern United States and implications for peach germplasm improvement. *J. Am.*

Soc. Hortic. Sci.

Serra, O., Donoso, J. M., Picañol, R., Batlle, I., Howad, W., Eduardo, I., et al. (2016). Marker-assisted introgression (MAI) of almond genes into the peach background: a fast method to mine and integrate novel variation from exotic sources in long intergeneration species. *Tree Genet. Genomes* 12, 1–13. doi: 10.1007/s11295-016-1056-1.

Shiran, B., Amirbakhtiar, N., Kiani, S., Mohammadi, S., Sayed-Tabatabaei, B. E., and Moradi, H. (2007). Molecular characterization and genetic relationship among almond cultivars assessed by RAPD and SSR markers. *Sci. Hortic. (Amsterdam).* 111, 280–292. doi: 10.1016/j.scienta.2006.10.024.

Sideli, G. M., Mather, D., Wirthensohn, M., Dicenta, F., Goonetilleke, S. N., Martínez-García, P. J., et al. (2023). Genome-wide association analysis and validation with KASP markers for nut and shell traits in almond (Prunus dulcis [Mill.] D.A.Webb). *Tree Genet. genomes.* 19, 13. doi: 10.1007/s11295-023-01588-9.

Silva, C., Garcia-Mas, J., Sánchez, A. M., Arús, P., and Oliveira, M. M. (2005). Looking into flowering time in almond (Prunus dulcis (Mill) D. A. Webb): The candidate gene approach. *Theor. Appl. Genet.* 110, 959–968. doi: 10.1007/s00122-004-1918-z.

Sjulin, T., and Dale, A. (1987). Genetic diversity of North American strawberry cultivars. *J. Am. Soc. Hortic. Sci.*, 375–385.

Socias i Company, R. (2011). Breeding self-compatible almonds. *Plant Breed. Rev.* 8, 313–338. doi: 10.1002/9781118061053.ch9.

Socias i Company, R. (2017). "Pollen-style (in)compatibility: development of autogamous cultivars," in *Almonds: Botany, production and uses* (Boston: CABI), 188–208.

Son, K. M., Kwon, S. Il, and Choi, C. (2012). Inbreeding, coancestry, and founding clones of apple cultivars released from Korea. *Hortic. Environ. Biotechnol.* 53, 404–409. doi: 10.1007/s13580-012-0012-8.

Sonneveld, T., Robbins, T. P., Bošković, R., and Tobutt, K. R. (2001). Cloning of six cherry self-incompatibility alleles and development of allele-specific PCR detection. *Theor. Appl. Genet.* 102, 1046–1055. doi: 10.1007/s001220000525.

Soto, A., Ruiz, K. B., Ziosi, V., Costa, G., and Torrigiani, P. (2012). Ethylene and auxin biosynthesis and signaling are impaired by methyl jasmonate leading to a transient slowing down of ripening in peach fruit. *J. Plant Physiol.* 169, 1858–1865. doi: 10.1016/J.JPLPH.2012.07.007.

Sutherland, B. G., Robbins, T. P., and Tobutt, K. R. (2004). Primers amplifying a range of Prunus S-alleles. *Plant Breed.* 123, 582–584. doi: 10.1111/j.1439-0523.2004.01016.x.

Swarup, S., Cargill, E. J., Crosby, K., Flagel, L., Kniskern, J., and Glenn, K. C. (2021). Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci.* 61, 839–852. doi: 10.1002/CSC2.20377.

Tamura, M., Ushijima, K., Sassa, H., Hirano, H., Tao, R., Gradziel, T. M., et al. (2000). Identification of self-incompatibility genotypes of almond by allele-specific PCR analysis. *Theor. Appl. Genet.* 101, 344–349.

Tavassolian, I., Rabiei, G., Gregory, D., Mnejja, M., Wirthensohn, M. G., Hunt, P. W., et al. (2010). Construction of an almond linkage map in an Australian population Nonpareil × Lauranne. *BMC Genomics* 11. doi: 10.1186/1471-2164-11-551.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939. doi: 10.1111/J.1365-313X.2004.02016.X.

Thudi, M., Palakurthi, R., Schnable, J. C., Chitikineni, A., Dreisigacker, S., Mace, E., et al. (2021). Genomic resources in plant breeding for sustainable agriculture. *J. Plant Physiol.* 257, 153351. doi: 10.1016/J.JPLPH.2020.153351.

Trainin, T., Brukental, H., Shapira, O., Attia, Z., Tiwari, V., Hatib, K., et al. (2022). Physiological characterization of the wild almond Prunus arabica stem photosynthetic capability. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.941504.

Trainotti, L., Tadiello, A., and Casadoro, G. (2007). The involvement of auxin in the ripening of climacteric fruits comes of age: the hormone plays a role of its own and has an intense interplay with ethylene in ripening peaches. *J. Exp. Bot.* 58, 3299–3308. doi: 10.1093/jxb/erm178.

Tsepilov, Y. A., Shin, S. Y., Soranzo, N., Spector, T. D., Prehn, C., Adamski, J., et al. (2015). Nonadditive effects of genes in human metabolomics. *Genetics* 200, 707–718. doi: 10.1534/genetics.115.175760.

Van De Wouw, M., Kik, C., Van Hintum, T., Van Treuren, R., and Visser, B. (2010). Genetic erosion in crops: concept, research results and challenges. *Plant Genet. Resour. Characterisation Util.* 8, 1–15. doi: 10.1017/S1479262109990062.

Van Ghelder, C., Lafargue, B., Dirlewanger, E., Ouassa, A., Voisin, R., Polidori, J., et al. (2010).

Characterization of the RMja gene for resistance to root-knot nematodes in almond: Spectrum, location, and interest for Prunus breeding. *Tree Genet. Genomes* 6, 503–511. doi: 10.1007/s11295-010-0268-z.

Van Zeist, W., and de Roller, G. J. (1992). The plant husbandry of aceramic Çayönü, SE Turkey. *Palaeohistoria* 33–34, 65–96.

Vargas, F. J., and Romero, M. A. (2001). Blooming time in almond progenies. *Options Méditerranéennes* 56, 29–34.

Velasco, D., Hough, J., Aradhya, M., and Ross-Ibarra, J. (2016). Evolutionary Genomics of Peach and Almond Domestication. *G3 (Bethesda).* 6, 3985–3993. doi: 10.1534/g3.116.032672.

Viruel, M. A., Messeguer, R., de Vicente J Garcia-Mas, M. C., Puigdom, P., Vargas P Arfis, nech F., Garcia-Mas, J., et al. (1995). A linkage map with RFLP and isozyme markers for almond. *Theor. Appl. Genet.* 91, 964–971. doi: https://doi.org/10.1007/BF00223907.

Wang, J. (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theor. Popul. Biol.* 107, 4–13. doi: 10.1016/j.tpb.2015.08.006.

Wang, Q., Cao, K., Li, Y., Wu, J., Fan, J., Ding, T., et al. (2023). Identification of co-expressed networks and key genes associated with organic acid in peach fruit. *Sci. Hortic. (Amsterdam).* 307, 111496. doi: 10.1016/j.scienta.2022.111496.

Wei, J., Wen, X., and Tang, L. (2017). Effect of methyl jasmonic acid on peach fruit ripening progress. *Sci. Hortic. (Amsterdam).* 220, 206–213. doi: 10.1016/J.SCIENTA.2017.03.004.

Weisdorf, J. L. (2005). From Foraging To Farming: Explaining The Neolithic Revolution. *J. Econ. Surv.* 19, 561–586. doi: 10.1111/J.0950-0804.2005.00259.X.

Wen, J., Berggren, S. T., Lee, C.-H., Ickert-Bond, S., and Yi, T.-S. (2008). Phylogenetic inferences in Prunus (Rosaceae) using chloroplast ndhF and nuclear ribosomal ITS sequences. *J. Syst. Evol.* 46, 322. doi: 10.3724/SP.J.1002.2008.08065.

Whalen, A., Gorjanc, G., and Hickey, J. M. (2020). AlphaFamImpute: High-accuracy imputation in full-sib families from genotype-by-sequencing data. *Bioinformatics* 36, 4369–4371. doi: 10.1093/bioinformatics/btaa499.

Willcox, G., Fornite, S., and Herveux, L. (2008). Early Holocene cultivation before domestication in northern Syria. *Veg. Hist. Archaeobot.* 17, 313–325. doi: 10.1007/s00334-007-0121-y.

Wood, M. N. (1925). *Almond varieties in the United States*. Washington D.C.: US Department of Agriculture.

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat.* 56, 330–338. doi: 10.1086/279872.

Xia, C., Wang, Y. J., Li, W. Q., Chen, Y. R., Deng, Y., Zhang, X. Q., et al. (2010). The Arabidopsis eukaryotic translation initiation factor 3, subunit F (AteIF3f), is required for pollen germination and embryogenesis. *Plant J.* 63, 189–202. doi: 10.1111/J.1365-313X.2010.04237.X.

Yahalom, A., Kim, T.-H., Roy, B., Singer, R., Von Arnim, A. G., and Chamovitz, D. A. (2008). Arabidopsis eIF3e is regulated by the COP9 signalosome and has an impact on development and protein translation. *The Plant journal.* 53, 300–311. doi: 10.1111/j.1365-313X.2007.03347.x.

Zaurov, D., Eisenman, S., Ford, T., Khokhlov, S., Kenjebaev, S., Shalpykov, K., et al. (2015). Genetic resources of almond species in the former USSR. *J. Am. Soc. Hortic. Sci.* 50, 18–29.

Zeinalabedini, M., Khayam-Nekoui, M., Grigorian, V., Gradziel, T. M., and Martínez-Gómez, P. (2010a). The origin and dissemination of the cultivated almond as determined by nuclear and chloroplast SSR marker analysis. *Sci. Hortic. (Amsterdam).* 125, 593–601. doi: 10.1016/j.scienta.2010.05.007.

Zeinalabedini, M., Khayam-Nekoui, M., Grigorian, V., Gradziel, T., and Martínez-Gómez, P. (2010b). The origin and dissemination of the cultivated almond as determined by nuclear and chloroplast SSR marker analysis. *Sci. Hortic. (Amsterdam).* 125, 593–601. doi: 10.1016/j.scienta.2010.05.007.

Zhong, R., Lee, C., Zhou, J., McCarthy, R. L., and Ye, Z. H. (2008). A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. *Plant Cell* 20, 2763–2782. doi: 10.1105/tpc.108.061325.

Zohary, D., and Hopf, M. (1993). *Domestication of plants in the old world*. 4th Editio. Oxford, UK: Clarendon Press.