UAB
Universitat Autònoma de Barcelona

Proteomic and peptidomic analyses of flower development in Arabidopsis

Raquel Álvarez Urdiola

UNIVERSITAT AUTÒNOMA DE BARCELONA

CENTER FOR RESEARCH IN AGRICULTURAL GENOMICS

DOCTORAL THESIS

Proteomic and peptidomic analyses of flower development in Arabidopsis

Raquel Álvarez Urdiola

Barcelona, 2023

UNIVERSITAT AUTÒNOMA DE BARCELONA

Animal Biology, Plant Biology and Ecology Department

Plant Biology and Biotechnology PhD Program

CENTER FOR RESEARCH IN AGRICULTURAL GENOMICS

(CRAG-CSIC-IRTA-UAB-UB)

Plant Development and Signal Transduction Program

Proteomic and peptidomic analyses of flower development in Arabidopsis

Dissertation submitted in fulfilment of the requirements for the degree of Doctor in Plant Biology and Biotechnology by the Universitat Autònoma de Barcelona.

Thesis directors
Dr José Luis Riechmann
Dr José Tomás Matus

PhD Candidate
Raquel Álvarez Urdiola

Barcelona, 2023

UAB
Universitat Autònoma de Barcelona

CRAG
CENTRE FOR RESEARCH IN AGRICULTURAL GENOMICS

This work was carried out at the laboratory of 'Gene regulatory networks in plant development', at the Centre for Research in Agricultural Genomics (CSIC-IRTA-UAB-UB) in Barcelona, under the supervision of Dr José Luis Riechmann and José Tomás Matus. Part of this dissertation was conducted during a three-month stay at the laboratory of Dr George Coupland at the Max Planck Institute for Plant Breeding Research (MPIPZ, Cologne, Germany).

"*Alone with our madness and favourite flower".*
Late echo – John Ashbery

"Las patas heridas,
las crines heladas,
dentro de los ojos
un puñal de plata".

Bodas de Sangre – Federico García Lorca

A Rafael, Aurora,
Fina y José

Abstract

The onset of flower formation and the process of flower development are excellent paradigms for developmental studies in plants as they are governed by complex regulatory networks. Extensive forward and reverse genetic analyses have led to the identification of many key regulatory genes that form part of these networks. Additional factors are now being characterized at the genome-wide level using multi-omics integrative methods. The genome-wide characterization of regulatory networks is key to understand, and eventually manipulate, the basis of plant development and physiology. However, and despite these advances, the emergent global, dynamic view of flower developmental processes is lacking an important component: the proteome. Current mass spectrometry methods now allow exploring in depth the composition of a proteome in its expression and complexity, its relationship with the transcriptome and even its dynamic posttranslational modifications. In recent years, it has also become evident that there is a substantial and still uncharted fraction of eukaryotic proteomes that is mainly composed of small, unannotated proteins and peptides (the 'non-conventional' peptidome), with functions yet to be discovered.

The Arabidopsis genome was sequenced 20 years ago. Since then, there have been plenty of public data concerning transcriptomes and their modulation throughout organ development, while also describing its plasticity in response to the environment. Conversely, the Arabidopsis proteome is far less comprehensively characterized. To fulfil this gap, a promising approach is the use of mass spectrometry methods for integrating its data with RNA sequencing. In this Thesis, the pAP1:AP1-GR *ap1 cal* Arabidopsis floral induction system was used to characterize genome expression at the proteome level throughout early Arabidopsis flower development, and its correlation to unbiased transcript expression data. Shotgun proteomic procedures (LC-MS/MS) and transcript profiling experiments (RNA-seq) were performed following a temporal series of five subsequent days after the activation of the flower developmental program. Almost 9,000 proteins and

around 23,000 transcripts were identified, of which 2,037 proteins and 8,125 transcripts showed significant abundance changes throughout the time-course. These experiments allowed to substantially expand the size or scope of the transcriptome (i.e., collection of genes) previously known to change its expression during the early stages of flower development; to identify RNA-protein pairs in which RNA and protein showed similar (correlated) or opposite (anti-correlated) expression trajectories and that are involved in different processes, such as photosynthesis, fatty acid metabolic processes, or amino acid biosynthesis; and, through the combined analysis of this novel transcriptomic dataset and previously published AP1 genome-wide binding data (ChIP-seq), identify novel putative AP1 direct targets.

Eukaryotic genomes contain many unannotated short open reading frames (sORFs) that, localized in different types of RNA molecules, including in long non-coding RNAs (lncRNAs), may encode and produce biologically functional peptides. Part of this Thesis is focused on the characterization of the Arabidopsis flower peptidome, using the floral homeotic mutants *apetala1, apetala2, apetala3, pistillata,* and *agamous* in comparison to the wild type. For peptide identification by LC-MS/MS, an extensive database of hypothetical novel Arabidopsis peptides was created. It comprised putative surf-encoded peptides (SEPs) from intergenic regions, UTRs, 'non-coding' RNAs and other transcripts. In total, 1,874 hypothetical peptides were detected by mass spectrometry, from which, 132 peptides were selected as candidates for further studies (60 of them were also predicted to be specifically expressed, or at least enriched, in one type of floral organ). Around 25% of the 132 peptide candidates belong to putative gene families in *A. thaliana*, and 103 have possible homologs in other plant species. In addition, different gene expression patterns for several peptide candidates were identified, with many of them showing specific expression in stamens during flower development.

Resumen

El inicio de la formación floral y el posterior desarrollo de las flores son paradigmas excelentes para los estudios de desarrollo en plantas, ya que se rigen por complejas redes de regulación. Gracias a análisis genéticos extensivos directos e inversos ha sido posible identificar multitud de genes reguladores clave para estos procesos que forman partes de dichas redes de regulación. Actualmente, hay una serie de factores adicionales que se están caracterizando a nivel del genoma gracias a métodos de integración de diferentes ómicas ('multi-omics'). La caracterización de las redes de regulación a nivel del genoma global es clave para entender, y eventualmente manipular, las bases del desarrollo y la fisiología de las plantas. Sin embargo, y a pesar de estos avances, la visión global y dinámica del proceso de desarrollo floral carece de un componente fundamental: el proteoma. Los métodos actuales de espectrometría de masas permiten explorar en profundidad la composición de un proteoma en su expresión y complejidad, su relación con el transcriptoma e incluso sus modificaciones postraduccionales. En los últimos años también se ha puesto de manifiesto que existe una parte sustancial de los proteomas eucariotas que no está anotada y está compuesta por péptidos y proteínas sin caracterizar (el peptidoma 'no convencional'), con funciones todavía por descubrir.

El genoma de Arabidopsis se secuenció hace 20 años. Desde entonces, diversos repositorios públicos han recogido información acerca de su transcriptoma y su modulación a lo largo del desarrollo, describiendo también su plasticidad en respuesta al ambiente. En cambio, la caracterización del proteoma de Arabidopsis ha sido mucho menos exhaustiva. En este respecto, es posible integrar la espectrometría de masas y la secuenciación de RNA. En esta Tesis, el sistema de inducción floral pAP1:AP1-GR *ap1 cal* se ha utilizado para caracterizar la expresión génica a nivel de proteoma a lo largo del desarrollo floral temprano de Arabidopsis, y su correlación con datos de expresión del transcriptoma. Se han combinado

métodos de secuenciación de proteínas (LC-MS/MS) y experimentos para la anotación del transcriptoma (RNA-seq) siguiendo una serie temporal de los cinco días posteriores a la activación del programa de desarrollo floral. Se identificaron casi 9000 proteínas y unos 23000 genes, de los cuales, 2037 proteínas y 8125 genes mostraron cambios significativos en su abundancia a lo largo de la serie temporal. Estos experimentos han permitido ampliar notablemente el tamaño de la colección de genes conocidos por tener cambios en sus niveles de expresión a lo largo de los estadios tempranos del desarrollo floral; identificar parejas RNA-proteína en las que ambas moléculas mostraban un patrón de abundancias similar (correlacionados), u opuesto (anti-correlacionados) y que están involucradas en diferentes procesos, como la fotosíntesis, el metabolismo de ácidos grasos o la biosíntesis de aminoácidos; y, gracias al análisis combinado de estos nuevos datos de transcriptómica y datos previamente publicados sobre la unión de AP1 en todo el genoma (ChIP-seq), identificar posibles dianas de AP1 nuevas.

Los genomas eucariotas contienen muchos marcos de lectura abiertos cortos (sORFs) que, localizados en diferentes tipos de moléculas de RNA, incluyendo RNA largos no codificantes (lncRNAs), pueden codificar y producir péptidos biológicamente funcionales. Parte de esta Tesis está enfocada a la caracterización del peptidoma floral de Arabidopsis, utilizando los mutantes homeóticos de floración *apetala1, apetala2, apetala3, pistillata* y *agamous* en comparación con las plantas de tipo silvestre. Para identificar péptidos por LC-MS/MS, se creó una extensa base de datos que incluye posibles péptidos codificados en sORFs (SEPs) en regiones intergénicas, UTRs, RNAs "no codificantes", y otros transcritos. En total se identificaron 1874 péptidos hipotéticos, de los cuales 132 fueron seleccionados como candidatos para otros análisis (además se predijo que 60 de ellos podrían estar expresados específicamente, o al menos enriquecidos, en alguno de los tipos de órganos florales). En torno al 25% de los 132 péptidos candidatos pertenecía a una posible familia de genes en *A. thaliana*, y 103 tenían al menos un homólogo en otras especies de plantas. Además, se encontraron diferentes patrones de expresión para muchos de los péptidos candidatos, en concreto, la mayoría mostró expresión específica en los estambres a lo largo del desarrollo floral.

Resum

L'inici de la formació floral i el posterior desenvolupament de les flors són paradigmes excel·lents en l'estudi del desenvolupament de plantes, ja que es regeixen per complexes xarxes de regulació. Gràcies a anàlisis genètiques extensives directes i inverses ha sigut possible identificar multitud de gens reguladors clau en aquets processos que formen part d'aquestes xarxes de regulació. Actualment, hi ha una sèrie de factors addicionals que s'estan caracteritzant a nivell del genoma gràcies a mètodes d'integració de diferents òmiques ('multi-omics'). La caracterització a nivell del genoma global es clau per a entendre, i eventualment, manipular les bases del desenvolupament i la fisiologia de les plantes. Tanmateix, i a pesar d'aquets avançaments, la visió actual i dinàmica d'aquets processos de desenvolupament manca d'un component fonamental: el proteoma. Els mètodes actuals d'espectrometria de masses permetran explorar en profunditat la composició d'un proteoma en la seva expressió i complexitat, la seva relació amb el transcriptoma, les modificacions postraduccionals dinàmiques i, fins i tot, la seva composició general. Als últims anys s'ha posat de manifest que existeix una part substancial dels proteomes eucariotes que no està anotada i que està composta per pèptids i proteïnes sense caracteritzar (el peptidoma 'no convencional'), amb funcions encara desconegudes.

El genoma d'Arabidopsis es va seqüenciar fa 20 anys. Des-de llavors, diversos repositoris públics han recogut informació sobre el seu transcriptoma i la seva modulació durant el desenvolupament, descriuen també la seva plasticitat en resposta a l'ambient. En canvi, la caracterització del proteoma d'Arabidopsis ha sigut molt menys exhaustiva. En aquest sentit, és possible integrar l'espectrometria de masses i la seqüenciació de RNA. En aquesta Tesi, el sistema d'inducció floral pAP1:AP1-GR *ap1 cal* ha sigut utilitzat per caracteritzar l'expressió gènica a nivell de proteoma al llarg del desenvolupament floral inicial d'Arabidopsis, i la seva correlació amb dades d'expressió del transcriptoma. S'han combinat mètodes de seqüenciació de proteïnes (LC-MS/MS) i experiments d'anotació de transcriptoma (RNA-seq)

a una sèrie temporal durant els cinc dies posteriors a l'activació del programa de desenvolupament floral. Es van identificar quasi 9000 proteïnes i uns 23000 gens, dels quals, 2037 proteïnes i 8125 gens van mostrar canvis significatius en la seva abundància al llarg de la sèrie temporal. Aquets experiments han permès ampliar notablement la mida de la col·lecció de gens coneguts per tenir canvis al seus nivells d'expressió al llarg dels estadis primerencs del desenvolupament floral; identificar parelles RNA-proteïna a les que ambdues molècules mostraven un canvi als seus nivells d'expressió similars (correlacionats) o oposats (anti-correlacionats) i que estan involucrades a diferents processos, com la fotosíntesi, el metabolisme d'àcids grassos o la biosíntesi d'aminoàcids; i, gràcies a l'anàlisi combinat d'aquets nous dades de transcriptòmica i dades prèviament publicats sobre la unió d'AP1 a tot el genoma (ChIP), identificar possibles dianes d'AP1 noves.

Els genomes eucariotes contenen molts marcs de lectura oberts corts (sORFs) que, localitzats a diferents tipus de molècules de RNA, inclouen RNA llargs no codificants (lncRNAs), poden codificar pèptids biològicament funcionals. Part d'aquesta Tesi està enfocada a la caracterització del peptidoma floral d'Arabdopsis, fent servir els mutants homeòtics de floració *apetala1, apetala2, apetala3, pistillata* i *agamous* en comparació amb les plantes de tipus silvestre. Per identificar pèptids, es va crear una extensa base de dades que va incloure potencial pèptids codificats en sORFs (SEPs) en regions intergèniques, UTRs, RNAs "no codificants" i altres transcrits. En total, es van identificar 1874 pèptids, dels quals 132 van ser seleccionats com candidats per altres anàlisis (a més, es va predir que 60 d'aquets estarien expressats específicament, o almenys enriquits, en algun dels tipus d'òrgans florals). Aproximadament 25% dels nous SEPs pertanyen a una possible família gènica en *A. thaliana*, i 103 té possibles homòlegs en altres especies de plantes. A més, es van trobar diferents patrons d'expressió per molts dels pèptids candidats, concretament, la majoria presentava expressió específica als estams durant el procés de desenvolupament floral.

Contents

List of Figures

List of Tables

List of Abbreviations

ACN	Acetonitrile
AGC	Auto Gain Control
AGI	Arabidopsis Gene Identifier
altORF	alternative ORF
BH	Benjamini & Hochberg
bp	base pair
CDS	Coding Sequence
CID	Collision-induced dissociation
CP	Conventional Peptide
DAP	Differentially Abundant Protein
DB	Database
DDA	Data Dependent Acquisition
DEG	Differentially Expressed Gene
DEX	Dexamethasone
dORF	downstream ORF
FDR	False Discovery Rate
FM	Floral Meristem
GC/MS	Gas Chromatography / Mass Spectrometry
GFP	Green Fluorescent Protein
GO	Gene Ontology
GWAS	Genome-Wide Association Studies
HCD	High-energy Collision Dissociation
HCT	AP1-High Confidence Target
IM	Inflorescence Meristem
JA	Jasmonic Acid
kNN	k-Nearest Neighbour
LB	Luria-Bertani medium
LC-MS/MS	Liquid Chromatograph Mass Spectrometry
LFC	Logarithmic Fold Change
lncRNA	long non-coding RNA

LRT	Likelihood Ratio Test
MAR	Missing At Random
ME	Module
miORF	Pri-miRNA-encoded ORF
miRNA	Micro RNA
MNAR	Missing Not At Random
mRNA	messenger RNA
MS	Mass Spectrometry
NA	Not Assigned value
NCP	Non-Conventional Peptide
NGS	Next Generation Sequencing
NMR	Nuclear Magnetic Resonance
NVG	Non-Variant Gene
NVP	Non-Variant Protein
PTM	Post-translational modification
qRT-PCR	Quantitative real-time Reverse Transcription-PCR
r	Pearson's coefficient
RA	Reliability Analysis
RD	Reliably Detected
RU	Reliably Undetected
SAM	Shoot Apical Meristem
SEP	short Open Reading Frame-Encoded Peptide
seq	Sequencing
sORF	short Open Reading Frame
SVG	Stage Variant Gene
SVP	Stage Variant Protein
TAIR	The Arabidopsis Information Resource
TbyS	TILLING by Sequencing
TILLING	Targeting Induced Local Lesions in Genomes
TF	Transcription Factor
TFA	Trifluoroacetic acid
TFE	Trifluoroethanol
TIS	Translation Initiation Site
T_m	Melting temperature
TPM	Transcripts Per Million

TUF	Transcript of Unknown Function
UD	Unreliably Detected
uORF	upstream ORF
UU	Unreliably Undetected
WGCNA	Weighted Gene Co-expression Network Analysis
WT	Wild Type
YEB	Yeast Extract Broth medium
ρ	Spearman's rank coefficient
ω (d_N/d_S)	Synonymous (d_S) and non-synonymous (d_N) substitution rate

Premises and hypothesis

This Thesis is built on the following premises:

1) Integrative multiomics analyses provide a wider interpretation of biological processes compared to approaches solely based on one type of omics.
2) Abundance levels of proteins and their corresponding mRNAs are not necessarily correlated.
3) Time-series analyses can provide a broad information on the fluctuating dynamics of regulatory networks controlling flower development.
4) The true extent of the peptidome of an organism is difficult to estimate due to the several possible origins for peptides and technical and experimental issues in peptide detection and identification.
5) Selection of a proper extraction method is key for peptide LC-MS/MS studies.

Based on the aforementioned premises, the following **hypothesis** is proposed for this Thesis:

"A more global understanding of the flower development process can be achieved from the combination of proteomics, peptidomics and transcriptomics studies."

The general aim of this work is to characterise the proteome of Arabidopsis in its expression and complexity, relationship with the transcriptome, and even in its composition, since it has become clear that plant genomes encode a substantial number of yet unknown peptides, and peptides play crucial roles in plant development and physiology.

Objectives

This Thesis combines genomic and proteomic technologies to advance towards the goal of a complete understanding of the genome-wide regulatory network of flower development in Arabidopsis, as well as to help understanding the functional information encoded in its genome, with a focus on the peptidome.

The specific objectives for this PhD Thesis are:

Aim 1.- *To establish a chronology of protein expression throughout (early) flower development and correlate these trajectories to unbiased transcript expression data*.

1.1 To perform shotgun proteomics experiments with the pAP1:AP1-GR *ap1 cal* floral induction system.
1.2 To develop a data analysis pipeline for the proteomics time-course data.
1.3 To perform transcript profiling experiments (RNA-Seq).
1.4 To correlate proteomics and transcriptomics data analyses.

Aim 2.- *To characterise the flower Arabidopsis peptidome (sORFs and hidden coding sequences in the Arabidopsis genome) and start deciphering its roles in flower development*.

2.1 To optimise a peptide extraction protocol and perform shotgun peptidomics experiments using the Arabidopsis floral organ identity mutants.
2.2 To develop a data analysis pipeline for the peptidomics data.
2.3 To identify novel, sORF-derived, unannotated peptides and analyse their intrinsic characteristics.
2.4 To characterise a group of selected peptides via transgenic lines expressing reporter constructs and loss-of-function mutants.

Chapter 1

General introduction (I)

Part of this chapter was published as:

Multi-omics methods applied to flower development.

Álvarez-Urdiola, R., Matus, J.T., Riechmann, J.L. (2023). Methods in Molecular Biology.

Chapter 1. General introduction (I)

1.1 Key points to understand flower development

The development of multicellular organisms depends on the capacity of cells to orchestrate a wide variety of gene expression programs, which result in the presence, absence, and differential accumulation of RNAs and proteins and allow the differentiation of organs and tissues. This capacity largely relies on the genome, in the form of *cis*-regulatory sequences that interact with transcription factors, co-regulators, and other types of regulatory proteins or RNAs, and in the structural organization of the genome, controlled by histones and their modifications, as well as by other epigenetic processes. These elements determine when, where, and how genes are expressed.

The onset of flower formation and the process of flower development in the Angiosperms constitute excellent paradigms for developmental studies in plants. Extensive genetic analyses have led to the identification of many key regulatory genes controlling these processes, and their corresponding gene regulatory networks are now being characterized at the genome-wide level using omics technologies.

Inflorescence development and architecture can widely vary among plant species, yet the basic organization of floral structure is extensively conserved. In angiosperms, flowers are formed when the shoot apical meristem (SAM) is transformed into an inflorescence meristem (IM) after the transition from vegetative to reproductive behaviours. This crucial shift from vegetative to reproductive growth is followed by the activation of a small number of floral meristem identity genes, such as *LEAFY* (*LFY*) and *APETALA1* (*AP1*), which specify floral meristems. These genes were originally identified in mutant screens (i.e., forward genetics approach) of plants with defects in early flower development. Floral meristems (FM) arise from the flanks of the IM and develop into flowers (CHAHTANE ET AL., 2023).

Flowers are frequently composed of four different classes of organs arranged in four whorls. The most exterior first and second whorls include the sepals and petals, respectively, while the internal third and fourth whorls represent the pollen-producing stamens and carpels, respectively. The appropriate development of a flower requires these whorls to be formed in a sequential manner, following a canonical pattern. In the thale or mouse-ear cress *Arabidopsis thaliana*, a plant model species belonging to the Brassicaceae family, sepal primordia are initiated first, followed by those determining petals and stamens. After that, carpels initiate and develop from the centre of the developing flower (SMYTH ET AL., 1990). Previous studies have identified several transcription factors and other regulatory molecules as responsible for initiating floral developmental programs in a partially overlapping manner. Understanding their regulatory networks has been a long-standing challenge in plant developmental genetics in relation to the specification of the distinct floral organs.

1.1.1 Floral organ identity: the ABC model

The organ identity genes, responsible for the formation of different organs in the four whorls, are activated by AP1 and LFY after the initiation of the floral meristems. The ABC model of flower organ identity describes how floral organs are specified by the domain-specific interaction of homeotic genes coding for different transcription factors and by their target genes. This model was proposed based on genetic studies in the *A. thaliana* floral homeotic mutants, *apetala1* (*ap1*), *ap2, ap3, pistillata* (*pi*) and *agamous* (*ag*), in which there is a replacement of one type of floral organ by another (BOWMAN ET AL., 1991; COEN & MEYEROWITZ, 1991). The *ap1* and *ap2* mutants show organ identity defects in the first and second whorls (determining sepals and petals). In *ap1*, sepals are transformed into bract-like organs, and petals are absent; while in *ap2*, petals are missing or transformed into stamens and sepals are transformed into carpel-like structures. The mutants *ap3* and *pi* are defective in the second and third whorls (petals and stamens): petals are replaced by sepals, and stamens, by carpels. Finally, in *ag*, stamens are replaced by petals, and carpels, by extra whorls of sepals and petals (BOWMAN ET AL., 1991) (**Figure 1.1**).

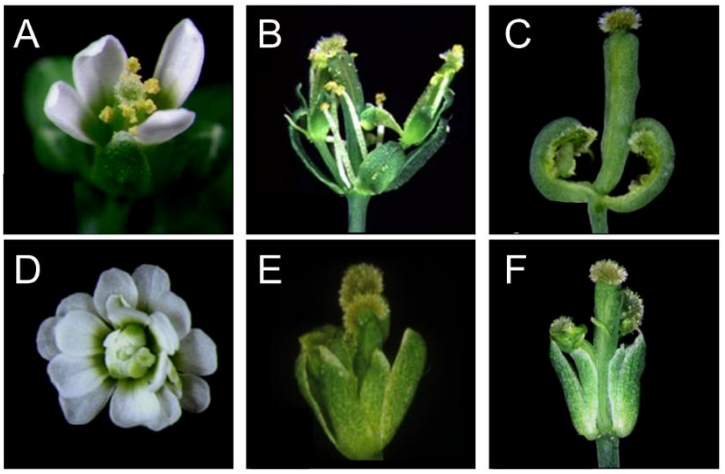


Figure 1.1. Floral homeotic mutants of Arabidopsis.
Mature flowers of **A)** *Landsberg erecta* (Ler, wild type flower), **B)** *ap1-1*, **C)** *ap2-2,* **D)** *ag-1*, **E)** *pi-1*, and **F)** *ap3-3*. Taken from (WELLMER ET AL., 2004).

As described by the ABC model, the activities of the genes being affected on each homeotic mutant can be assigned to three different functions, namely 'A' (represented by *AP1* and *AP2* genes), 'B' (embodied by *AP3* and *PI*), and 'C' (*AG*) with each function required for organ specification in different meristematic regions. A-function genes specify sepals, and together with B-function genes determine petals, while B- and C-function genes act together leading stamen development. The C-function genes alone control carpel formation (COEN & MEYEROWITZ, 1991).

Modifications and expansions of the ABC model – but still maintaining its basic tenets – have been developed through the years in order to accommodate newly identified genes and gene functions as well as the floral diversity that exists among angiosperm species (e.g., (PAJORO, BIEWERS, ET AL., 2014; THOMSON & WELLMER, 2019)). The original model was extended to the ABCDE model by (THEIßEN, 2001; THEIßEN & SAEDLER, 2001), with the inclusion of D-function genes, such as *SEEDSTICK* (*STK*), *SHATTERPROOF1* (*SHP1*) and *SHP2,* whose encoded proteins interact with E-class proteins to control ovule development (COLOMBO ET AL., 2010; FAVARO ET AL., 2003; PINYOPICH ET AL., 2003). On the other hand, E-function genes, like *SEPALLATA1* (*SEP1*), *SEP2*, *SEP3*, and *SEP4*, are involved in the specification of all types of flower organs (DITTA ET AL., 2004; PELAZ ET AL., 2000).

As A-function mutants could only be found in *A. thaliana*, and A- and E-function genes are both involved in specifying floral meristems, a modified (A)BCD model was proposed, with (A) incorporating both A- and E-functions (CAUSIER ET AL., 2010; F. WU ET AL., 2017). Similarly, C- and D-functions in the Angiosperms are traced back to a combined C/D-function provided by *AG-like* genes in extant gymnosperms and stem group seed plants (GRAMZOW ET AL., 2014). Hence, the model could be modified again into an (A)B(C) model with (C) encompassing C- and D-function genes. The (C) function can specify reproductive organ identity, and its expression distinguishes reproductive from non-reproductive organs (THEIßEN ET AL., 2016) (**Figure 1.2**).

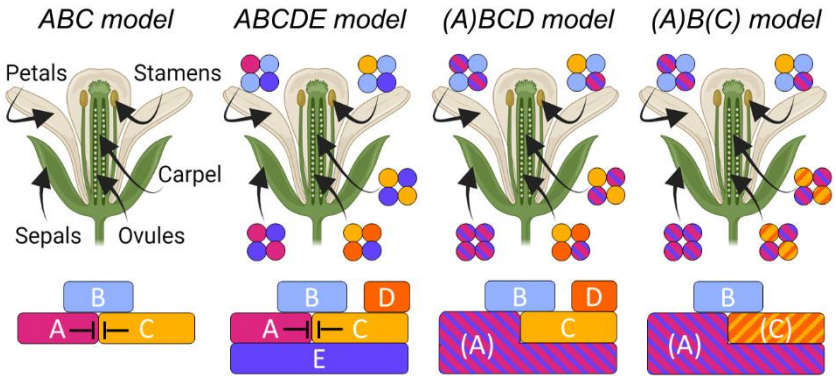


Figure 1.2. Proposed models for organ identity determination in *A. thaliana.*
The upper part of the figure depicts the tetrameric protein complexes formed by different classes of homeotic genes described in the floral quartet model. Each combination functions in specific whorls of the flower to specify floral organ identity. The colours of the proteins (circles) indicate the classes to which they belong. "(A)" represents the combination of A- and E-class genes, and "(C)", the combination of C- and D-class genes. Adapted from (THOMSON & WELLMER, 2019).

1.1.2 Molecular control of flower development

The majority of (A)B(C) genes code for MADS-domain transcription factors (MTFs) with the exception of *AP2* which codes for an Ethylene Response Factor (ERF) type transcription factor (TF). MADS-domain proteins harbour DNA-binding, nuclear localization and protein-protein interaction domains required to fulfil their roles as (A)B(C) proteins. The floral quartet model describes how the flower organ identity is specified during development by

tetrameric protein complexes composed of these MADs-domain proteins. These floral quartet complexes (FQCs) are asumed to function as transcription factors by binding to the DNA of their target genes, activating or represing them to control the emergence and development of the respective floral organs (THEIßEN, 2001). Homeotic and other flower development genes can also enhance or repress each other's expressions, resulting in a complex and stage-dependent transcriptional regulatory network (PAJORO, MADRIGAL, ET AL., 2014) (**Figure 1.3**).

The MADS-domains of the two dimers of each tetrameric complex bind to proximate CArG-box sequences (CArG: C-A-rich-G; consensus: 5'-$CC(A/T)_6GG$-3') to induce the looping of chromatin (THEIßEN, 2001; THEIßEN & SAEDLER, 2001). First, a single CArG box and its flanking regions are recognised by a MTF dimer via a combination of base and shape readout. Attractive or repulsive forces between the dimerization interfaces of two interacting MTFs facilitate or impede dimerization. The distance between two neighbouring CArG boxes and whether both are directed to the same site of the DNA double helix determine whether FQCs formation is favoured or not. In addition, the ability to form tetramers facilitates cooperative binding of a second MTF dimer while looping the DNA in between both binding sites (KÄPPEL ET AL., 2023) (**Figure 1.4**).

The possible interaction of FQCs with chromatin acting as 'pioneer transcription factors' to regulate the expression of their target genes has been described by (THEIßEN ET AL., 2016). FQCs would act as sequence-specific transcription factors with (half-) nucleosome-like properties that help to establish a permissive or repressive chromatin modification at CArG-box-containing promoters. After being incorporated to chromatin, in a gene-activating case, the FQCs would recruit histone-modifying factors, leading to the recruitment of the basal transcriptional machinery. Finally, the presence of at least one transactivation domain (TAD) in a DNA-bound FQC recruits the basal transcription machinery and eventually initiates transcription at the transcriptional start site (TSS) (KÄPPEL ET AL., 2023) (**Figure 1.4**).

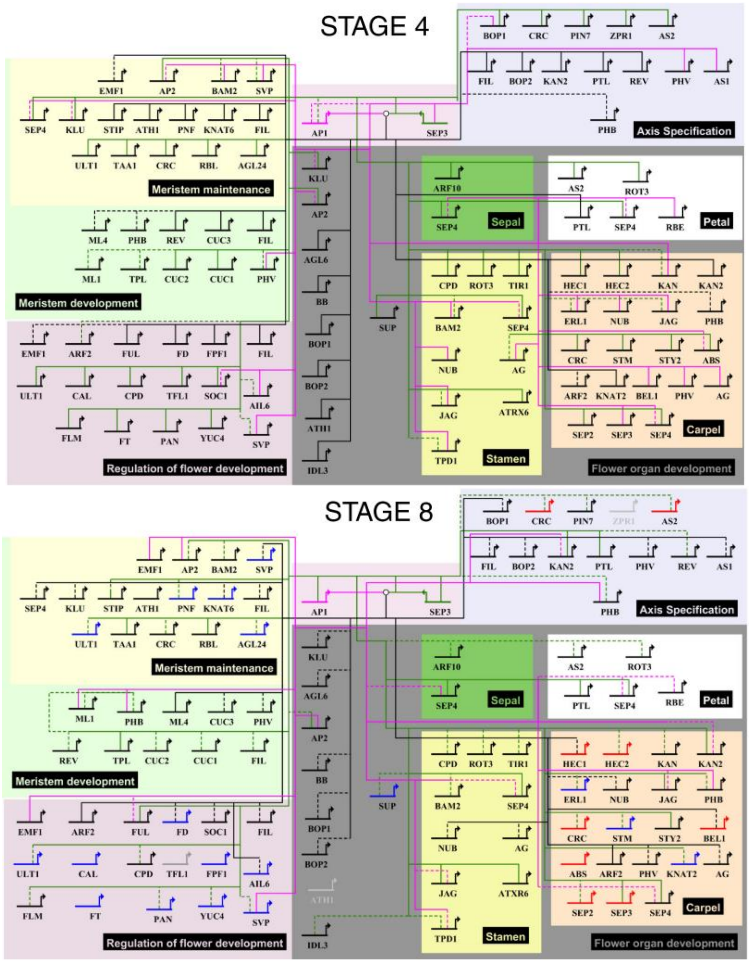


Figure 1.3. Putative target gene networks at stage 4 and stage 8 of flower development reflecting preferential binding of AP1 and SEP3 at different time points.

Representative Gene Ontology (GO) categories are included: meristem development, meristem maintenance, regulation of flower development, axis specification, and floral organ development (sepal, petal, stamen, and carpel development). Only genes that belong to these categories and that were found to be preferentially bound by either APETALA1 or SEPALLATA3 on a comparison of floral stages 4 and 7/8 are included. Black line indicates common targets, while pink line indicates AP1-specific targets, and green line indicates SEP3 targets. Dashed lines are used to indicate gene with significant (FDR <0.001) TF-binding peak, while solid lines for genes with higher peak respectively at stage 4 or stage 8. Grey: genes not bound at the specific stage. Red: upregulated genes. Blue: downregulated genes. Taken from (PAJORO, MADRIGAL, ET AL., 2014).

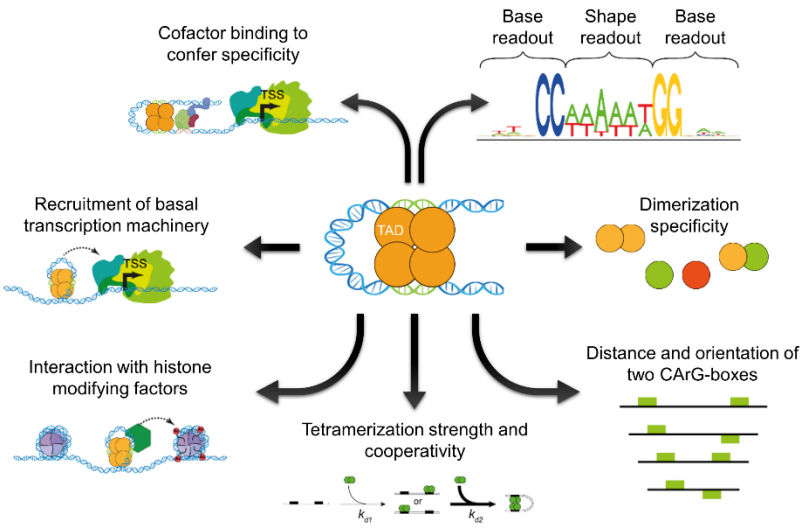


Figure 1.4. Possible mechanisms involved in FQC target gene recognition. Based on (GOSLIN ET AL., 2023; KÄPPEL ET AL., 2023).

Another important aspect for the specificity of the FQCs is the cofactor binding (GOSLIN ET AL., 2023). The specificities of the different floral organ identity MTF tetramers could be in part a result of their interactions with different combinations of additional transcription factors on the promoters of target genes (NAGY & NAGY, 2020) (**Figure 1.4**).

The floral quartet model is well supported by experimental evidence. For instance, mass spectrometry analyses demonstrated the existence of all the major binary interactions proposed in the floral quartet model and provided clues towards deciphering the specificity of their interaction with DNA (SMACZNIAK ET AL., 2012), even though their exact stoichiometry remains unknown. As the MADS-box genes involved in flower development show very specific and restricted expression patterns, what genes are expressed where determines the proteins that could interact from the quartets. Moreover, the induction of DNA looping by floral quartets has been demonstrated *in vitro* and *in vivo* (MELZER ET AL., 2009; MENDES ET AL., 2013), although it remains unclear whether it is a prerequisite for floral quarter activity. However, the model of FQCs evicting nucleosomes to activate or repress chromatin modifications (THEIßEN ET AL., 2016) is not well supported by experimental evidence yet.

1.1.3 APETALA 1 as a main orchestrator of flower meristem initiation and development

In Arabidopsis, flower development is initiated by LFY. *LFY* encodes a TF and is up-regulated by the SUPPRESSOR-OF-OVEREXPRESSION-OF-CO 1 (SOC1) and AGAMOUS-LIKE 24 (AGL24) MADS-domain proteins, which are induced throughout the inflorescence meristem by environmental and endogenous cues. Auxin phytohormone also helps in the induction of *LFY* expression by defining floral meristem initiation sites. *LFY* is expressed specifically in floral primordia because its induction in the SAM is repressed by the TERMINAL FLOWER1 (TFL1) inflorescence identity protein. In the floral primordium, LFY induces *AP1* and its paralog *CAULIFLOWER* (*CAL*), which regulate *LFY* with positive feedback, while repressing *SOC1, AGL24* and *TFL1*. Thus, the floral fate of the new meristem is stabilised (BOWMAN ET AL., 1993) **(Figure 1.5)**.

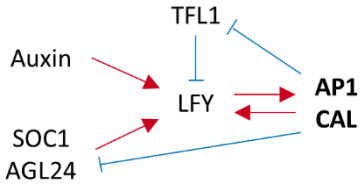


Figure 1.5. Interactions between major floral regulators.
Red arrows depict activation and blue barred lines indicate repression.

In addition to its early-stage role during the specification of floral meristems, AP1 function is subscribed within the (A)B(C) model, as it promotes both sepal and petal identity (THEIßEN ET AL., 2016). In *ap1* mutants, the sepals are transformed into leaf-like structures with petals failing to develop. In the axils of these leaf-like structures, secondary flowers arise that repeat the same pattern as the primary ones (BOWMAN ET AL., 1993; MANDEL ET AL., 1992).

The AP1 network had been extensively delineated through genetic studies. However, to better understand the regulatory networks that underlie the events that take place after the activation of AP1, it was necessary to study the downstream AP1 targets. The result was a highly interconnected network with AP1 acting as a transcriptional orchestrator controlling the expression of a wide variety of TFs and other types of genes (KAUFMANN ET AL., 2010). AP1 acts predominantly as a transcriptional repressor during the earliest

stages of flower development, whereas, at more advanced stages, it predominantly activates regulatory genes required for floral organ formation. In addition, AP1 also acts as a 'pioneer transcription factor' (PAJORO, MADRIGAL, ET AL., 2014), directly binding condensed chromatin exerting both positive and negative effects on transcription.

1.1.4 The ap1 cal floral induction system

The floral meristem identity gene *AP1* and its paralog *CAL* control the onset of Arabidopsis flower development in a partially redundant manner (FERRÁNDIZ ET AL., 2000). In an *ap1 cal* double mutant background, the AP1/LFY positive feedback is absent and TFL1 is not repressed by AP1/CAL in the nascent floral meristem. Consequently, young flower primordia cannot maintain *LFY* expression and start expressing *TFL1* (**Figure 1.5**). As a result, plants do not transition to flowering and, instead, exhibit massive over-proliferation of undifferentiated inflorescence-like meristems, leading to a cauliflower-like appearance (AZPEITIA ET AL., 2021; BOWMAN ET AL., 1993; KEMPIN ET AL., 1995).

The *ap1 cal* background allowed to create a floral induction system based on the expression of AP1 fused to the binding domain of the rat glucocorticoid receptor (GR). At first, under the control of the 35S promoter (WELLMER ET AL., 2006), but later, of the endogenous promoter of *AP1* (Ó'MAOILÉIDIGH ET AL., 2023). The activation of the AP1-GR fusion protein by applying dexamethasone (DEX) to the cauliflower structure triggers flower formation synchronously throughout the meristematic tissue (**Figure 1.6**).

The *ap1 cal* floral induction system has served for the study of early flower development in Arabidopsis at genomic and transcriptomic level (KAUFMANN ET AL., 2010; PAJORO, MADRIGAL, ET AL., 2014; WELLMER ET AL., 2006). The integration of transcriptomics and regulomics (i.e., ChIP-seq studies) has been used in combination with this system to explore the time-scaled regulatory networks of AP1 (KAUFMANN ET AL., 2010), SEP1 (PAJORO, MADRIGAL, ET AL., 2014), and LFY (GOSLIN ET AL., 2017), for example. However, proteomics or metabolomics methods have never been used in the AP1 induction system.

In this Thesis, I extend the analysis to the proteomic level, also comparing the data with unbiased transcriptomic data (RNA-sequencing).

Figure 1.6. The pAP1:AP1-GR *ap1 cal* floral induction system.
A) Wild-type-like Arabidopsis flower developed after treatment with DEX. **B)** Flowers developed after treatment of inflorescence-like meristems with "mock" solution. **C)** *pAP1:AP1-GR ap1 cal* mutant cauliflower-like inflorescence meristem before DEX induction. **D)** Inflorescence meristem 6 days after the induction of flower development triggered by DEX. Based on (Ó'MAOILÉIDIGH ET AL., 2023).

1.2 Integrative genome-wide analyses and their association to proteomic data to understand plant biology

For the last fifteen years, the technological advances, and lower costs of genome-wide approaches, together with the enormous increase of computational biology tools to process large biological datasets (often referred as omics), are causing a shift in the way developmental studies in plants are approached. Several studies of reproductive organ development have used genomic analyses of transcription factors and global gene expression changes for modelling complex gene regulatory networks (reviewed in (MATEOS ET AL., 2017; PAJORO, BIEWERS, ET AL., 2014; WELLMER & RIECHMANN, 2010; WILS ET AL., 2017)).

As a result, hundreds of target genes of the floral homeotic factors in *A. thaliana* have been identified through a combination of genome-wide binding analyses and transcriptomics studies (KAUFMANN ET AL., 2010; PAJORO, MADRIGAL, ET AL., 2014; WELLMER ET AL., 2006). Nevertheless, the emergent global and dynamic view of developmental processes requires an important component: the proteome, in its expression, complexity and relationship with the transcriptome. Thus, to assert the whole comprehension of a

network from a global perspective requires the integration of several types of omics data, including proteomics approaches (DECOURCELLE ET AL., 2015; KOEHLER ET AL., 2015; LE SIGNOR ET AL., 2017; MERGNER ET AL., 2020).

Proteomics became a valuable tool to uncover the molecular background of many biological processes including plant stress responses and developmental and signalling processes. The last advancements of this approach have been made possible by significant improvements in methods of protein extract preparation, separation of proteins and peptides, mass spectrometry instrumentation and downstream bioinformatics analyses (TAKÁČ ET AL., 2017).

Genomics and transcriptomics are closer to the genotype of the studied organisms, whereas proteomics and metabolomics are closely related to their phenotype. Through these technologies, research has described in depth the hierarchical levels of plant organization and functioning, improving the odds to predict the behaviour of whole plants (phenome) as a response to genetic perturbations and/or environmental changes (DO AMARAL & SOUZA, 2017).

1.2.1 Genomics and transcriptomics methods as the most utilized

DNA sequencing-based technologies are the most advanced of the omics technologies in terms of standardized protocols, analytical tools, and public repositories for data sharing. They provide unique opportunities to obtain high quality data from small amounts of tissues or individual cells, addressing a wide range of biological questions, including the understanding of plant biology (MARDIS, 2017; VAN DIJK ET AL., 2018).

Genome-wide analyses dependent on high-throughput technologies are revealing the complexity and scope of regulatory networks that can be fluctuant in time and largely plastic by the effects of the environment and that are governed by transcription factors (MATEOS ET AL., 2017; T. YU ET AL., 2019; Q. G. ZHU ET AL., 2018), microRNAs (LUO ET AL., 2018; SHI ET AL., 2017), movable factors (X. LIU ET AL., 2018), hormones (SAHA ET AL., 2016) and chromatin-modifying proteins (ENGELHORN ET AL., 2018).

The progression of genomics has played a noteworthy role in the field of flower development research, primarily through the use of gene expression profiling (transcriptomics; first DNA microarrays and subsequently RNA-seq and related methods) (RICH-GRIFFIN ET AL., 2020; WELLMER ET AL., 2004, 2006) and of genome-wide DNA binding studies (ChIP-Seq) (KAUFMANN ET AL., 2010).

Nevertheless, high-throughput genomics and transcriptomics might still fall sort of proving a full network description or understanding in the context of biological function. This is because mRNA abundance is not necessarily a reliable indicator of protein quantity and activity (YANSHENG LIU ET AL., 2016). Combining data from genomics and transcriptomics with proteomics (MERGNER ET AL., 2020) or metabolomics (GARCIA-MOLINA ET AL., 2020; MATUS, 2016) provides molecular information to further genetic and epigenetic changes and variations with phenotypic alterations or differences.

1.2.2 Proteomics: from the sidelines to the mainstream

The analysis of the proteome of eukaryotic cells is challenging due to the substantial diversity in the properties of the individual proteins that compose it (e.g., abundance, stability, molecular weight, structure, hydrophobicity, hydrophilicity, or variety of post-translational modifications –PTMs–, among others), compared to the relative simplicity of DNA molecules. This large heterogeneity represents a significant hurdle for achieving 'genome-wide' coverage in proteomics experiments and complicates the proteomics methodologies and procedures. Yet, proteomics approaches provide an important contribution to understanding gene function and cell organismal biology.

Along with an enhancement of throughput, sensitivity and resolution of analytical technologies in mass spectrometry (MS), computational methods have emerged focusing on the abundance and diversity of proteins in a complex sample (ASLAM ET AL., 2017; GROSSMANN ET AL., 2010; MERGNER & KUSTER, 2022; TAKÁČ ET AL., 2017). Current proteomics methods can identify thousands of proteins in a sample, including information on their posttranslational modifications.

In plants, MS-based proteomics approaches have been applied for the measurement of differential protein expression or the detection of PTMs (Navrot et al., 2011; Z. Zhang et al., 2017) in different tissues and biological processes (reviewed in (MERGNER & KUSTER, 2022)). Deep proteome studies have led to the development of proteome atlases of the major plant organs for different plant species (ABRAHAM ET AL., 2013; DUNCAN ET AL., 2017; M. KUMAR ET AL., 2022; MARX ET AL., 2016; MERGNER ET AL., 2020; SZYMANSKI ET AL., 2017). Besides, cell type-specific proteome studies are crucial for a better understanding of the unique biological functions and properties of individual cell types in a tissue (DAI & CHEN, 2012), as well as subcellular plant proteomics and predictions (BERNHOFER ET AL., 2018; BRUCE, 2000; EMANUELSSON ET AL., 2001). As the proteome is in constant flux, several proteome studies are based on temporal series (BASSAL ET AL., 2020) during developmental processes (FENG ET AL., 2022), or stress responses (JAIN ET AL., 2021; NIU ET AL., 2021).

1.2.3 Multiomics approaches to maximize the power of data-driven research

In omics-based analyses, collecting as much information as possible is especially relevant to elaborate accurate biological models. The power of omics could be enhanced by combining them with other experimental methods, such as cell biology, biochemistry, molecular biology, and also other omics (**Figure 1.7**). Numerous studies are based in the combination of datasets from a single omics, obtained with the same or different techniques in various parts of an organism, developmental stages, or related to different transcription factors (D. CHEN ET AL., 2018; VALENTIM ET AL., 2015; J. WANG ET AL., 2017). Nevertheless, the possibility of combining results from more than one type of omics has gained prominence in the past few years as a way to explore different aspects of plant biology (KOEHLER ET AL., 2015; LE SIGNOR ET AL., 2017; LEHMANN ET AL., 2021; MATUS, 2016; MERGNER ET AL., 2020; G. ZHU ET AL., 2018). These integration studies are usually referred as multi-omics, trans-omics, or integrated omics in the current literature.

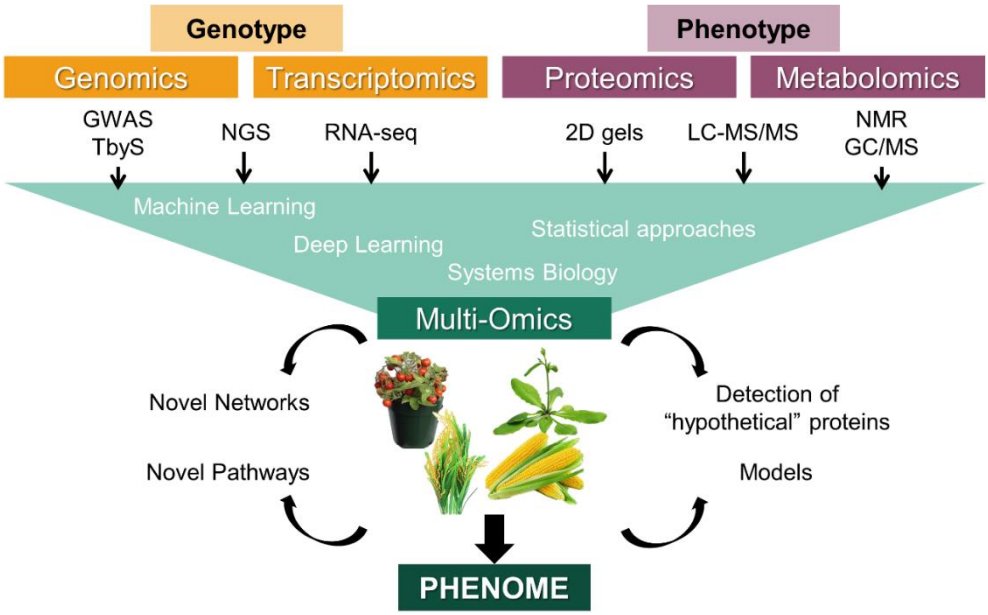


Figure 1.7. Integrated omics workflow.
Datasets can be integrated using machine learning and statistical approaches to produce findings in the form of novel pathways and networks, adding information to previously known processes, the development of new biological models, or the detection at the experimental level of proteins and peptides encoded in newly annotated ORFs.

The integration of omics using statistical, or machine learning approaches could lead to a better understanding of both known and unknown pathways, draw more complex regulatory networks or propose novel data-driven models, thanks to the combination of 'closer to genotype'-datasets (i.e., genomics and transcriptomics) with those 'closer to phenotype' (i.e., proteomics and metabolomics). Successful implementation of more than two omics datasets is very rare (Misra et al., 2019), although there are relevant cases described in crops (e.g., (Decourcelle et al., 2015; Koehler et al., 2015; Yinghao Liu et al., 2022)). In this Thesis, I am focusing on the integration of two types of omics: transcriptomics and proteomics.

The correlation between mRNA expression levels and the abundance of their matching proteins has been exhaustively studied in different processes and species during the last years (reviewed in (D. KUMAR ET AL., 2016; YANSHENG LIU ET AL., 2016; MANZONI ET AL., 2018)). While the genome is more or less static through an organism's life, its proteome and transcriptome vary rapidly, albeit in a controlled manner, as a response to different environmental perturbations and growth conditions. These changes are not only due to transcript and protein expression levels, but also to posttranscriptional (e.g., alternative splicing) and posttranslational (e.g., phosphorylation) control. Thus, to properly understand developmental or environment-responsive cell processes, it is crucial to comprehend their proteome expression patterns as a complement to their transcriptome levels (D. KUMAR ET AL., 2016).

In the case of Arabidopsis, there are only a few studies that combine transcriptomics and proteomics to analyse developmental processes, such as embryogenesis (HUANG ET AL., 2022), seed germination (BAI ET AL., 2021), leaf development (OMIDBAKHSHFARD ET AL., 2021), and floral transition (X. WANG ET AL., 2020). In addition, other studies in Arabidopsis have focused on the photoperiodic control of its proteome (SEATON ET AL., 2018; UHRIG ET AL., 2021).

In other plants, combined transcriptome-proteome analyses have already been used to study petal shape in peonies (Y. WU ET AL., 2018), carotenoid synthesis in maize (DECOURCELLE ET AL., 2015), and fruit development and ripening in fruit trees such as orange (J. H. WANG ET AL., 2017) and pear (P. WANG ET AL., 2023) trees; as well as reproductive development, in particular, male reproductive development in cabbage (HAN ET AL., 2018; JI ET AL., 2018; KELLER ET AL., 2018; XING ET AL., 2018), female reproductive development in peanut (ZHAO ET AL., 2015), and flower development in general in species as jujube (R. CHEN ET AL., 2017) and loquat (JING ET AL., 2020) (**Table 1.1**).

Table 1.1. Combined transcriptomics and proteomics studies in plants.

Aim of study	Correlation between RNA and protein levels	Reference
I. Development		
A. thaliana		
Characterization of protein changes during seed germination	Higher correlation between RNA levels at a timepoint and protein levels at the next timepoint than at the same timepoint	(Bai et al., 2021)
Analysis of leaf development	Protein changes showed correlation with changes at transcriptome level, but with a certain delay	(Omidbakhshfard et al., 2021)
Study of early embryogenesis proteome	Overall positive correlation	(Huang et al., 2022)
Study of transcriptome and proteome of floral transition	Weak correlation between RNA and protein levels, except for 55 genes which were DEGs and DAPs	(X. Wang et al., 2020)
Maize		
Study of the correlation between RNA and protein abundance during leaf development	Significant positive correlations between RNA and protein levels	(Ponnala et al., 2014)
Cabbage		
Characterization of a recessive male sterile mutant	Similar changing trend (positive correlation) for most of the detected RNA-protein pairs	(Ji et al., 2018)
Analysis of Ogura cytoplasmic male sterility	Generally low correlation, except for some DEGs and DAPs	(Xing et al., 2018)
Analysis of Ogura cytoplasmic male sterility	Poor correlation	(Han et al., 2018)
Jujube		
Mapping of the jujube floral organ	Positive correlation between RNA and protein levels	(R. Chen et al., 2017)
Loquat		
Analysis of flower development	Positive correlation between DEGs and DAPs	(Jing et al., 2020)
Pear tree		
Proteogenomics atlas (fruit development)	Overall positive correlation	(P. Wang et al., 2023)
Orange tree		
Analysis of the differences among cultivars during fruit development and ripening	Positive correlation between RNA and protein levels	(J. H. Wang et al., 2017)

Pomegranate		
Understanding the molecular mechanisms under petaloidy in pomegranate	The correlation between DEGs and DAPs was higher than the correlation between all genes and all proteins detected	(Huo et al., 2023)
Watermelon		
Quantitative transcriptomic and proteomic analysis of fruit development and ripening	Low correlation	(Y. Yu et al., 2022)
II. Stress		
Cucumber		
Understanding post-germinative development under salinity and drought	Good correlation between RNA and protein levels of DEGs and DAPs	(Du et al., 2021)
Maize		
Multi-omics analysis of pathogen-induced cell death	Low when comparing all RNA and protein pairs, stronger when dividing the dataset into correlation modules	(Barghahn et al., 2023)
Tomato		
Analysis of transcriptome and proteome adaptation during heat stress response	Low correlation	(Keller et al., 2018)
Soybean		
Study of roots grown under heat stress	Low correlation	(Valdés-López et al., 2016)
Cotton		
Study of genetic regulation of salt tolerance	Low correlation	(Peng et al., 2018)
III. Others		
A. thaliana		
Study of the photoperiodic control of the proteome	Stronger correlations of transcript and protein abundance for the arrhythmic transcripts	(Seaton et al., 2018)
Characterization of the diurnal dynamics of the rosette proteome / phosphoproteome	Most proteins with diurnal changes in abundance fluctuated independently of their transcript levels	(Uhrig et al., 2021)
Sweet cherry		
Creation of a proteogenomics atlas	Low correlation	(Xanthopoulou et al., 2022)

1.2.4 Limitations in multiomics studies

Within integrative omics studies, the degree of correlation between transcript and protein levels (and between changes in transcript and in protein levels) is still a lingering issue (BISHOP & HAWLEY, 2022; YANSHENG LIU ET AL., 2016) as, whereas some studies conclude that there is not a strong correlation, in others such correlation is more apparent (**Table 1.1**).

In this regard, a general aspect of label-free quantitative proteomics (and LC-MS/MS based metabolomics), which can hinder the subsequent data analysis and its comparison with other omics data, is the high rate of missing values. Statisticians defined three types of missing values depending on the nature of the missingness: i) Missing Completely At Random (MCAR) and ii) Missing At Random (MAR) values, which are due to minor errors or stochastic fluctuations and to conditional dependencies respectively; and iii) Missing Not At Random (MNAR) values, which have a targeted effect (LAZAR ET AL., 2016). These not assigned values (NAs) can be imputed by different methods, that must be chosen depending on their nature.

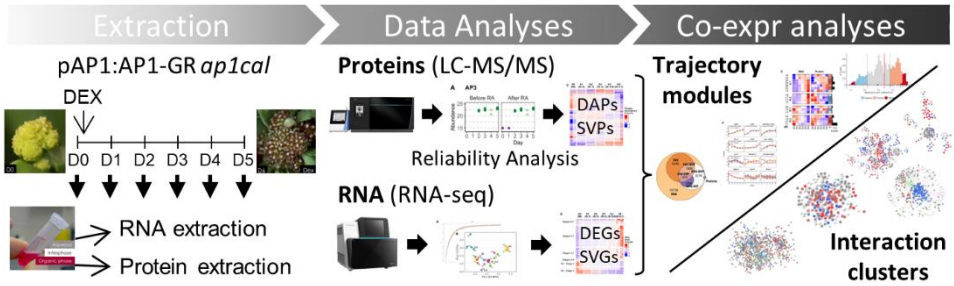
As there are many types of NAs that coexist in most quantitative datasets, hybrid strategies of imputation may be a better approach (JIN ET AL., 2021; LAZAR ET AL., 2016). Despite the optimization of imputation methods for proteomics, the sensitivity of extraction and quantification techniques highly differ from those used in transcriptomics analyses. Moreover, the lack of correlation among omics data could be also derived from the difficulties to obtain truly comparable datasets. However, the observed differences might also be caused by posttranslational regulation of protein levels (VOGEL & MARCOTTE, 2013), or by their different expression and degradation kinetics, as longer protein half-lives buffer changes in mRNA levels (CSÁRDI ET AL., 2015; OLIVA-VILARNAU ET AL., 2020; RAJ ET AL., 2006; TANIGUCHI ET AL., 2011).

Time-course studies may be an approach for addressing this gap, as successive analyses at different time points could allow the discovery of correlative behaviours of protein and mRNA levels through time (BAI ET AL., 2021; OMIDBAKHSHFARD ET AL., 2021; TARAZONA ET AL., 2018). In this regard, the use of the AP1-GR floral induction system (Ó'MAOILÉIDIGH ET AL., 2023), in

combination with proteomics, offers an opportunity to explore this and other questions in Arabidopsis early flower development.

Chapter 2

Chronology of transcriptome and proteome expression during early flower development



Part of this chapter will be published as:

Chronology of transcriptome and proteome expression during early flower development.

Álvarez-Urdiola, R., Matus, J.T., González, V.M., Bernardo-Faura, M., Riechmann, J.L. Manuscript in preparation.

Chapter 2. Chronology of transcriptome and proteome expression during early flower development

2.1. Background

The *Arabidopsis thaliana* flower developmental program has represented a proxy to understand the early steps of organ development in plants. In fact, the onset of flower formation is a key regulatory event during the life cycle of all angiosperms, and it is under a tight and widely-conserved genetic control (Theißen et al., 2016; Thomson & Wellmer, 2019). The identification of the roles of many transcription factors through forward and reverse genetic analyses has allowed the understanding of their contribution in flower initiation and development and other related developmental processes (e.g., fertilization and fruit formation) via gene regulatory networks (Wils et al., 2017). However, a comprehensive view of a regulatory network requires the integration of more than one type of omics data (e.g., (Mergner et al., 2020)).

The characterization of the proteome as a complement of the transcriptome is essential for understanding the different developmental and ambient-responsive cellular processes, as transcriptome and proteome composition can vary rapidly as a response to developmental perturbations and growth conditions (D. Kumar et al., 2016). The integration of mass spectrometry and RNA-sequencing (RNA-seq) has been used to study the correlation, or lack thereof, between transcriptome and proteome data in various organisms (e.g., (Edfors et al., 2016; Gyuricza et al., 2022; Hoogendijk et al., 2019; L. Jiang et al., 2020; Lindeboom et al., 2018; Sidhaye et al., 2023; D. Wang et al., 2019)). Very few studies have specifically addressed this issue to characterize developmental processes in plants, although there are examples of the combination of omics to analyse the development of leaves (Omidbakhshfard et al., 2021; Ponnala et al., 2014), flowers (R. Chen et al., 2017; Jing et al., 2020; X. Wang et al., 2020), and fruits (J. H. Wang et al.,

2017; P. WANG ET AL., 2023; Y. YU ET AL., 2022), as well as seed germination (BAI ET AL., 2021) and embryogenesis (HUANG ET AL., 2022). The apparent lack of correlation between transcript and protein levels found in some of those studies signals the existence of complex control mechanisms for both types of molecules, such as post-translational regulation of protein levels, and different stability, or expression and degradation kinetics of RNA and proteins (CSÁRDI ET AL., 2015; RAJ ET AL., 2006). A possible approach to explore if a higher correlation exists is the time-course study of a process, as there might be a temporal shift between mRNA and protein level changes (BAI ET AL., 2021; HOOGENDIJK ET AL., 2019; HUANG ET AL., 2022; OMIDBAKHSHFARD ET AL., 2021; P. WANG ET AL., 2023).

In this regard, the APETALA1 (AP1)-based floral inducible system has been used to study the early stages of flower development in *A. thaliana* through genomic approaches (Ó'MAOILÉIDIGH ET AL., 2023). The MADS-domain transcription factor (TF) AP1 is a key regulator of floral meristem identity and activation of flower development in Arabidopsis (LILJEGREN ET AL., 1999; MANDEL ET AL., 1992; NG & YANOFSKY, 2001). The integration of the transcriptome, cistrome, and epigenome associated to this TF led to a better understanding of early flower development (KAUFMANN ET AL., 2010; PAJORO, MADRIGAL, ET AL., 2014; WELLMER ET AL., 2006). However, these studies were conducted using microarray setups, whereas in this work a non-biased transcriptomics analysis (RNA-seq) was performed. Specifically, the floral induction system pAP1:AP1-GR *ap1 cal* was used as a model to understand non-biased abundance changes for transcripts (RNA-seq) and proteins (LC-MS/MS) in a temporal sequence after the activation of the early flower development programme. For the transcript-protein comparison to be possible, an imputation guideline was developed for dealing with different types of proteomic missing data existing in the quantitative MS dataset. Gene and protein expression was analysed on a genome-wide scale, identifying transcript-protein pairs with significant expression changes for both molecules at different stages of flower development. The differences in expression patterns from mRNA and proteins strongly suggest the existence of complex regulatory mechanisms for protein and transcript levels.

2.2 Results

2.2.1 Integrated transcriptome and proteome analyses in Arabidopsis early flower development

An APETALA1-based floral induction system (pAP1:AP1-GR *ap1 cal* line) (Ó'MAOILÉIDIGH ET AL., 2013, 2023) was used to characterize proteomics (LC-MS/MS) and transcriptomics (RNA-seq) changes during early flower development. In this system, dexamethasone (DEX) treatment activates the AP1 protein fused to a glucocorticoid receptor, causing the simultaneous transformation of the inflorescence-like meristems of *ap1 cal* plants into floral primordia and initiating the normal flower development process (Ó'MAOILÉIDIGH ET AL., 2023). DEX-treated plants were compared to mock-treated samples to study whether and how mRNA levels were correlated to proteome changes during early flower development. Samples were collected at one-day intervals after floral induction, encompassing six time points up to day 5 (as in (WELLMER ET AL., 2006)) which included up to stages 6-7 of flower development (SMYTH ET AL., 1990) (**Figure 2.1**). More than 74,000 peptidic fragments from almost 9,000 proteins were identified in at least one sample, and around 23,000 transcripts (84% of the Arabidopsis genome) were quantified in the RNA-seq experiments. Overall, 8,708 protein-coding genes were identified at both transcript and protein level, although only 7,003 pairs corresponded to quantifiable proteins. There were 95 proteins that did not have their matching transcript in the RNA-seq dataset, although expression of 88 of these genes was detected in previous microarray analyses (WELLMER ET AL., 2006).

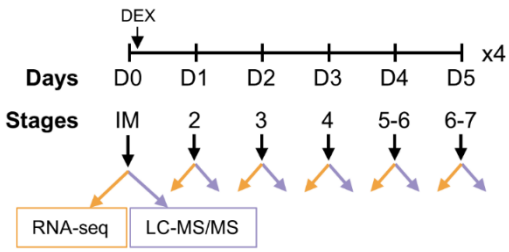


Figure 2.1. Experimental setup.
Inflorescence samples of four biological replicates of 40 to 80 plants each were collected immediately after DEX-solution application (D0), and at 1, 2, 3, 4 and 5 days (D1-5) after the treatment.

2.2.2 Set up of a reliability analysis to deal with missing values in the proteomics dataset

The proteomics data corroborated a substantial number of annotated open-reading frame borders based on the detection of 1,972 N-terminal and 1,560 C-terminal peptides (**Figure 2.2A**), of which 1,761 and 1,499, respectively, were unique peptides (discarding peptide sequences that only differ by post-translational modifications -PTMs-). N-terminal peptides often showed cleavage of the initiator methionine, and N-terminal acetylation was strongly dependent on the amino acid adjacent to the initiator methionine, as previously described (MERGNER ET AL., 2020) (**Figure 2.2B, C**). The mass spectrometry data covered, on average, around 21% of each protein sequence, enabling the detection of 75,244 unique peptidic fragments for 8,924 proteins.

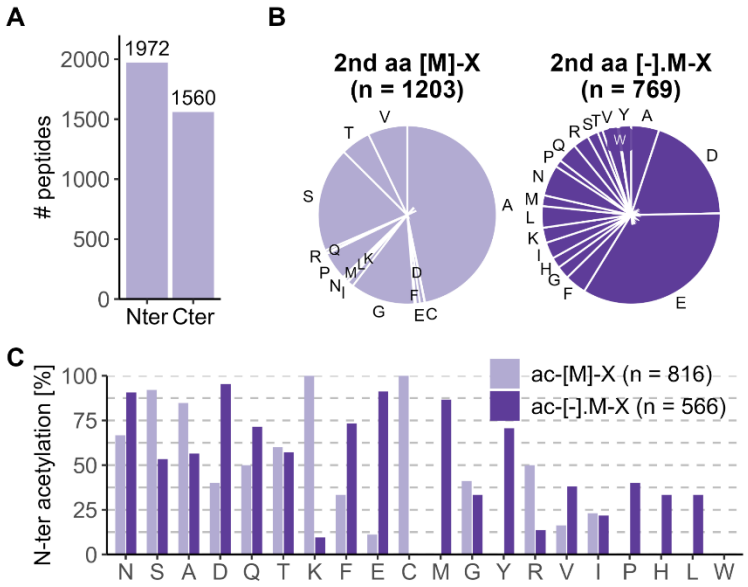


Figure 2.2. Proteomics results overview.
A) Number of identified amino (N-ter) or carboxy (C-ter) terminal peptides of proteins. **B)** Frequency of amino acids following the initiator methionine in N-ter peptides with ([M]-X) or without ([-].M-X) cleavage of the initiator methionine. X denotes the amino acid after the start codon. **C)** Frequency of protein N-ter acetylation for amino acids in B.

As missing mass spectrometry detection data could represent low abundance (below detection threshold) or simply no protein presence (i.e., Missing Not At Random -MNAR- values) instead of technical artifacts (i.e., Missing At Random -MAR- values), a pipeline with a series of rules was elaborated to deal with non-assigned (NA) values in the proteomics dataset, taking advantage of the characteristics of the experimental design, that is, successive timepoints and replicates. In such pipeline, it was considered that the reliability of detection of a protein would depend on the number of missing values per timepoint (n = 4 biological replicates) in the dataset. A protein at a given timepoint was classified as Reliably Detected (RD), Unreliably Detected (UD), Unreliably Undetected (UU) or Reliably Undetected (RU) depending on the number of replicates of that timepoint (day) in which the protein showed NAs, and the number of NAs in the immediately adjacent timepoints (**Table 2.1; Figures 2.3A, B, 2.4;** *see* also **Materials and Methods** section **2.4.5**). In total, 7,033 proteins (out of the initial set of 8,924) were considered as 'quantified' (RD or UD in at least one timepoint), whereas the remaining 1,891 MS-identified proteins were discarded for further analyses because they were classified as RU or UU at all timepoints (**Figure 2.5A**; the peptide-based coverage of 1,685 of the 1,891 discarded proteins was lower than 3 peptides per protein**, Figure 2.5B**, pointing out the limitations of the mass spectrometry technique to measure accurately the expression levels of some proteins). In each timepoint, about 5,000 proteins were classified as RD, and a total of 3,176 proteins were classified as RD or UD for all timepoints (**Figure 2.3C**). The highest number of RU proteins corresponded to day 0 (D0) timepoint (**Figure 2.3C**). Finally, in the last step of the pipeline all the remaining NA values were imputed using the kNN method.

To check for the appropriateness of the Reliability Analysis, a group of 69 flower-'marker' proteins was selected on the basis that their corresponding genes are known as up- or down-regulated in floral organs and/or throughout flower development (KAUFMANN ET AL., 2010; PAJORO, MADRIGAL, ET AL., 2014; WELLMER ET AL., 2006), as well as seven 'supermarker' proteins which met this requirement but also are well-characterized transcription factors related to flower initiation and development (**Sup Table 2.1**).

Approximately 60% of marker and all 'supermarker' proteins were retained as 'quantified' after the Reliability Analysis (RA), as they had at least one timepoint classified as RD or UD (**Figure 2.4A**).

Table 2.1. Description of the Reliability Analysis.
Supporting neighbour: 0, 1 or 2 NAs.
Unsupporting neighbour: 3 or 4 NAs.

	Initial/Final timepoint (D0/D5)	**Intermediate timepoint (D1-D4)**
Reliably Detected (RD)	0 or 1 NA	
Unreliably Detected (UD)	2 or 3 NAs + supporting neighbour	
Unreliably Undetected (UU)	2 or 3 NAs + unsupporting neighbour	2 or 3 NAs + unsupporting neighbour OR 4 NAs + supporting neighbour
Reliably Undetected (RU)	4 NAs	4 NAs + unsupporting neighbour

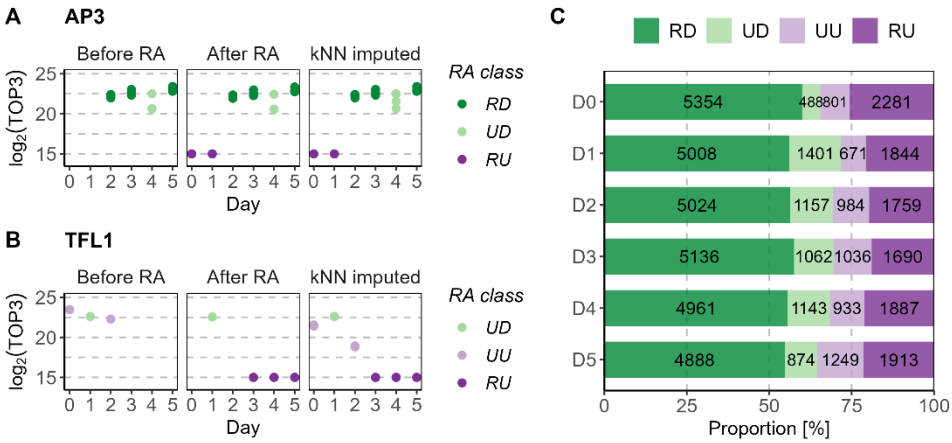


Figure 2.3. Imputation of missing values considering their biological context.
Log_2(TOP3) abundances through time of AP3 (**A**) and TFL1 (**B**) before and after the Reliability Analysis (RA), and after kNN imputation. **C)** Proportion of proteins considered as RD, UD, UU, and RU for each time point. For A-C: RD: Dark green, UD: Light green, UU: Light purple, RU: Dark purple.

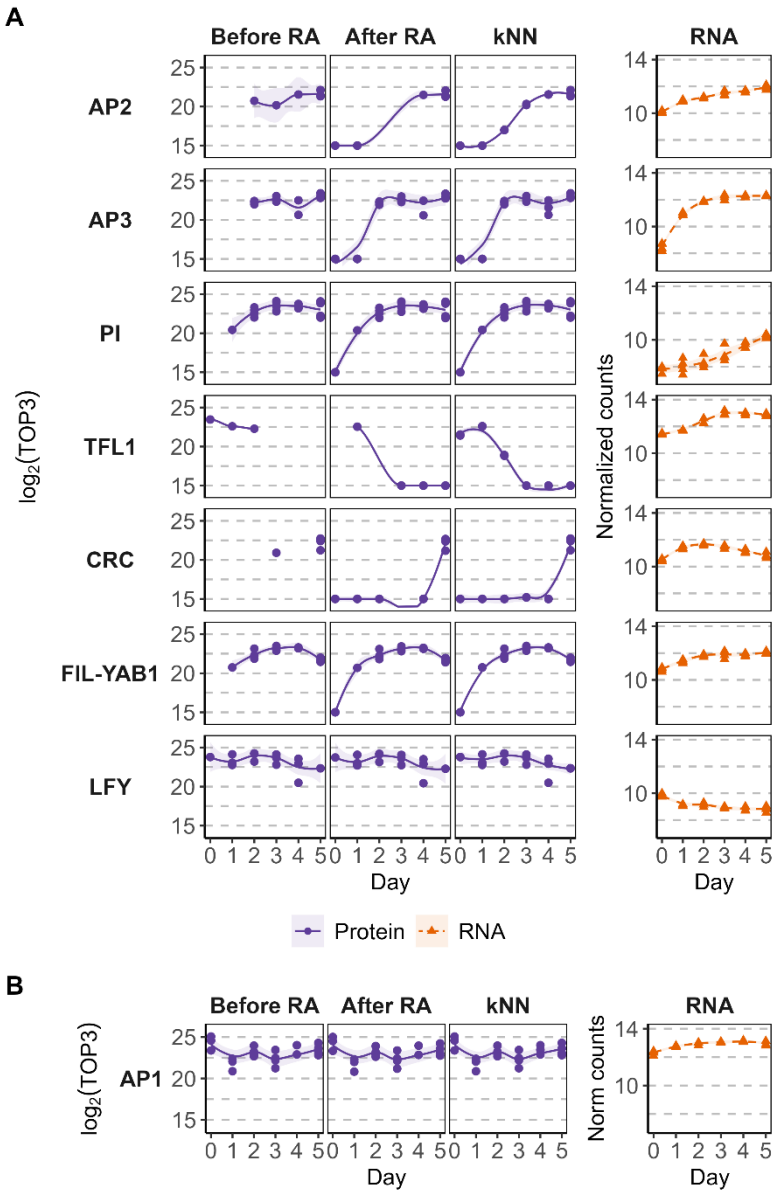


Figure 2.4. Expression of the 'supermarker' proteins through the time-course.

Representation for the 7 supermarker proteins (**A**) and AP1 (**B**) of the log_2(TOP3) protein abundance values through time before the Reliability Analysis (RA), after the RA, and after kNN imputation, and the normalized RNA counts.

The effect of the RA and kNN imputation was also analysed by Principal Component Analysis (PCA) and hierarchical clustering. PCA was performed with the subset of proteins without missing values before performing the RA and the imputation, and with all the proteins classified as 'quantified' after those data processing steps (**Figure 2.6A**). After the RA, the variability observed (33% of which could be explained by PC1 and PC2, **Figure 2.6B**) was discretely grouped by timepoint, most clearly in the case of D0 and D1 but also for D2-D5 replicates (with the exception that replicate 3 of D5 was closer to D4 replicates). Hierarchical clustering performed before and after the RA and kNN imputation showed that after the classification and imputation, replicates all clustered together by day, with adjacent days also clustering together (**Figure 2.6C**). This clearer separation through time and according to the flower developmental stages demonstrated the robustness of the LC-MS/MS data followed by a Reliability Analysis and imputation approach.

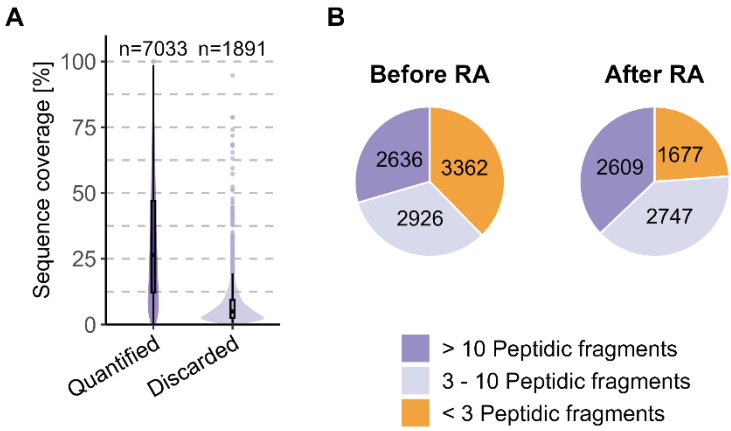


Figure 2.5. Proteomics sequence coverage.
A) Distribution of peptide-based sequence coverage of proteins which were Reliably or Unreliably Detected in at least one timepoint – day – (quantified) and those that were Reliably or Unreliably Undetected at every timepoint (discarded). **B)** Pie charts showing percentage and number of proteins identified by < 3, 3-10 or > 10 peptidic fragments before and after the Reliability Analysis.

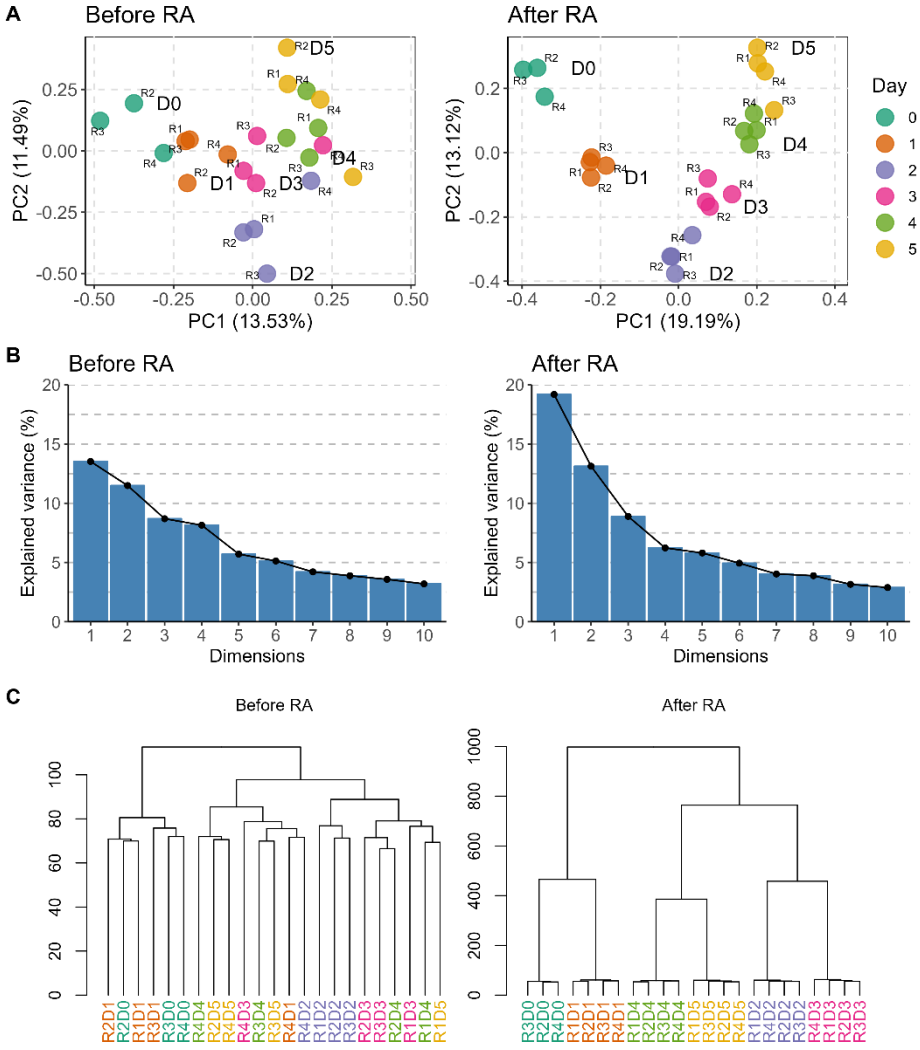


Figure 2.6. Effects of the reliability analysis in the data.
A) PCA of proteins without NAs before the RA and all proteins after the RA and kNN imputation. **B)** Percentage of variances explained by each principal component for the proteomics data (visualization of the eigenvalues). **C)** Hierarchical clustering of all proteins before and after RA and kNN imputation. R1D0 was discarded in all analyses because of its great differences with the rest of the data (only 165 proteins were quantified in this sample, *see* **Materials and Methods** section **2.4.5**).

2.2.3 Stage-variant proteins showed different abundance patterns over time

To determine which of the 7,033 quantified proteins showed significantly altered levels throughout early flower development, an ANOVA analysis was performed, resulting in the classification of a total of 2,037 proteins as stage-variant proteins (SVPs) (false discovery rate -FDR- = 5%), among which 1,430 were considered as RU at least at one timepoint (**Table 2.2, Sup Table 2.2**). SVPs presented different expression patterns that can be summarized as: i) increased expression over time (groups A.1 and A.2), ii) reduced expression over time (groups B.1 and B.2), iii) transient expression at middle timepoints (groups C.1 and C.2), and iv) gap of expression at middle timepoints (groups D.1 and D.2) (**Figure 2.7**).

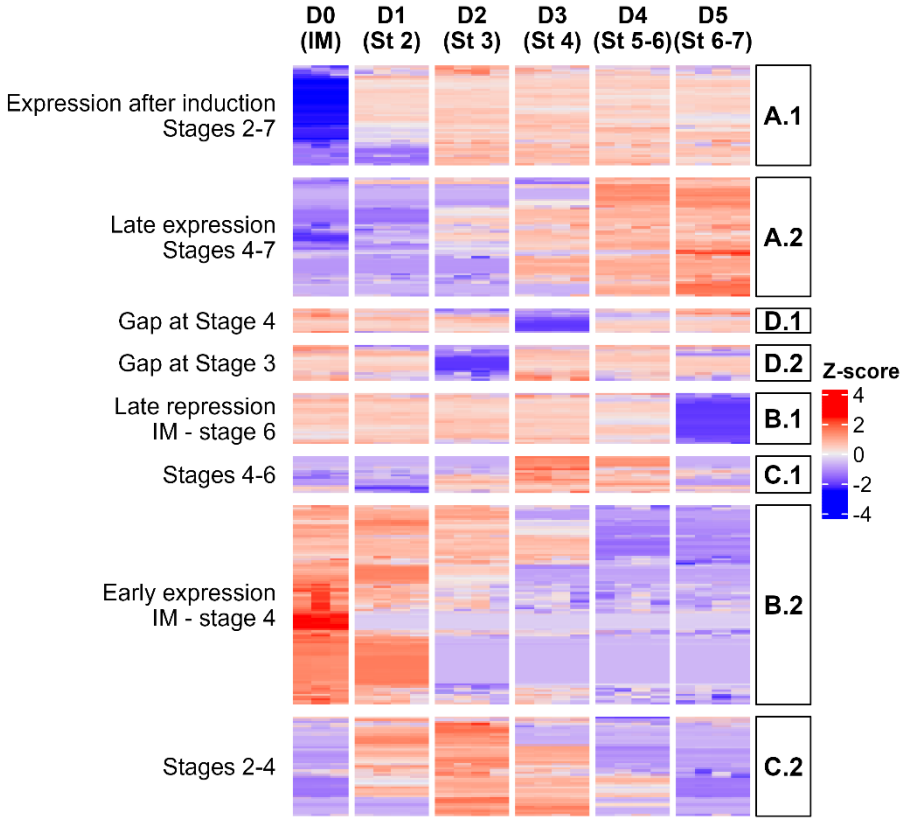


Figure 2.7. Stage-variant proteins (SVP).
Heatmap of the 2,037 SVPs through the time course. Colour scale represents Z-scored TOP3 values.

Table 2.2. Summary of stage variant proteins depending on their classification in the Reliability Analysis.

	RD or UD in all time points	RU or UU at some time points	
SVPs	490	1,547	*2,037*
NVPs	2,686	2,310	*4,996*
	3,176	*3,857*	***7,033***

2.2.4 Patterns of gene expression changes throughout the time course

In the RNA-seq experiment, 23,088 genes were identified with more than ten counts across all samples. A PCA comprising all these genes separated the early flower developmental stages by timepoint (~37% of the variability could be explained by PC1 and PC2, **Figure 2.8A, B**). Samples clustered following a trajectory along PC1 that reflects the time factor, with later timepoints placed more distant relative to D0 (**Figure 2.8A**). A moderated Likelihood Ratio Test (LRT) was applied in order to get a statistical metric for ranking genes according to the differences in their expression profiles over time. There were 8,125 transcripts in the dataset with a significant variation through time, from now own called Stage-Variant Genes (SVGs) (LRT with adjusted p-value ≤ 0.01). These SVGs can be considered as 'related to' or 'influenced by' AP1 expression (**Sup Tables 2.3, 2.4**).

The total 8,125 genes defined as SVGs showed four different transcript accumulation patterns: i) increment in expression through time, ii) higher expression during mid-term stages (D1-4), iii) down-regulated expression during mid-term stages with high expression levels at D0 and D5, and iv) decrease in expression over time (**Figure 2.8C**).

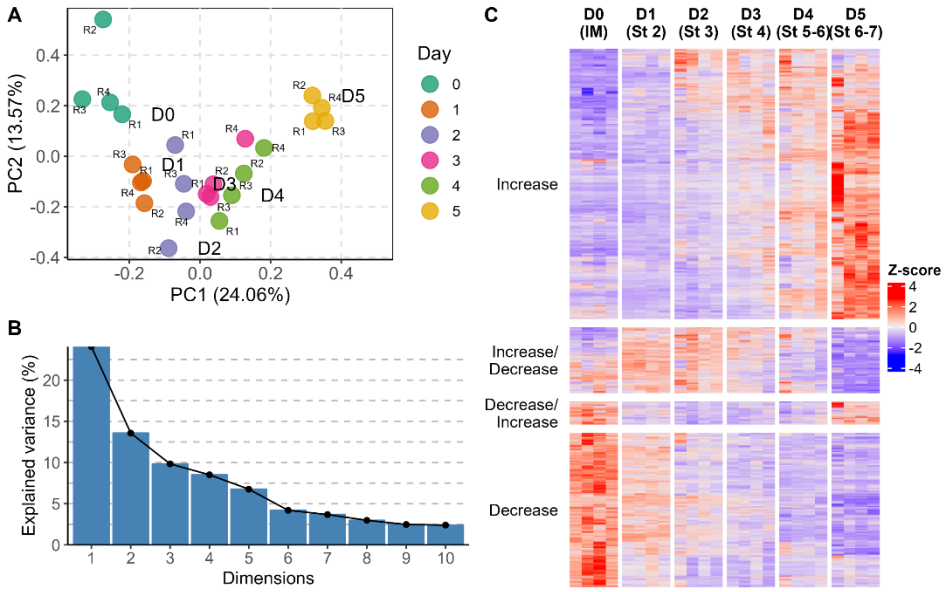


Figure 2.8. Stage-Variant Genes (SVGs).
A) PCA of the RNA-seq data, showing each biological replicate (R1 to R4) and coloured by timepoint (D0 to D5). Samples clustered according to PC1, except for replicate 4 D3, which clustered closer to samples from D4. The later the timepoint, the more distant relative to D0. Distances between D2 and D1, and D5 and D4 were substantial, while distances between D2, D3 and D4 were smaller. **B)** Percentage of variances explained by each principal component for the RNA-seq data (visualization of the eigenvalues). **C)** Heatmap displaying the expression patterns of the SVGs (Z-scored RNA counts, n = 8,125). Colour scale represents Z-scored normalized RNA counts.

2.2.5 RNA-seq results expand previously published transcriptome data and identify novel AP1 targets

The RNA-seq results were compared to those obtained in a previous microarray study conducted with the same experimental time-course design but using a different AP1-based floral induction system (p35S:AP1-GR *ap1 cal*; (WELLMER ET AL., 2006)). To make the RNA-seq data results more comparable to those of the microarray (a ratiometric gene expression platform), a differential expression analysis between subsequent stages (i.e., D1 vs D0, D2 vs D1, D3 vs D2, D4 vs D3, and D5 vs D4) was performed as in (WELLMER ET AL., 2006). In this analysis, and using the same criteria for differential expression (no Logarithmic Fold Change –LFC– cut-off, and an

adjusted p-value < 0.05) a total of 5,150 genes were classified as differentially expressed genes (DEGs), compared to the 1,653 DEGs identified by Wellmer *et al.* (**Figure 2.9A, Sup Table 2.5**). That is, the RNA-seq expanded by at least three times the scope of the transcriptome previously identified as changing during early flower development. Furthermore, with the LRT approach an even higher number of variable genes was identified (8,125 SVGs versus 5,510 DEGs) as the LRT statistic is more sensitive than pairwise comparisons to slight changes in expression levels between subsequent days (**Figure 2.9A**).

As observed previously from the microarray data, the RNA-seq results showed that, between D1 and D5 and on every *day-to-previous day* comparison, the number of up-regulated genes was higher than that of the down-regulated genes – likely corresponding to the initiation of organ primordia and potentially representing the activation of genes involved in floral organ development –, whereas for the first timepoint after the induction (D1 vs D0), there was a preponderance of gene downregulation (**Figure 2.9B**). Interestingly, substantially more gene expression changes were detected by RNA-seq in the first and last time points (D1 vs D0, and D5 vs D4 comparisons, respectively) than in the intermediate timepoints (D2 vs D1, D3 vs D2, D4 vs D3 comparisons) (**Figure 2.9B**). Finally, I compared the LFC of those genes whose estimated LFCs were supported by enough statistical confidence in the RNA-seq and microarray results (adjusted p-value < 0.05) at every *day-to-previous day* comparison. Over 86% of the DEGs that were quantified in both the RNA-seq, and microarray experiments were either overexpressed, or else underexpressed in both analyses at every *day-to-previous day* comparison (**Figure 2.9C**).

Focusing on the expression levels over time of the 5,150 RNA-seq DEGs, a time-dependent clustering analysis revealed three main kinds of trajectories during the early stages of flower development captured by the time-course, showing either an increasing (A) or decreasing (B.1, B.2) pattern of abundance through time or an increment tendency up to stages 2-3 (D3-D4) followed by a decrease thereafter (C.1, C.2) (**Figure 2.9D**). These were the same main trajectories identified for the microarray DEGs in Wellmer *et al.*, as well as for the 8,125 RNA-seq SVGs (**Figure 2.8C**). In addition, in the case

of the SVG classification, a set of genes that were repressed at first and then activated was identified as a separated group (**Figure 2.8C**), whereas for the RNA-seq DEGs, individual genes with those trajectories could be visualized but were not grouped together (**Figure 2.9D**).

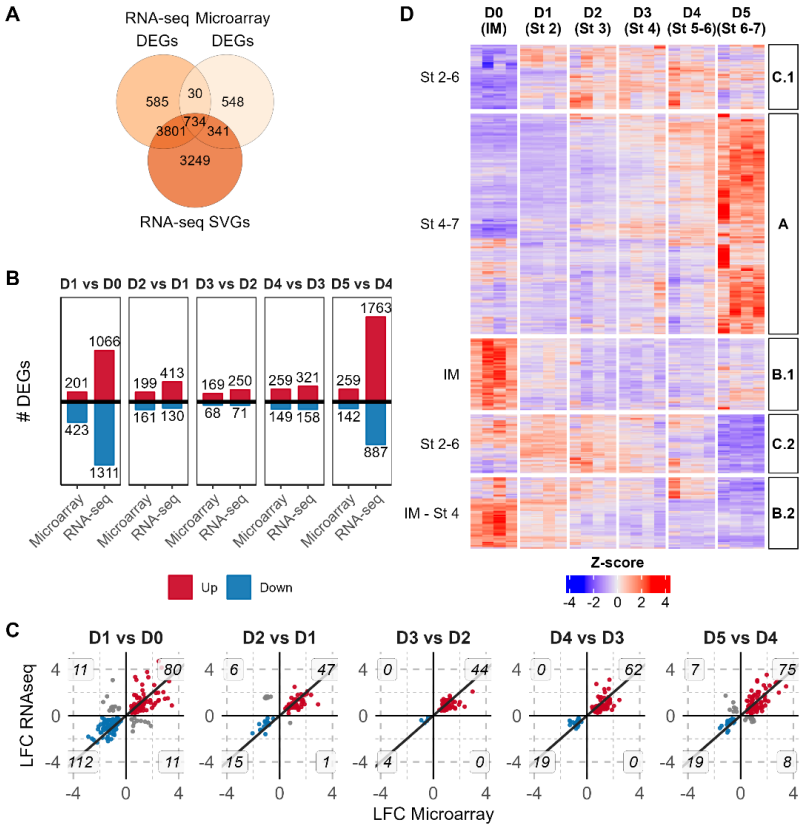


Figure 2.9. DEGs during early flower development in pAP1:AP1-GR *ap1 cal* plants. Comparison with (WELLMER ET AL., 2006) p35S:AP1-GR *ap1 cal* microarray results.

A) Venn diagram showing the number of microarray DEGs and RNA-seq DEGs and SVGs and the overlap between the datasets. **B)** Bar plots showing the number of up- and down-regulated DEGs in RNA-seq and microarray results at each *day-to-previous-day* comparison (adj. p-value < 0.05). **C)** Microarray – RNA-seq data comparisons for each *day-to-previous-day* combination (adj. p-value < 0.05). The diagonal line represents y = x. Grey dots indicate those DEGs with opposite trajectories in the two datasets. Up-regulated genes are coloured in red, and down-regulated genes are coloured in blue. The number of pictured DEGs is indicated in each quadrant. **D)** Heatmap of the RNA-seq DEGs (n = 5,150). Colour scale represents Z-scored normalized counts values.

A combination of genome-wide DNA binding by AP1 (ChIP-seq) and gene expression profiling (microarray data) was used in a previous study to identify AP1 direct target genes, which was conducted with a 35S:AP1-GR *ap1 cal* line and a 12-hour time-course after floral induction (KAUFMANN ET AL., 2010). In that study, 249 AP1-high confidence targets (HCTs) were identified. From those, 247 were detected as expressed in the RNA-seq data reported here, and 183 were within the group of genes classified as SVGs (**Sup Tables 2.3, 2.4**).

Since the RNA-seq dataset substantially expanded the scope of the transcriptome identified as changing during early flower development, the possibility that it could help identify novel AP1 direct targets was explored. In the RNA-seq experiment, the D1 versus D0 comparison was the closest one to the experimental design used in (KAUFMANN ET AL., 2010) (12-hour time-course), and it was therefore used for the analysis (i.e., all other timepoints were excluded). Among the 2,377 DEGs identified in D1 vs D0 time comparison (**Figure 2.9B**), there were 81 of the HCTs defined in (KAUFMANN ET AL., 2010), including key flowering time genes that are downregulated by AP1, such as *FD, TFL1, SPL9,* and *SPL15*, other downregulated HCTs as *AGL20, SAP, LSH1, LSH2,* and LSH4, and flower development HCT genes that are upregulated, for instance *LFY, SEP3, GA2ox1, RGA-like2, ATHB1,* and *AP2*. These results validated the appropriateness of using the RNA-seq dataset to identify novel targets by combining it with the previous ChIP-seq dataset, despite the differences in the AP1-GR lines that were used in both studies (pAP1:AP1-GR *ap1 cal* vs 35S:AP1-GR *ap1 cal*), in the experimental design (time-course in days vs time-course in hours), and in the method used to detect gene expression (RNA-seq vs microarrays).

The criteria for classifying a gene as an AP1 HCT in (KAUFMANN ET AL., 2010) included (i) containing one or more AP1 ChIP-seq binding sites within 3 kb upstream of the 5' end and 1 kb downstream of the 3' end of the gene (which defined a set of 2,298 putative AP1 targets), and (ii) showing robust differential expression in the time-course (> 1.8-fold) (which restricted the set of 2,298 genes to 249 HCTs). All the 81 HCTs that were detected in the RNA-seq D1 vs D0 comparison were above an absolute LFC of 0.29. Therefore, this LFC value was used as a threshold to search for novel AP1-

HCTs in the RNA-seq data. The 1,782 D1 vs D0 DEGs that showed robust expression changes (with an absolute LFC > 0.29 and adjusted p-value < 0.05) were compared to the list of 2,298 putative AP1 targets identified in (KAUFMANN ET AL., 2010). In total, this comparison defined a set of 311 putative AP1-HCTs, the 81 indicated above and 230 that were newly identified from this RNA-seq-based analysis (**Sup Table 2.3**). The latter included flowering time genes *SVP* and *AGL24* (known to be regulated by AP1 but not identified as HCTs in (KAUFMANN ET AL., 2010)) or *SPL5*, all downregulated, or genes that participate in flower development such as *SEP2*, *SEP4*, *BLH11*, *CUC1*, or *PIN1*, upregulated.

In summary, all these results have substantially expanded the identification of genes whose expression changes during early flower development and the set of putative AP1 high confidence targets, in addition to corroborating the previous findings indicating that AP1 acts predominantly as a transcriptional repressor during the earliest stage of flower development, and predominantly as a transcriptional activator afterwards, and to providing further support for previously identified AP1 HCTs.

2.2.6 Correlation between RNA and protein levels during early flower development

The dynamic range of protein and transcript expression, as determined by MS and RNA-seq spanned six and four orders of magnitude, respectively (**Figures 2.10, 2.11A**). Protein evidence was underrepresented for low-abundance transcripts (ANOVA with Tukey post-hoc test; p-value ≤ 0.001), as described in other RNA-protein comparison studies (HOOGENDIJK ET AL., 2019; MERGNER ET AL., 2020), and the median expression levels for transcripts were similar within days (**Figure 2.11A**).

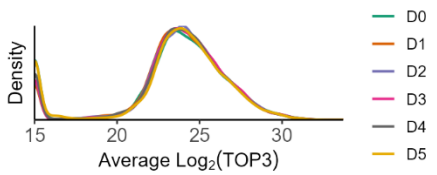


Figure 2.10. Density plot of protein abundance expressed as the average Log_2 TOP3 for each time point.

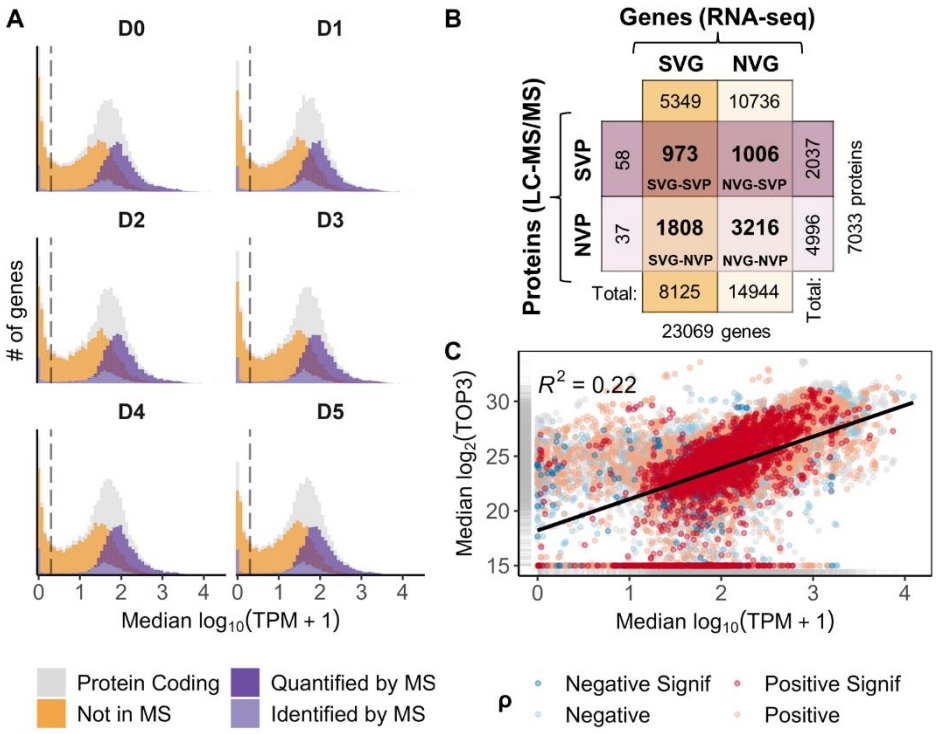


Figure 2.11. Gene and protein classification depending on abundance through time.

A) Histogram of RNA expression range. Grey: all detected protein-coding transcripts; orange: protein-coding transcripts not detected as protein in MS; dark purple: protein coding transcripts quantified as protein; purple: transcripts corresponding to a protein identified by LC-MS/MS that was discarded because it was classified as UU or RU at every timepoint (i.e., not quantified). Dashed line indicates TPM = 1. **B)** Schema illustrating the number of expressed genes, Stage Variant Genes (SVG), quantified proteins and Stage Variant Proteins (SVP) identified. The sum of gene-protein pairs differs from the number of genes and proteins identified separately because there are cases of the same AGI associated to more than one Uniprot code and *vice versa*. **C)** Scatter plot of protein abundances and RNA expression levels for all RNA-protein pairs at every timepoint. Coloured by RNA-protein correlation (Spearman's rank coefficient, ρ). Positive if ρ ≥ 04. Negative if ρ ≤ -0.4. Significant if p-value < 0.05.

Amongst the 7,003 quantified transcript-protein pairs, there were: i) 973 pairs with stage-dependent variation during early flowering development at both RNA and protein levels (SVG-SVP), ii) 1,006 pairs non-variant at the

RNA level, but stage-variant for proteins (NVG-SVP), iii) 1,808 pairs stage-variant at transcript level, and non-variant for proteins (SVG-NVP), iv) and 3,216 pairs which presented non-variable levels for both molecules (NVG-NVP) (**Figure 2.11B, Sup Table 2.6**). The seven supermarker proteins, 37 of the marker proteins and 43 of the AP1-bound HCTs defined in (KAUFMANN ET AL., 2010) were found as quantified at both transcript and protein levels. These three subsets were significantly enriched in SVG-SVP pairs (Fishers' t-test, p-value ≤ 0.001), especially the group of supermarkers, from which six out of seven were classified as SVG-SVP. In addition, SVG-NVP pairs were proportionally more abundant in the AP1-targets group in comparison with the markers group (Fishers' t-test, p-value ≤ 0.05) (**Figure 2.12**).

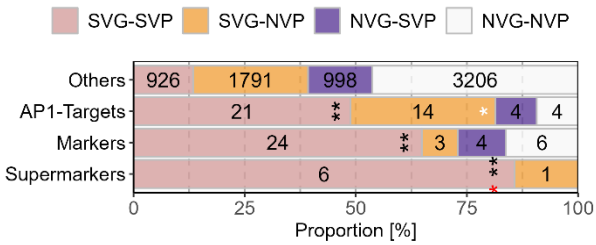


Figure 2.12. Relative distribution of absolute numbers of transcript-protein pairs in selected classes across the expression categories: SVG-SVP, SVG-NVP, NVG-SVP, and NVG-NVP.
There are two markers which are also AP1-targets (both SVG-SVP), and three supermarkers which are also AP1-targets (two SVG-SVP and one SVG-NVP). Fisher's t-test results (asterisks): black = Significantly enriched when compared with the summation of the rest of subsets (p-value < 0.001); red = Significantly enriched when compared with the AP1-targets-subset (p-value < 0.05); white = Significantly enriched when compared with the markers-subset (p-value < 0.05).

To provide a measure of similarity among developmental stages and to check whether there is a shift between mRNA and protein levels at different timepoints, Pearson's correlation coefficient (r) was calculated for the gene expression and protein abundance of all pairwise timepoint combinations (e.g., protein D0 vs protein D0-D5, protein D0 vs RNA D0-D5, etc.). Correlations were computed on protein and transcript level for all RNA-protein pairs and the SVG-SVP pairs (**Figure 2.13A**) separately. Pearson's coefficients were slightly higher for the SVG-SVP group. For both SVG-SVP

pairs and all transcript-protein pairs, the correlation RNA-protein seemed to be moderately shifted at D3 and afterwards in the time course, as protein levels at D4 correlated equally well with RNA levels at D3, and protein levels at D5 correlated equally well with RNA levels at D4. PCA for average *Z*-scored values (*Z*-scored independently) showed that RNA-protein levels clustered according to the timepoint, being this correlation more obvious at D2, D3 and D4 (**Figure 2.13B**). D0, D1 and D5 presented the greater differences between RNA and protein levels; in fact, D5 protein levels correlated better with RNA levels at D4 than at D5. Although the distribution along the PCA is similar for both groups (all RNA-protein pairs and SVG-SVP pairs), the percentage of variability that could be explained by PC1 and PC2 is higher for the SVG-SVP pairs (67% against the 43% for all RNA-protein pairs).

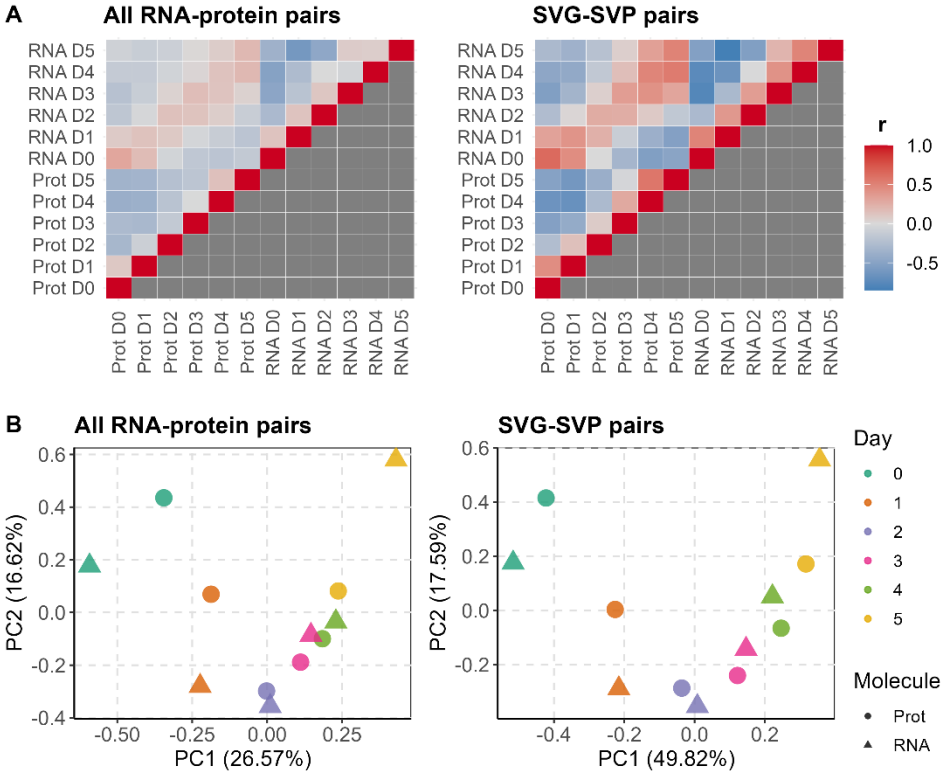


Figure 2.13. RNA-protein comparisons.
A) Pearson's correlation coefficient (r) matrix of D0-D5 after floral induction on the transcriptome and proteome level using all RNA-protein pairs and SVG-SVP pairs separately. **B)** PCA for average Z-scored values of all RNA-protein pairs (Z-scored independently) and SVG-SVP pairs.

In general, there is a relatively good correlation between RNA levels and protein abundances during the early stages of flower development. Nevertheless, the differences observed, especially in the limits of the time-course (i.e., D0 and D5), can be explained by the time lag. This is relevant because D1 (versus D0) and D5 (versus D4) are by far the days where there were more RNA expression changes (in terms of DEGs) (**Figure 2.9B**); on D1 with a preponderance of downregulation, which probably does not correlate as well with protein levels as when it is upregulation (as factors such as protein half-life or degradation intervene), and on D5 it is new upregulation than would be translated into proteins partly on D6, according to the detected time lag.

The correlation of mRNA and protein levels through time was also measured by calculating the Spearman's rank correlation coefficient (ρ) for each RNA-protein pair. In total, there were 2,540 RNA-protein pairs with a positive correlation ($\rho \geq 0.4$, as defined in (AKOGLU, 2018)), and almost 6% of these pairs had a significant and highly positive correlation ($\rho \geq 0.8$ and p-value ≤ 0.05). In contrast, 975 RNA-protein pairs presented a negative correlation ($\rho \leq -0.4$), and around 1.5% of them with a significant and highly negative correlation ($\rho \leq -0.8$ and p-value ≤ 0.05) (**Figure 2.14, Sup Table 2.6**). Moreover, the mRNA-to-protein abundance correlation was very different for the SVG-SVP, SVG-NVP, NVG-SVP and NVG-NVP subsets (**Figure 2.15A, Table 2.4**).

RNA-protein pairs that vary at both molecule levels (SVG-SVP) presented the strongest positive correlation, with a median ρ of 0.6, a 63% of pairs with positive correlation and less than an 8% of pairs with negative correlation. For the other subsets, SVG-NVP, NVG-SVP and NVG-NVP, 44%, 30% and 26% of the RNA-protein pairs showed a positive correlation, as opposed to 10%, 16% and 16% of pairs with a negative correlation in each group respectively (**Table 2.4**, **Figure 2.15A**). To sum up, around 36% of the total RNA-protein pairs presented a positive correlation between their RNA and protein expression levels, and the correlation between RNA and protein expression levels was higher for those RNA-protein pairs with differential expression over time for both molecules.

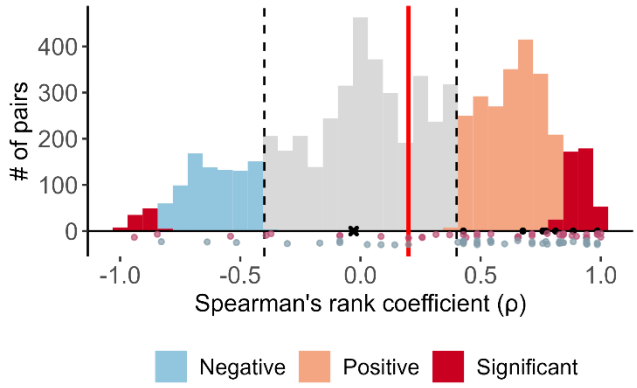


Figure 2.14. Correlation between each RNA-protein pair for the complete dataset.
Correlation analysis of protein-to-RNA abundance (non-*Z* scored) across samples measured as Spearman's rank correlation coefficient (ρ) for each RNA-protein pair. Red line represents the median correlation. Dashed lines indicate the limits to consider positive and negative correlations. The points represent ρ for: supermarkers (black), markers (pink) and AP1-targets (grey). The black cross represents ρ for AP1.

Table 2.3. Spearman's rank coefficient (ρ) among subsets were highly variable.
Spearman correlation between RNA and protein levels of each subset depended on the overall expression pattern of the molecules. Significant (sig.): adjusted p-value (BH) < 0.05.

	Median ρ	**Positive (ρ > 0.4)**	**Negative (ρ < -0.4)**	**Uncorrelated**	
SVG-SVP	0.6	615 (217 sig.)	76 (15 sig.)	282	*973*
SVG-NVP	0.31	794 (96 sig.)	197 (18 sig.)	817	*1,808*
NVG-SVP	0.13	300 (42 sig.)	167 (28 sig.)	539	*1,006*
NVG-NVP	0.08	831 (74 sig.)	535 (36 sig.)	1,850	*3,216*
	0.2	*2,540 (429 sig.)*	*975 (97 sig.)*	*3,488*	***7,003***

A total of 80% of both flower-markers and AP1-targets (HCTs) in the SVG-SVP subset presented a positive correlation, although there were some exceptions with a significant and highly negative correlation (e.g., WUSCHEL RELATED HOMEOBOX13 -WOX13- and AT4G27980) or without positive nor negative ρ (e.g., LUTEIN-DEFICIENT1 -LUT1- and PYRABACTIN RESISTANCE1-LIKE1 -PYL1-) (**Figure 2.15A, B**).

Besides, I inspected the time course trajectories (mRNA and protein) of the seven supermarkers, which showed a positive ρ above the median, and compared them to their previously published expression patterns (WELLMER ET AL., 2006) and found them to be in good agreement (**Figures 2.4, 2.15B**).

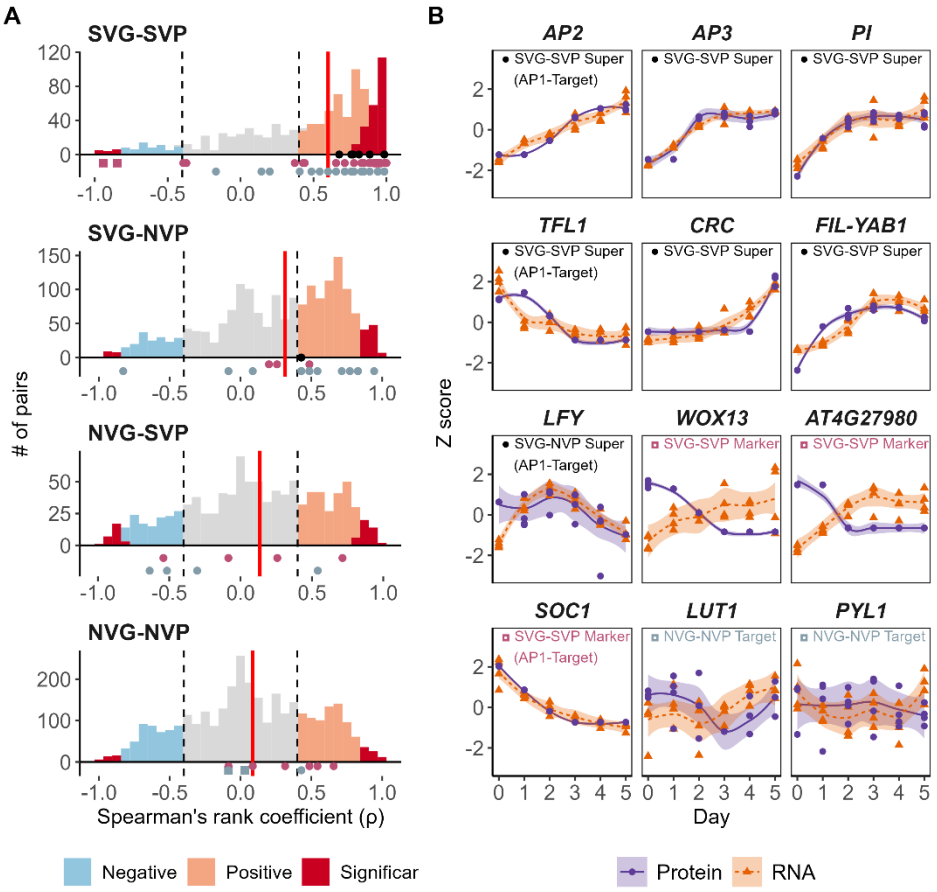


Figure 2.15. Correlation and trajectory patterns for gene-protein pairs. **A)** Spearman's rank correlation coefficient (ρ) between RNA and protein levels of each pair depending on the SV – NV classification. Red lines: median ρ of each subset. Dashed lines indicate the limits to consider positive and negative correlations. Points signal ρ for: supermarkers (black), markers (pink) and AP1-targets (grey). Squares represent the ρ for the markers and AP1-targets depicted in **B**. **B)** Z-scored abundances of RNA and protein levels (Z-scored separately) of selected proteins. The seven supermarkers (ρ ≥ 0.4) (SVG-SVP: AP2, AP3, PI, TFL1, CRC, FIL-YAB1; SVG-NVP: LFY), two markers with ρ ≤ -0.8 (SVG-SVP: WOX13, AT4G27980), one marker and AP1-target with ρ ≥ 0.8 (SVG-SVP: SOC1) and two AP1-targets with non-significant ρ (NVG-NVP: LUT1, PYL1).

2.2.7 RNA-protein pairs clustered in various expression pattern modules

In order to elucidate transcript-protein dynamics of the complete dataset, unbiased clustering based on the correlation of mRNA and protein expression patterns was performed. Weighted gene co-expression network analysis (WGCNA) using the SVG-SVP, NVG-SVP and SVG-NVP transcript-protein pairs separately resulted in 18, 18 and 25 co-expression eigen-modules (MEs), respectively, ranging in size from 10 to 485 gene-protein pairs (**Figure 2.16**).

Combined expression patterns of SVG-SVP pairs were categorized in four groups: i) increasing mRNA and protein levels (A.1), ii) increasing mRNA and decreasing protein levels (A.2), iii) decreasing mRNA and increasing protein levels (B.1), and iv) decreasing mRNA and protein levels (B.2), with groups A.1 and B.2 (that is, those in which RNA and protein levels change in the same direction) comprising the vast majority of SVG-SVP pairs (and of MEs) and with most of the different MEs in those groups showing a high correlation (i.e., substantial ρ values) (**Figure 2.16A**). In groups A.2 and B.1 in which mRNA and protein levels are anticorrelated, a few MEs also showed relevant ρ values (ME11, 13 and 17 in A.2 and ME19 in B.1), although the number of gene-protein pairs encompassed by those MEs is small (64 versus 824 in the most significant MEs of A.1 and B.2). These observations support the idea that there is a relatively good correlation between RNA and protein level changes during early flower development, and also identify a few specific and small subgroups of genes in which the changes are anticorrelated (*see* section **2.2.8**).

NGV-SVP pairs grouped in clusters with patterns that were similar to those observed for the complete list of SVPs (**Figures 2.7, 2.15B**), that is: i) increased protein abundance over time (NVG-SVP ME01, 05, 07, 09), ii) reduced levels over time (NVG-SVP ME02, 03, 04, 11, 14), iii) transient proteins expression at middle timepoints (NVG-SVP ME06, 10, 12, 13, 15, 16, 17), and iv) transient proteins with a gap in their expression at intermediate timepoints (NVG-SVP ME08, 18). Last, SVG-NVP grouped pairs also showed the same trajectory patterns as the complete set of SVGs (**Figures 2.8C, 2.16C**): i) increased expression through time (SVG-NVP ME02, 05, 06, 10, 13,

25, 26, 27, 29, 30), ii) higher expression during mid-term stages (D1-4) (SGV-NVP ME09, 14, 17, 29), iii) high expression at D0 and D5, but down-regulated expression during mid-term stages (SVG-NVP ME15, 22, 24, 28), and iv) expression reduction over time (SVG-NVP ME01, 04, 07, 08, 11, 12, 23).

The correlation between protein and RNA expression levels was different for each one of the modules, and, as indicated above (**Figure 2.16A**), it was especially high for the modules in which both molecules behaved similarly (SVG-SVP A.1 and B.2 modules).

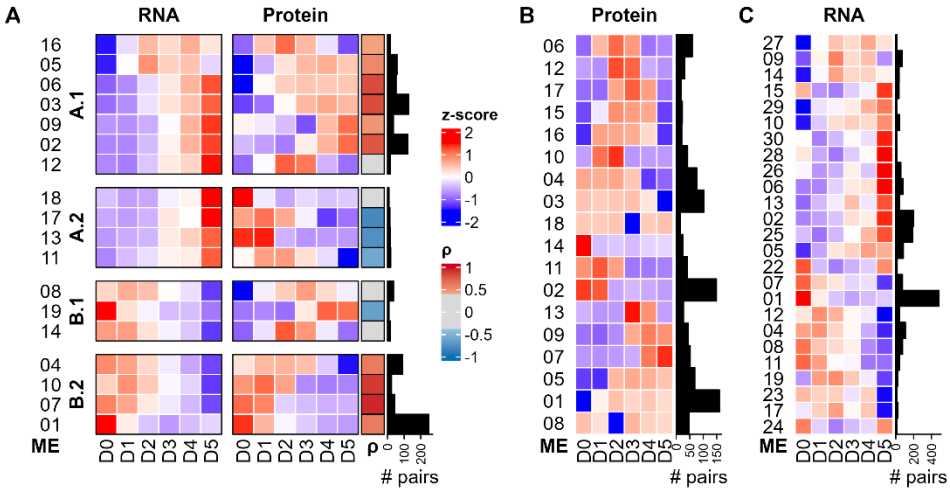


Figure 2.16. Trajectory patterns for gene-protein pairs.
Trajectory clustering (WGCNA) for SVG-SVP (18 modules) (**A**), NVG-SVP (18 modules) (**B**), and SVG-NVP (25 modules) (**C**). The right bar graph in each panel indicates the number of gene-protein pairs included in each module. The average ρ values for gene-protein pairs included in each SVG-SVP (**A**) modules are included. This value is not included for the NVG-SVP (**B**) and SVG-NVP (**C**) modules because it is between -0.4 and 0.4 in all cases (no-correlation, 'grey').

2.2.8 Modules with opposite patterns for mRNA and protein levels were enriched in hormone responsive pathways

Gene Ontology (GO) and KEGG enrichment analyses were performed to retrieve the functional biological processes that accompany early flower development (**Sup Tables 2.11, 2.12**). Interestingly, a high percentage of gene-protein pairs with decreasing levels of RNA combined with increasing protein abundance (i.e., SVG-SVP ME19) are known to correspond to proteins

localized to the chloroplasts, whereas gene-protein pairs with increasing levels of RNA combined with decreasing protein abundance (i.e., SVG-SVP ME11, 13, 17 and 18) contain proteins involved in fatty-acid metabolic process related with acetyl-CoA and jasmonic acid (JA) pathways.

Six out of ten genes included in the SVG-SVP ME19, with decreasing levels of RNA and increasing protein abundance, are expressed in chloroplasts (*AT1G79460, AT3G07310, AT1G29070, AT4G17300, AT2G29180* and *AT5G23040*). Among these, GA2 (AT1G79460) and a putative phosphoserine aminotransferase (AT3G07310) are included in the gibberellic acid signalling pathway, PRPL34 (AT1G29070) is a structural constituent of the ribosome, AT2G29180 (thylakoid membrane protein) positively regulates transcription, NS1 (AT4G17300) is also related with chloroplast transcription, as it acts as a ligase on tRNA (asparaginyl-tRNA aminoacylation for amino acid activation) and CDF1 (AT5G23040) is a thylakoid membrane chaperone required for chloroplast biogenesis and development. All these proteins are related with cellular response to lipids and gibberellins, although nor this module nor any other modules were significantly enriched in gibberellin-related pathways (adjusted p-values > 0.05). Gibberellin indirectly promotes chloroplast biogenesis to maintain the chloroplast population of expanded cells, yet the relationship between chloroplast biogenesis with cell division and cell expansion remains poorly understood (X. JIANG ET AL., 2012).

On the other hand, gene-protein pairs with an increasing pattern in their RNA levels combined with a decrease in protein abundance (i.e., SVG-SVP ME11, 13, 17 and 18) are enriched in proteins involved in fatty-acid metabolic processes related with acetyl-CoA and jasmonic acid (JA) pathways. These pathways are required for proper flower developmental processes such as flower maturation (REEVES ET AL., 2012). JA induces the expression of *WOX13* (included in SVG-SVP ME13) orthologous *BpWOX9* and *BpWOX10 in Broussonetia papyrifera* (TANG ET AL., 2017). Besides, WOX13 interacts with other member of SVG-SVP ME13: RAD-LIKE3 (AT4G36570) (TRIGG ET AL., 2017), a probable transcription factor assigned as a member of the MYB-related family, whose members show considerable response to JA signalling (ALI & BAEK, 2020; ZHAI ET AL., 2015). *TEOSINTE BRANCHED/CYCLOIDEA/PCF*

(TCP) genes, such as *TCP15* also found in this module, controls the biosynthesis of JA (SCHOMMER ET AL., 2008). In addition, other gene-protein pairs with this specific expression pattern regulate JA-dependent and JA-independent responses, such as the calmodulines CML11, 16 and 19 and the CYTOCHROME P450 family members AT4G12300 and AT1G13080 (LEON ET AL., 1998).

Other modules with increasing levels of mRNA (SVG-NVP ME02 and ME06), or of mRNA and proteins (SVG-SVP ME09) were enriched in auxin metabolic processes, whereas modules with decreasing levels of mRNA (SVG-NVP ME01), or of mRNA and protein (SVG-SVP ME01) were enriched in cytokinin-responsive processes.

In summary, the main result from the GO analysis is that modules with opposite patterns for mRNA and protein levels were enriched in hormone responsive pathways, although determining the possible functional significance of this observation and the molecular mechanisms that would underlie divergent RNA and protein trajectories would require additional studies.

2.2.9 Physically interacting proteins had different RNA-protein expression levels through time

To investigate possible functions of, or functional relationships within, the different MEs (that is, of or within the various groups with different RNA-protein trajectories), the interaction network of all MS-detected proteins and other Arabidopsis proteins was analysed by by collecting information about known and predicted protein-protein interactions (ppi) from STRING (SZKLARCZYK ET AL., 2017).

The final ppi network (6,403 nodes, 66,350 edges) was divided into five main clusters with more than 30 proteins, and 55 smaller protein groups, ranging from two to 23 proteins (**Figure 2.17, Sup Tables 2.7-2.9**). No clear association between RNA-protein trajectory patterns and the interaction clusters was found, as each of the interaction clusters contained proteins that presented different RNA-protein patterns (**Figures 2.17**), except for the

interaction cluster 2, which had a central hub composed by proteins whose RNA-protein levels were downregulated during the time course and whose main KEGG pathway was 'photosynthesis and carbon fixation' (**Sup Table 2.10**).

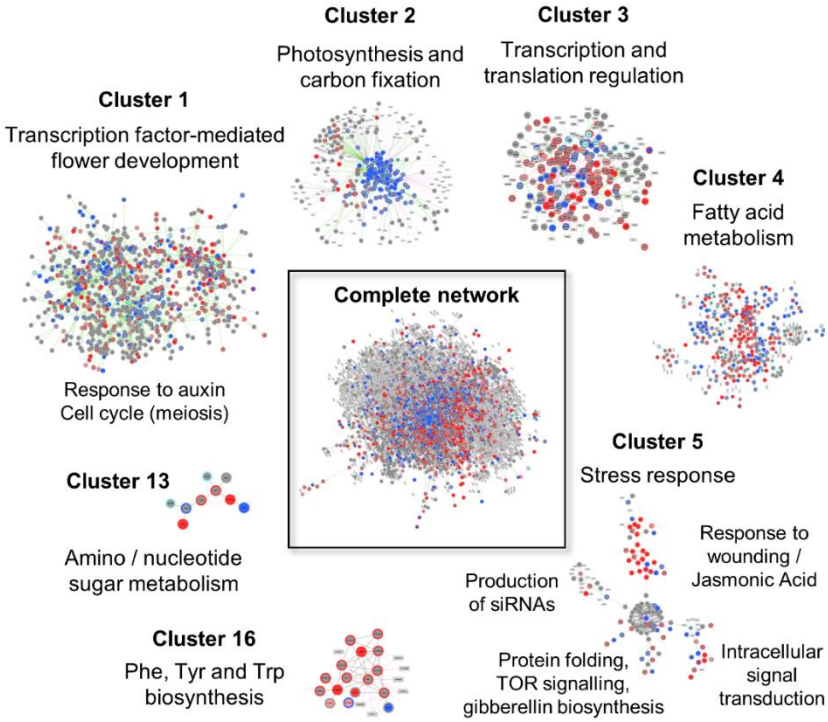


Figure 2.17. Protein-protein interaction clusters.
Network depicting physical interactions and co-expression between proteins included in the dataset and other proteins in *A. thaliana* (IntAct, STRING). This figure includes the five largest interaction clusters (Clusters 1 to 5), as well as two interaction clusters that only contain proteins from a specific metabolic pathway (Clusters 13 and 16). The main KEGG pathways for each cluster are annotated (**Sup Table 2.10**). Clusters 1, 3 and 5 contained proteins involved in developmental processes and stress responses, whereas clusters 2, 4, 13 and 16 were enriched in proteins related to metabolic pathways. Node legend: outer line represents RNA levels and inner circle, protein levels. Blue: decreasing trajectories; red: increasing trajectories; salmon: trajectories with a maximum peak (increase – decrease); light blue: trajectories with a minimum (decrease – increase); grey: non-variant. Squares represent proteins not included in the MS results.

2.3 Discussion

In this Chapter, the early flower development process was analysed by comparing gene and protein expression profiles in a pAP1:AP1-GR *ap1 cal* inducible line. Despite the inherent complications to combine datasets (RNA-seq and LC-MS/MS) that are different in their generation, acquisition, and analysis, I identified several groups of genes with various cases of protein-RNA expression patterns of positive, negative, or neutral correlation.

A major concern in label-free quantitative proteomics that hinders the subsequent data analysis and its comparison with other omics data is the high rate of missing values. Thanks to the 'Reliability Analysis' workflow designed in this work, it was possible to distinguish the nature of the data missingness, and to treat the not-assigned values (NAs) of the LC-MS/MS results accordingly. The highest number of Reliably Undetected proteins corresponded to D0 and D5 (**Figure 2.3C**), when proteins whose expression is regulated by AP1 have not been expressed yet, or are strongly downregulated, respectively (KAUFMANN ET AL., 2010; PAJORO, MADRIGAL, ET AL., 2014; WELLMER ET AL., 2006). After the Reliability Analysis and NA imputation, replicates clustered better together by day (**Figure 2.6**), demonstrating the robustness and reproducibility of the LC-MS/MS followed by a Reliability Analysis approach.

The RNA-seq data corroborated previous findings stating that AP1 acts predominantly as a transcriptional repressor during the earliest stages of flower development, whereas, at more advanced stages, predominantly as an activator (KAUFMANN ET AL., 2010; WELLMER ET AL., 2006), but more significantly, the RNA-seq data triplicated the number of differentially expressed genes (DEGs) identified during early flower development in Arabidopsis (5,150 DEGs vs 1,653 DEGs described in (WELLMER ET AL., 2006)). In addition, with the likelihood ratio test (LRT) approach, the total number of variable genes was even higher (8,125 stage variant genes – SVGs –), as this approach is more sensitive than pairwise comparisons to slight changers in expression levels between subsequent days. Furthermore, it was possible to identify 230 novel putative AP1-high confidence targets (HCTs) based on their differentially expression data and previous ChIP-seq data from

(KAUFMANN ET AL., 2010), including flowering time genes (e.g., *SVP* and *AGL24*) and genes that participate in flower development (e.g., *SEP2, SEP4, BHLH11, CUC1* and *PIN1*) that were down- and up-regulated during the D1 vs D0 time comparison in the RNA-seq data, respectively.

Multiomics studies provide a wider interpretation of a process than a research based solely on one kind of molecule. In this study, the correlation of mRNA and protein levels through time of each RNA-protein pair was measured by calculating the Spearman's rank correlation coefficient (ρ). In total, there were 2,540 RNA-protein pairs with a positive correlation, 975 RNA-protein pairs presented a negative correlation, and 3,488 were considered as not significantly correlated in either way.

The expression patterns of AP1 high-confidence targets (KAUFMANN ET AL., 2010) at the mRNA and protein levels were positively correlated (e.g., for the case of *SUPPRESSOR OF CONSTANS OVEREXPRESSION 1; SOC1*), except some cases of discordancy, as some examples of RNA-protein comparisons with opposite expression patterns between both molecules were also found. This was the case of *WUSCHEL RELATED HOMEOBOX13* (*WOX13*) (COSTANZO ET AL., 2014; H. LIN ET AL., 2013) and *AT4G27980* (Y. WANG ET AL., 2008), two of the marker proteins whose mRNA levels increased, as in (WELLMER ET AL., 2006), whereas their protein levels decreased through time, as in (Y. WANG ET AL., 2008), showing a significantly negative Spearman's rank correlation coefficient ($\rho \leq 0.8$, p-value ≤ 0.05).

In the analysis for this Thesis, almost 50% of total mRNA-protein pairs showed no correlation between their individual abundances, such as the AP1-targets *LUTEIN-DEFICIENT 1* (*LUT1*) (TIAN ET AL., 2004) and *PYRABACTIN RESISTANCE 1 - LIKE 1* (*PYL1*) (YIN ET AL., 2016), both NVG-NVP (**Figure 2.15B**). The observed apparent lack of correlation between mRNA and protein levels could be related to the methods of detection and quantification that were used, but also to biologically relevant processes, such as post-translational and post-transcriptional regulatory events, etc. PCA for averaged *Z*-scored values of SVG-SVP pairs (*Z*-scored independently) revealed that D0, D1 and D5 had the lower mRNA-protein correlations (**Figure 2.13B**). This observation was somehow expected given the

difference in average half-life of mRNA and proteins and the variations in transcriptional and translational kinetics, specially at the beginning of the induction. In addition, protein levels at D5 also correlated with mRNA levels at D4, while following the same trend, protein levels at D4 were slightly closer to mRNA levels at D3 than at D4 (**Figure 2.13B**). This highlights the usefulness of time-series analysis to compare gene and protein expression and relates with the low correlations found in many similar studies following single sampling timepoints (as in (HUANG ET AL., 2022; MERGNER ET AL., 2020; P. WANG ET AL., 2023)).

In some cases, there was a correlation between the behaviour in expression of gene-protein pairs and their functions and protein-protein interactions. Gene-protein pairs with decreasing levels of RNA and increasing levels of protein abundance were mostly chloroplast-related genes (SVG-SVP ME19). Gene-protein pairs with increasing RNA levels and decreasing protein abundances are related with jasmonate synthesis and metabolism. Jasmonic acid, and its derivative metabolites, are important for plant growth and development processes, including senescence, growth inhibition, flower development and leaf abscission (ZOU ET AL., 2020), as well as, plant response to abiotic and biotic stresses (GRIFFITHS, 2020). A network of plant hormones such as jasmonic acid with miRNA-transcription factors have a role in flower senescence, and probably in floral organ abscission (RUBIO-SOMOZA & WEIGEL, 2013).

These differences in RNA levels and protein abundances reflect the existence of possible regulatory processes, such as positive and/or negative feedback loops or posttranscriptional and posttranslational modifications, affecting both molecules differently.

2.4 Materials and methods

2.4.1 Plant lines and plant growth conditions

The pAP1:AP1-GR *ap1 cal* (Ó'MAOILÉIDIGH ET AL., 2023) plants were grown on a soil:vermiculite:perlite mixture at 21 °C under long day conditions (16 h light, 8 h darkness), after a 4-day period of stratification at 4 °C in darkness.

2.4.2 Tissue collection

For RNA-seq and LC-MS/MS experiments, 4-week-old pAP1:AP1-GR *ap1 cal* plants were used. Four biological replicates were generated for each time point. For each replicate, from around 80 (D0) to 40 plants (D5) were needed to obtain 300-500 µg of total protein. Inflorescence tissue was collected using jeweler's forceps as previously described (WELLMER ET AL., 2006). For induction, inflorescences were treated with a DEX-induction solution (2 µM DEX, 0.01% (v/v) ethanol, and 0.01% Silwet L-77). Using plastic pipettes, the solution was directly applied onto the inflorescences so that the cauliflower-like structures were completely drenched. First induction was performed 8 h after lights on, and daily inductions, at 4 h after lights on. Samples were collected immediately after solution application (D0), as well as at 1, 2, 3, 4 and 5 days (D1-5) after the first treatment.

2.4.3 Protein extraction

Protein and RNA extractions had common initial steps (as described in (ÁLVAREZ-URDIOLA, MATUS, ET AL., 2023)). Tissue was ground in liquid nitrogen. For each timepoint, ~0.25 g of plant material was used. Ground material was resuspended in 1 mL of Trizol and incubated on ice for 5 min. Then, 200 µL of chloroform were added and properly mixed by vortexing. After a 5-min incubation on ice, samples were centrifuged at 4 °C for 15 min at maximum speed. Upon centrifugation, three phases are formed, the aqueous phase contains RNA (~550 µL, transparent), the interphase, DNA (white), and the organic phase, proteins, and lipids (~450 µL, pink). After the aqueous phase was transferred to a new microcentrifuge tube (*see* **RNA extraction**), 300 µL of ethanol 100% (v/v) were added to the organic phase to continue with protein extraction and the mix was incubated on ice.

Samples were centrifuged for 10 min at 2,000 x g to separate DNA from proteins. The supernatant was placed in a clean 2 mL microcentrifuge tube, 1 mL of pure isopropanol was added, and samples were incubated at room temperature for 10 min. After a 10-min centrifugation at 4 °C at 12,000 x g, the supernatant was discarded. The pellet was resuspended in 2 mL of a solution of 0.3 M guanidine in ethanol 95% (v/v) for washing and sonicated during 5 min. Samples were centrifuged at 4 °C for 5 min at 8,000 x g. This washing procedure was repeated twice. The final pellet was stored at -20 °C and washed with ethanol 90% (v/v) before the final resuspension in an acetonitrile 70% (v/v) buffer for LC-MS/MS.

2.4.4 RNA extraction

The organic phase (*see* **Protein extraction**) was transferred to a clean microcentrifuge tube and mixed vigorously with one volume of pure isopropanol. After 15 min of incubation on ice, samples were centrifuged at 4 °C for 10 min and the supernatant was discarded. Each pellet was resuspended in 750 µL of LiCl 3 M, incubated on ice for 10 min and centrifuged at 4 °C for 10 min at maximum speed. The supernatant was discarded, and each pellet was gently washed with 500 µL of ethanol 85% (v/v). The last centrifugation was performed at 4 °C for 10 min at maximum speed and supernatant was discarded. Each pellet was resuspended in 21 µL of diethylpyrocarbonate (DEPC)-treated water after drying. Samples were quantified with a NanoDrop 1000 Spectrophotometer.

2.4.5 Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS)

Proteomics experiments were conducted in collaboration with Dr. Eduard Sabidó and Dra Eva Borrás from the proteomics facility at the Center for Genomic Regulation (CRG).

Sample preparation. Samples were reduced with dithiothreitol (30 nmol, 37 °C, 60 min) and alkylated in the dark with iodoacetamide (60 nmol, 25°C, 30 min). The resulting protein extract was first diluted to 2M urea with 200 mM ammonium bicarbonate for digestion with endoproteinase LysC (1:10 w:w, 37 °C, o/n, Wako, cat # 129-02541), and then diluted 2-fold with 200

mM ammonium bicarbonate for trypsin digestion (1:10 w:w, 37 °C, 8h, Promega cat # V5113). After digestion, peptide mix was acidified with formic acid and desalted with a MicroSpin C18 column (The Nest Group, Inc) prior to LC-MS/MS analysis.

Chromatographic and mass spectrometric analysis. Samples were analysed using an LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EASY-nLC 1000 (Thermo Fisher Scientific (Proxeon), Odense, Denmark). Peptides were loaded directly onto the analytical column and were separated by reversed-phase chromatography using a 50 cm column with an inner diameter of 75 μm, packed with 2 μm C18 particles spectrometer (Thermo Scientific, San Jose, CA, USA). Chromatographic gradients started at 95% buffer A and 5% buffer B with a flow rate of 300 nL/min for 5 minutes and gradually increased to 22% buffer B and 78% A in 79 min and then to 35% buffer B and 65% A in 11 min. After each analysis, the column was washed for 10 min with 10% buffer A and 90% buffer B. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in acetonitrile.

The mass spectrometer was operated in positive ionization mode with nanospray voltage set at 2.4 kV and source temperature at 275 °C. Ultramark 1621 for the was used for external calibration of the FT mass analyzer prior the analyses, and an internal calibration was performed using the background polysiloxane ion signal at m/z 445.1200. The acquisition was performed in data-dependent acquisition (DDA) mode and full MS scans with 1 micro scans at resolution of 120,000 were used over a mass range of m/z 350-1500 with detection in the Orbitrap mass analyzer. Auto gain control (AGC) was set to 1E5 and charge state filtering disqualifying singly charged peptides was activated. In each cycle of data-dependent acquisition analysis, following each survey scan, the most intense ions above a threshold ion count of 10,000 were selected for fragmentation. The number of selected precursor ions for fragmentation was determined by the 'Top Speed' acquisition algorithm and a dynamic exclusion of 60 seconds. Fragment ion spectra were produced via high-energy collision dissociation (HCD) at normalized collision energy of 28% and they were acquired in the ion trap mass analyzer. AGC was set to 1E4, and an isolation window of 1.6 m/z and a maximum

injection time of 200 ms were used. All data were acquired with Xcalibur software v4.1.31.9. Digested bovine serum albumin (New England Biolabs cat # P8108S) was analysed between each sample to avoid sample carryover and to assure stability of the instrument and QCloud (CHIVA ET AL., 2018) has been used to control instrument longitudinal performance during the project.

Data Processing. Acquired spectra were analysed using the Proteome Discoverer software suite (v2.0, Thermo Fisher Scientific) and the Mascot search engine (v2.5 Matrix Science) (PERKINS ET AL., 1999). The data were searched against a UniProt *A. thaliana* database plus a list of common contaminants (BEER ET AL., 2017) and all the corresponding decoy entries. For peptide identification a precursor ion mass tolerance of 7 ppm was used for MS1 level, trypsin was chosen as enzyme, and up to three missed cleavages were allowed. The fragment ion mass tolerance was set to 0.5 Da for MS2 spectra. Oxidation of methionine and N-terminal protein acetylation were used as variable modifications whereas carbamidomethylation on cysteines was set as a fixed modification.

False discovery rate (FDR) in peptide identification was set to a maximum of 5%. Peptide quantification data were retrieved from the 'Precursor ion area detector' node from Proteome Discoverer (v2.0) using 2 ppm mass tolerance for the peptide extracted ion current (XIC). Protein abundance in each condition was estimated using the average of the three most intense peptides per protein group (TOP3) (SILVA ET AL., 2006). The raw proteomics data have been deposited to the PRIDE repository (PEREZ-RIVEROL ET AL., 2022) with the dataset identifier PXD038980. For subsequent statistical analysis, median normalisation was performed by subtracting from each logged value the sample median and adding the global dataset median. Replicate 1 of Day 0 (R1D0) highly differed from the rest (**Figure 2.18**), so it was removed from the dataset, as well as the 165 proteins that were only detected in this sample.

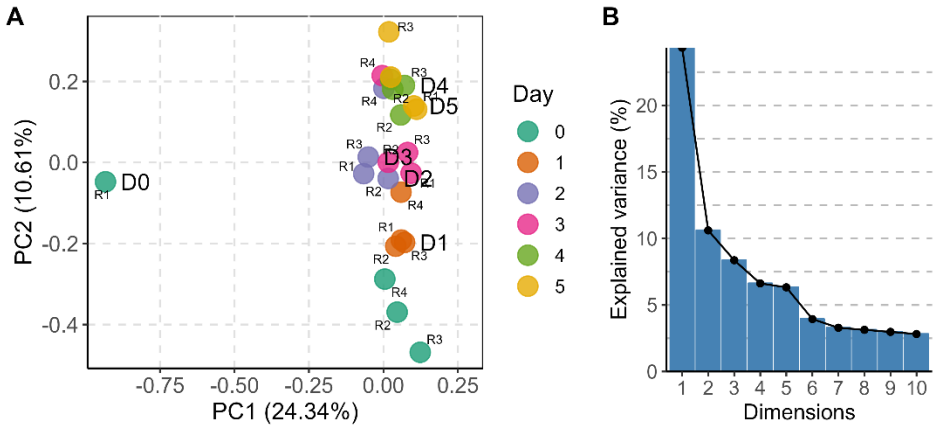


Figure 2.18. Inter-sample variability of the proteomics data before D0R1 removal.
A) PCA of proteins without NAs before the RA and before R1D0 removal. **B)** Percentage of variances explained by each principal component for the proteomics data before R1D0 removal (eigenvalues).

Not Assigned values: Reliability analysis. For the *Reliability Analysis*, each timepoint for a protein was classified as reliably or unreliably detected or undetected depending on its number of NAs and the number of NAs of its immediately adjacent days (neighbours). Days 0 and 5 were considered as Reliably Undetected when all replicates were NAs, and days 1 – 4, besides that, must had at least one neighbour with two or more NAs. Those were considered as MNAR missing values, and NAs were replaced by the minimum of detection of the dataset (Deterministic Minimum Imputation method (MELETH ET AL., 2005)). Days with one or no NAs were defined as Reliably Detected and their abundance values were kept. Finally, days with two or more NAs were classified as Unreliably Detected when they had at least one neighbour with two or less NAs, keeping their quantification values; otherwise, they were classified as Unreliably Undetected, and its quantification values were replaced by NAs in all replicates. All those proteins which were Reliably or Unreliably Undetected in every timepoint were discarded. The remaining NA values were estimated by k-Nearest Neighbour (kNN) imputation (k = 10) (TROYANSKAYA ET AL., 2001) (**Figures 2.3A, B, 2.4**).

2.4.6 RNA-seq experiments

The 24 samples were sequenced on an Illumina HighSeq 2000 machine. Cleaned reads together with the transcriptome of *A. thaliana* (TAIR10) were used to quantify gene expression at transcript level, in counts (regularized-logarithm transformation with DESeq2) and Transcripts Per Million (TPMs) using the software Salmon (v0.12.0). The quantification data were grouped so that genes instead of transcripts were analysed, using *tximport* package in R. Genes that had less than ten counts across all the samples were removed to facilitate further analyses (*DESeq2* package in R). Size factors, corrected by library size, and dispersions were estimated using DEseq function from the package *DESeq2* in R. Dispersion estimates for all genes were obtained considering the information for each gene separately.

2.4.7 Representative proteins and genes: Markers, Supermarkers and AP1-targets

A group of 69 proteins were selected as markers on the basis of the detection of expression of their corresponding genes in previous time-course experiments performed using AP1 floral induction systems, including gene expression profiling using DNA microarrays (KAUFMANN ET AL., 2010; WELLMER ET AL., 2006), and unpublished data (our laboratory; Bustamante et al.). Marker proteins corresponded with up- or down-regulated genes in the microarray experiments (absolute FC ≥ 2 for the first replicate when comparing days 1 and 0, and BH ≤ 0.05) or in the RNA-seq (absolute FC ≥ 2 for all replicates when comparing day 2 and 0, and day 4 and 0, and FPKM > 1). The seven supermarker proteins, with similar characteristics as the markers, are transcription factors controlling different aspects of flower development (AP2, AP3, PI, TFL1, CRC, LFY, FIL-YABI1). Out of the 249 AP1 high confidence targets (HCTs) defined in (KAUFMANN ET AL., 2010), 247 were quantified in the RNA-seq experiment reported in this thesis.

2.4.8 Data analysis

Genome and proteome annotations. Araport11 gene identifiers (AGI codes: AT (*A. thaliana*); 1, 2, 3, 4, 5, M, C (chromosome number, M for

mitochondrial, C for chloroplast); G (gene), 00000 (five-digit code for position on chromosome)) were mapped to the UniProt *A. thaliana* reference proteome (taxon identifier 3702; UP000006548; downloaded in 2018) based on protein sequence. N- and C-terminal peptide sequences were extracted from the Mascot.txt file and filtered for zero missed cleavages (**Figure 2.2A**). N-terminal peptides were divided into groups with (n = 1,203) or without (n = 769) cleavage of the initiator methionine. Then, the frequency of the 20 genetically encoded amino acids at the position after the start codon was calculated and displayed as a pie chart (**Figure 2.2B**) for both groups. The percentage of acetylated N-terminal peptides with the same amino acids in the second position was calculated for both groups and represented as bar plot (**Figure 2.2C**).

Protein and RNA level variation through time. An ANOVA analysis was performed for the normalised proteome dataset, followed by a Tukey post-hoc test. Proteins were considered as stage variant (SVPs) if their Benjamini & Hochberg (BH) adjusted p-value, which can be interpreted as False Discovery Rate (FDR) (BENJAMINI & HOCHBERG, 1995), was lower than 0.05. For the RNA-seq dataset, a moderated likelihood ratio test (LRT) was applied to get a statistic for ranking genes according to the difference in expression profiles among timepoints. LRT is a test of significance for differences of any level of the factor. Genes with an adjusted p-value (FDR) lower than 0.01 were considered as stage variant (SVGs). The log2 fold-change (LFC) in expression between subsequent stages were calculated for all transcripts. The p-values were adjusted for multi-hypothesis testing using the BH procedure (FDR). Transcripts with a LFC with an adjusted p-value lower than 0.05 at any *day-to previous day* comparison were considered as differentially expressed genes (DEGs). No LFC cut-off was applied.

Gene-protein correlations. 7,003 quantified proteins in the MS proteome dataset had their correspondent gene in the RNA-seq transcriptome dataset. Protein-gene pairs were grouped in 4 subsets: stage variant at RNA and protein levels (SVG-SVP, n = 973), non-variant genes – stage variant proteins (NVG-SVP, n = 1006), stage variant genes – non-variant proteins (SVG-NVP, n = 1808) and non-variant pairs (NVG-NVP, n = 3216). Pearson's correlation coefficient (r_s; p-value ≤ 0.05) was used to find correlations between protein

levels and corresponding genes within and between all timepoints, using a square matrix (**Figure 2.13**). The Spearman's rank correlation coefficient (ρ) of each gene-protein pair individually was used for correlating transcriptome and proteome levels in each subset (SVG-SVP, SVG-NVP, NVP-SVP, NVG-NVP) (**Figures 2.14, 2.15A**). The slopes were estimated by ranged major-axis (RMA) regression, which allows errors in both variables and is symmetric, using the R package *lmodel2* (CSÁRDI ET AL., 2015) (**Figure 2.19**).

PCA of proteome and transcriptome data was performed in R for each normalised dataset separately (proteome before and after the Reliability Analysis, and transcriptome), but also for their intersection (n = 7003 transcript-protein pairs), and for the SVG-SVP subset (n = 973).

Transcript-protein co-expression network analysis. Transcriptome and proteome dynamics were evaluated by means of weighted gene co-expression network analysis (WGCNA) (LANGFELDER & HORVATH, 2008). Normalized RNA-seq counts and protein abundances data (after Reliability Analysis) were separately z-score transformed for each subgroup, and WGCNA was performed with a soft-power of 6 signed network. Modules were defined by dynamic tree cut with a minimum size of 10 and deep split of 4. To reduce the final number of modules, those with a similitude superior to 0.9 were merged, leading to the final number of modules that were considered.

Protein-protein interaction network. Arabidopsis protein-protein interactions were downloaded from STRING (March 2021, https://stringdb-static.org/download/protein.links.detailed.v11.0/), and IntAct (March 2021, https://www.ebi.ac.uk/intact/). In addition, it was checked which ppi between the proteins quantified by MS were annotated in The Arabidopsis Information Resource (TAIR, https://arabidopsis.org), finding 598 interactions (**Sup Table 2.13**). Interaction clusters were determined using the *GLay* clustering tool (SU ET AL., 2010) of ClusterMaker package (MORRIS ET AL., 2011) for Cytoscape (SHANNON ET AL., 2003).

Proteins from transcript-protein pairs with similar expression patterns (from the same interaction module), on average, showed higher STRING co-expression scores (**Figure 2.20A**). To determine if among the protein dataset

there were high confidence previously-described physical interactions, the STRING data from interactions with high co-expression (above median) were combined with IntAct-registered (HERMJAKOB ET AL., 2003) ppi. There were 129 self-interacting proteins, 70 ppi between proteins in the same module, and 2,656 ppi between proteins from different trajectory modules (**Figure 2.20B**).

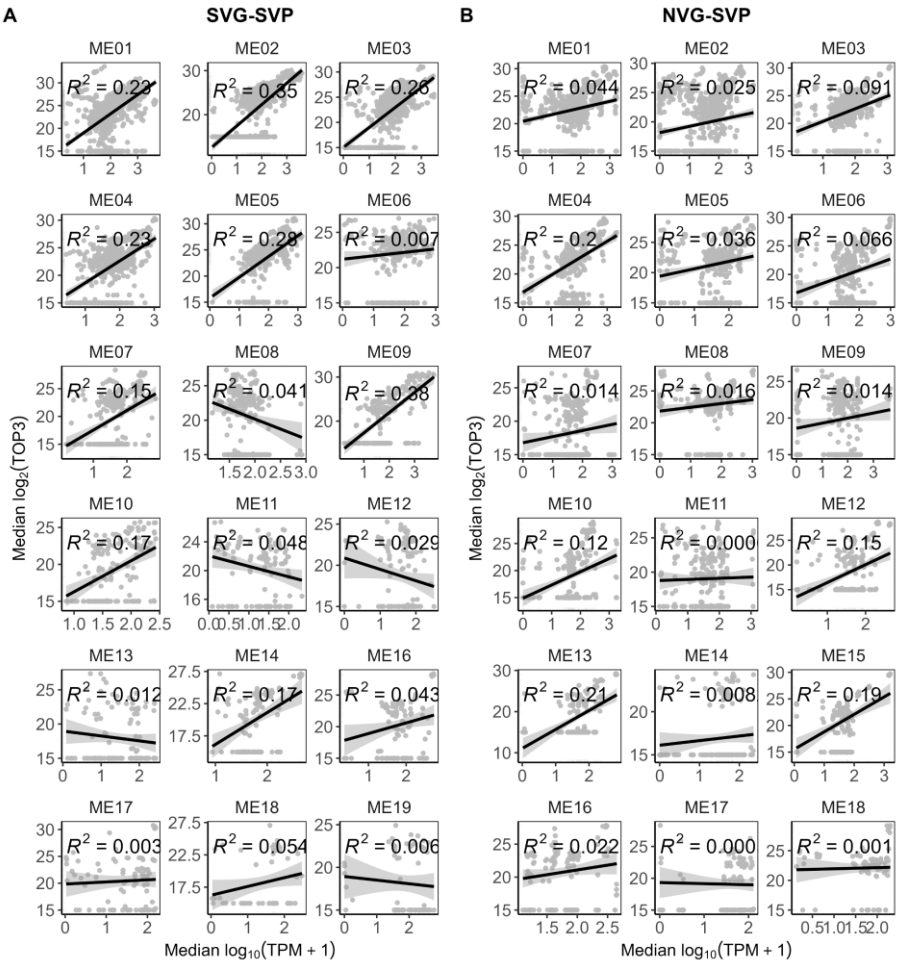


Figure 2.19. Correlation between RNA and protein levels.
Scatter plot of the median logarithmic representation of TOP3 abundances for proteins and TPM for RNA molecules for all RNA-protein pairs at every time-point for all the modules for the different groups: SVG-SVP (**A**), NVP-SVG (**B**), SVG-NVP (**C**), and NVG-NVP (**D**). Colours represent some of the marker and supermarker proteins. *Cont. in next page.*

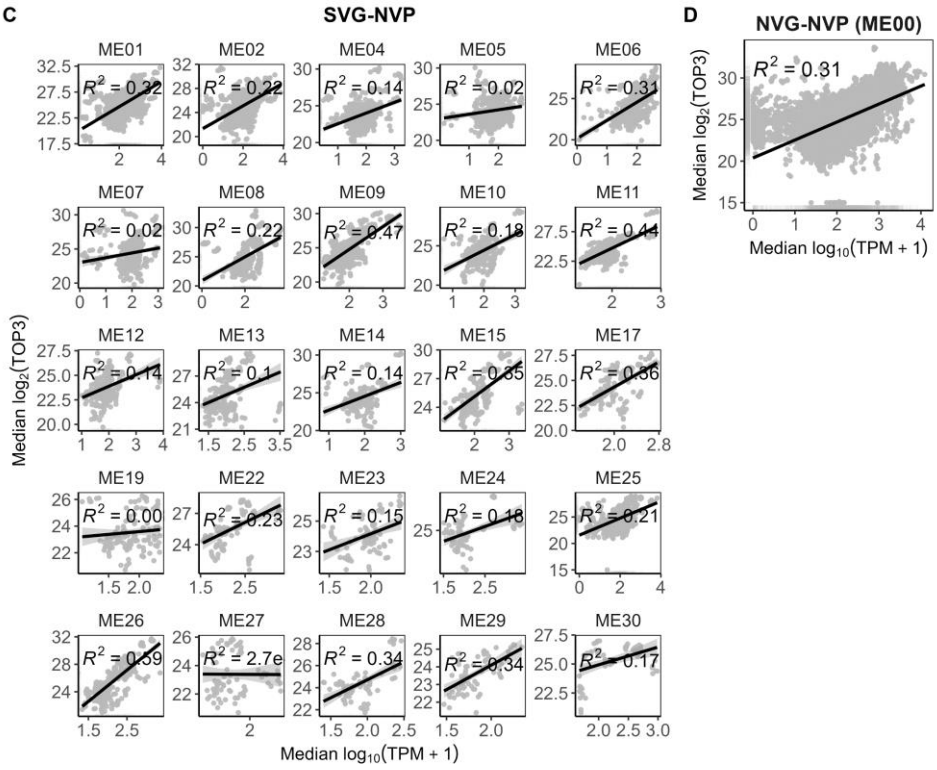


Figure 2.19. Correlation between RNA and protein levels (Cont.).

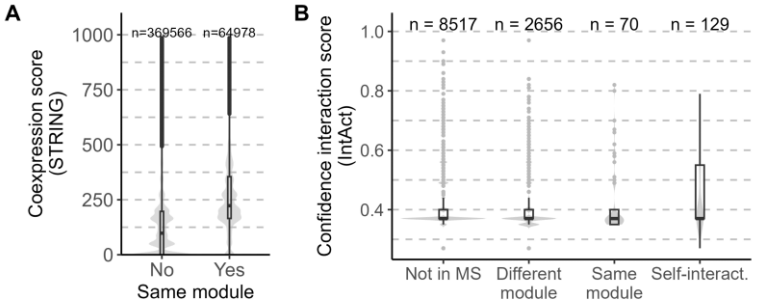


Figure 2.20. STRING and IntAct interaction scores.

A) STRING co-expression scores for each ppi expressed in the same module (yes) or not (no). Interacting proteins from the same module have a significantly higher STRING co-expression score (t-test, p-value < 0.001). **B)** IntAct confidence of interaction score for each one of the ppi between a protein in the dataset and other which is not included (Not in MS), proteins from different or the same module and reported polymers (self-interaction).

Function data analysis. Gene ontology (GO) (G. YU ET AL., 2012) and KEGG term enrichments (KANEHISA & GOTO, 2000) were performed using clusterProfiler (G. YU ET AL., 2012). Enrichment was determined with fisher exact tests followed by Bonferroni-Yekutieli multiple testing correction. The composition of the interaction clusters and the trajectory modules was also analysed in the sense of whether they contained proteins from the same annotated family according to TAIR, but there was no significant enrichment in members of any specific family for the clusters. Family annotations were downloaded from TAIR (downloaded on the 18th of March 2021: gene_families_sep_29_09_update.txt,).

2.4.9 Comparison with previous studies

RNA-seq differential expression results were compared to those found in (WELLMER ET AL., 2006). A filter on adjusted p-values (< 0.05) was applied to the list of common genes to keep only those with significant values.

Novel AP1-high confidence targets (HCTs) were defined using the RNA-seq data (D1 vs. D0 DEGs, adjusted p-value < 0.05, absolute LFC < 0.29) and ChIP-seq data from (KAUFMANN ET AL., 2010).

2.4.10 Data availability statement

The LC-MS/MS proteomics data for this project have been deposited at the ProteomeXchange Consortium with the dataset identifier PXD038980.

The RNAseq transcriptomics data are available at GEO with the dataset identifier GSE217606.

Chapter 3

General introduction (II)

Part of this chapter will be published as:

The 'non-conventional' plant peptidome: a new layer on flower development regulatory mechanisms.

Álvarez-Urdiola, R., Riechmann, J.L. Manuscript in preparation.

Chapter 3. General introduction (II)

3.1 The plant peptidome

Peptides play multiple and diverse roles in plants, acting as signalling molecules in cell-to-cell interactions or long-distance communication, affecting stress or external stimuli responses, controlling development, morphogenesis, growth, fertilization, symbiosis with nitrogen-fixing bacteria, or by virtue of their antimicrobial activities (BREIDEN & SIMON, 2016; GRIENENBERGER & FLETCHER, 2015; MATSUBAYASHI, 2011, 2014; TAKAHASHI ET AL., 2019; TAVORMINA ET AL., 2015). They can be classified according to how they are generated and to their sequence, structural, and functional characteristics, and are generally defined – albeit somewhat arbitrarily – as of less than 100 amino acids long (TAVORMINA ET AL., 2015).

The vast majority of the plant peptides that have been characterized to date are produced through the processing of larger, non-functional precursor polypeptides, which result in the mature peptide upon removal of an N-terminal signal sequence (NSS; that directs the precursor to the secretory pathway) and/or of other amino acid segment(s) (**Table 3.1**) (TAVORMINA ET AL., 2015). These precursor-derived peptides ('conventional' peptides) functionally correspond, to a large extent, to small signalling peptides (SSPs) and antimicrobial peptides (AMPs), and structurally can be sub-grouped into two major classes, post-translationally modified (PTM) peptides and cysteine-rich (Cys-rich) peptides, each of them containing several gene families; in addition, several non-functional precursor derived peptides are not cysteine-rich and are not known to be post-translationally modified (**Table 3.1**). The presence of signature sequences or motifs (NSSs, Cys residues) and of sequence similarity within gene families has facilitated the identification of these 'conventional' peptides within and across plant species (MATSUBAYASHI, 2018)

Functional or bioactive peptides can also be generated through the proteolytic processing of otherwise functional proteins, resulting in so-called cryptides, defined by having a biological activity that is distinct to that of the protein that the cryptide originates from (SAMIR & LINK, 2011). Only a few examples of cryptides with relevant roles in plants have been reported to date, which are for instance related to the defence response and other stresses (CHEN ET AL., 2014; CHIEN ET AL., 2015; LYAPINA ET AL., 2019; TAVORMINA ET AL., 2015; YUAN ET AL., 2019). In fact, the proteolytic degradation of proteins generates peptides that can be localized intracellularly or extracellularly and, in plants, the composition of this protein "degradome" – of which chloroplasts are a major source (KMIEC ET AL., 2018; MAMAEVA ET AL., 2020) – is affected during stress responses or upon treatment with plant stress-related hormones (FESENKO, AZARKINA, ET AL., 2019; FILIPPOVA ET AL., 2019). Whether multiple cryptides with specific functions exist in the plant peptide "degradome", or if it is rather the existence of a pool of peptide degradation products and changes in its abundance or composition what may be perceived by the cells as part of stress signalling, is an open question, although a potential cryptide with antimicrobial activity has been detected in *P. patens* upon methyl jasmonate treatment (FESENKO, AZARKINA, ET AL., 2019).

In addition to the peptides that are generated through the processing of non-functional or functional precursors, peptides can also be produced through the direct translation of short/small open reading frames (sORFs/smORFs) (**Table 3.1**). This is the case, for instance, of Arabidopsis *ROTUNDIFOLIA4* and *DEVIL1*, which were identified in activation-tagging (gain-of-function) genetic screens and are the founding members of the RTFL/DVL gene family, involved in organogenesis (GUO ET AL., 2015; IKEUCHI ET AL., 2011; NARITA ET AL., 2004; VALDIVIA ET AL., 2012; WEN ET AL., 2004). Likewise, *KISS-OF-DEATH* (*KOD*) was identified through a promoter trap screening and the encoded small peptide was shown to act as an inducer of programmed cell death in embryo development and during stress (BLANVILLAIN ET AL., 2011), and *BRICK1* (*BRK1*; identified in a mutant screen in maize) is an essential component of the complex that controls the spatiotemporal dynamics of actin nucleation and therefore affecting morphogenesis (CHIN ET AL., 2021; DJAKOVIC ET AL., 2006; FRANK & SMITH, 2002; LE ET AL., 2006). More recently,

ZENGDA SMALL PEPTIDE 1 (ZSP1) was identified in a search of Arabidopsis small genes that lacked functional annotation and was shown to affect organ size via the cytokinin pathway (ZENG ET AL., 2022).

However, the fact is that until relatively recently, the coding potential of eukaryotic sORFs at the genome-wide level had mostly been overlooked. This was due to traditional assumptions (e.g., a monocistronic nature of eukaryotic mRNAs, or that short peptides would be unlikely to fold into stable -and functional- structures), to computational constraints for *de novo* sORF identification and annotation in genome sequences, and -particularly- to experimental limitations for determining whether these sequences are in fact translated. However, the development of high-throughput methods to identify translating RNAs (ribosome profiling; Ribo-seq and Polyribo-seq) (HSU ET AL., 2016; INGOLIA, 2016; INGOLIA ET AL., 2014; INGOLIA ET AL., 2009; INGOLIA ET AL., 2011) evidenced an unanticipated complexity to mammalian proteomes and revealed that translation outside of conserved or standard/annotated reading frames is pervasive on cytosolic transcripts (INGOLIA ET AL., 2014; INGOLIA ET AL., 2011). These observations were quickly extended to other eukaryotic organisms, including plants (BAZIN ET AL., 2017; HSU ET AL., 2016; JUNTAWONG ET AL., 2014), and demonstrated also through mass spectrometry (MS) proteomic studies (MENSCHAERT ET AL., 2013; SLAVOFF ET AL., 2013; VANDERPERRE ET AL., 2013).

As a result and contrary to what was previously considered, it is now well established that small and long non-coding RNAs (ncRNAs and lncRNAs) and transcripts of unknown function (TUFs), pseudogene transcripts, 5'- and 3´-UTRs of mRNAs, antisense transcripts, unannotated intergenic regions, primary miRNA transcripts (pri-miRs), ribosomal RNAs, and introns and circular RNAs, might contain translatable sORFs encoding non-precursor-derived peptides, which are generally referred to as sORF-encoded peptides (SEPs), 'non-conventional peptides' (NCPs), microproteins, or micropeptides (and also usually defined as shorter than 100 amino acids in length) (**Figure 3.1**). These terms therefore generally refer to a class of peptides and proteins that are "born small" (SCHLESINGER & ELSASSER, 2022), in contraposition to 'conventional' peptides derived from the processing of larger precursor polypeptides (**Table 3.1**).

In addition, mRNAs can also be polycistronic by containing ORFs that, although being internal (completely or partially) and out-of-frame to the main/annotated coding sequence (CDS), can be translated, mostly into peptides or proteins that are also small. These have often been referred to as alt-ORFs (for 'alternative') and alt-proteins (alt-Prots) (BRUNET ET AL., 2018; CARDON ET AL., 2021; LEBLANC ET AL., 2022; SAMANDI ET AL., 2017) (in the literature, however, there is overlap but not complete coincidence between the categories defined as 'sORF' and 'alt-ORF' peptides; *see* (BRUNET ET AL., 2020; COUSO & PATRAQUIM, 2017; MUDGE ET AL., 2022) for more on terminology and classifications).

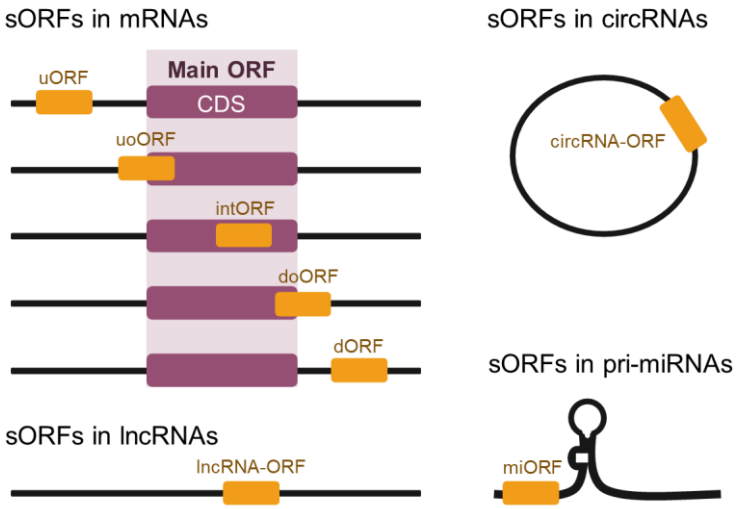


Figure 3.1. Overview of main sORF classes with respect to the type of RNA in which they reside.
Messenger RNAs might contain, in addition to the main, canonical ORF (CDS; coding sequence), sORFs that are located: in the 5´-UTR (upstream ORF; uORF); upstream but overlapping the CDS in a different reading frame (upstream overlapping ORF; uoORF); internal to the CDS in a different reading frame (internal ORF; intORF), internal and in a different reading frame but extending downstream of the CDS (downstream overlapping; doORF); or fully downstream in the 3´-UTR (downstream ORF; dORF). Translatable sORFs can also be located in lncRNAs (lncRNA-ORF), in circular RNAS (circRNA-ORF) or in primary miRNA transcripts (miORF), as well as in other types of RNAs or genetic elements (not pictured). Upstream sORFs (uORFs) and lncRNA sORFs constitute the most abundant classes identified in Ribo-Seq experiments.

The terms 'cryptic proteins' and 'ghost proteins' have also been used to refer to non-annotated or non-canonical proteins, or proteins encoded in lncRNAs, and therefore encompassing -but not being equal to- SEPs/NCPs: a vast majority of cryptic proteins are small, but not all of them (RUIZ CUEVAS ET AL., 2021; ZHENG ET AL., 2023). In a sense, it could be argued that SEPs/NCPs/microproteins represent the low end of the spectrum of 'canonical' proteins (SCHLESINGER & ELSASSER, 2022), even though they might frequently display some 'non-canonical' characteristics (*see* below), and although their origin (i.e., the type of RNA molecules they are derived from) is much more varied and continues to expand. It has recently been found, for instance, that plant and animal positive-sense single-stranded RNA viruses encode functional SEPs in their negative-sense, replication-intermediate RNA, previously thought to be devoid of coding capacity (GONG ET AL., 2023). In any case, even the known the 'low end' of the proteome spectrum is much less understood than the 'standard' proteins: more than 70% of the genes encoding proteins smaller than 50 amino acids that have been already annotated in the Arabidopsis genome still lack functional information (**Figure 3.2**).

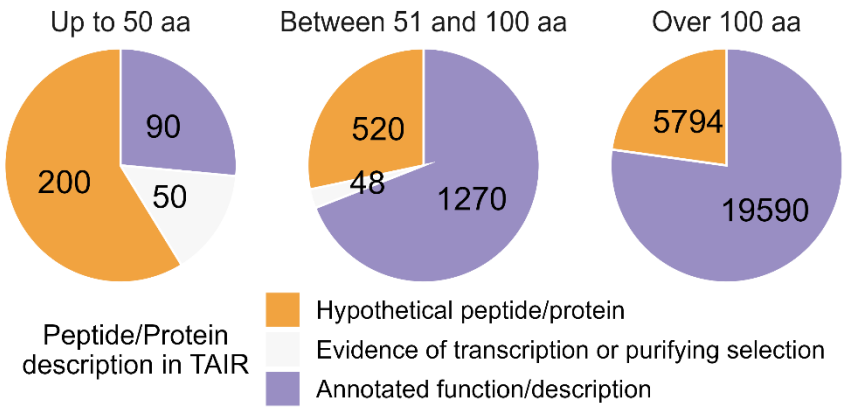


Figure 3.2. Proportion of Arabidopsis peptides and proteins with functional annotation in TAIR.
Peptides/proteins classified as 'hypothetical' or 'evidence of transcription of purifying selection' lack functional annotation. There are 340 peptides of up to 50 aa, 1,838 peptides from 51 to 100 aa, and 25,384 proteins of over 100 aa annotated in TAIR (file: Araport11_pep_20220914_representative_gene_model.gz; data downloaded from https://www.arabidopsis.org/).

Table 3.1. The plant peptidome: summary classification of functional peptides.
Arabidopsis thaliana (At), Brassica oleracea (Bo), Coffea canephora (Cc), Glycine max (Gm), Ipomoea batatas (Ib), Medicago truncatula (Mt), Nicotiana tabacum (Nt), Petunia hybrida (Ph), Oryza sativa (Os), Phaseoulus vulgaris (Pv), Populus tremula and P. tremuloide (Pt), Solanum lycopersicum (Sl), S. nigrum (Sn), S. tubersosum (St), Triticum aestivum (Ta), Vigna unguiculata (Vu), Vitis vinifera (Vv), Zea mays (Zm).

PEPTIDE TYPE **Family or Class/Type**	Representative Peptides (*Species*)	Size (aa)	Number of members (*Species*)	Functions	References
I - Precursor-derived peptides ('conventional' peptidome)					
Non-functional precursor					
	PTM peptides				
CEP (C-terminally encoded peptide)	CEP (*At, Mt*), ZmCEP1 (*Zm*)		15 (*At*), 4 (*Mt*)	Plant organogenesis and response to abiotic stress. Signalling (root-to-shoot).	(Ogilvie et al., 2014; Ohyama et al., 2008; Roberts et al., 2013; Tabata et al., 2014; Xu et al., 2021; Zhou et al., 2019)
CIF (Casparian strip integrity factor)	CIF1-2 (*At*)	83 (*At*)	2 (*At*)	Peptide hormone required to form the casparian strip.	(Doblas et al., 2017; Nakayama et al., 2017)
CLE (CLAVATA3/ESR-related)	CLV3 (*At*), CLEs	12-14	32 (*At*), 104 (*Ta*)	Plant growth. Signalling.	(Fletcher, 2020; Goad et al., 2017; Whitewoods, 2021; Willoughby & Nimchuk, 2021)
GLV/RGF/CLEL (GOLVEN/ROOT MERISTEM GROWTH FACTOR/CLE-like)	GLV1-3 (*At*)	15-20 (*At*)	11 (*At*)	Plant growth (root gravitropism). Signalling.	(Buhler et al., 2023; Fernandez et al., 2013; Furumizu & Sawa, 2021; Jourquin et al., 2023; Stegmann et al., 2022; Whitford et al., 2012; Xu et al., 2023)
HYPSYS	HYPSIS I and II (Nt)	15-20 (Nt, Sl, St, Ph, Sn, Ib, Pt, Cc)	2 (*Nt*)	Defence signalling.	(Pearce et al., 2009; Pearce et al., 2001; Pearce et al., 2007; Ryan & Pearce, 2003; Zhang et al., 2020)

IDA/IDL (INFLORESCENCE DEFICIENT IN ABSCISSION/IDA-like)	IDA, IDL (*At*)	77 (*At*)	6 (*At*)	Control of floral organ abscission and lateral root emergence.	(Santiago et al., 2016; Vie et al., 2015; Wang, Wu, Jiang, et al., 2023)
PIP/PIPL/TOLS (PAMP-INDUCED SECRETED PEPTIDE/PIP-like)	PIP, PIPL, TOLS2 (*At*)	72-86 (*At*)	5 (*At*)	Innate immune response and response to abiotic stress (signalling). Lateral root development.	(Hou et al., 2014; Najafi et al., 2020; Toyokura et al., 2019; Vie et al., 2015; Zhou et al., 2022)
PSK (Phytosulfokine)	AtPSK (*At*)	77-87 (*At*)	6 (*At*)	Plant growth. Plant immunity. Signalling.	(Ding et al., 2023; Matsubayashi et al., 2006; Sauter, 2015; Stuhrwohldt et al., 2015)
PSY (PEPTIDE CONTAINING SULFATED TYROSINE)	PSY1 (*At*)	75 (*At*)		Cellular proliferation and expansion. Seedling development.	(Amano et al., 2007; De Giorgi et al., 2021; Ogawa-Ohnishi et al., 2022)
SCOOP (SERINE-RICH ENDOGENOUS PEPTIDE)	(PRO)SCOOP1-14 (*At*), EWR1 (ENHANCER OF VASCULAR WILT RESISTANCE, *At*)	69-140 (*At*)	23 (*At*)	Plant growth and pathogen defence.	(Guillou et al., 2022; Gully et al., 2019; Hou et al., 2021)
	Cys-rich peptides				
CYSTM (CYSTEIN-RICH TRANSMEMBRANE MODULE)	CYSTM3 (*At*)	57 (*At*)	13 (*At*)	Response to stress. Signalling.	(Xu et al., 2018)
EPF/EPFL/STOMAGEN (EPIDERMAL PATTERNING FACTOR-like)	EPF1 (*At*), EPFL2 (*At*), EPFL9 (STOMAGEN, *At*)	45 (*At*)	12 (*At*)	Plant growth and organogenesis (gynoecium and fruit growth with ovule initiation). Signalling.	(Bessho-Uehara et al., 2016; Hara et al., 2007; Kawamoto et al., 2020; Qi et al., 2020; Sugano et al., 2009)
LURE	AtLURE (*At*)	~90 (*At*)	7 (*At*)	Plant reproduction (pollen tube attractants). Signalling.	(Okuda et al., 2009; Zhong et al., 2019)

NCR (Nodule-specific cysteine-rich)	NCRs (*Mt*), NFS1-2 (*Mt*)	43-47 (*Mt*)	3 (*Mt*)	Nitrogen-fixing symbiosis.	(Horvath et al., 2023; Pan & Wang, 2017; Van de Velde et al., 2010)
PCP-B (POLLEN COAT PROTEIN B)	PCP-B (*At, Bo*)	76-126 (*At, Bo*)	4 (*At*)	Pollination. Signalling.	(Liu et al., 2021; Wang et al., 2017)
RALF/RALFL (RAPID ALKALINIZATION FACTOR/RALF-like)		25-105 (*At*)	> 60 (*At*)	Plant development, immunity response, pollen tube perception, and rupture (Polytubey block).	(Lan et al., 2023; Zhong et al., 2022)
WIP (WOUND INDUCED POLYPEPTIDES)	AtWIP1-5 (*At*), WIPs (*Gm*)	83-95 (*At*), ~90 (*Gm*)	5 (*At*), 38 (*Gm*)	Immune response and symbiotic interactions. Signalling.	(Yu et al., 2018)
Non-PTM, Non-Cys-rich-peptides					
CTNIP/SCREW (SMALL PHYTOCYTOKINES REGULATING DEFENSE AND WATER LOSS)	CTNIP1-5 / SCREWs (*At*)	60-70 (*At*)	5 (*At*)	Stress response (stomatal closure). Signalling.	(Liu et al., 2022; Rhodes et al., 2022)
GRI (GRIM REAPER)	GRI (*At*)	60-70 (*At*)		Response to abiotic stress (programmed cell death induced by extracellular reactive oxygen species - ROS-) and plant development (flowers and seeds).	(Wrzaczek et al., 2009)
PEP (PLANT ELICITOR PEPTIDE)	PEP1 (*At*)	23 (*At*)	8 (*At*)	Defence response. Signalling.	(Bartels & Boller, 2015; Hander et al., 2019; Huffaker et al., 2006)
SYS (SYSTEMIN)	SYS (*Sl*)	18 (*Sl*)		Defence response.	(Ryan & Pearce, 1998, 2003; Zhang et al., 2020)

Functional precursor					
Cryptides	SUBPEP (*Gm*), CAPE1 (*Sl*), INCEPTIN (*At, Pv, Os, Vu, Zm*)	11-13 (*Vu*)		Defence response. Signalling.	(CHEN ET AL., 2014; CHIEN ET AL., 2015; PEARCE ET AL., 2010; SCHMELZ ET AL., 2006)
II - Non-precursor-derived peptides ('non-conventional' peptidome)					
sORF (small genes, intergenic)					
BRK1 (BRICK1)	BRK1 (*Zm*), HSPC300 (*At*)	84 (*Zm*)		Morphogenesis (actin nucleation). Component of the actin reorganization complex.	(CHIN ET AL., 2021; DJAKOVIC ET AL., 2006; FRANK & SMITH, 2002; LE ET AL., 2006)
FIS (FLOODING INDUCIBLE GENES)	FIS1-3 (*Gm*)	70-80 (*Gm*)	3 (*Gm*)	Response to abiotic stress.	(NANJO ET AL., 2011)
KOD (KISS OF DEATH)	KOD (*At*)	25 (*At*)		Programmed cell death in embryogenesis, stress.	(BLANVILLAIN ET AL., 2011)
RTFL/DVL (ROTUNDIFOLIA4-LIKE/DEVIL)	ROT4 (*At*), DVL1 (*At*)	40-144 (*At*)	22 (*At*)	Organogenesis, cell proliferation, nodule development.	(GUO ET AL., 2015; IKEUCHI ET AL., 2011; NARITA ET AL., 2004; VALDIVIA ET AL., 2012; WEN ET AL., 2004)
ZSP1 (ZENGDA SMALL PEPTIDE 1)	ZSP1 (*At*)	57 (*At*)		Organ size (cytokinin pathway).	(ZENG ET AL., 2022)
lncRNA ORFs					
ENOD40	ENOD40-I/A, ENOD40-II/B (*Mt*)	13, 27 (*Mt*)	2 (*Mt*)	Symbiotic nodule development.	(KERESZT ET AL., 2018; ROHRIG ET AL., 2002; SOUSA ET AL., 2001)
IMA (IRONMAN)/FEP (Fe-UPTAKE-INDUCING PEPTIDE)	IMA1-8 (*At*)	~50 (*At*)	8 (*At*)	Plant response to abiotic stress (iron transport). Signalling.	(GRILLET ET AL., 2018; HIRAYAMA ET AL., 2018)

OSIP108 (OXIDATIVE STRESS-INDUCED PEPTIDE 108)	OSIP108 (*At*)	10 (*At*)		Oxidative stress tolerance.	(De Coninck et al., 2013)
PLS (POLARIS)	PLS (*At*)	36 (*At*)		Root growth, vascular development (hormonal crosstalk).	(Casson et al., 2002; Chilley et al., 2006; Liu et al., 2010; Moore et al., 2015)
Zm401p10	Zm401p10 (*Zm*)	89 (*Zm*)		Anther development.	(Ma et al., 2008; Wang et al., 2009)
Zm908p11	Zm908p11 (*Zm*)	97 (*Zm*)		Pollen germination and tube growth.	(Dong et al., 2013)
pri-miRNA sORFs/mirPEPs					
miPEPs: miPEP156a,c; 160b; 162; 163; 164a; 165a; 167a,b,c; 169; 171b,d; 172b,c; 319a; 395c; 396a; 858a (*At* and other species)		5-50 (*At, Bo, Mt, Gm, Vv*)	>18 (*At*)	Plant growth and morphology (flowering, root, leaf and flower development).	(Gautam et al., 2023; Lauressergues et al., 2015; Lauressergues et al., 2022; Ormancey et al., 2023; Sharma et al., 2020)

Regardless of the terms that are used, however, what has become increasingly clear over the past ten years is that SEPs/NCPs -the 'non-conventional' peptidome- constitute an important part of the eukaryotic proteome. This 'non-conventional' peptidome is still poorly defined and annotated and largely uncharacterized, but it is already apparent that SEPs/NCPs can carry out important biological functions (reviewed in: BRUNET ET AL., 2020; HELLENS ET AL., 2016; HSU & BENFEY, 2018; KUTE ET AL., 2021; MAKAREWICH & OLSON, 2017; MUDGE ET AL., 2022; ORR ET AL., 2020; PLAZA ET AL., 2017; SCHLESINGER & ELSASSER, 2022; VITORINO ET AL., 2021; WRIGHT ET AL., 2022). In fact, the legume gene *early nodulin 40* (*ENOD40*) was first considered as representing a 'non-translatable' RNA (CRESPI ET AL., 1994), but is arguably the first case of a lncRNA that was found to act through NCPs encoded in sORFs (ROHRIG ET AL., 2002; SOUSA ET AL., 2001). *ENOD40* participates in the initiation of symbiotic nodule primordia, and two *ENOD40* peptides (ENOD40-I and ENOD40-II) as well as a structured RNA region of the transcript are required for its activity, through binding to sucrose synthase and re-localizing the RNA binding protein RBP1, respectively (KERESZT ET AL., 2018). Other plant NCPs derived from what could otherwise be considered (or were first considered) as lncRNAs are: POLARIS (PLS), involved in the auxin-cytokine-ethylene crosstalk in Arabidopsis and required for correct root growth and leaf vascular patterning (CASSON ET AL., 2002; CHILLEY ET AL., 2006; LIU ET AL., 2010; MOORE ET AL., 2015); the Arabidopsis IRON MAN peptides (IMA; also called FEP, for FE-UPTAKE-INDUCING PEPTIDE), that control iron transport (GRILLET ET AL., 2018; HIRAYAMA ET AL., 2018); maize Zm908p11, which functions in pollen germination and pollen tube growth (DONG ET AL., 2013); and maize Zm401p10, which is essential for anther development (MA ET AL., 2008; WANG ET AL., 2009) (for a recent 'consensus statement' on lncRNA definitions and functions, *see* (MATTICK ET AL., 2023)).

The distinction between a 'non-conventional' SEP/NCP peptidome that is largely derived from highly heterogeneous types of genetic elements, on one hand, and otherwise 'canonical' but small proteins is nevertheless further blurred. For instance, somewhere in between are microProteins (with capital P; miPs), a term originally coined in plants (STAUDT & WENKEL, 2011) to

specifically refer to small (5-15kDa) proteins that show sequence homology and are evolutionary related to larger, multidomain proteins -in particular, transcription factors (TFs)-, but that instead contain a single domain, specifically a protein-protein interaction domain (BHATI ET AL., 2018; BHATI ET AL., 2021; BHATI ET AL., 2020; EGUEN ET AL., 2015; MAGNANI ET AL., 2014; STAUDT & WENKEL, 2011). MicroProteins would thus be able to disrupt or modulate the formation of protein complexes by their 'target' proteins (MAGNANI ET AL., 2004; STAUDT & WENKEL, 2011). The miPs that have been functionally characterized to date usually function through interactions with the TFs that they are evolutionary related to (homotypic interactions) (BHATI ET AL., 2021), although an example of a heterotypic miP interaction with non-homologous TFs has been reported recently (WU ET AL., 2020). Thus, miPs have already been shown to be involved -through modulating the interactions of regulatory TFs- in photomorphogenic development (WU ET AL., 2020; YADAV ET AL., 2019), axillary meristem formation (ZHANG ET AL., 2018), shoot apical meristem development (KIM ET AL., 2008; XU ET AL., 2019), flowering time (GRAEFF ET AL., 2016; RODRIGUES ET AL., 2021), floral meristem termination (BOLLIER ET AL., 2018), or jasmonic acid signalling (HONG ET AL., 2020), and this variety of physiological roles will continue to expand, as plant genomes are thought to encode for hundreds of miPs (BHATI ET AL., 2020; MAGNANI ET AL., 2014; STRAUB & WENKEL, 2017). Importantly, however, it seems that during plant evolution miPs appeared after their homologous TFs, suggesting that they evolved from the TFs by domain loss (MAGNANI ET AL., 2014), whereas sequences generating SEPs/NCPs have been proposed as raw material for *de novo* gene birth (RUIZ-ORERA & ALBA, 2019; RUIZ-ORERA ET AL., 2018; RUIZ-ORERA ET AL., 2020) (*see* below).

This chapter will primarily focus on the 'non-conventional' peptidome in plants, but with the background of the state of knowledge on this topic in animals (and in particular humans), in which a vast majority of the studies in this emerging field have been conducted so far. Specific types of 'conventional' plant peptides will be mentioned, but for many of those, in-depth reviews are available elsewhere (*see* **Table 3.1**). Several key questions should be considered with respect to the non-conventional plant peptidome. Where does it originate from, from what types of genetic elements? What is

the nature and extent of its composition? How conserved is it across plant species? And, most importantly, what are the physiological functions of the peptidome, the specific functions carried out by this potentially large number of novel peptides, and how do SEPs/NCPs operate at a molecular and mechanistic level in plants? Answers to these questions are also emerging from mammalian studies that may help guide plant research on this topic.

3.2 Uncovering SEPs/NCPs: finding the needles in the haystack

The discovery and identification of functional sORFs embedded in eukaryotic genomes relies on three different methodological approaches: (i) ribosome and polysome profiling (Ribo-Seq and Poly-Ribo-Seq) for evidence of sORF translatability, (ii) mass spectrometry (MS)-based proteomics for direct SEP detection, and (iii) computational analyses for sORF prediction (Álvarez-Urdiola, Borràs, et al., 2023; Makarewich & Olson, 2017; Mohsen et al., 2023; Peeters & Menschaert, 2020; Prensner et al., 2023; Schlesinger & Elsasser, 2022).

3.2.1 Evidence of sORF translatability: ribosome and polysome profiling

Ribo-seq consists on the deep sequencing of ribosome-protected RNA fragments (ribosome footprints, of about 30 nt in length), whereby the periodicity of ribosome footprints (ribosomes decipher mRNA every three nucleotides) is used to identify bona-fide translation interactions (Hsu et al., 2016; Ingolia, 2016; Ingolia et al., 2014; Ingolia et al., 2009; Ingolia et al., 2011). Poly-Ribo-Seq is a modification of Ribo-Seq in which polysomes are enriched for the ribosome footprinting (Aspden et al., 2014). Sequencing of the ribosome footprints reveals the abundance and positions of ribosomes on a given transcript, providing a genome-wide view of active translation that can also be used to uncover previously unrecognized or unannotated translatable ORFs. In fact, Ribo-Seq provided the first large-scale experimental evidence that 'noncanonical' translation events existed in eukaryotic cells, and indicated that (thousands of) sequences annotated as non-coding RNAs, pseudogenes and UTRs could be an important source of

novel peptides (Aspden et al., 2014; Bazzini et al., 2014; Chothani et al., 2022; Duffy et al., 2022; Fields et al., 2015; Hartford & Lal, 2020; Ingolia et al., 2014; Ji et al., 2015; Martinez et al., 2020; Raj et al., 2016; Ruiz Cuevas et al., 2021; Ruiz-Orera et al., 2014; van Heesch et al., 2019). At present there is a variety of computational methods to analyse the Ribo-Seq data and infer potential coding sORFs, and it is important to note that different data processing pipelines may produce substantially different results in terms of the overall number, stringency, identity, and specific characteristics of the sORFs that are identified as translated, and that different methods may have different capacity for identifying certain classes of sORF (for an extensive discussion of this topic, *see* (Prensner et al., 2023)). It is also important to note that the bioinformatic tools that are used for translated ORF detection through Ribo-Seq depend on transcript information, either from the annotated genome or from RNA-Seq experiments, and therefore that the scope of the Ribo-Seq results is also determined by the datasets used for the analysis. Furthermore, it should be kept in mind that sORF translation may not result in the production of a stable and functional SEP. For instance, the translation of 5´-UTR sORFs (or upstream ORFs; uORFs) may often function to regulate the translation of the downstream main ORF of the mRNA (Schlesinger & Elsasser, 2022) (*see* below).

3.2.2 Direct SEP detection: Mass spectrometry

Mass spectrometry (MS)-based methods can be used to directly detect SEPs encoded by sORFs that are predicted from the sequence of the genome or the transcriptome, or by sORFs identified in Ribo-Seq experiments, and thereby to confirm the protein-coding nature of the corresponding sequences and transcripts (e.g., Aspden et al., 2014; J. Chen et al., 2020; Chothani et al., 2022; Duffy et al., 2022; Koch et al., 2014; Lu et al., 2019; Ma et al., 2014; Mackowiak et al., 2015; Martinez et al., 2020; Martinez et al., 2023; Ouspenskaia et al., 2022; Ruiz Cuevas et al., 2021; Slavoff et al., 2013; van Heesch et al., 2019; Vanderperre et al., 2013; Zhu et al., 2018). Although MS methods for peptide detection are still limited in sensitivity with respect to Ribo-Seq and have their own experimental limitations, including the possibility of producing high false-positive rates, (e.g., (Prensner et al.,

2023)), the fact is that over the past few years they have revolutionized our understanding of the peptidome, in particular because of the power provided by the combination of MS methods with Ribo-Seq or translatomics (i.e., three- or six-frame translation of transcriptomic or genomic sequences), in what are called peptidogenomic approaches (for review, *see* FABRE ET AL., 2021; NESVIZHSKII, 2014; SCHLESINGER & ELSASSER, 2022; SONG ET AL., 2023).

However, the detection by MS of novel peptides derived from sORFs presents specific challenges that should be taken into consideration (for a more detailed description on methodologies, focused on plant peptidomics, *see*: (ÁLVAREZ-URDIOLA, BORRÀS, ET AL., 2023)). First, an efficient and high-quality peptide-specific extraction protocol is key to improve the identification and sequence coverage of low-abundance SEPs by MS, as well as the use of methods for the separation and enrichment of peptides from proteins prior to the LC-MS/MS analyses (CAO ET AL., 2023; CARDON ET AL., 2020; KHITUN & SLAVOFF, 2019; MA ET AL., 2016). In the end, it is the combination of extraction, enrichment, and processing (i.e., protease cleavage prior to MS) methods what will determine the identification of a particular set of peptides in any given sample (ÁLVAREZ-URDIOLA, BORRÀS, ET AL., 2023; FABRE ET AL., 2021). Second, additional difficulties lie in under-sampling (i.e., identification of only a subset of the peptides) by conventional data acquisition methods, and in that SEPs detection is stochastic due to their size and expression characteristics, as suggested for example in a study to optimize a SEP discovery MS workflow using human samples (MA ET AL., 2014).

For peptide identification from tandem mass spectra there are two approaches that could be used: database search and *de novo* sequencing. In the database search method, all potential peptide sequences included in a specified database are retrieved for each spectrum, and each peptide-spectrum match is scored via a scoring function calculated by database search engines; in contrast, *de novo* sequencing extracts peptide sequences directly from tandem mass spectra using specific algorithms (ÁLVAREZ-URDIOLA, BORRÀS, ET AL., 2023; FABRE ET AL., 2021). The database search method is widely used for proteomics and peptidomics and can be based on canonical (annotated) protein databases (e.g., UniProt) or, if the purpose of the study is the identification of novel SEPs, customized databases containing

putative SEPs defined by bioinformatic or transcriptomic analyses (i.e., RNA-sequencing or Ribo-Seq). In fact, current integrated peptidomics pipelines include different database creation strategies (Álvarez-Urdiola, Borràs, et al., 2023), from the use of Ribo-Seq data (e.g., Aspden et al., 2014; J. Chen et al., 2020; Duffy et al., 2022; Koch et al., 2014; Menschaert et al., 2013; Raj et al., 2016; van Heesch et al., 2019) to the three-frame translation of transcriptomics datasets (e.g., Chothani et al., 2022; Duffy et al., 2022; Guruceaga et al., 2020; Lu et al., 2019; Ma et al., 2018; Ma et al., 2014; Slavoff et al., 2013; Vanderperre et al., 2013; Wright et al., 2016). Strategies based on the six-frame translation of the genome sequence have also been used, for instance in yeast (He et al., 2018), Drosophila (Zheng & Zhao, 2022), humans (Zhu et al., 2018) and plants (S. Wang et al., 2020), although it is a challenging approach because searching very large databases reduces the sensitivity of peptide identification by introducing more false positives, as the likelihood of obtaining high-scoring random matches is increased (Nesvizhskii, 2010, 2014). In comparison to database search, the *de novo* method for peptide identification is less powerful and mature, but in plants it has been used for specific peptide characterization or as a complement to database search (e.g., Culver et al., 2021; Gemperline et al., 2016; Jorge & Balbuena, 2021; Ye et al., 2016). An issue that is still not satisfactorily resolved in MS shotgun proteomics is the large number of unassigned spectra, i.e., where the originating peptide cannot be identified despite the spectra being of reasonable quality (Chick et al., 2015). Several factors might contribute to the prevalence of unassigned spectra: from the corresponding peptidic sequences not being present in the search databases to naturally occurring posttranslational modifications (PTMs), or chemical modifications that might have occurred during sample processing, or other experimental issues (Chick et al., 2015). The identification of PTMs, however, is relevant for improving the understanding of this hidden part of the proteome, as PTMs may play important roles in the yet to be discovered biological functions of SEPs.

3.2.3 Prediction of sORFs: in silico approaches

Bioinformatic approaches have been used (and continue to be developed) to distinguish coding and non-coding sequences and predict sORFs and SEPs from eukaryotic genomes and transcriptomes, including lncRNAs (e.g., Z. CHEN ET AL., 2023; FRITH ET AL., 2006; HANADA ET AL., 2010; HANADA ET AL., 2007; LADOUKAKIS ET AL., 2011; LIN ET AL., 2011; MACKOWIAK ET AL., 2015; TONG ET AL., 2020; TONG & LIU, 2019; ZHANG ET AL., 2022; Y. ZHANG ET AL., 2021; ZHAO, MENG, KANG, ET AL., 2022; ZHAO, MENG, & LUAN, 2022; ZHAO ET AL., 2023; ZHU & GRIBSKOV, 2019). These computational tools and analyses for sORF prediction can be divided into two categories (alignment-based and alignment-free) and be based on detecting sequence conservation and purifying selection, sequence similarity, codon pattern, or in the use of machine learning and deep learning.

Sequence conservation, determined by analysing the occurrence of synonymous and non-synonymous codon substitutions, is frequently used to detect coding regions and assess their protein-coding potential, on the basis that as synonymous substitutions do not lead to amino acid sequence changes, they occur more frequently in coding regions. In the case of SEPs, the short length of the aligned sequences and the limited number of possible changes pose a difficulty for obtaining statistical significance in these analyses, but a tool such as PhyloCSF takes a phylogenetic approach by analysing a multispecies nucleotide sequence alignment to determine whether it is likely to represent a conserved protein-coding region, based on a formal statistical comparison of phylogenetic codon models (LIN ET AL., 2011). PhyloCSF has been used extensively to detect sORFs in multiple eukaryotic genomes, and in particular in combination with Ribo-Seq for either the Ribo-Seq data to provide support for the sORFs identified through PhyloCSF comparative genomics (MACKOWIAK ET AL., 2015) or, conversely, for PhyloCSF to support translated sORFs detected by Ribo-Seq (e.g., BAZZINI ET AL., 2014; JI ET AL., 2015; LI ET AL., 2016; MARTINEZ ET AL., 2020). Other tools specifically test for the coding potential of sORFs without the need for sequence alignments, such as sORF finder (HANADA ET AL., 2010), which is based on the distinct hexamer composition in coding versus non-coding sequences and has been used in plant and animal genomes (CRAPPE ET AL.,

2013; HANADA ET AL., 2007); MiPepid (ZHU & GRIBSKOV, 2019), a machine learning tool developed specifically for the prediction of micropeptides directly from DNA sequences that is based on nucleotide patterns (4-mer features); CPPred (TONG ET AL., 2020; TONG & LIU, 2019), which estimates transcript coding potential by using multiple features derived from RNA and protein sequences and improves distinguishing between coding and non-coding RNAs; or the more recently developed DeepCPP (Y. ZHANG ET AL., 2021), a deep neural network for RNA coding potential prediction; csORF-finder (ZHANG ET AL., 2022); and, specifically tailored for the identification of sORFs in plant lncRNAS, sORFplnc (ZHAO ET AL., 2023), sORFPred (Z. CHEN ET AL., 2023), lncPepid (ZHAO, MENG, & LUAN, 2022), and ISPL (ZHAO, MENG, KANG, ET AL., 2022). The development of computational tools to predict translatable sORFs and characterize their coding potential is a very active area of current research, but in any case, the results are computational predictions that require experimental verification through Ribo-Seq, MS, or functional screening approaches (*see* below).

Another strategy for SEP identification is based on sequence similarity with previously identified proteins. This approach would miss on species-specific candidates and orphan genes and, in general, it is not well suited for global NCP searches because their levels of homology and conservation tend to be lower than those of canonical proteins (*see* below), even though some NCPs have been found to be highly conserved in animals (e.g., KOH ET AL., 2021). In plants, sequence similarity has been extensively used to identify families of conventional precursor-derived peptides across different species. For instance, a search for RGF/GLV/CLEL-family peptides (initially discovered as signalling peptides involved in root development in Arabidopsis) led to the identification of hundreds of homologs in all major extant land plant lineages (except hornworts) (FURUMIZU & SAWA, 2021) (**Table 3.1**). Likewise, a BLAST approach was used to identify in Medicago members of several SSP gene families (e.g., CLE, CEP, RGF/GLV/CLEL, IDA, PSK, PSY, CIF, EPF; **Table 3.1**), generating a database that was then used in MS data analyses to detect secreted peptides (PATEL ET AL., 2018) (*see* also below).

A variety of databases and online repositories have been created to store and make available information on peptides and sORFs detected through Ribo-

Seq, MS and/or bioinformatic approaches. Examples include repositories devoted to sORF-encoded peptides in Arabidopsis (e.g., ARA-PEPs (HAZARIKA ET AL., 2017)), in multiple plants (e. g., PsORF (Y. CHEN ET AL., 2020)), or in animals (e. g., SmProt (Y. LI ET AL., 2021), sORFs.org (OLEXIOUK ET AL., 2018), or OpenProt (BRUNET ET AL., 2021)). These repositories might in turn be used to ensemble the search databases that are required in MS experiments. Peptide databases that are literature- or sequence similarity-based are also available (e. g., PlantPepDB (DAS ET AL., 2020)).

sORFs and SEPs that are discovered through these prospective, genome-wide approaches can then be specifically confirmed and further investigated through low-throughput molecular biology methods (epitope-tagging and expression, subcellular localization studies, antibody generation, *in vitro* translation experiments, etc.).

3.3 The non-conventional eukaryotic peptidome: lessons from animals

Studies in mammals (mouse, human) have demonstrated that SEPs can be present in the cell at concentrations that are within the range of typical cellular proteins and that they can exhibit different and specific subcellular localizations (e.g., PRENSNER ET AL., 2021; SLAVOFF ET AL., 2013; VAN HEESCH ET AL., 2019). SEPs can be structural or regulatory components of macromolecular complexes, participate in signalling cascades, or act in an autonomous fashion (SCHLESINGER & ELSASSER, 2022). Specific human SEPs have already been found to play significant roles in cancer, metabolism, mitochondrial processes, muscle physiology, development, DNA repair, apoptosis or immunology (for an extensive summary of functions already determined for human SEPs, *see* SCHLESINGER & ELSASSER, 2022; WRIGHT ET AL., 2022). Moreover, the massive and widespread transcription of the eukaryotic genome and the pervasive translation of lncRNAs habilitate sORFs and the resulting small peptides or microproteins as raw materials for *de novo* gene origin and evolution (RUIZ-ORERA & ALBA, 2019; RUIZ-ORERA ET AL., 2018; RUIZ-ORERA ET AL., 2020; SCHLOTTERER, 2015). In fact, recent results demonstrate that there has been *de novo* birth of (functional) microproteins

in the human (SANDMANN ET AL., 2023; VAKIRLIS ET AL., 2022) and Drosophila (ZHENG & ZHAO, 2022) lineages, as well as in *Oryza* (rice) (Zhang et al., 2019). For instance, in rice the *de novo* gene *GSE9* evolved from a previous non-coding region of wild rice *Oryza rufipogon* through the ORF acquisition of a start codon and contributes to grain shape difference between the indica and japonica rice varieties (*GSE9* codes for a small protein, 107 aa long) (R. CHEN ET AL., 2023). Furthermore, using random sequence libraries it has been shown in *E. coli* that randomly generated sORFs can confer beneficial effects to cells and that new functions can emerge *de novo* from these sORFs (BABINA ET AL., 2023; NEME ET AL., 2017).

3.3.1 General observations

Several general – and to some extent intriguing – observations that can be deduced from the current findings on the mammalian peptidome include the following:

(i) Translatable sORFs are abundant in lncRNAs, and lncRNAs can be an important source of SEPs, as determined by Ribo-Seq and MS. For instance, ribosome profiling of the human heart resulted in the identification of 1,577 noncanonical ORFs, of which 339 (22%) were sORFs from lncRNAs, and also determined that over 20% of the heart lncRNAs (169 out of 783) were translated; furthermore, over 40% of those lncRNA SEPs were confirmed by MS (VAN HEESCH ET AL., 2019) (in this study the most abundant class of sORFs, 69%, were uORFs). Likewise, a Ribo-seq study of human neural cultures detected 706 ncRNAs (mostly lncRNAs) whose expression was altered by neuronal activity, and 128 (18%) of those showed active translation of novel sORFs, with a subset being also verified by MS (DUFFY ET AL., 2022). Specific lncRNAs whose physiological functions are carried out by the encoded NCP (e.g., D'LIMA ET AL., 2017; KONDO ET AL., 2007; MAGNY ET AL., 2013; MATSUMOTO ET AL., 2017; MISE ET AL., 2022; NELSON ET AL., 2016; ZHANG ET AL., 2017), or by both the encoded NCP and the RNA itself (e.g., ANDERSON ET AL., 2015; LEE ET AL., 2021; LIN ET AL., 2014; SENIS ET AL., 2021; YU ET AL., 2017) have in fact been identified. It may then be that for perhaps many lncRNAs the emerging question would not be if they have a physiological role as non-coding RNAs, but rather whether they function solely through the

encoded SEP or whether the RNA and the encoded microprotein or peptide have distinct and independent functions (DUFFY ET AL., 2022). Although the exact proportion of sORFs/SEPs that are derived from lncRNAs may vary among different genome-wide studies, it consistently represents a substantial fraction in random sampling experiments of all possible sORFs (frequently around 25% (J. CHEN ET AL., 2020; OUSPENSKAIA ET AL., 2022), but may reach up to 40% (HUANG ET AL., 2021)). In this context, it is also noteworthy that out of a list of 42 human SEPs already characterized as functionally or physiologically significant (WRIGHT ET AL., 2022), 55% are derived from lncRNAs (**Figure 3.3**). The mean length for lncRNA-encoded sORFs in humans has been estimated in ~54 aa (NEVILLE ET AL., 2021).

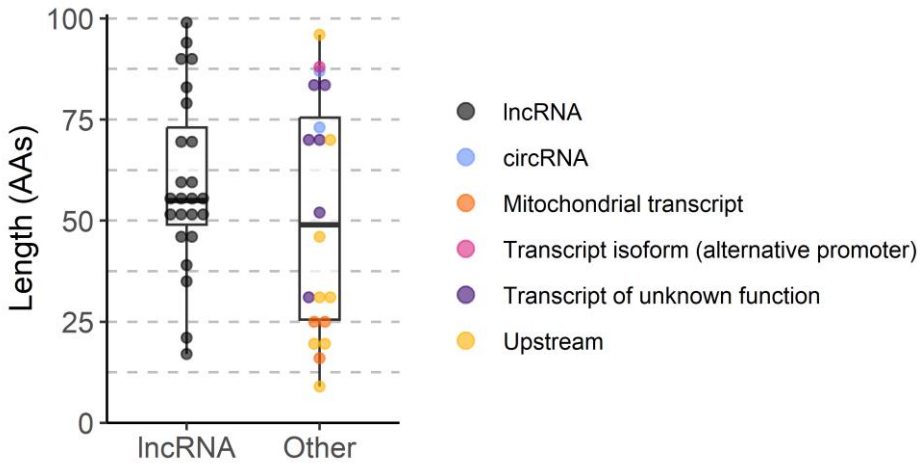


Figure 3.3. Human sORF-encoded non-canonical peptides that have been functionally or physiologically characterized.
Boxplot depicting the size and number of characterized human NCPs according to the type of RNA in which the corresponding sORF resides (lncRNAs, circRNAs, mitochondrial transcripts, transcript isoforms, transcripts of unknown function, and uORFs). Data from (WRIGHT ET AL., 2022).

(ii) uORFs are a major source, and perhaps the primary source, of translatable sORFs. For instance, in a recent ultra-high-depth RNA- and Ribo-Seq study that encompassed six human primary cell types and five human tissues as well as a tailored data analysis pipeline to generate a high resolution map of human RNA translation, 7,767 high-confidence translated

sORFs were detected, of which 5,308 (68%) were located in the 5′-UTRs of known protein-coding transcripts (CHOTHANI ET AL., 2022). The second major class of translated sORFs identified in the study was that of novel sORFs in annotated lncRNAs (1,652 sORFs, 21%), with the remaining being 3′-UTR sORFs (807 sORFs, 10%). Using previously available MS datasets, a total of 614 of the corresponding SEPs were detected (8% of the 7,767 sORFs) although, interestingly, the lncRNA-encoded peptides were detected much more frequently than the 5′-UTR peptides (286, or 17% of lncRNA sORFs, versus 281, or 5.3% of 5′-UTR sORFs) despite the fact that the level of translation of lncRNA sORFs was generally lower than that of 5′-UTR sORFs (CHOTHANI ET AL., 2022). These observations might suggest that, overall, lncRNA sORFs are a more probable source of biologically functional SEPs than the 5′-UTR sORFs. Although uORF-encoded microproteins with critical roles in cellular processes have already been identified, as for example MP31, Kastor, Pollucks, SEHBP, and EMBOW (120 aa) (*see* below, and (Y. CHEN ET AL., 2023; HUANG ET AL., 2021; KOH ET AL., 2021; MISE ET AL., 2022)), a general assumption is that many uORFs may simply function to downregulate the expression of the downstream main ORF.

(iii) A 'traditional' characteristic for predicting protein-coding ORFs is the presence of an ATG start codon. However, it is now apparent that non-AUG translation initiation of SEPs is extended, and that sORFs show a trend towards a much-increased use of near-cognate or alternative start codons relative to canonical ORFs (CAO & SLAVOFF, 2020; CHU ET AL., 2015). Various MS-based (e.g., MA ET AL., 2016; MA ET AL., 2014; MENSCHAERT ET AL., 2013; RUIZ CUEVAS ET AL., 2021; SLAVOFF ET AL., 2013; VANDERPERRE ET AL., 2013; Q. ZHANG ET AL., 2021) and Ribo-Seq (e.g., J. CHEN ET AL., 2020; CHOTHANI ET AL., 2022; DUFFY ET AL., 2022; MARTINEZ ET AL., 2020; RUIZ CUEVAS ET AL., 2021) studies have indicated that up to 35-75% of the identified sORFs/SEPs would initiate with non-AUG start codons.

(iv) In general, sORFs/SEPs are less evolutionary conserved than standard ORFs/proteins and have lower conservation scores (e.g., J. CHEN ET AL., 2020; FESENKO, KIROV, ET AL., 2019; FESENKO ET AL., 2021; RUIZ-ORERA ET AL., 2018; SANDMANN ET AL., 2023; VAN HEESCH ET AL., 2019; WRIGHT ET AL., 2022), which is also in agreement with the concept that ncRNA sORFs may

facilitate *de novo* gene evolution. For instance, out of 3,877 microprotein-encoding sORFs from mouse adipocytes, 991 (25.5%) showed homology to rat sequences, but only approximately 250 (6.5%) to more distant species such as human, dog, or pig (MARTINEZ ET AL., 2023). Likewise, an analysis of over 7,000 Ribo-Seq human sORFs only identified 273 (4%) as showing high similarity to mouse sequences (MARTINEZ ET AL., 2020) and, in another study, approximately 68% of the sORFs identified as translated in human brain were not detected in other species, including primates (DUFFY ET AL., 2022). These observations were further strengthened by a recent study on the conservation and evolutionary origin a of a set of 7,264 high-confidence human sORFs, which found that a vast majority were evolutionary young (6,506, 90%) as they lacked significant protein homology outside of primate mammals, and identified 222 as being human specific (SANDMANN ET AL., 2023).

(v) sORFs with limited sequence conservation or a *de novo* origin can produce functional microproteins that participate in crucial cellular and biological processes, i.e., functionality is not limited to highly conserved SEPs (e.g., SANDMANN ET AL., 2023; VAN HEESCH ET AL., 2019). For example, a subset of 124 of the sORFs that showed evidence of translation in the human brain had been previously identified as causing growth phenotypic changes when knocked-out in human induced pluripotent stem cells (iPSCs) and in a leukemia cell line (*see* below) and, strikingly, 101 (81%) of those sORFs were human-specific, lending support to the idea that newly evolved, species-specific SEPs can acquire important functions (DUFFY ET AL., 2022). It has also been shown that novel, adaptive transmembrane NCPs can emerge from thymine-rich non-genic regions in yeast (VAKIRLIS ET AL., 2020).

(vi) The mechanisms of action of SEPs/NCPs are varied. Because of their reduced size, frequent presence of intrinsically disordered regions or of transmembrane (TM) helices, and other physicochemical characteristics, it is assumed that SEPs would primarily act by interacting with proteins and other cellular components and modifying or modulating their functions.

SEPs containing single-pass TM α-helices are, for example, a group of micropeptides that interact with SERCA calcium transporters and regulate

muscle relaxation and contractility (e.g., myoregulin –46 aa, MLN–, DWORF –34 aa–, or Sarcolamban –28 aa, SCL–, among others (ANDERSON ET AL., 2015; ANDERSON ET AL., 2016; MAGNY ET AL., 2013; NELSON ET AL., 2016)) or neural differentiation and cellular homeostasis in pancreatic β cells (e.g., pTUNAR/BNLN, 48 aa (M. LI ET AL., 2021; SENIS ET AL., 2021)). Myomixer/Minion/Myomerger (84 aa) is a membrane-localized micropeptide that is involved in myoblast fusion during skeletal muscle development, perhaps through the interaction with Myomaker (a transmembrane protein) and/or other proteins (BI ET AL., 2017; QUINN ET AL., 2017; ZHANG ET AL., 2017). Other examples are provided by SPAR (90 aa), which localizes to the lysosomes and regulates mTORC1 activation (MATSUMOTO ET AL., 2017), and by Kastor (53 aa) and Pollucks (40 aa), which insert in the outer mitochondrial membrane and directly interact with voltage-dependent anion channel (VDAC) affecting spermatogenesis and fertility (MISE ET AL., 2022). It could be that the activity of many membrane proteins is regulated by interactions with TM micropeptides.

Beyond membrane compartments, SEPs have also been found to interact directly with a variety of proteins in other subcellular contexts. For instance, the intrinsically disordered NoBody micropeptide (68 aa) is a component of the mRNA decapping complex via direct interaction with EDC4 and localizes to the cytoplasmic ribonucleoprotein granules called P-bodies (D'LIMA ET AL., 2017; NA ET AL., 2020); and MP31 (31 aa) interacts with lactate dehydrogenase inhibiting its activity in mitochondria and having a tumour-suppressing role (HUANG ET AL., 2021).

Short NCPs can also function as signalling molecules in the control of metabolic homeostasis (MOTS-c, 16 aa, (LEE ET AL., 2015)), act in a non-cell-autonomous manner in development (*pri* peptides, 11 or 32 aa, (KONDO ET AL., 2007)), or have cytoprotective activity (humanin, 24 aa, (LEE ET AL., 2013)) despite being devoid of characteristic N-terminal signal sequences for secretion.

Transcriptional regulation and gene expression can also be affected by NCP activity. For example, SEHBP (46 aa) is a mammalian conserved SEP that interacts with chromatin associated proteins, localizes to distinct loci in the

genome and can affect transcription, perhaps playing a role in epigenetic regulation (KOH ET AL., 2021); and the lncRNA-derived GATA3-interacting cryptic protein (GT3-INCP, 120 aa) is detected in the nucleus, binds DNA, and interacts with the GATA3 transcription factor, facilitating GATA3 binding to the common cis regulatory elements and coregulating genes associated with estrogen response/cell proliferation (ZHENG ET AL., 2023). EMBOW is an overlapping uORF microprotein (120 aa) that interacts with WD40-repeat protein WDR5 and regulates its binding to other partners, thus affecting cell cycle and gene expression (Y. CHEN ET AL., 2023).

In summary, and as these examples illustrate, there is an extensive functional and molecular mechanistic diversity among SEPs, which will undoubtedly increase as more of them are identified and characterized.

3.3.2 Unanswered questions

Beyond the results and observations summarized above and once that the existence of an extensive (and still largely unannotated and uncharacterized) eukaryotic peptidome is accepted, several outstanding issues remain to be addressed:

– **The SEP/NCP-coding capacity of any eukaryotic genome is still unclear, but probably large.** Many Ribo-Seq experiments in humans or mouse have each revealed thousands of translated sORFs, but estimates of the actual number that exist in the genome vary from the thousands to the tens of thousands (PRENSNER ET AL., 2023). In addition, and from an experimental point of view, the overlap among the sets of sORFs/SEPs identified in different Ribo-Seq studies (or among different MS studies) can be limited. This is undoubtedly the result of both experimental aspects and the fact that sORF/SEP expression can be tissue, cell-type or condition dependent. Among the experimental aspects are differences in Ribo-Seq protocols, depth of sequencing and, in particular, data processing pipelines (CHOTHANI ET AL., 2022; PRENSNER ET AL., 2023). However, even when only high-confidence sORF sets resulting from the in-depth analyses of multiple Ribo-Seq samples are compared, the results are still more additive than overlapping. For instance, two high-confidence human sets of 7,264 (MUDGE

ET AL., 2022) and 7,767 (CHOTHANI ET AL., 2022) Ribo-Seq sORFs, both derived from multiple tissues and cell types, showed only 1,702 (22%) sORFs in common (although that percentage increased to 70% when additional filtering criteria were introduced such that the number of sORFs that were compared was reduced to 2,475 (PRENSNER ET AL., 2023)). Furthermore, the extensive presence of potentially translatable sORFs requires experimental demonstration of their capacity to actually produce stable (detectable) SEPs/NCPs in the cell, and although MS-based evidence is accumulating for some organisms, in particular human and mouse (J. CHEN ET AL., 2020; CHOTHANI ET AL., 2022; DUFFY ET AL., 2022; MARTINEZ ET AL., 2020; MARTINEZ ET AL., 2023; OUSPENSKAIA ET AL., 2022; PRENSNER ET AL., 2021; SLAVOFF ET AL., 2013; VAN HEESCH ET AL., 2019; ZHU ET AL., 2018), even in those cases their peptidome is still far from completely defined. Particularly relevant is the fact the three available approaches for sORF/SEP genome-wide detection identify sORFs/SEPs in different orders of magnitude: usually hundreds to low thousands in the case of MS proteomics, thousands to tens of thousands in Ribo-Seq experiments and up to hundreds of thousands in computational predictions. This, together with the inherent -but distinct- limitations of each methodology, inevitably leads to discordances in sORF/SEP number estimations and hampers the overlapping between sORF/SEP sets obtained through the different approaches (BRUNET ET AL., 2020; PRENSNER ET AL., 2023; RATHORE ET AL., 2018). As illustrated above, in studies that combine Ribo-Seq and MS-proteomics, only a minority of the Ribo-Seq identified sORFs are also detected as SEPs, due to both the lower sensitivity of MS-based detection versus Ribo-Seq and that some sORFs might generate unstable and undetectable peptides. In summary, the 'non-conventional' peptidome has substantially expanded the limits of the eukaryotic proteome, but where those limits reside for any organism is still unclear.

– **Relatively few SEPs/NCPs have been functionally characterized.** Experimental evidence for the biological functionality of a vast majority of the predicted or identified SEPs/NCPs is still lacking in any organism. However, as the sORFeomes and peptidomes of human and mouse become established, systematic, large-scale genetic, functional, or molecular screenings are starting to address this issue. For instance, a screen of 553

noncanonical ORFs (primarily from lncRNAs) in human cancer lines determined that a majority of them could induce gene expression changes when expressed and that this biological effect was mediated by the corresponding protein/peptide and not by the RNA; furthermore, a CRISPR/Cas-9 loss-of-function viability screen showed that many affected cell survival (PRENSNER ET AL., 2021). Similarly, in another CRISPR-based knock-out screen of 2,353 noncanonical CDSs, including 1,098 uORFs and 613 lncRNA ORFs, that was performed in human induced pluripotent stem cells (iPSCs) and a leukemia cell line, disruption of the translatable ORF resulted in consistent growth defects in over 400 of the cases (J. CHEN ET AL., 2020). In another study, the combination of Ribo-seq, a CRISPR/Cas9 knockout pooled screen, and large scale computational analysis of molecular/clinical data for breast cancer to analyse 758 lncRNA-encoded ORFs, led to the identification of 28 sORFs that could be clinically relevant, and it was further demonstrated that one of these lncRNA-encoded microproteins is an integrated component of the transcriptional regulatory network that drives aberrant transcription in cancer (ZHENG ET AL., 2023). At a smaller scale, a screening of SEPs of human vascular muscle cells and a gain- and loss-of-function approach identified NCPs with regulatory functions in those cells and potentially linked to atherosclerosis (LI ET AL., 2023). The specific interaction of SEPs with other cellular proteins can also be taken as an indication of SEP functionality, and the identity of the interacting partners help identify the biological process that the SEP might be involved in. Accordingly, methods have been developed to identify cellular SEP interactors (e.g., DITTMAR ET AL., 2019; KOH ET AL., 2021; SANDMANN ET AL., 2023) and used in medium-size screens. For example, a MS-based interactome screen was conducted for a set of 266 selected human SEPs revealing interactions for the vast majority of them with proteins involved in a variety of cellular processes, including with proteins essential for cell survival (SANDMANN ET AL., 2023). Interestingly, most of the SEPs included in this study were either recently evolved (showing that the capacity of a SEP to interact may be present at its evolutionary origin or appear shortly afterwards, i.e., that *de novo* originated proteins can quickly become functional) or very short in length, between 3 and 15 amino acids

(questioning if a clear-cut lower size limit for SEP functionality exists) (SANDMANN ET AL., 2023).

The sampling of the human peptidome that these various studies represent further demonstrates – far and beyond the individual cases of biologically active human SEPs/NCPs that have already been characterized – that the eukaryotic peptidome constitutes an important source of unrecognized small proteins with important biological roles in physiology, development, and disease, and that in humans it could be a potential target for the development of novel therapies. The full repertoire of their functions and molecular mechanisms of action remains to be established.

These questions and issues are also very pertinent to plants.

3.4 The non-conventional plant peptidome: current status

The 'non-conventional' or sORF-derived plant peptidome is largely undefined and unexplored. However, as in mammals and yeast, the existence of novel, uncharacterized small peptides has been inferred from transcriptome data (e.g., in *Populus* (YANG ET AL., 2011)), and in particular as Ribo-Seq has been used to demonstrate extensive translation of open reading frames, including novel sORFs, in species such as Arabidopsis (BAZIN ET AL., 2017; HSU ET AL., 2016; KURIHARA ET AL., 2020), maize (LIANG ET AL., 2021), wheat (GUO ET AL., 2023) or tomato (WU ET AL., 2019) (for review, *see* FUJITA ET AL., 2019; HSU & BENFEY, 2018; KAGE ET AL., 2020) (**Table 3.2**). Initial experiments in Arabidopsis showed ribosome association with some noncoding RNAs (JIAO & MEYEROWITZ, 2010; JUNTAWONG ET AL., 2014) and that uORFs could be translated (JUNTAWONG ET AL., 2014; LIU ET AL., 2013), and ribosome profiling of Arabidopsis roots and shoots identified actually translated sORFs in noncoding transcripts, at least some of which could produce stable SEPs *in planta* as determined by epitope tagging (HSU ET AL., 2016).

Table 3.2. Plant Ribo-Seq studies and translated sORF detection.

Species	Tissue / Process	Identified sORFs	sORF type	Average SEP Length (aa)	Experiment	Reference
Arabidopsis	Root and shoot	208 (<100 aa)	173 uORFs; 9 dORFs; 26 sORFs in ncRNAs	26 aa (uORFs); 48 aa (sORFs)	Ribo-Seq ORFs were selected without length cut-off; 3 uORFs and 1 ncRNA-derived sORF > 99 aa were also detected.	(HSU ET AL., 2016)
Arabidopsis	Root / Pi response	197 (<100 aa)	lncRNA sORFs	36 aa	sORFs were selected by translational efficiency (TE) and ribosome release score (RRS), witout length cut-off; 30 ORFs > 99 aa were included for a total set of 227 lncRNA-derived sORFs. MS evidence for 19 of these sORFs.	(BAZIN ET AL., 2017)
Arabidopsis	Seedling / Blue light response	1,613 (<50 aa)	1378 uORFs; 32 dORFs; 203 sORFs in ncRNAs	21 aa (uORFs); 30 aa (sORFs)		(KURIHARA ET AL., 2020)
Tomato	Root	1,540 (<100 aa)	1290 uORFs; 250 sORFs in novel transcripts	25 aa (uORFs); 47 aa (sORFs)	A small subset of the peptides encoded by these uORFs and sORFs (16 and 12, respectively) were detected by MS. 68 sORFs showed a predicted signal peptide and could represent secreted peptides.	(WU ET AL. 2019)
Wheat	Grain / Development	2,737 (<100 aa)	1041 uORFs; 274 dORFs; 655 internal ORFs; 767 lncRNA sORFs	39 aa (uORFs); 67 aa (sORFs)	Ribo-Seq ORFs were selected without length cut-off, for a total of 1254 uORFs, 367 dORFs, 825 internal ORFs and 914 lnc RNA ORFs. Approximately 22% of the ORFs use non-AUG start codons.	(GUO ET AL. 2019)

A subsequent Ribo-Seq study of the root translatome in response to phosphorous (Pi) limitation largely expanded those initial observations, identifying 1,140 lncRNAs as ribosome-associated (50% of all lncRNAs detected) and in particular 225 sORFs with a higher potential of being functionally translated (BAZIN ET AL., 2017). Analysis of previously obtained proteomic datasets provided MS evidence for some of these sORFs, demonstrating peptide stability in the plant, and it was also determined that translation of some of the sORFs was upregulated or downregulated by Pi deficiency, suggesting that the encoded SEPs could be of physiological importance (BAZIN ET AL., 2017).

Some of those novel, translated plant sORFs identified through Ribo-Seq were shown to be evolutionary conserved, but in many instances homologs were detected only in closely related species. For instance, 31 of the 225 lncRNA sORFs identified in the Pi-starvation study were detected in Brassicaceae outside of the Arabidopsis genus, of which 9 were broadly conserved in angiosperms (BAZIN ET AL., 2017), and 15 of the 19 single-exon sORFs detected in ncRNAs of roots and shots (HSU ET AL., 2016) showed at least one homolog outside of *A. thaliana* (6 were detected only in Brassicaceae, and 9 were also detected in other plants). Similarly, a Ribo-Seq analysis of the translatome of tomato roots revealed 1,540 sORFs (<100 aa long), of which 1,290 were uORFs and 250 sORFs detected in novel transcripts (WU ET AL., 2019). Further analysis of a subset of 157 of those 250 sORFs (selected by being single-exon sORFs) indicated that a majority of them (96, or 61%) were specific to the Solanaceae, including 18 unique to tomato and 78 shared by tomato and either wild tomato or potato, whereas a total of 139 had homologs in at least one other plant genome (including non-Solanaceae species) (WU ET AL., 2019). Similarly, in a study of the Arabidopsis seedling translatome and its response to blue light, 203 sORFs were identified in non-coding intergenic or antisense RNAs: 55% of them were conserved in *A. lyrata* and 20% in *B. napus* (KURIHARA ET AL., 2020).

Although the level and degree of sORF/SEP evolutionary conservation that is detected varies among these different studies, it is apparent that some sORFs/SEPs can be highly conserved across plants (but perhaps a minority),

whereas others may be relevant for the evolution of lineage- or species-specific characteristics, paralleling what has been observed in animals.

In bread wheat, a recent study of the translatome during grain development identified 2,737 unannotated sORFs, including uORFs (1,041; 38%) and sORFs in lncRNAs (767; 28%) (GUO ET AL., 2023). A large number of these sORFs (1,883) showed differential expression dynamics at the translational level throughout grain development, and analysis of the corresponding SEP sequences indicated that a third of them harboured potential signal peptides, altogether suggesting that these sORFs might encode true functional peptides. Considering that the translatome of only one particular developmental process was characterized in the experiments (grain development at 5, 10, and 15 days after anthesis), it seems reasonable to expect that the total number of potential sORFs and SEPs in bread wheat will eventually be substantially large. These results also provide a strong indication of the potential relevance of the non-conventional peptidome in flower and fruit development.

In addition to Ribo-Seq studies, computational analyses had previously suggested that several thousands of novel, potentially coding sORFs could exist in the intergenic regions of the Arabidopsis genome (HANADA ET AL., 2007). In fact, it was found that when overexpressed, some of those novel sORF sequences could induce developmental alterations in plant growth, development, or cause lethality, raising the possibility that (many) sORFs with coding potential but that are still uncharacterized in plant genomes might be associated with morphogenesis and other developmental and physiological processes (HANADA ET AL., 2013; HIGUCHI-TAKEUCHI ET AL., 2020). Homologs for some of these computationally identified sORFs were detected in rice (OKAMOTO ET AL., 2014). But whether the phenotypic effects reported in those gain-of-function studies were caused by the sORF RNA or by a derived SEP was not determined, and neither was a loss-of-function approach pursued. However, for a specific sORF of that set it was subsequently found that it acts as a hormone-like peptide -AtPep3- involved in salinity stress tolerance (NAKAMINAMI ET AL., 2018), illustrating that non-conventional peptides identified through genome-wide approaches can play physiological roles in plants, much like it is being discovered in animals.

Initial experiments in moss (*Physcomitrella patens*) (FESENKO, KIROV, ET AL., 2019; FESENKO ET AL., 2021), maize (LIANG ET AL., 2021; S. WANG ET AL., 2020), Eucalyptus (JORGE & BALBUENA, 2021), pear (WANG, WU, SHI, ET AL., 2023) and Arabidopsis (S. WANG ET AL., 2020) have attempted the analysis of the global plant peptidome through MS-based approaches (**Table 3.3**). In addition, in the case of Arabidopsis the MS-based characterization of its global proteome also allowed the identification of a small number of SEPs (CASTELLANA ET AL., 2008; MERGNER ET AL., 2020); in soybean, MS analysis was used to obtain coding evidence for lncRNAs, identifying 153 NCPs derived from 179 lncRNAs (LIN ET AL., 2020); and in *Populus* a few tens of novel sORFs predicted through computational analyses of its transcriptome were confirmed by proteomics data (YANG ET AL., 2011). From these limited experiments, as well as the Ribo-seq studies described above, it appears that key observations on the characteristics of sORFs/SEPs from animals, such as the relevance of lncRNAs as a source of SEPs, the limited sequence conservation of SEPs, and the -extended- use of near-cognate or alternative start codons, also apply to plants.

In moss, a genome-wide bioinformatic analysis resulted in the identification of 70,095 novel potentially coding sORFs that were: AUG-initiated, single exon, 10-100 aa long, and located on either annotated transcripts (uORFs, internal ORFs, or dORFs; 63,109, or 90%), lncRNAs (5,745, 8%), or intergenic regions (unannotated transcripts; 1,241, 2%) (FESENKO, KIROV, ET AL., 2019). These sequences were then used as search database in an MS analysis that included samples from three different types of moss cells, which led to confirming the translation and peptide accumulation for 46 of the sORFs (36 located in annotated transcripts, 1 intergenic, and 9 in lncRNAs -20%-). The degree of sORF evolutionary conservation was low: 5,034 (7%) of the total sORFs were conserved among the transcriptomes of at least 1 out of 10 plant species, as well as 5 (11%) out the 46 translated sORFs, with three of these five corresponding to lncRNA-sORFs. Furthermore, functional analysis of four of the translated lncRNA-sORFs revealed that knocking them out affected moss growth and development, and that phenotypic alterations were also caused by their overexpression (FESENKO, KIROV, ET AL., 2019). Thus, this study provided evidence that different types of sORFs are translated in plants and demonstrated that some of them encode functional SEPs.

Table 3.3. Analysis of the global plant peptidome through MS-based approaches.

Species	Tissue/Process	Aim of study	Identified peptides	Reference
Moss (*Physcomitrella patens*)	Protonemata	Identification of translated sORFs in plant cells	828 peptide sequences: 46 high confidence SEPs (17 in gametophores, 29 in protonemata, 14 in protoplasts) (14-99 aa)	(Fesenko, Kirov, et al., 2019)
Maize (*Zea mays*)	Maize inbred line B73 leaves (three-leaf stage)	Large-scale discovery of novel peptides	1,993 novel SEPs	(S. Wang et al., 2020)
Maize (*Z. mays*)	Inbred line B73 seeds	Identification and characterization of small peptides	2,695 small peptides (up to 100aa)	(Liang et al., 2021)
Arabidopsis	Columbia-0 leaves (four-leaf stage)	Large-scale discovery of novel peptides	1,860 novel SEPs	(S. Wang et al., 2020)
Pear (*Pyrus bretscneiden*)	Twenty-four different tissues/samples covering all major organs	Global protein expression patterns. Discovery of novel proteins and peptides	607 novel SEPs (up to 100 aa)	(Wang, Wu, Shi, et al., 2023)
Soybean (*Glycine max, Glycine soja*)	Various tissues and conditions	lncRNA discovery	153 unique novel small peptides encoded by 179 lncRNA genes	(Lin et al., 2020)

Note: Only examples of studies that attempted the global characterization of the plant peptidome and to expand the annotation of the corresponding genome are listed. For more comprehensive lists of peptidogenomic and proteogenomics experiments in plants, *see*: (Álvarez-Urdiola, Borràs, et al., 2023; Song et al., 2023).

In a subsequent and more comprehensive analysis of *P. patens* lncRNAs, 175,272 sORFs were identified computationally, approximately 50% of which were AUG-initiated and the rest initiating from the near-cognate codons UUG or CUG. These lncRNA-sORFs were assessed for conservation across nearly 500 plant species: approximately 86% were not conserved, whereas 22,524 sORFs (13%) were moss-specific, and 645 were highly conserved, suggesting that a large pool of potential SEPs encoded by lncRNAs exists in plants but that a vast majority of them would be lineage- or species-specific (FESENKO ET AL., 2021). Mirroring SEP characteristics already identified in animals, putative transmembrane domains, signal peptides, or 'consensus disorder prediction' motifs were identified in subsets of the lncRNA-sORFs (4,978, 9,472, and 8,595, respectively), and evidence of translation for 195 sORFs was obtained from various moss MS datasets. Altogether, these first analyses of the moss 'non-conventional' peptidome support the idea that lncRNAs could be an important source of functional SEPs in plants, as is the case in animals. The limits of sequence conservation of putative SEPs among different plant species also highlight the importance of species-specific MS analyses for the characterization of the plant peptidome, and are also in agreement with the idea that sORFs/SEPs are raw materials for *de novo* gene origin and evolution.

In the case of maize two different peptidogenomics studies are available. The first one was based on a six-frame translation of the maize genome and reported the identification of 2,837 peptides by MS, 1,993 of which were derived from 'not-annotated' sequences (i.e., were identified as NCPs in the study) and 844 were derived from annotated proteins/peptides (identified as conventional peptides, CPs) (S. WANG ET AL., 2020). Ribo-seq analyses provided further evidence for 732 (37%) of the identified NCPs and, interestingly, a certain NCP enrichment was detected within genomic regions associated with phenotypic variation and domestication selection, suggesting that NCPs could potentially be involved in the genetic regulation of complex traits and domestication in this species (S. WANG ET AL., 2020). In the second study, Ribo-seq and RNA-seq data were used to generate a search database of 9,388 sORFs, which comprised uORFs (2,907), dORFs (3,445), and also uoORFs, intORFs, and doORFs (*see* **Figure 3.1** for nomenclature),

but only 49 sORFs derived from non-coding transcripts (i.e., the database was essentially based on alternative translation of annotated mRNAs), and that led to the identification by MS of 2,695 NCPs (LIANG ET AL., 2021). However, the overlap between the sets of peptides identified in the two studies was very limited: it consists of only a few CPs, and no NCP was independently identified by both studies (**Figure 3.4**). Moreover, Liang et al. (LIANG ET AL., 2021) analysed the MS data from (S. WANG ET AL., 2020) using their custom translatome database of 9,388 sORFs, identifying 158 NCPs, 66 of which were also among the 2,695 NCPs that they had reported (i.e., a 2.4% overlap when the translatome database was confronted with the two different MS datasets). This limited overlap is not necessarily surprising. First, the database and approach used for MS peptide search were very different in the two studies, and in fact in (S. WANG ET AL., 2020) it was reported that a vast majority (1,652, 83%) of the 1,993 NCPs identified by the six-frame genome translation could be assigned to lncRNAs through transcriptomic analyses, whereas lncRNA-sORFs were largely absent from the translatome database used in (LIANG ET AL., 2021). In addition, it is well-known that the combination of experimental protocols used for peptide extraction, enrichment, and MS analysis will influence the set of SEPs that are identified in the experiment (*see* above, and (FABRE ET AL., 2021)). Last, the two studies utilized different sample types, six tissues in (LIANG ET AL., 2021) versus only seedling leaves in (S. WANG ET AL., 2020).

In any case, the comparison of the two studies makes clear that the real size and scope of the maize peptidome (or, in fact, of the peptidome from any plant) are still undefined. It is worth noting that even for the much better characterized human peptidome and in studies that not only identify SEPs but that also include large-scale functional analyses (summarized above, J. CHEN ET AL., 2020; PRENSNER ET AL., 2021) the hit overlap is relatively limited (15-25%), indicating that the functional sORFs that those studies identified represent only a fraction of those encoded by the human genome. Similarly, in a study that used extensive Ribo-Seq profiling and three different human cell lines, > 7,500 sORFs were detected, but only ~1,500 (20%) in at least two of the three cell lines, and only ~480 (6.4%) in the three of them (MARTINEZ ET AL., 2020).

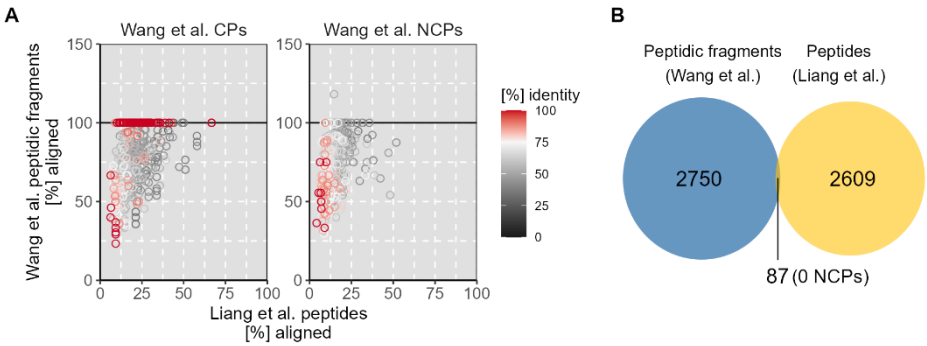


Figure 3.4. Genome-wide non-canonical peptide identification in maize.
Analysis of the overlap between two different published studies for maize peptide identification by MS. A) Scatter plot representing the differences between the aligned sequences of the MS peptidic fragments identified by Wang et al. (S. WANG ET AL., 2020), as derived from conventional peptides (CPs) and non-conventional peptides (NCPs), and the SEPs identified by Liang et al. (LIANG ET AL., 2021). BLASTp was used to compare the amino acid sequences of the peptides identified in the two studies. B) Venn diagram showing the overlap between the datasets of (S. WANG ET AL., 2020) and (LIANG ET AL., 2021). The intersection in the diagram corresponds to peptides (peptidic fragments) identified in (S. WANG ET AL., 2020) whose sequences align in 100% of their length and with more than 90% of identity to peptide sequences from (LIANG ET AL., 2021). The overlap between the two datasets is limited to CPs, as no NCP was identified by both studies.

These observations all further highlight the technical challenges and the complexity of defining and characterizing the peptidome in eukaryotes and, importantly, also point to a substantial level of cell-/tissue-/condition-specificity.

The six-frame genome translation peptidogenomics strategy was also applied to Arabidopsis (leaf tissue), which resulted in the identification of 1,860 NCPs; of those, 666 (36%) were derived from intergenic regions (i.e., potential ncRNAs), 154 (8%) from UTRs, 651 (35%) from out-of-frame exons (i.e., intORFs), and the rest from introns and junctions (S. WANG ET AL., 2020).

A large-scale proteogenomic atlas or pear has recently been developed through the integration and correlation of transcriptome and proteome data from 24 tissues and/or developmental stages, including seedling tissues, floral organs and tissues, fruit tissues, and fruit developmental stages (WANG, WU, SHI, ET AL., 2023). Although the main purpose of the study was not the identification of NCPs per se, as neither a small peptide MS strategy nor a customized putative sORF search database were employed, the annotation of the pear genome was improved through the identification of 4,294 'new protein-encoding events', of which 607 were of no more than 100 codons and therefore represented small ORFs. Some of those small ORFs were fully localized in intergenic regions (206), or in introns (17) or in the opposite strand (49) of annotated genes, and could therefore represent different types of sORFs/SEPs (WANG, WU, SHI, ET AL., 2023).

In addition to these broad peptidogenomic approaches, 'targeted' peptidomic experiments that in general address known families of plant peptides, in particular secreted signaling peptides/small proteins (SSPs), have been used to confirm peptide presence *in planta* and to associate peptides to specific physiological or developmental processes. These studies do not attempt to characterize the whole-genome 'non-conventional' plant peptidome, but can nevertheless identify new members of the corresponding gene families. Examples include: SSPs affecting root growth in *Medicago truncatula* (Patel et al., 2018); Arabidopsis SSPs (OHYAMA ET AL., 2008), including potential auxin-responsive SSPs (LUO ET AL., 2019); rice SSPs induced by the blast fungus *Magnaporthe oryzae* that could be involved in immunity – which also led to the discovery of an additional 51 unannotated SSPs – (P. WANG ET AL., 2020); or cysteine-rich, potential antimicrobial peptides (AMPs) in *Capsicum* (CULVER ET AL., 2021), among others.

In summary, knowledge on the 'non-conventional' plant peptidome is starting to accumulate and it appears that most or all of the overall findings that have been pioneered by research on animal, particularly human, SEPs will also apply to plants, and that newly identified SEPs will be found to play important roles in plant development and physiology.

3.5 The 'non-conventional' plant peptidome and flower development: the tip of the iceberg?

Precursor-derived peptides have long been known to play important roles in flower and fruit development and physiology (**Table 3.1**, **Figure 3.5**), from CLAVATA3 (CLV3), which is expressed in the shoot apical and floral meristem stem cell reservoirs and forms part of the network that maintains stem cell homeostasis (FLETCHER, 2020; FLETCHER ET AL., 1999), to RALF peptides that control an intergeneric hybridization barrier on Brassicaceae stigmas (LAN ET AL., 2023). Moreover, it is starting to become clear that non-precursor-derived peptides and in particular novel SEPs/NCPs identified through genome-wide analyses, that is, the 'non-conventional' peptidome, should also be taken into consideration to understand flower, fruit, and seed development.

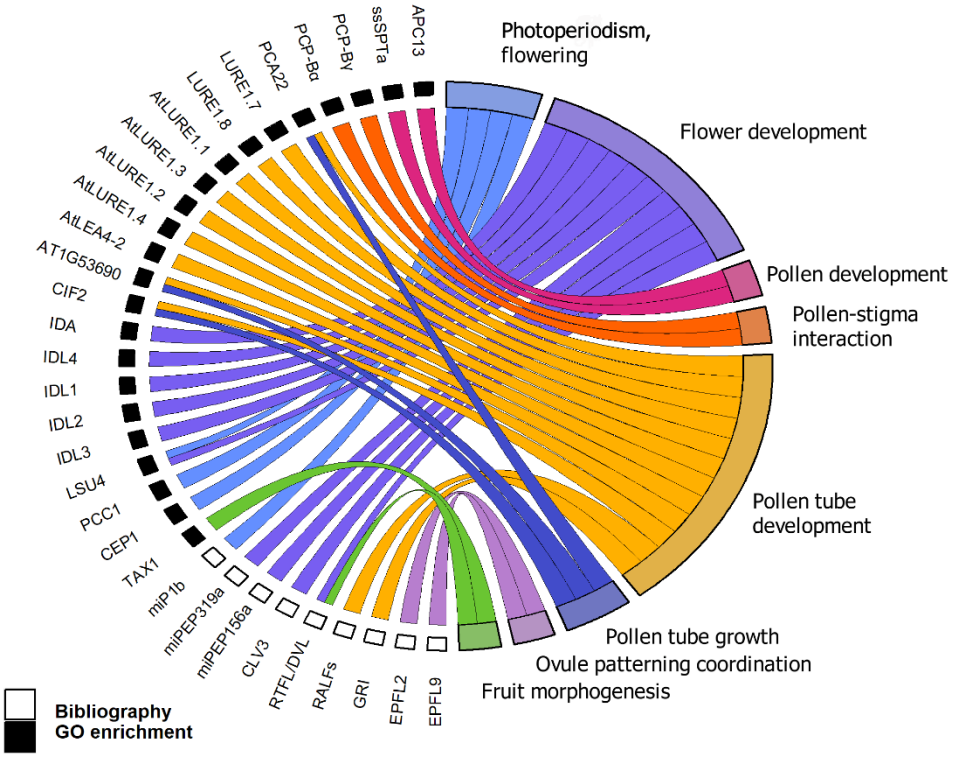


Figure 3.5. Arabidopsis peptides with functions related to flowering and flower and fruit development.
GO enrichment results of peptides annotated in **Sup Table 3.1** (in black) and other peptides described in the literature that are known to have a role in flower development (in white). GO results in **Sup Table 3.2**.

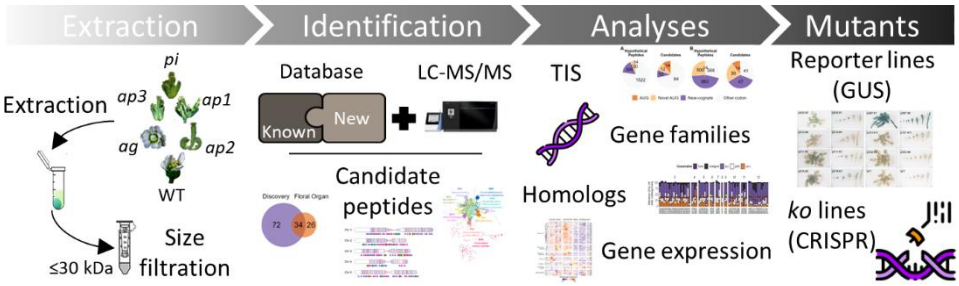
As summarized above, a recent Ribo-seq study of bread wheat grain development identified a large number of sORFs as differentially expressed during the process (GUO ET AL., 2023), and although there was no functional characterization of any of those sORFs, nor a demonstration by proteomics or other methods of the accumulation of the corresponding SEPs, it is reasonable to expect that some of them will indeed produce functional SEPs. Interestingly, a rice *de novo* gene (*GSE9*) was recently shown to contribute to grain shape differences between *indica* and *japonica* varieties, and to have been evolved from a previous non-coding region of wild rice (*Oryza rufipogon*) through the acquisition of a start codon (R. CHEN ET AL., 2023). Although the GSE9 protein is slightly larger (107 aa) than the arbitrary upper size limit for SEPs, it otherwise fulfils many of the characteristics outlined above: it contains intrinsic disordered regions, is predominantly localized in the plasma membrane, and shows no significant similarity with proteins from other eukaryotic species, as befits a *de novo*, sORF-generated, functional gene (R. CHEN ET AL., 2023). These studies suggest that SEPs might play specific functional roles in monocot grain physiology.

In maize, lncRNA-sORF encoded SEPs that play a role in anther development and pollen tube growth have been identified. *Zm908* is expressed predominantly in mature pollen grains and encodes a 97 aa-long SEP (Zm908p11) that functions in maize pollen germination and tube growth. Transgenic analyses in tobacco demonstrated that the peptide is necessary for *Zm908* function, and it was also found that it interacts with maize profilin 1, suggesting that Zm908p11 could be involved in the actin dynamics that are essential for pollen tube growth (DONG ET AL., 2013). *Zm401* is expressed primarily in the anthers (tapetal cells as well as microspores) in a developmentally regulated manner, and a knockdown of this gene led to aberrant development of the microspore and tapetum, and finally male sterility (MA ET AL., 2008). Zm401p10 peptide accumulates in the nucleus and its overexpression in maize retarded tapetal degeneration and caused microspore abnormalities (WANG ET AL., 2009).

Last, in the proteogenomic analysis of pear described above, 69 (10%) of the 607 'new coding event' small ORFs identified by MS were detected in style tissue, 18 of which were style-specific. Eight of those style-specific SEPs (49 to 88 aa in length) were expressed and purified as recombinant proteins and tested in pollen tube growth *in vitro* assays: four promoted pollen tube growth whereas one inhibited it, demonstrating that the newly identified SEPs could be biologically functional (WANG, WU, SHI, ET AL., 2023) and suggesting that the plant 'non-conventional' peptidome could play important roles in flower and fruit development and physiology.

Chapter 4

Arabidopsis non-conventional peptidome as related to flower development



Part of this chapter will be published as:

The Arabidopsis floral peptidome.

Álvarez-Urdiola, R., Matus, J.T., González, V.M., Bernardo-Faura, M., Riechmann, J.L. Manuscript in preparation.

Chapter 4. Arabidopsis 'non-conventional' peptidome as related to flower development

4.1 Background

The transcriptional and post-transcriptional regulation of flower development, as summarised in **Chapter 1**, has been characterized in the last decade using genomics and transcriptomics approaches (GREGIS ET AL., 2013; KAUFMANN ET AL., 2010; PAJORO, MADRIGAL, ET AL., 2014; YANT ET AL., 2010). These methods, alone or in combination with more traditional genetic studies, have validated the complex and highly interconnected gene regulatory network of the most innovative process that allowed angiosperms to rapidly expand during plant evolution. However, a wider view of these processes requires the study of the proteome, as recent studies have shown that translational regulation is determinant in developmental programs and that protein levels can vary despite mRNA levels being constant, and *vice-versa* (Y. GUO ET AL., 2023) (*see* **Chapter 2**). Over the past few years, it has also become evident that there is a substantial but uncharted fraction of the eukaryotic proteomes that is mainly composed of small proteins (peptidome), with roles and functions yet to be discovered (*see* **Chapter 3**).

The sources of plant peptides are numerous, either reliant on the processing of a polypeptide precursor or encoded by a short Open Reading Frame (sORF). Contrary to what was previously thought, long non-coding RNAs (lncRNAs), transcripts of unknown function (TUFs), 3'UTR´s, 5'UTR´s, intergenic regions, junctions, introns and primary miRNA transcripts (pri-miRs) might contain translatable sORFs (HANADA ET AL., 2013; HAZARIKA ET AL., 2017; LAURESSERGUES ET AL., 2022; S. WANG ET AL., 2020). Although peptidomics approaches have a lower sensitivity for detecting SEPs than RNA-based methods to detect potentially translating sORFs (ASPDEN ET AL.,

2014), a few mass spectrometry (MS)-based studies have been conducted in monocot and dicot plants for identifying novel alternative sORFs (LIANG ET AL., 2021; MERGNER ET AL., 2020; S. WANG ET AL., 2020).

Functional proteomics or gene-editing approaches are now available to characterize peptide roles. In Arabidopsis, well-characterized peptides have been found to be involved in organogenesis and development (GHORBANI ET AL., 2015; P. GUO ET AL., 2015; VALDIVIA ET AL., 2012). Besides the experimental validation for revealing the biological function of the peptidome, the biological function of new SEPs can also be studied using sequence features and their conservation across species (KIM ET AL., 2018). For a coding sequence (CDS), a non-synonymous substitution rate that is significantly lower compared to the synonymous substitution rate indicates that the sequence has experienced purifying selection or functional constraint (HANADA ET AL., 2007). Nevertheless, these criteria are not always applicable, as some non-conserved sORFs could evolve as newly coding ORFs with relevant roles (YEASMIN ET AL., 2018) or possess functions unrelated to their conservation (LAURESSERGUES ET AL., 2022).

The computational predictions and functional peptide characterizations recently available (GHORBANI ET AL., 2015; SLAVOFF ET AL., 2013; VANDERPERRE ET AL., 2013) motivated our group to explore the nature and true extent of the Arabidopsis peptidome, with the goal of understanding the potential role of non-conventional peptides in developmental programs. This study is aimed at continuing the understanding of the molecular mechanisms involved in the process of floral development in *A. thaliana* by the characterization of its sORF-encoded peptidome. The objective was to find novel functional peptides potentially encoded in lncRNAs, TUFs, and intergenic regions of the Arabidopsis genome, and upstream, downstream, or alternative ORFs (uORFs, dORFs, altORFs) of annotated Arabidopsis genes. Specifically, I addressed whether these SEPs could be involved in flower development by virtue of their differential expression in the Arabidopsis floral homeotic mutants. For this Thesis, a combination of transcriptomics, proteomics, and genetic techniques was used, including liquid chromatography with tandem mass spectrometry (LC-MS/MS) guided by a reference database composed of hypothetical and canonical SEPs and proteins.

4.2 Results

4.2.1 Detection of novel SEPs by mass spectrometry

Inflorescences of wild-type (WT, Ler-0 ecotype) and floral homeotic mutants (*ap1, ap2, ap3, pi* and *ag*) (**Figure 4.1**) were collected and peptides were extracted using size-selection by a 30K-ultrafiltration method followed by reverse phase chromatography (as described in (ÁLVAREZ-URDIOLA, BORRÀS, ET AL., 2023)) of four independent biological replicates for each genotype. For the identification of novel SEPs in the peptide samples, a database-guided mass spectrometry approach was used. The custom database that was generated was composed of ~100,000 non-redundant sequences that included, in addition to the annotated peptides and proteins from The Arabidopsis Information Resource (TAIR, Araport11; www.arabidopsis.org), potential peptides encoded (i) by lncRNAs (CNTdb 2.0; http://cantata.amu.edu.pl/) (SZCZEŚNIAK ET AL., 2019) and other transcripts (TAIR 'non-coding' –'nc'–) (potential peptides were directly inferred from the three-frame translation of those transcripts), (ii) in intergenic regions as identified by *in silico* analyses (HANADA ET AL., 2007, 2013), and (iii) poly-Ribo-seq identified sORFs present up- and down-stream of the main ORF of annotated genes or in alternative ORFs (HSU ET AL., 2016) (**Dataset S4.1**).

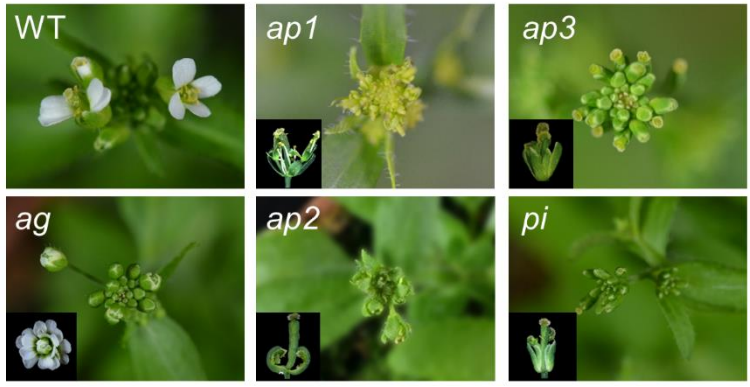


Figure 4.1. Floral phenotypes of the lines used in this study.
Landsberg erecta (WT – Ler-0), *ag, ap1, ap2, pi* and *ap3* inflorescences and mature flowers (bottom left of each panel) are shown.

In these LC/MS-MS experiments, 5,608 proteins (longer than 100 aa) and 2,084 peptides (of up to 100 aa) were identified in the Arabidopsis flower homeotic mutants and wild type plants. Among the identified peptides, I distinguished between those already annotated and described in TAIR (210 peptides; referred to as 'canonical peptides' in the text below), and those annotated as "hypothetical proteins" in TAIR or those from other sources in the custom database (1,874 peptides; collectively referred to as 'hypothetical peptides' below) (**Table 4.1, Dataset S4.2**).

Table 4.1. Number of identified peptides and proteins from each database.

	Araport11 (Canonical)	Araport11 (Hypothetical)	Hsu et al.	Hanada et al.	CNT db2.0	TAIR 'nc'
Proteins (> 100 aa)	5,387	122	-	-	62	37
Peptides (≤ 100 aa)	210	22	21	42	1,224	565
	Canonical peptides	**Hypothetical peptides**				

The dynamic range of protein and peptide abundance spanned six orders of magnitude. In all genotypes, the average intensity abundance of the peptides was slightly higher than that of the proteins for Araport11, CNTdb2.0 and TAIR 'nc' sequences (**Figure 4.2**). This corroborated that the peptide extraction worked properly in the sense that the samples were enriched in small peptides rather than in proteins. The LC-MS/MS data resulted in the detection of the N-terminal aminoacidic sequences of 682 peptides and 1,892 proteins, and of the C-terminal sequences of 321 peptides and 889 proteins, altogether corroborating a substantial number of annotated open-reading frame borders from Araport11 and of predicted sORFs borders for other sequences of the customized database (**Figure 4.3A**). Moreover, 28 of the smallest predicted peptides (with 10 aa) were detected as a single aminoacidic sequence containing both, N- and C-terminal ends (**Figure 4.3A**). N-terminal peptides often showed cleavage of the initiator methionine, especially for those sequences corresponding to hypothetical peptides. N-terminal acetylation was strongly dependent on the amino acid adjacent to the initial amino acid (**Figure 4.3B-E**).

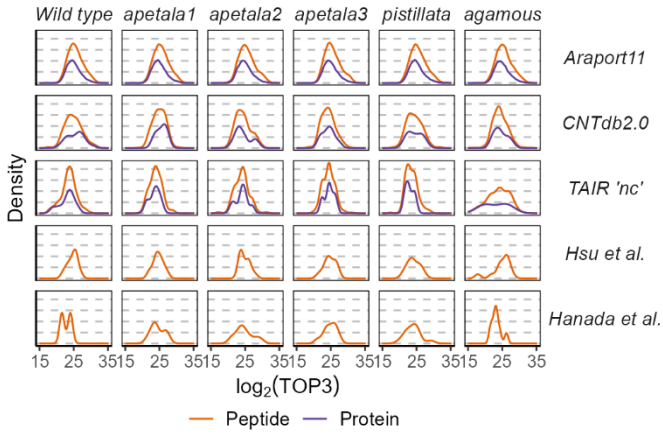


Figure 4.2. Dynamic range of protein and peptide expression in the different genotypes.
Density plot of protein abundance expressed as the average Log_2 TOP3 abundance for each genotype depending on their origin.

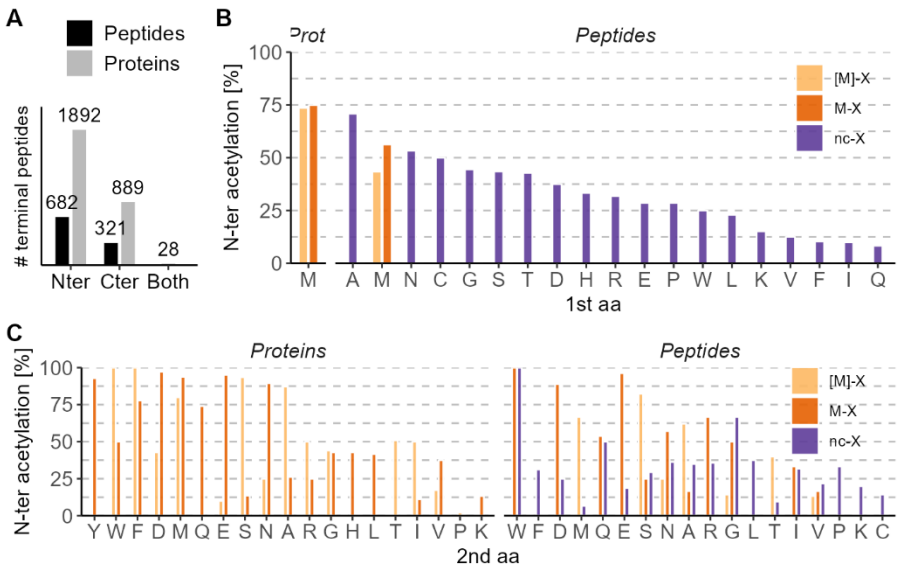


Figure 4.3. Amino acid composition of the sequences detected by mass spectrometry.
A) Bar graph indicating the number of identified N-terminal (N-ter) or C-terminal (C-ter) peptides (black) or proteins (grey). **B)** Frequency of N-terminal acetylation for sequences starting with M or not (1st aa). **C)** Frequency of N-terminal acetylation depending on the amino acid which follows the initiator (2nd aa).
M: methionine, nc: non-conventional. [M]-X, [nc]-X: missing first amino acid, M-X and nc-X: not missing first amino acid. *Cont. in next page.*

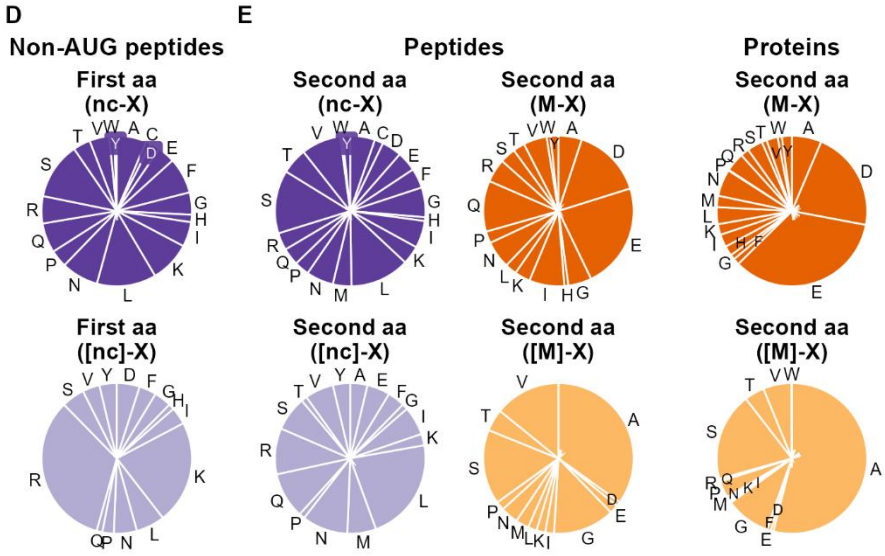


Figure 4.3. Amino acid composition of the sequences detected by mass spectrometry (Cont.).
D) First amino acid for those sequences beginning with a non-canonical initiation codon (different from AUG -M-). **E)** Second amino acid for the different detected sequences, depending on the cleavage of the initiator amino acid.

In most cases, the number of fragments detected for each peptide and protein was lower than three. This was the expected distribution due to the small size of the peptides (of up to 100 aa) and the exclusion of most proteins thanks to the size-filtration during peptide extraction. Nevertheless, the LC-MS/MS data covered, on average, ~30% and ~12% of each peptide and protein sequence, respectively, enabling the detection of unique amino acid sequences for 2,084 peptides and 5,608 proteins. (**Figure 4.4**). The median length of the peptides differed depending on their source, that is, on the type of genetic element from which their sequences were derived. The median peptide length was also affected by the cut-off that was established for the generation of each part of the database (i.e., 30 aa for peptides from Hanada et al., and 10 aa for peptides from Hsu et al., CNTdb 2.0 and TAIR 'nc'). The detected hypothetical peptides were in general smaller than the canonical peptides (**Figure 4.5A, B**).

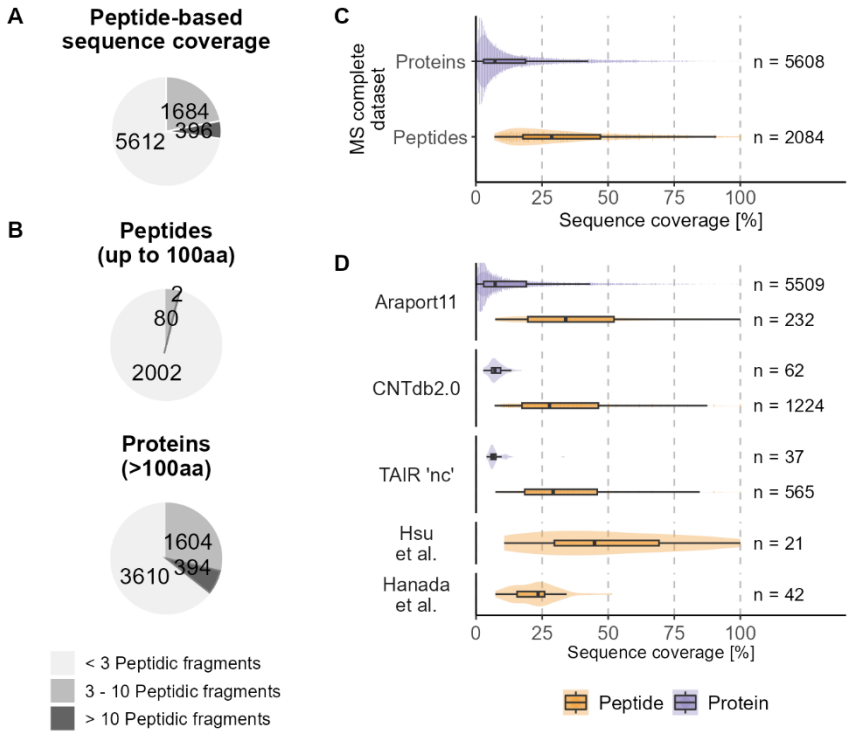


Figure 4.4. Sequence coverage in LC-MS/MS results.
Pie charts showing percentage of total sequences **(A)** and peptides and proteins separately **(B)** identified by < 3, 3-10 or > 10 peptide fragments. Distribution of peptide-based sequence coverage of all peptides and proteins **(C)** depending on their source **(D)**.

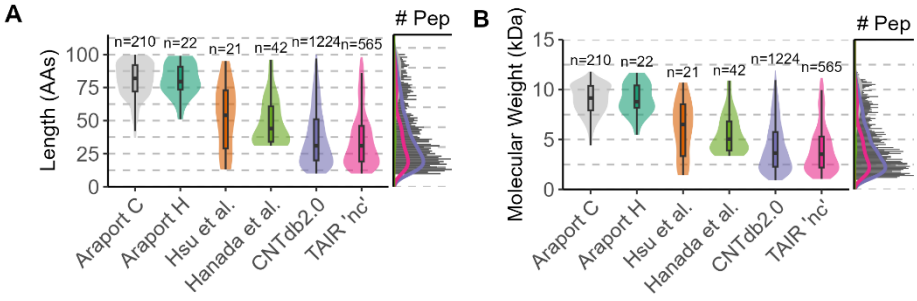


Figure 4.5. Canonical and Hypothetical peptides in *A. thaliana*.
A) Length distribution of peptides (AAs) included in the different databases represented as a violin plot for each database and as a histogram for the complete dataset. Lines in the histogram depicts the density distribution of the peptides of each database. **B)** Molecular weight distribution of peptides (kDa). Coloured by their source.

4.2.2 Overlap with previous a peptidomics study

A recent LC-MS/MS study of the non-conventional peptidome in maize leaves also included Arabidopsis leaf samples (S. WANG ET AL., 2020). BLASTp was used to investigate the possible overlap between the set of peptides identified in this Thesis and the peptide fragments identified by (S. WANG ET AL., 2020). I compared the sequences of the 2,084 peptides that were identified by LC-MS/MS in inflorescence tissues with the 2,363 conventional peptidic fragments (CPs) and 1,860 non-conventional peptidic fragments (NCPs) detected by Wang et al. in leaves (**Sup Table 4.1**).

The BLASTp results indicated that the two datasets were largely different, and very few *bona fide* identity matches were retrieved: most of the BLASTp-aligned peptide fragments from Wang et al. covered less than half the sequence of their corresponding match in the LC-MS/MS dataset, and less than 20% of the aligned sequences had an identity greater than the 75% (**Figure 4.6A**). Nevertheless, there were 91 peptides that had at least 90% identity between datasets, and also more than 90% of the sequence from Wang et al. aligned with the peptide sequence that was identified by LC-MS/MS. From these, 68 peptide pair matches corresponded to canonical Araport11 peptides and CPs from Wang et al.; one corresponded to a hypothetical Araport11 peptide and a CP from Wang et al.; 20 to Hsu et al

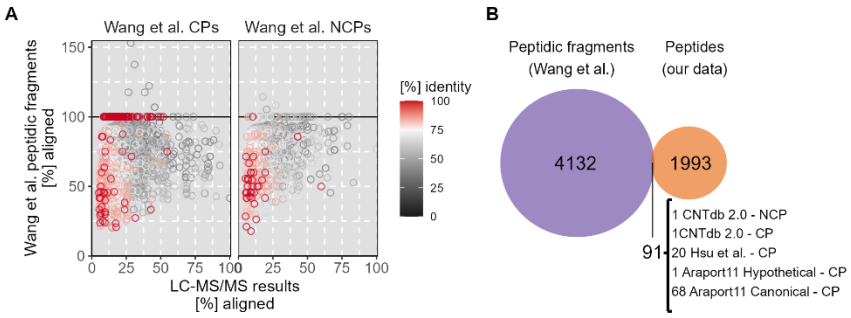


Figure 4.6. Data comparison (BLASTp results).
A) Scatter plot representing the differences between the aligned percentage of the total length of the peptidic fragments identified by MS by Wang et al. and SEPs identified in the LC-MS/MS peptidomics study, coloured by the percentage of identity. **B)** Venn diagram representing the intersection of those peptides with more than the 90% of identity between the datasets and more than the 90% of the Wang et al. peptidic sequence aligned.

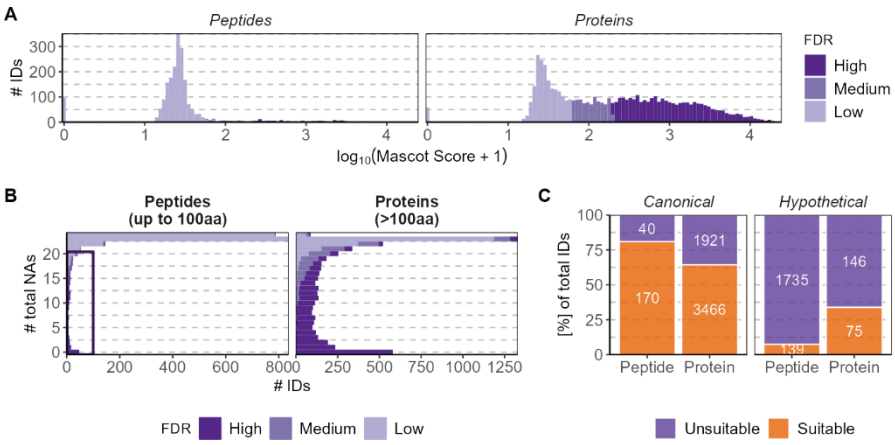
uORFs, dORFs and altORFs and CPs from Wang et al.; and two CNTdb 2.0 peptides, one paired with a CP from Wang et al. and the other with a NCP (**Figure 4.6B**). In summary, the Arabidopsis non-conventional peptidome identified in leaf tissue by Wang et al. is mostly non-overlapping with that identified in this Thesis from floral tissues.

4.2.3 Identification of over a hundred novel peptides specific to floral buds

Most identified peptides had a low Mascot Score (below 90) which, although considered as a low confidence of detection, is related to the fact that – as a consequence of their reduced length – many peptides were detected through a single peptidic fragment (**Figure 4.7A**). Another indicator of confidence can be derived from the total number of identifications and not-assigned values (NAs) for each peptide (**Figure 4.7B**). The criteria to select a final list of candidate peptides for further analyses were defined on the basis of the number of NAs for each peptide (**Sup Table 4.2**).

The main goal was to find new peptides encoded in sORFs and 'nc' RNAs, and with a potential role in floral organ development. Two different selection pathways were established: i) genotype-independent peptide discovery, and ii) genotype-dependent selection of peptides with a floral organ-specific accumulation pattern. These selection pathways were used in parallel, as there were peptides that would meet both (*see* below). On one hand, hypothetical peptides with less than 21 NAs (out of a total of 24 samples: 4 biological replicates for each of the 6 genotypes) were classified as genotype-independent discovery peptides. With this criterium, 106 discovery peptide candidates were selected, half of which had a high or medium confidence of detection (Mascot score) (**Figure 4.7B**). On the other hand, to predict organ-specific peptides and proteins, I considered their quantification in the different mutants at both complete peptide or protein and single peptidic fragment detection levels (i.e., raw spectra). Peptides and proteins, as well as their individual detected peptidic fragments, were classified as suitable for the floral organ classification analysis when they had less than three NAs in at least one genotype (quantified for that genotype/s) (**Figure 4.7C**). These criteria to select peptides based in the number of NAs, in both pathways,

were actually very conservative, given the stochastic nature of peptide detection in MS experiments (*see* below).



D

Discovery candidate	< 21 NAs / 24 total samples
Suitable for the floral organ analysis	< 3 NAs / 4 samples in a genotype

Figure 4.7. Selection criteria depending on the number of NAs.
A) Distribution of Mascot Scores for the peptides and proteins in LC-MS/MS results. **B)** Histogram of the number of IDs depending on their number of NAs in the dataset (0-24NAs). 'Discovery' candidates (less than 21 NAs) are framed in panel. **A-B**) panels are coloured by their FDR (high, medium, or low). **C)** Proportion and total number of peptides and proteins that were 'Suitable' and 'Unsuitable' for the floral organ classification. Canonical peptides are significantly enriched in 'Suitable' IDs when compared to canonical proteins (Fisher's p-value = 1.68e-07). Hypothetical peptides are significantly enriched in Unsuitable IDs when compared to hypothetical proteins (Fisher's p-value = 2.2e-16).

To make use of the different genotypes used in the experiment, I took advantage of the combinatorial nature of the (A)B(C) model of flower development (similarly to what was done in a previous work to predict organ-specific transcript expression (WELLMER ET AL., 2004)). Potential sepal-specific peptides were those quantified in *ap3, pi* and *ag* mutants, but not in *ap1* nor *ap2*. Petal-specific peptides would be identified by exclusively being present in *ag*. Stamen-specific peptides would be detected in *ap1* and/or *ap2*, but not in *ap3, pi* or *ag*. Finally, carpel-specific peptides would be those found in *ap3* and *pi* and absent from *ag* samples, irrespectively of their

quantification in *ap1* and *ap2* (**Figure 4.8A**). To be considered as organ-specific, proteins needed to be quantified in *Ler-0* as well, however this was not a requirement for peptides (up to 100aa) (**Figure 4.9**). This criterion was different between proteins and peptides to avoid discarding potentially interesting peptides with low abundances that were not quantified in the wild type samples.

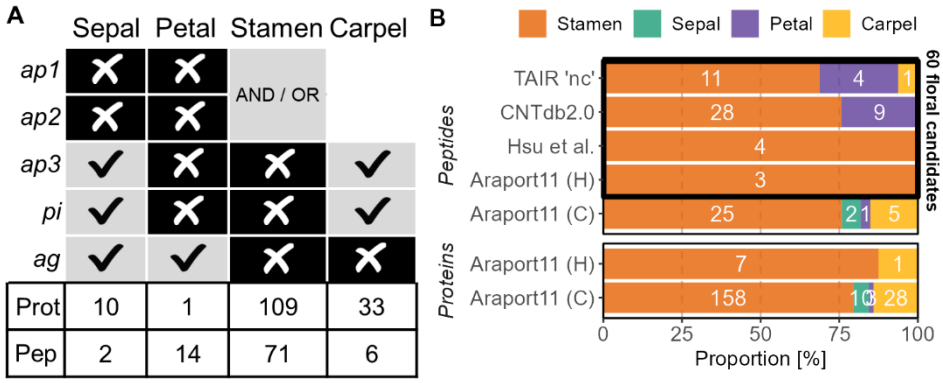


Figure 4.8. Selection of possible organ-specific proteins and peptides.
A) Criteria to select peptides and proteins specific to a certain type of floral organ. Tick: quantified (0, 1 or 2 NAs) in the indicated genetic background; cross: unquantified (3 or 4 NAs) in the indicated genetic background. **B)** Proportion of peptides and proteins identified in LC-MS/MS which were associated to each one of the floral organs divided according to their source. H: hypothetical, C: canonical. The squared section comprehends the 60 'floral organ' peptide candidates.

Using this set of criteria, 60 floral organ peptide candidates were selected (**Figure 4.8B**), from which 34 had been also retrieved as genotype-independent discovery peptide candidates, for a total of 132 peptides initially selected for further consideration (**Figure 4.10A, B**). The selection of candidates was not length-dependent (**Sup Table 4.3**, **Figure 4.10C**). In addition, the 132 candidates were evenly distributed among the complete genome of *A. thaliana* independently of their genetic element of origin (annotated ORFs -CDS-, intergenic regions, altORFs, uORFs, dORFs or ncRNAs) (**Figure 4.10D**).

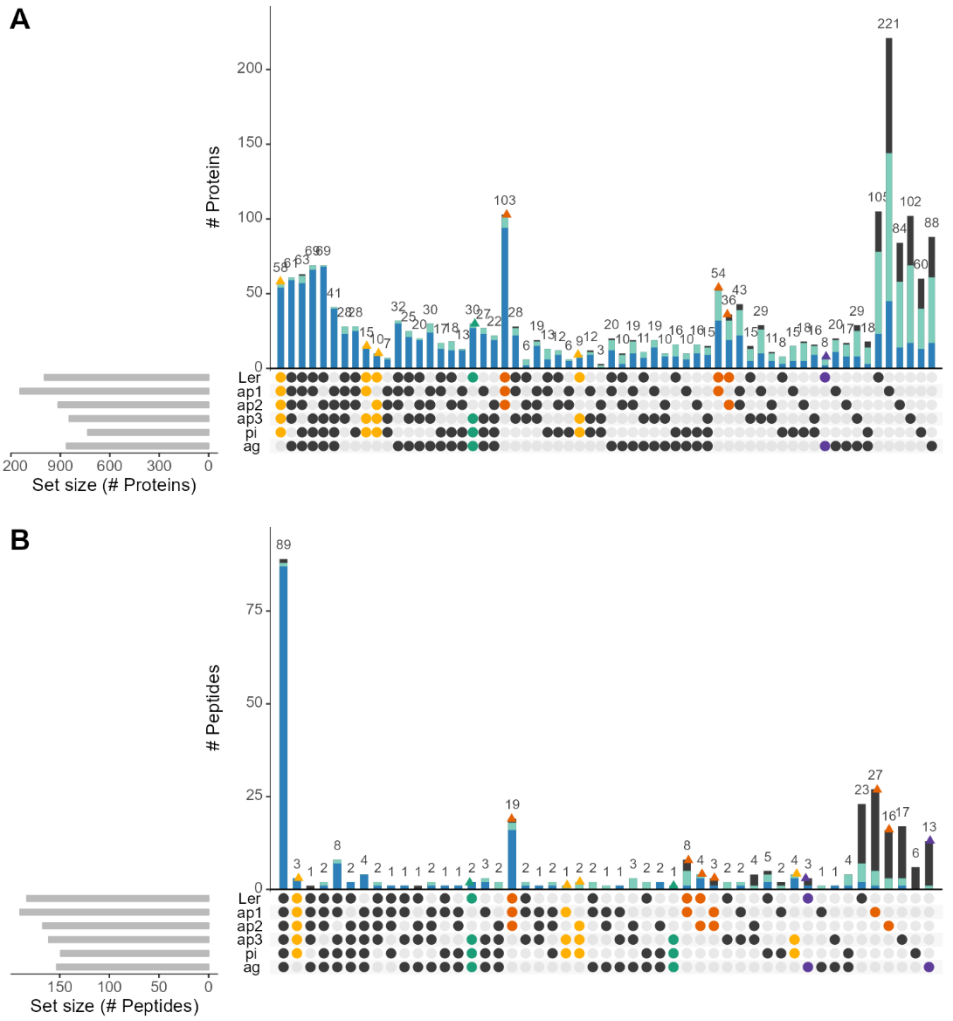


Figure 4.9. Number of possible organ-specific proteins and peptides.

Upset plots to visualize the intersections between proteins (**A**) and peptides (**B**) in each genetic background (*Ler, ap1, ap2, ap3, pi, ag*). Rows (left, horizontal bar graph) correspond to the total proteins (**A**) and peptides (**B**) detected in genetic background, and columns (top, vertical bar graph) correspond to the intersections. For each column, the filled in circles signal the genetic backgrounds that are part of the intersection. Vertical bars are coloured depending on the confidence of detection (low: grey, medium: light blue, high: dark blue), and circles are coloured by the organ assignation (stamen: orange, carpel: yellow, sepal: green, petal: purple). Besides the represented proteins and peptides, there were 2,257 proteins identified in all genotypes.

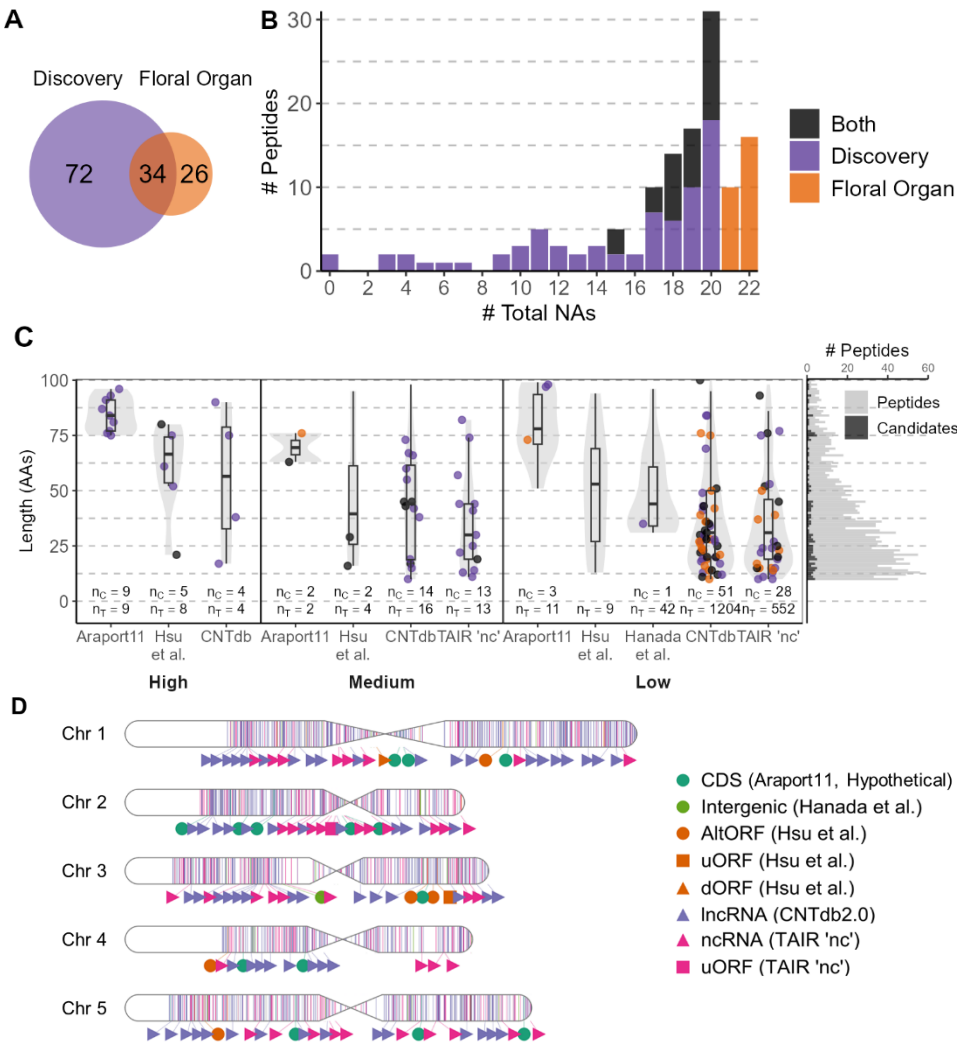


Figure 4.10. General information about the candidates.

A) Venn diagram indicating the number of peptide candidates selected through each method. **B)** NA distribution of the 'Discovery' (purple), and 'Floral Organ' (orange) peptide candidates. In black: peptides which are 'discovery' and 'floral' peptide candidates. **C)** Violin plots (and boxplots) showing the size distribution for the detected hypothetical peptides (up to 100 aa; grey shadow), with the selected peptide candidates superimposed with filled circles according to whether they had been identified as 'floral' (orange), 'discovery' (purple) or both (black), according to their confidence of detection (high, medium, low) and their source (Araport11, Hsu et al., Hanada et al., CNTdb2.0, TAIR 'nc'). Numbers indicate the total number of candidates (n_C) and the total number of hypothetical peptides detected (n_T) for each group. **D)** Genome-wide distribution of all hypothetical peptides detected by MS (coloured by their origin). Candidates are also separately shown with different shapes and colours depending on their origin and ORF type.

The obtained results about organ-specific proteins and peptides were compared to those obtained at transcript level by (WELLMER ET AL., 2004). In the LC-MS/MS results, most of the possible floral-organ proteins and canonical peptides were associated to stamens, followed by carpels, and then petals and sepals (**Figure 4.8B**), as in Wellmer et al. at transcript level (**Table 4.2**). In the case of the floral organ candidate peptides (hypothetical peptides), the proportion of peptides assigned to each organ was slightly different. As in the case of the proteins and canonical peptides, the highest number of possible floral-organ hypothetical peptides were associated with stamens. However, the proportion of possible petal-specific peptide candidates was higher than expected, and there was only one putative carpel-specific peptide candidate and none in the case of the sepals (**Figure 4.8B**).

Table 4.2. LC-MS/MS identified peptides and proteins classified as organ-specific in comparison to the organ-specific transcripts identified by (WELLMER ET AL., 2004).

	Araport11 (C)	*Araport11 (H)*	*Hsu et al.*	*Hanada et al.*	*CNTdb 2.0*	*TAIR 'nc'*
Carpel	32	1	0	0	0	1
Petal	4	0	0	0	9	4
Sepal	12	0	0	0	0	0
Stamen	183	10	4	0	28	11
Unassigned	3,456	71	3	1	43	30
Unsuitable	1,961	64	14	41	1,206	556

	Wellmer et al. (and in LC-MS/MS)	*Wellmer et al. (in total)*
Carpel	89	260
Petal	4	18
Sepal	4	13
Stamen	242 (C) + 8 (H)	1,162

The accuracy of the organ-specific classification criteria was also checked by performing a Gene Ontology (GO) enrichment analysis of the proteins and peptides annotated in Araport11 that through the candidate selection process were classified as organ-specific. The groups of peptides and proteins that were classified as specific for each organ type were indeed enriched in peptides and proteins known to be related with the development of that organ (**Figure 4.11A**). Moreover, a correlation network was created

based on the LC-MS/MS abundances of proteins and peptides, and a new GO enrichment analysis of the abundance modules calculated using the Random Matrix Theory was performed (**Figure 4.11B**). The module ME01 included 192 peptides and proteins that were classified as stamen-specific peptides, and it is enriched in pollen exine formation AGIs according to the GO results (**Table 4.3, Sup Table 4.4**).

Table 4.3. Number of peptides and proteins forming the modules of the correlation network.

Module	*Carpel*		*Petal*		*Sepal*		*Stamen*		*Unassigned*		TOTAL
	Pep	Prot	Pep	Prot	Pep	Prot	Pep	Prot	Pep	Prot	
ME01	4	1			1		27	165	23	327	**548**
ME02		16	11						7	137	**171**
ME03	2	12	3	3	1	10	3		32	485	**551**
ME04							26		3	309	**338**
ME05							15		10	165	**190**
ME06									28	146	**174**
ME07									7	123	**130**
ME08									3	89	**92**
ME77									22	215	**237**
ME78									3	34	**37**

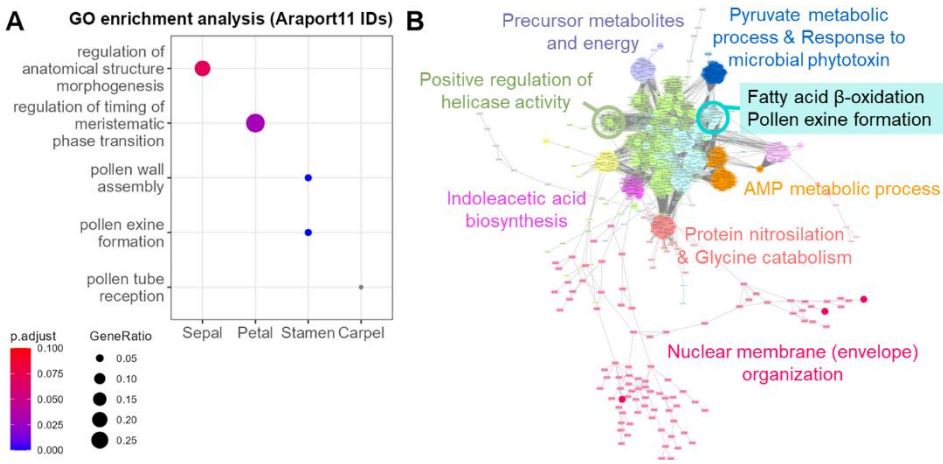


Figure 4.11. Validation of the 'floral organ' classification criteria. **A)** GO enrichment analysis: main category for those Araport11 peptides and proteins that were associated to each organ type. **B)** Correlation network for peptides and proteins identified in LC-MS/MS results. The main GO categories of each correlation module (ME) are indicated in the graph. Coloured by correlation ME. Circles represent peptides in MS and rectangles, proteins.

4.2.4 Translation initiation sites of the identified peptides

From the 1,874 hypothetical peptides identified by LC-MS/MS, only 131 of the corresponding sORFs were predicted to start with an AUG (14 annotated in Araport11, 117 in other sources). Furthermore, the sORFs of 292 peptides began with a near-cognate codon, that is, triplets that differ from AUG by only one nucleotide (e.g., AUC or AAG, **Figure 4.12A, D**). In addition to the initial annotation and prediction, it should be noted that the mass spectrometry results identified the complete N-terminal fragment for 551 of the 1,874 hypothetical peptides. Of this particular subset, 156 TIS were AUG or near-cognate (69 and 87, respectively), and 394 were other codons (similar to (CAO & SLAVOFF, 2020; NA ET AL., 2018)).

For those hypothetical peptides for which the fragment identified by mass spectrometry did not correspond to the N-terminus, their putative translation initiation sites (TIS) were searched for and re-annotated on the basis of the specific amino acid sequences (internal peptidic fragments) identified by LC-MS/MS (*see* **Materials and Methods** section **4.4.4**)

In summary, the sORFs of 26% of the total 1,874 hypothetical peptides, and 34% of the selected candidates held putative canonical start codons (AUG). The sORFs of another 45% of the total hypothetical peptides commenced with a near cognate codon, as was the case for the 36% of the selected candidates. Ten candidates had more than one TIS in a sequential arrangement (e.g., AUG-AAG or AAG-AUG). Finally, no potential AUG and near-cognate codon TIS was identified for about 30% of the hypothetical peptides and selected candidates (**Figure 4.12B, C, E**). That is, it appears that for the SEPs identified in this study, in addition to AUG-mediated translation initiation, near-cognate codons are also frequently used, and that the possibility of non-AUG, non-near-cognate initiation also exists, since for a subset of the hypothetical peptides a clear 'conventional' TIS could not be identified despite the fact that for some of those peptides the MS peptidic fragment corresponded to the peptide N-terminus.

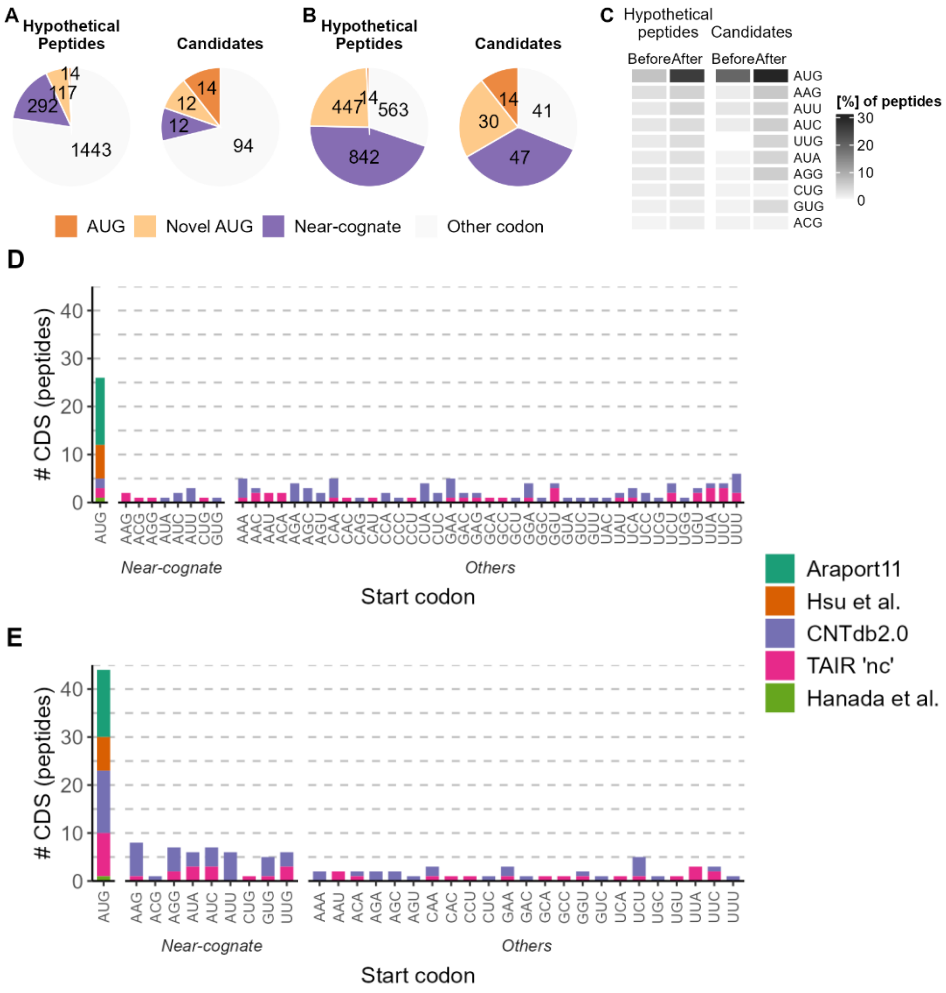


Figure 4.12. TIS of the hypothetical peptides identified by LC-MS/MS.

A-B) Pie chart depicting the number of peptides with each kind of TIS among all the hypothetical peptides identified by LC-MS/MS (n = 1,874) and for the selected candidates (n = 132) before (**A**) and after (**B**) the re-annotation based on LC-MS/MS spectra**. C)** Percentage of peptides whose start codon is AUG or a near-cognate codon before and after the re-annotation. **D-E)** Bar plot representing the number of peptides from each source (colours) that began with each possible codon for the 132 candidates before (**D**) and after (**E**) the re-annotation.

In addition to the identification of putative TIS, I categorised thirteen candidates as putative precursors of small-secreted peptides (SSPs) based on the mass spectrometry results (i.e., the peptidic fragment that was detected lacked a tryptic beginning and did not correspond to the TIS of the sequence). Two candidates (#014 and #101, which correspond to AT5G43695.1 and AT2G05215.1_29 respectively) were confirmed as carriers of potential secretory signals using the online tool SignalP (**Figure 4.13, Sup Table 4.3**).

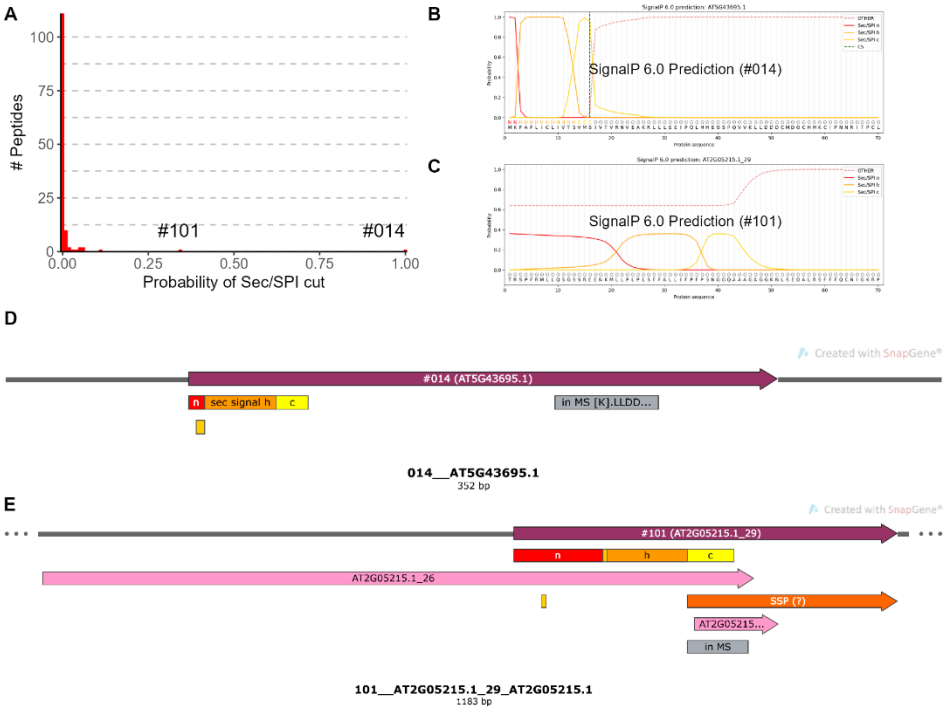


Figure 4.13. Identification of putative secretory signals in candidate peptides.

A) Probability of containing a Sec/SPI secretory signal. SignalP 6.0 results for candidates #014 (**B**) and #101 (**C**). Schematic map of candidates with putative secretory signals #014 (**D**) and #101 (**E**). The dark purple arrow represents the CDS of each candidate in their corresponding RNA. In grey, the peptidic sequence identified by mass spectrometry. In pink, other putative SEPs identified in the same RNA. The orange arrow in E represents the fragment that I identified as the possible SSP. The secretory signal is indicated in red (secretory signal n), orange (secretory signal h) and yellow (secretory signal c). Pink arrows represent the CDS of other peptides from the MS database that could be encoded by the same transcript as candidate #101.

4.2.5 Several SEPs belong to putative peptide families in A. thaliana

The custom database described above (*see* **4.2.1**) without the proteins and peptides annotated in Araport11 was searched against itself using BLASTp in order to identify peptide families and to determine whether the SEPs identified through LC-MS/MS formed part of them (*see* **Materials and Methods** section **4.4.5** for the strategy used to filter the results of the BLAST analysis; the strategy took into account the length of each peptide sequence query).

Among the ~100K peptides and proteins in the database, there were 14,366 families with two to eleven members (**Table 4.4, Sup Table 4.5**). Some families were exclusively formed by members encoded in the same transcript, including peptides from Hsu et al., CNTdb 2.0 and TAIR 'nc'. Around 55% of the families were comprised by peptides with the same origin (**Figure 4.14**). Besides, the 85% of the families were formed exclusively by putative peptides encoded in lncRNAs and TUFs (CNTdb 2.0, TAIR 'nc' or a combination of both sources). Out of the 1,874 hypothetical SEPs in the LC-MS/MS results, 515 were associated to at least one of these families, and there were 15 families with two members detected in the LC-MS/MS results (**Table 4.4, Sup Table 4.5**).

Table 4.4. Putative peptide families in *A. thaliana*.
Number of putative peptide families with member(s) detected or not in the LC-MS/MS results. Families include two to eleven peptides encoded in one to eleven different transcripts. The members of families encoded in more than one transcript can be encoded in overlapping loci, or in completely separated loci (e.g., in different chromosomes).

	Families encoded in more than one transcript	*Families encoded in a single transcript*
0 peptides in LC-MS/MS	12,248	1,356
1 peptide in LC-MS/MS	664	83
2 peptides in LC-MS/MS	13	2

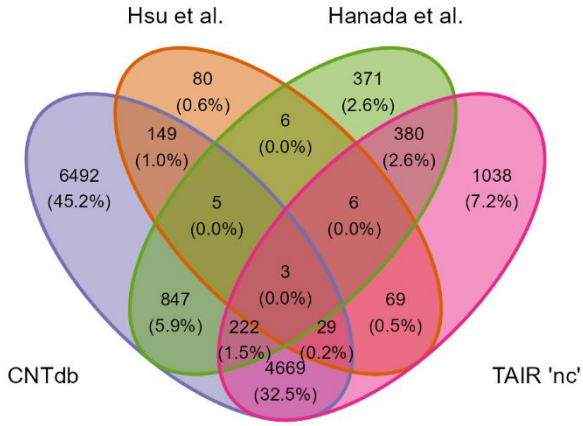


Figure 4.14. Peptides grouped in families by BLASTp have multiple origins.
Venn Diagram representing the origin (Hanada et al., Hsu et al., CANTATA db2.0 or TAIR 'nc') of the members of each family. The number (and percentage) of families formed by members of each origin is indicated.

In the case of the 132 selected candidates, I searched for homologous sequences along the complete *A. thaliana* genome with a BLASTn analysis, to identify putative families at genome level that were not detected at peptide level using the custom database (e.g., peptides annotated in Araport11). In this case, there were 68 candidates with homologous sequences within the Arabidopsis genome (**Sup Tables 4.3, 4.6**).

4.2.6 Amino acid sequences of SEPs are conserved across species

When inspecting the conservation of the 132 selected candidate peptides, putative homologs were found for 103 of them in the genome of at least one of other twelve plant species, namely *A. lyrata, Brassica oleracea, Camelina sativa, Vitis vinifera, Citrus clementina, Cucumis melo, Glyine max, Medicago truncatula, Populus trichocarpa, Solanum lycopersicum, Oryza sativa* and *Zea mays* (**Table 4.5, Sup Table 4.7**). The putative homologs of the peptide candidates were evenly distributed in the genome of the different analysed species, as it was already shown for the candidates in *A. thaliana* (**Figure 4.10D**). For some species, the putative homologs for the peptide candidates were larger than 100 aa (up to 200 aa) (**Figure 4.15**). I also explored if any

transcripts and/or peptide sequences for the putative homologs were already listed in the transcriptomes and proteomes of the twelve species used in the homology study (**Table 4.5, Sup Tables 4.3, 4.7**). Whereas the identified homologs for a majority of the selected candidates (76 out of 132) were localized in annotated transcripts in at least one of the corresponding species, many others (56) were identified only from the corresponding genome sequence. This was expected given that transcriptome depth (and in particular identification of lncRNAs) and quality of the genome annotation varies greatly among species (**Figure 4.16**).

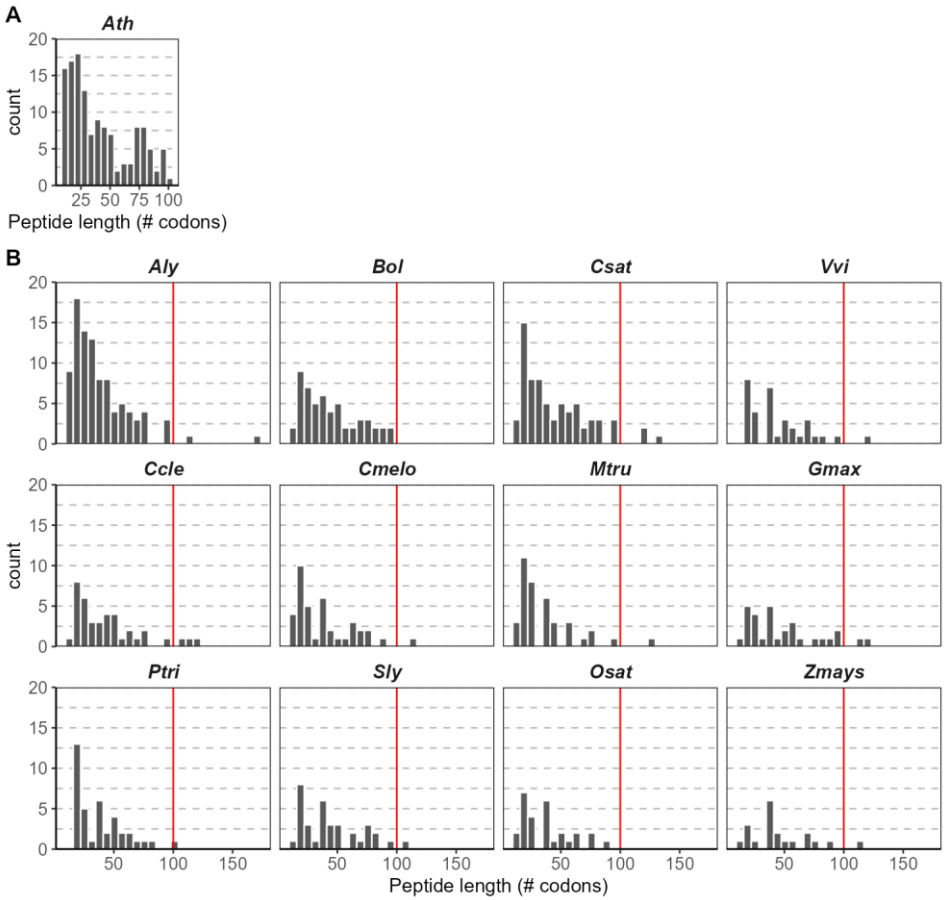


Figure 4.15. Length distribution of the 132 selected candidates and their putative homologs.
Histogram representing the length distribution of the selected candidates (**A**) and their putative homologs (**B**). Red line: 100 aa of length.

Table 4.5. Species for the homology analysis.
Twelve species selected depending on their evolutionary distance to *A. thaliana* and their available genome information, indicating the number of candidates with putative homologs in each species, and the number of possible transcripts and peptides for those putative homologs in the different species that are already annotated in their corresponding databases.

Species	*Monocot / dicot*	*Family*	*Reference genome*	*# Homologs (tBLASTn)*	*# Transcripts (BLASTn)*	*# Peptides (BLASTp)*
A. lyrata	Dicot	Malvids	*A. lyrata subsp. lyrata (v.1.0)*	95	63	27
B. oleracea	Dicot	Malvids	*B. oleracea var. oleracea (BOL)*	54	44	19
C. sativa	Dicot	Malvids	*C. sativa (Cs)*	70	56	25
V. vinifera	Dicot	Rosids	*V. vinifera (12X)*	33	31	18
C. clementina	Dicot	Malvids	*C. clementina (Citrus_clementina_v1.0)*	39	31	18
C. melo	Dicot	Cucurbitales	*Melon_v.4*	39	30	18
G. max	Dicot	Fabids	*G. max (assembly Glycine_max_v2.1)*	30	28	18
M. truncatula	Dicot	Fabids	*M. truncatula (MtrunA17r5.0-ANR)*	39	33	19
P. trichocarpa	Dicot	Fabids	*P. trichocarpa (assembly Pop_tri_v3)*	39	29	15
S. lycopersicum	Dicot	Asterids	*S. lycopersicum (SL3.0)*	35	30	19
O. sativa	Monocot	Poales	*O. sativa Japonica Group (IRGSP-1.0)*	28	21	11
Z. mays	Monocot	Poales	*Z. mays (Zm-B73-REFERENCE-NAM-5.0)*	20	17 + 3 ncRNAs	10

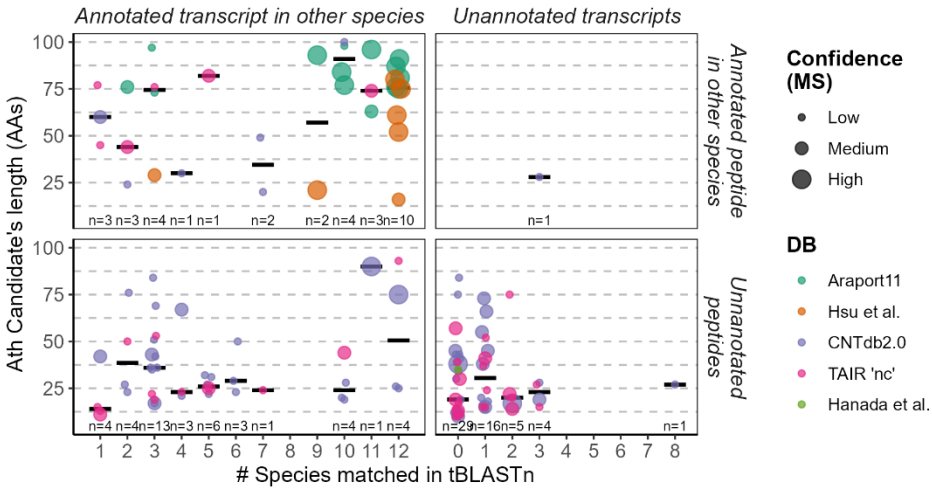


Figure 4.16. Distribution of the selected candidate peptides according to their length and whether their homologs are identified in the transcriptome or proteome of the corresponding species.

Diagrams reflect the length of the candidate peptide (AAs) and the number species in which homologs were identified. Coloured depending on the database of origin of each candidate. Point size depicts the confidence (FDR) of detection of the candidate in the LC-MS/MS results. Black lines represent the median values for each group. A candidate is described as annotated transcript/peptide in other species if at least one of its assigned putative homologs is present in the transcriptome and/or proteome of its corresponding species.

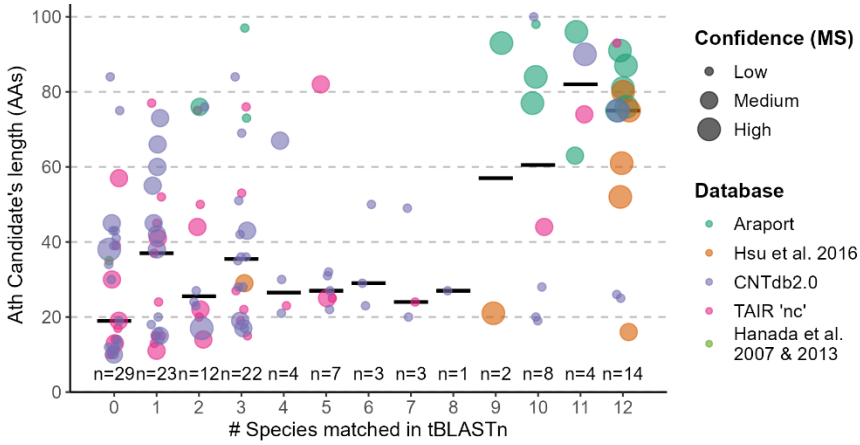


Figure 4.17. Distribution of the selected candidate peptides according to their length and the number of species in which they may have a putative homolog.

Coloured depending on the origin of each candidate. Point size depicts the confidence of detection of the candidate in the LC-MS/MS results. Black lines represent the median values for each group; n indicates the number of peptide candidates per group.

A substantial number of the selected candidates (29) seemed to be specific to *A. thaliana*; eighteen were also identified in *A. lyrata;* and fourteen also in *B. oleracea* and *C. sativa*, for a total of 61 candidates that were apparently specific to the Brassicaceae. There were also fourteen candidates with possible homologs in the twelve species. In contrast, the number of candidates present in four to eight species was smaller (**Figure 4.17, Sup Tables 4.3, 4.7**).

When non-synonymous substitutions (d_N) and synonymous substitutions (d_S) were compared, the resulting d_N/d_S ratios (ω) were very variable among different candidates and their putative homologs. *A. lyrata* followed by *C. sativa* and *B. oleracea* were the species with higher rates of positive selection for the putative homologs (ω > 1) (**Figure 4.18A**). For those candidates with putative homologs in at least three species, the frequency of site categories (negative [pi(-)], neutral [pi(N)], positive [pi(+)]) was calculated for each alignment. Almost 40% of the candidates presented pi(-) ≤ 60% of the total length of the candidate, and another 5% had pi(-) ≤ 60% of the aligned fraction of the candidate (**Figure 4.18B, Sup Tables 4.3, 4.7**). This conservation of the peptide candidates could be interpreted as an indicator of their translation and also of a possible common functionality in different plant species.

Fourteen of the candidates showed homologs in the twelve species analysed, whereas another twelve could be deemed as relatively conserved, as homologs were detected in ten or more species (i.e., 26 peptide candidates out of the set of 132). Of those fourteen, only one candidate with putative homologs in the twelve species had also a pi(-) ≤ 60% value (#002: AT1G47278.2). This candidate was detected with high confidence in LC-MS/MS and corresponds to the AT1G47278.2 locus, which is annotated in Araport11 as hypothetical and was also detected in the RNA-seq experiments described in **Chapter 2**.

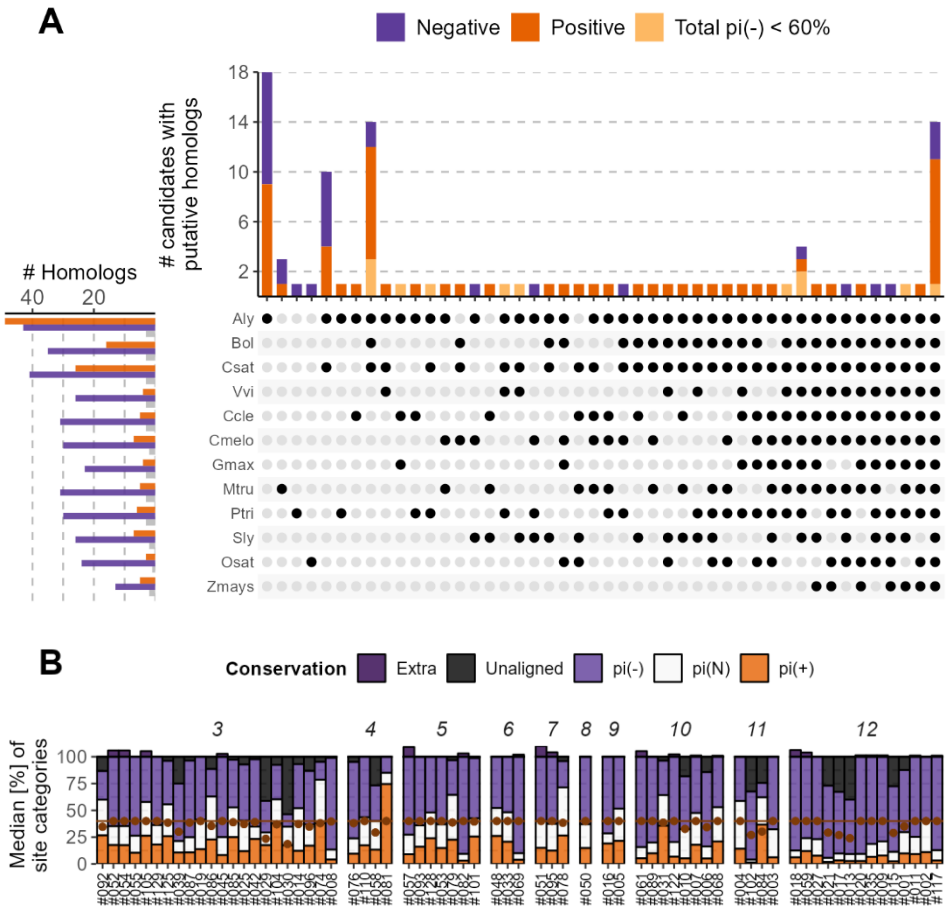


Figure 4.18. Sequence conservation of selected peptides.
A) Upset plot representing the number of candidates with each species combination of putative homologs. Coloured according to the general sequence selection of the homologs: all negative (ω<1) in purple, at least one homolog with positive (ω>1) in orange, and a total pi(-) of the sequence ≤ 60% in light orange. The bar graph on the left represents the number of putative homologs of each species with positive (ω>1, orange), negative (ω<1, purple) or indetermined (grey) selection in different species. **B)** Proportion of conserved and non-conserved positions for each candidate with putative homologs in at least three species. The candidates are grouped by their number of putative homologs (3-12) and sorted by size (10-100 from left to right in each group). The brown line signals the 40% threshold for a candidate to be considered as totally conserved and the brown dots the 40% threshold to consider the aligned part of the candidate as conserved. The length difference between the largest putative homolog compared to the candidate is displayed in dark purple ('extra' in the colour legend).

4.2.7 SEPs identified in floral buds show differential gene expression patterns across tissues

The RNA expression levels of the selected candidates in different tissues at various developmental stages were evaluated through quantitative real time PCR (qRT-PCR). RNA samples were obtained from inflorescences of the homeotic mutants and wild type Ler-0 plants, pAP1:AP1-GR *ap1 cal* inflorescences at various time-points after flower development induction (samples described in **Chapter 2**), young and mature siliques of wild type plants, and seedlings and rosette leaves of wild type plants (Ler-0).

Among the 53 candidates that were classified as tissue variant genes in this experiment (TVGs, ANOVA p-value ≤ 0.05), five different general expression patterns were observed, encompassing 46 of those candidates (**Figure 4.19**). As expected from the samples that were used for the LC-MS/MS experiment (floral tissues), the differential expression that was observed by qRT-PCR mostly consisted of flower-specific expression (44 out of the 46 candidates). In some cases, a certain enrichment of expression in the homeotic mutants associated with the presence of sepals (*ap3, pi*, and *ag*; 7 candidates) or of stamens (*ap1* and *ap2*; 9 candidates) was observed, and four candidates (#008, #010, #043 and #048) seemed to be enriched in mature flowers and siliques. In addition, two candidates (#020 and #122) showed higher expression levels in seedlings and leaves. Differential gene expression suggests that the corresponding selected candidates might participate or play a role in specific tissues or developmental processes or stages, suggesting a plausible role during fruit development.

Forty candidates had a flower-specific expression pattern; from which seven showed certain enrichment in the homeotic mutants associated with the presence of sepals (*ap3, pi* and *ag*), and nine presented an enrichment in those homeotic mutants associated with the presence of stamens (*ap1* and *ap2*). Besides, candidates #020 and #122 showed higher expression levels in seedlings and leaves, while candidates #008, #010, #043 and #048 seemed to be enriched in mature flowers and siliques, suggesting a plausible role during fruit development (**Figure 4.19**).

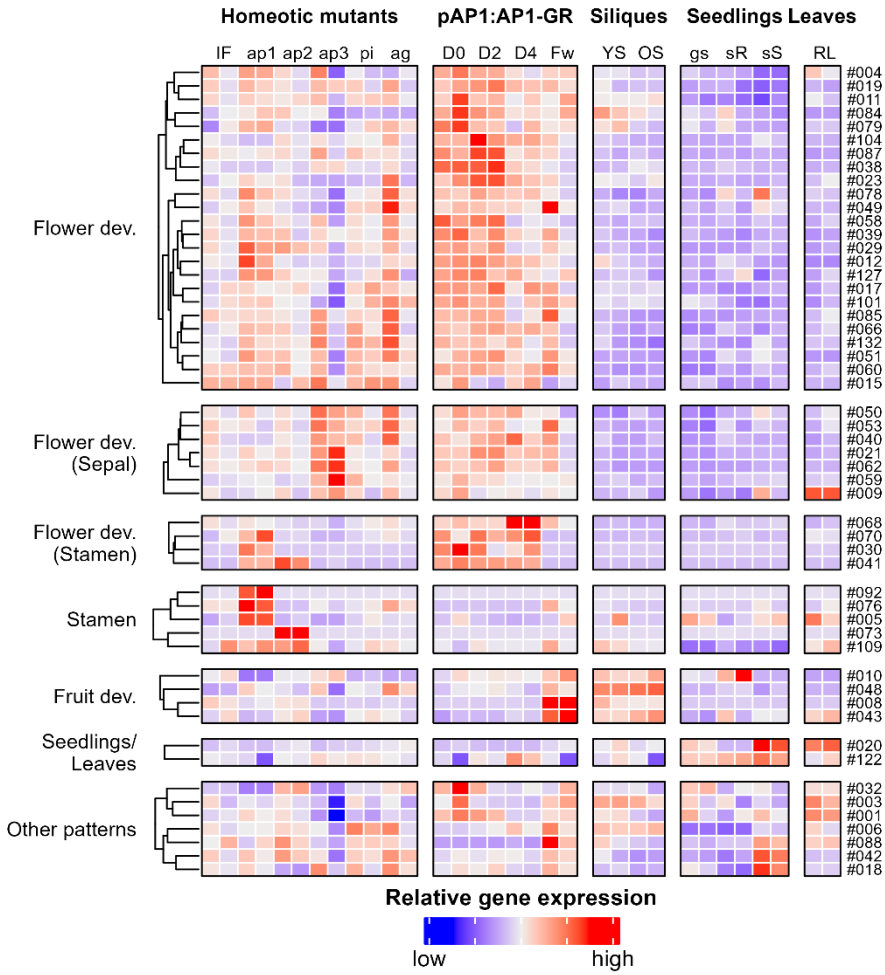


Figure 4.19. Gene expression of the candidates.
Heatmap depicting the results of the qPCR chip. Coloured by their z-scored relative gene expression (purple: low; orange: high). Samples: inflorescences of WT plants and the homeotic mutants (IF, *ap1, ap2, ap3, pi, ag*), inflorescences of pAP1:AP1-GR *ap1 cal* inflorescences 0, 2 and 4 days after flower development induction with dexamethasone (D0, D2, D4), WT mature flowers (Fw), young siliques (YS), old siliques (OS), rosette leaves (RL), germinated seeds (gs), seedling roots (sR) and seedling shoots (sS). Seven candidates were classified as differentially expressed by the ANOVA analysis but did not show a specific expression pattern ('Other patterns' in the figure).

The transcripts of 553 hypothetical peptides (out of the 1,874 detected by LC-MS/MS) were quantified at RNA level in the RNA-seq data described in **Chapter 2**. In addition, the transcripts of another 5,810 proteins and peptides were also detected at RNA level in the RNA-seq. In total, 2,310 stage variant genes (i.e., genes with expression changes in at least one stage; SVGs) were detected among the different time-points analysed by RNA-seq (moderated Likelihood Ratio Test – LRT – with an adjusted p-value ≤ 0.01, **Figure 4.20A**). These corresponded to 2,222 canonical proteins and peptides, 39 hypothetical proteins and 49 hypothetical peptides, of which ten were discovery candidates (**Figure 4.20B, C**).

The genes defined as SVGs showed three different transcript accumulation patterns for both the total dataset of 2,310 proteins and peptides (**Figure 4.20A**) and the 49 hypothetical peptides (**Figure 4.20B**): i) higher expression at later time-points, ii) increment in expression during mid-term time-points (D1-D4), and iii) higher expression at the initial time-points. In the case of the ten discovery candidates classified as SVGs, there were eight that could be included in the described generic patterns: i) three with higher expression at later time-points (#012, #016, and #068), ii) two showing higher levels at intermediate time-points (#046, and #087), and iii) three with higher levels at the initial time-points (#020, #029, and #039). However, candidates #007 and #109 described a U-like pattern, being down-regulated during mid-term time-points and with high expression levels at D0 and D5, as it was described in **Chapter 2** (section 2.2.4) (**Figure 4.20C**).

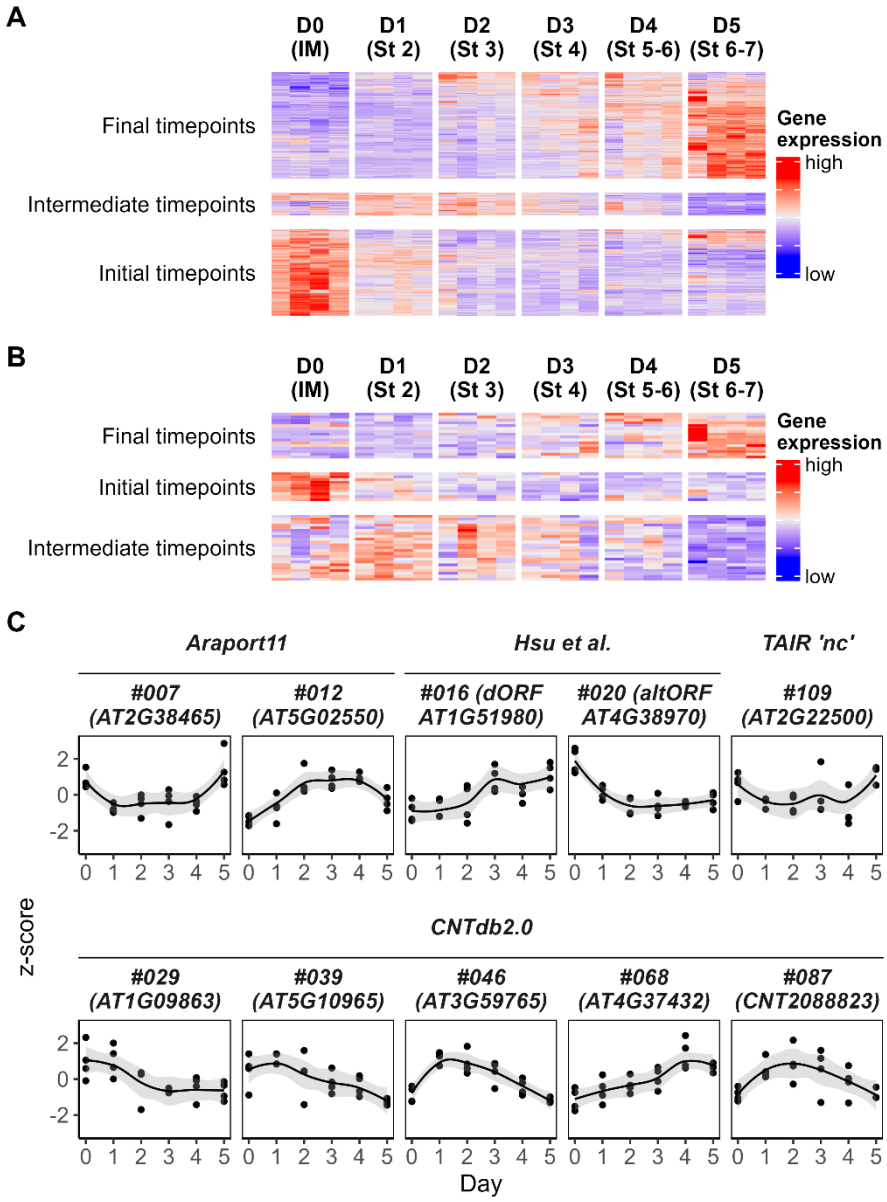


Figure 4.20. RNA expression during early flower development in *A. thaliana*.

A-B) Z-score representation of the 2,310 SVGs (LRT adjusted p-value ≤ 0.01) corresponding to all peptides and proteins (**A**) or to the 49 hypothetical peptides (**B**) detected in the homeotic mutants at protein level and at RNA level in pAP1:AP1-GR *ap1 cal* plants (samples of inflorescence meristem after DEX induction) (*see* **Chapter 2**). **C)** Z-scored RNA abundance of the 10 SVG candidates through time.

4.2.8 Expression patterns for selected candidates determined by reporter gene fusions

The initial set of 132 peptide candidates that were selected through the 'genotype-independent' ('discovery') and 'organ-specific' selection pipelines was further narrowed down to a set of 37, a more manageable number for subsequent molecular genetic studies. This additional selection step was based on all the available data for each peptide, including the analyses described above on putative translation initiation sites, homolog identification, sequence conservation, gene families, expression patterns, etc. (*see* **Materials and Methods** section **4.4.7**). Green fluorescent protein – β-glucuronidase (GFP-GUS) translational fusion constructs were generated using the putative protomer regions (1.5 kb upstream from the putative peptide translation initiation site) (pXXX:GFP-GUS constructs). Independent transgenic reporter lines were obtained for 20 of the selected candidate constructs.

Histochemical GUS assays were performed with three independent lines for each of the 20 candidate reporter constructs (**Figure 4.21**). Consistent with the fact that the candidate peptides were identified from developing inflorescences, the transgenic pXXX:GFP-GUS lines showed GUS staining in floral tissues. Ten of the 20 candidate reporter constructs were classified as possible stamen-specific peptides, and another two as possible petal-specific. In the GUS assays, the most frequent pattern was staining in developing stamens and anthers: staining in anthers during stamen formation at early stages of development (p004, p005, p010, p019, p022, p025, p043, p050, p061, p062, p077, p121, p128, and p131); in anthers up to more advanced stages (p008, p011); or during the complete anther development process (p101). In addition, p018:GFP-GUS plants were stained in anthers and at the tip of the sepals, and p087:GFP-GUS plants showed GUS staining in mature anthers, pedicels, and carpels. Only one of the twenty reporter gene fusions (p032) failed to show staining in floral tissues.

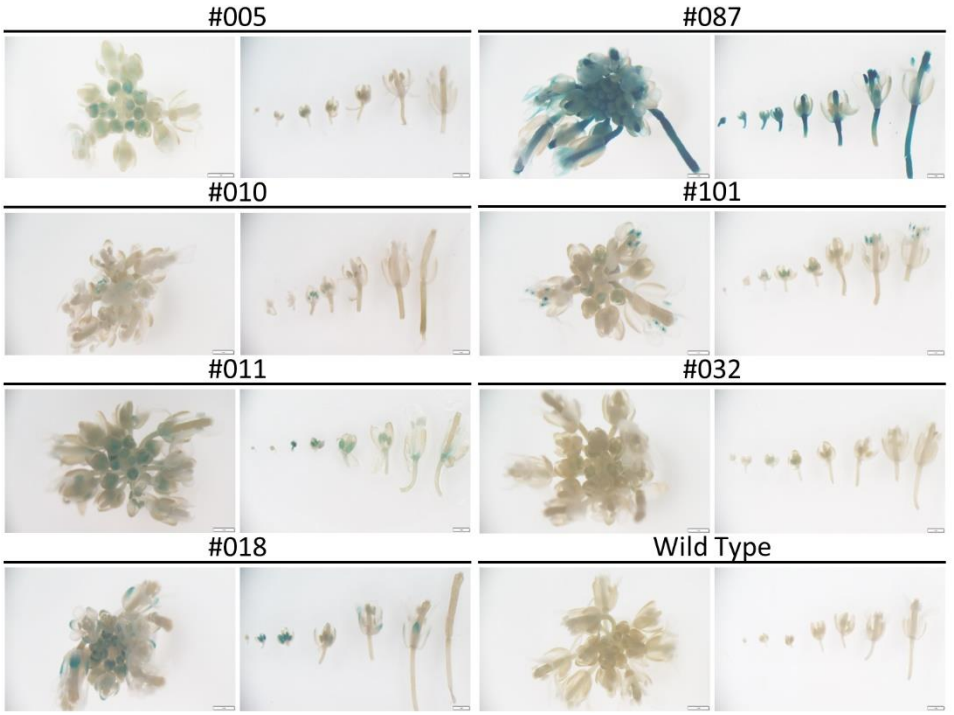


Figure 4.21. Examples of GUS staining patterns of pXXX:GFP-GUS lines in floral tissues.

4.2.9 Future perspectives: characterization of knock-out lines of selected SEPs

Whereas the results from the various analyses and assays described above raise the strong possibility that at least some of the identified novel peptides play a role in flower development or physiology, demonstrating and elucidating those roles might require genetic loss-of-function and gain-of-function approaches. Thus, a total of 21 potential SEPs were selected for generating *knock-out* lines using CRISPR-Cas9 technology (**Table 4.6**); thirteen candidates from these LC-MS/MS experiments and that showed different expression patterns according to the results from the GUS staining assays (above), and eight additional potential candidates that were identified in previous analyses by the research group (based on computational predictions, transcriptomics, and 5'- and 3'-RACE) and that also showed defined and particular expression patterns in reporter gene fusion experiments (**Figure 4.22**).

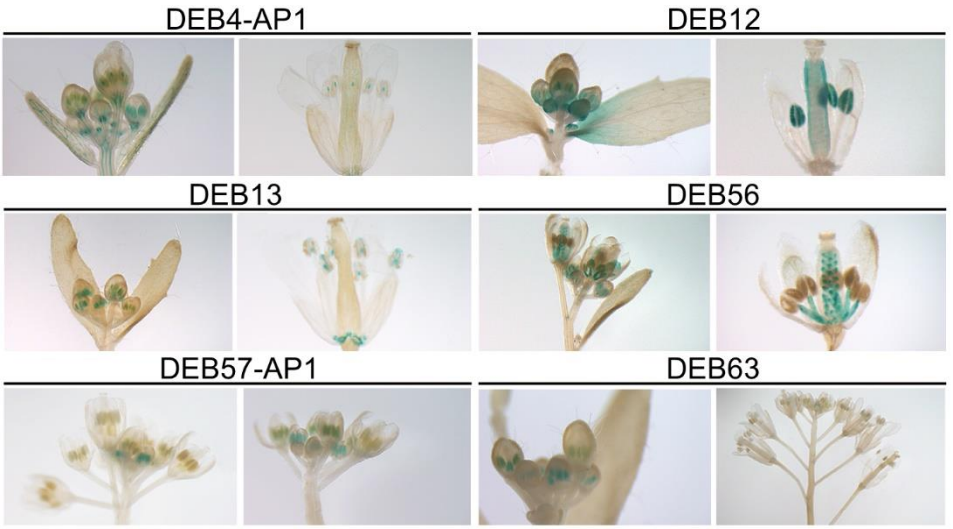


Figure 4.22. GUS staining patterns of pXXX:GFP-GUS lines of putative SEPs identified by computational predictions, transcriptomics and 3'- and 5'-RACE. Data from Dr Thilia Ferrier.

I managed to generate the plasmids carrying two guide RNAs per candidate and the Cas9 cassette for the 21 SEPs (**Figure 4.23**), and transformed wild type plants (Columbia ecotype, Col-0) by the floral dip method.

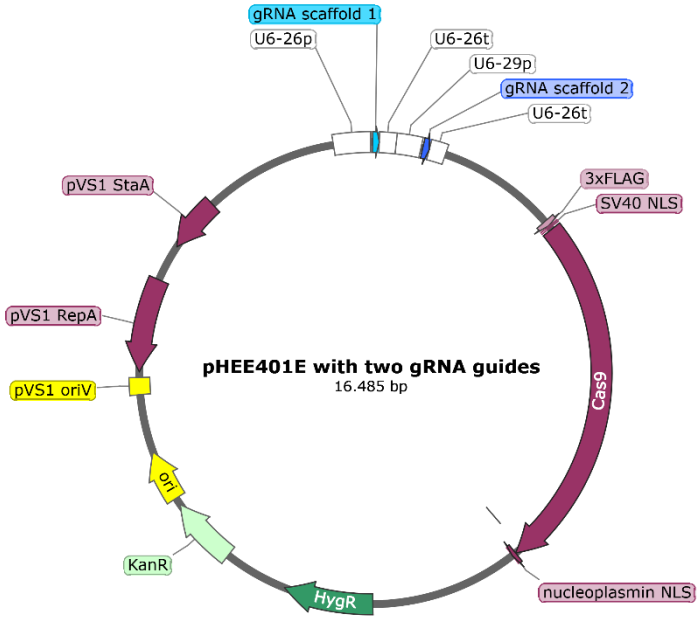


Figure 4.23. Map of the construction used to generate *knock-out* mutant lines.

The different *knock-out* lines are currently being generated. Once the T3 homozygous lines carrying the mutation while lacking the Cas9 gene are obtained, they will be evaluated by inspecting flowering or developmental traits. Additional functional studies could be complemented with the generation of overexpression lines for the different candidates.

Table 4.6. Data about the peptide candidates that were selected for the generation of loss-of-function mutant lines.

ID	Candidate	ORF type	Source	Putative TIS	Consecutive putative TIS	Possible SSP	Length (aa)	GUS staining
#004	AT2G20480.1	CDS	Araport11	AUG	ACG-AUG-AAG		63	Developing anthers
#005	AT2G21195.1	CDS	Araport11	AUG			93	Developing anthers
#008	AT2G46360.1	CDS	Araport11	AUG			97	Developing anthers
#010	AT4G23885.1	CDS	Araport11	AUG	AAG-AUG	LC-MS/MS	77	Developing anthers
#011	AT4G35980.1	CDS	Araport11	AUG			87	Developing anthers
#022	AT3+0 15349459-15349566	intergenic	Hanada et al.	Novel AUG	AUG-AAG		35	Developing anthers
#043	CNT2086293_4	lncRNA-ORF	CNTdb2.0	Near-cognate (AUC)			43	Developing anthers
#050	CNT2086628_21 (AT1G05833)	antisense lncRNA-ORF	CNTdb2.0	Near-cognate (AAG)			27	Developing anthers
#062	CNT2087373_8 (AT2G08335)	antisense lncRNA-ORF	CNTdb2.0	Near-cognate (AAG)			23	Developing anthers
#077	CNT2088303_1	lncRNA-ORF	CNTdb2.0	Novel AUG			16	Developing anthers
#087	CNT2088823_20	lncRNA-ORF	CNTdb2.0	Novel AUG			28	Anthers, pedicels, carpels
#101	AT2G05215.1_29	lncRNA-ORF	TAIR 'nc'	Novel AUG		SignalP6.0, LC-MS/MS	82	Developing anthers
#121	AT4G05205.1_14	lncRNA-ORF	TAIR 'nc'	Near-cognate (GUG)			39	Developing anthers

Table 4.6. Cont. Consecutive putative TIS: candidates that had more than one TIS in a sequential arrangement. Possible SSP: putative small-secreted peptides according to LC-MS/MS spectra or SignalP6.0 online tool. # Ath homologs: Number of homologs in the *A. thaliana* genome. # Species: Number of species with at least one putative homolog for the candidate.

ID	Mascot Confidence	Candidate type	Assigned organ	Family in Ath	# Ath homologs	# Species (homologs)	RNA-seq	qPCR
#004	Medium	Both	Stamen	FALSE	0	11	NVG	Flower
#005	High	Discovery	-	FALSE	0	9	NVG	Stamen
#008	Low	Discovery	-	FALSE	0	3	-	Flower → Silique
#010	High	Discovery	-	TRUE	1	10	NVG	Flower → Silique
#011	High	Discovery	-	FALSE	0	12	NVG	Flower
#022	Low	Discovery	-	FALSE	4	0	-	-
#043	Low	Discovery	-	TRUE	2	0	NVG	Flower → Silique
#050	Low	Discovery	-	FALSE	2	8	-	Carpel, Sepal
#062	Low	Floral	Petal	FALSE	1	2	-	Sepal
#077	Low	Floral	Stamen	FALSE		1	-	-
#087	Low	Both	Stamen	FALSE	0	3	SVG	Sepal
#101	Medium	Discovery	-	TRUE	1	5	NVG	Stamen, Petal
#121	Low	Floral	Stamen	TRUE	0	0	-	-

Table 4.6. Cont. Consecutive putative TIS: candidates that had more than one TIS in a sequential arrangement. Possible SSP: putative small-secreted peptides according to LC-MS/MS spectra or SignalP6.0 online tool.

ID	Candidate	ORF type	Source	Putative TIS	Consecutive putative TIS	Possible SSP	Length (aa)	GUS staining
DEA15	AT3G19274	novel CDS	*In silico*	Novel AUG			50	-
DEB4	AT1G31319	novel CDS	*In silico*	Novel AUG	AUA-AUG		35	Vascular tissue
DEB12		intergenic	*In silico*	Novel AUG	AUG-AUG-AUA-ACG		43	Stamens and carpels
DEB12 Alt	Alternative sORF of DEB12	intergenic	*In silico*	Novel AUG			66	Stamens and carpels
DEB13		intergenic	*In silico*	Novel AUG		SignalP 6.0	65	Anthers and sepal dehiscence junction
DEB56		intergenic	*In silico*	Novel AUG			60	Stamen filaments, ovules, seedling stipules
DEB57		intergenic	*In silico*	Novel AUG			55	Anthers
DEB63	AT5G66607	*novel CDS*	*In silico*	Novel AUG	AUG-AUA		37	Developing anthers

4.3 Discussion

The first large-scale experimental evidence of non-canonical translation in eukaryotic cells was provided by ribosome profiling. Thanks to this technique, thousands of sequences annotated as non-coding RNAs, pseudogenes or UTRs have been redefined as an important source of novel peptides in plant species such as Arabidopsis (BAZIN ET AL., 2017; HSU ET AL., 2016; KURIHARA ET AL., 2020; LIANG ET AL., 2021), maize (LIANG ET AL., 2021), tomato (H. Y. L. WU ET AL., 2019) or wheat (Y. GUO ET AL., 2023). Despite these advances, the evaluation of the coding potential of the sequences identified through ribosome profiling is computation-wise (MAKAREWICH & OLSON, 2017), meaning that sORF translation may not result in the production of a stable and functional SEP. To solve this issue, it is also possible to use MS-based methods to confirm the protein-coding nature of a sORF.

In the last years, several efforts have been conducted for the characterization of the Arabidopsis proteome and peptidome using MS-based methods to identify novel sORFs and SEPs (e.g., (MERGNER ET AL., 2020; S. WANG ET AL., 2020)). In this work, almost 2,000 unannotated peptides were identified thanks to the consideration and application of several key aspects for peptide identification. To begin with, the selection of an adequate peptide extraction method is crucial. The processing of cellular proteins increases the complexity of the peptidome, deteriorating the signal-to-noise ratio. Besides, protein and peptide separation strategies are important for the identification and quantification of low-abundance peptides and for increasing their overall sequence coverage (BARASHKOVA & ROGOZHIN, 2020; KULJANIN ET AL., 2017). In this Thesis, two extraction methods were compared, namely ultrafiltration and ammonium sulphate precipitation, while also testing different kinds of filtration devices for separating peptides of various sizes. Finally, the best method resulted in the 30K-ultrafiltration. However, it is important to design and test several methods that adapt to each special need.

In a typical MS/MS data analysis, more than 60% of the spectra remains unassigned, even after database improvements for guided identification and the use of *de novo* identification algorithms. Although some of these could be attributed to the low quality of the unassigned spectra, a portion can still be

classified as high-quality (PATHAN ET AL., 2017). Good quality spectra could be further analysed and reinspected for modifications, mutations, and sequence variants using peptidogenomics for new database generation (S. WANG ET AL., 2020), or by mass-tolerant database searches for PTM considerations (CHICK ET AL., 2015). To try to overcome this issue in the experiments sorted here, the initial search database was expanded during the process of spectra identification by considering noncanonical peptides derived from lncRNAs, and in addition by performing a more comprehensive search using a Semi-tryptic approach (*see* **Materials and Methods** section **4.4.3 Reference database**).

The database extension considered that lncRNAs can be an important source of SEPs as they encode translatable sORFs, as shown by Ribo-Seq and mass spectrometry experiments (HSU ET AL., 2016; H. Y. L. WU ET AL., 2019). For instance, ribosome profiling of the human heart resulted in the identification of 1,577 noncanonical ORFs, of which 339 (22%) were sORFs from lncRNAs. Furthermore, over 40% of those lncRNA SEPs were confirmed by mass spectrometry (VAN HEESCH ET AL., 2019). Although the exact proportion of sORFs/SEPs that are derived from lncRNAs may vary among different studies, it consistently represents a substantial fraction, frequently around 25% (J. CHEN ET AL., 2020; OUSPENSKAIA ET AL., 2021). In this context, it is also noteworthy that out of a list of 42 human SEPs that have already been already characterized as functionally or physiologically relevant, 55% are derived from lncRNAs (WRIGHT ET AL., 2022). In plants, 153 lncRNA-encoded peptides have been identified by LC-MS/MS in soybean (X. LIN ET AL., 2020). As a consequence, the research community is beginning to contemplate the revision of the classification criteria of lncRNAs, due to the presence of translatable sequences shorter than 100 aa despite their definition of "non-coding" (X. LIN ET AL., 2020; PALOS ET AL., 2022). My results are also an example of the overlooked coding potential of lncRNAs, intergenic regions and other transcripts. In this study, almost 60% of the identified peptides derived from lncRNAs annotated in CNTdb 2.0, and another 26% from other transcripts classified as "non-coding" in TAIR. Nevertheless, the number of unassigned spectra did not decrease significantly, leaving the door open for further analyses using an even wider database or other identification approaches.

Most of the hypothetical peptides identified in the LC-MS/MS analysis were detected in only one out of the 24 total samples (**Figures 4.7B, 4.24A**). This paucity of SEP detection was also observed in a study in which samples of 24 different tissues of pear were analysed with the objective of creating a proteomics atlas, and that also resulted in identifying 608 novel peptides (P. WANG ET AL., 2023) (**Figure 4.24B**). Although the number of novel peptides identified by P. Wang et al. was lower than that obtained in this Thesis (the pear samples were not processed to enrich for peptides and small proteins, and a specific data analysis pipeline for SEP detection was not developed), it is noteworthy that more than 75% of the peptides were identified exclusively in one or two of the samples. This points to the influence of chance in the detection of peptides that are found in smaller amounts in the samples, and for that reason also in problems in mass spectrometry when it comes to reproducibility of replicates. In the pear study, 124 peptides were identified only in floral organs (**Figure 4.24C**). Among them, the number of peptides identified exclusively in stamens is the highest, as was the case in the results reported in this chapter (**Figures 4.8, 4.24D**). Thus, it appears that the general conclusions that are or might be obtained from the study of the non-conventional peptidome in Arabidopsis will extend to other flowering plants.

It is worth to note the existence of a high number of peptides with non-AUG translation initiation sites (TIS). In spite of the "traditional" feature for predicting protein-coding ORFs through the presence of an ATG start codon, it is now apparent that non-AUG translation initiation cases might be abundant, and that sORFs show a trend towards an increased use of near-cognate or alternative start codons relative to canonical ORFs (CAO & SLAVOFF, 2020). There are several MS-based (e.g., (MA ET AL., 2014; SLAVOFF ET AL., 2013; VANDERPERRE ET AL., 2013; S. WANG ET AL., 2020)) and Ribo-Seq (e.g., (J. CHEN ET AL., 2020; Y. GUO ET AL., 2023; INGOLIA ET AL., 2011; LI & LIU, 2020; MARTÍNEZ ET AL., 2020)) studies that also indicated that 35-50% of the identified sORFs use non-AUG start codons. According to my results, around 45% of the total novel peptides identified by LC-MS/MS had near-cognate codons as TIS, while another 29% corresponded to peptides beginning with other codons that also differ from AUG. It remains unclear whether the amino acid corresponding to the non-AUG start codon is incorporated at the TIS or

methionine is still incorporated. According to (NA ET AL., 2018) methionine seems to be incorporated at almost all non-canonical TISs identified by LC-MS/MS.

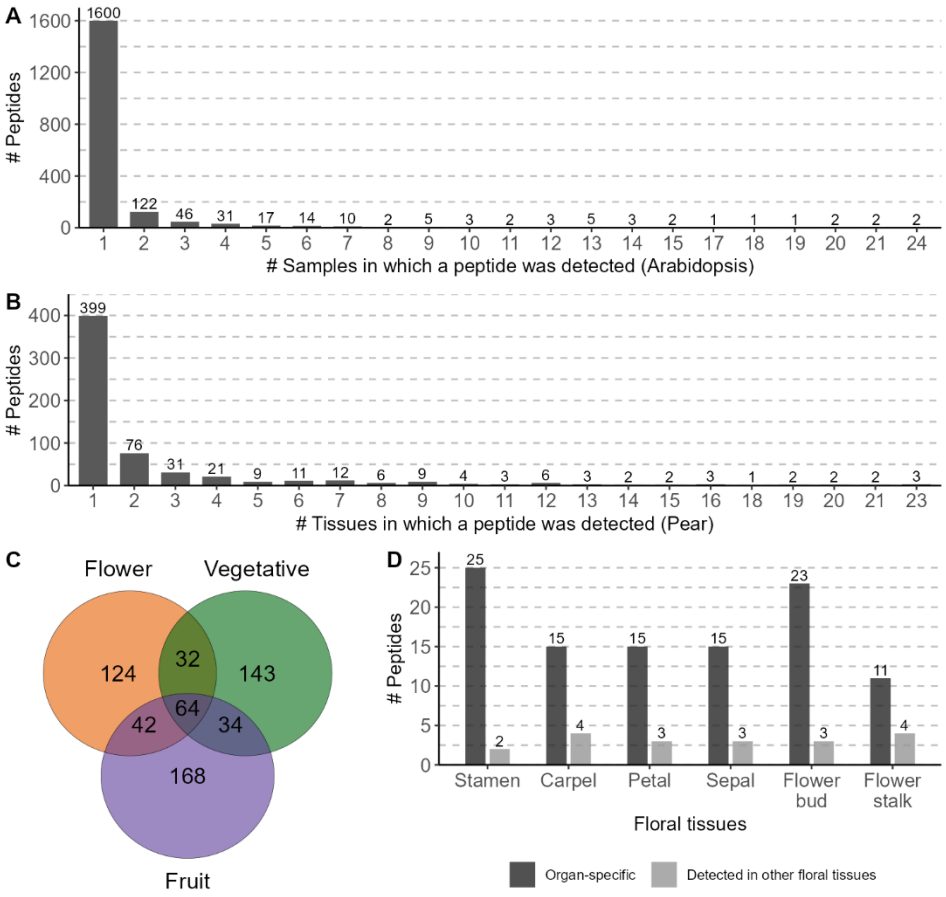


Figure 4.24. Comparison of the peptidomics results with a proteogenomics study in pear.
A) Bar graph depicting the number of peptides identified in 1 to 24 of the samples (4 biological replicates x 6 genotypes) in the LC-MS/MS of inflorescences in Arabidopsis. **B)** Bar graph representing the number of peptides identified in 1 to 24 of the samples (1 replicate x 24 tissues) in a LC-MS/MS analysis in pear. **C)** Venn diagram indicating the number of peptides identified in pear in floral tissues (flower), vegetative tissues (vegetative) and fruit tissues (fruit). **D)** Number of peptides identified in exclusively in floral tissues in pear. Dark grey: organ-specific peptides (identified exclusively in one sample). Light grey: peptides detected in more than one floral tissue. Data of panels B-D extracted from (P. WANG ET AL., 2023).

Despite the variation in the absolute frequencies of AUG and non-AUG initiation codons, there is an increasing trend of near-cognate start codons in the novel peptidome relative to main ORFs and annotated ORFs. In the case of putative TISs with non-AUG and non-near cognate start codons, there are at least three possibilities to consider: i) the identification of the peptide was incorrect (e.g., a false positives), ii) the identification of the peptide was correct, but it was not possible to elucidate the real TIS (e.g., there is a splicing region or an intron that was not identified), and iii) the identification of the peptide was correct and it truly starts with a codon that differs from AUG and near-cognate codons. This third case would be much less frequent than AUG- or near-cognate codon-initiated translation, but there are studies that have demonstrated the existence of non-AUG and non-near-cognate translation initiation events for SEPs (e.g., (CAO & SLAVOFF, 2020; NA ET AL., 2018)).

Amino acid sequences of certain SEPs are conserved across species, but, in general, sORFs/SEPs are less evolutionary conserved than standard proteins (e.g., (J. CHEN ET AL., 2020; FESENKO ET AL., 2019, 2021; RUIZ-ORERA ET AL., 2018; VAN HEESCH ET AL., 2019; WRIGHT ET AL., 2022)). Their lower conservation scores are also in agreement with the concept of lncRNA-derived sORFs facilitating *de novo* gene evolution. Among the peptide candidates from *A. thaliana* identified in this Thesis, 103 peptides seemed to have putative peptide homologs in other plant species, and around the 40% of those peptides had a positive or neutral selection of a good fraction of their amino acidic sequence, which might be related to their function.

Whereas SEPs that show evidence of conservation across multiple and distant species are (more) likely to have specific biological functions, it is also apparent that limited conservation does not exclude the production of functional peptides (LAURESSERGUES ET AL., 2022; VAN HEESCH ET AL., 2019; YEASMIN ET AL., 2018). Some plant peptides identified through classic and molecular genetic approaches are known to play significant roles in various processes (development, stress, signalling, etc.), however, the plant peptidome is largely undefined and experimental evidence for the biological functionality of the vast majority of the predicted or identified SEPs is still lacking (HSU & BENFEY, 2018; TAVORMINA ET AL., 2015).

For the set of identified floral-specific peptides, I hope to find specific phenotypes for the loss-of-function mutants being currently generated. Alternatively, we will also generate overexpression lines for some of the candidates to characterize them at functional level.

4.4 Materials and Methods

4.4.1 Plant lines, growth conditions, and tissue collection

The mutant strains used in this study were *ap1-1/-*, *ap2-2/-*, *ap3-3/+*, *pi-1/+*, and *ag1-1/+* (BOWMAN ET AL., 1989, 1991, 1993; JACK ET AL., 1992). Plants of the accession *Landsberg erecta* (Ler-0) were used as wild type reference, except for the generation of *knock-out* lines that Columbia (Col-0) plants were used. In addition, pAP1:AP1 *ap1cal* D0, D2 and D4 inflorescence material (*see* **Chapter 2**) was used for the qRT-PCR analysis. Plants were grown, after a 1-week period of stratification at 4 °C in darkness, on a soil:vermiculite:perlite mixture at 21 °C under long day conditions (16h light, 8h darkness), or in plates of 0.5 x Murashige and Skoog (MS) salt mixture with vitamins, and 0.8% plant agar after being surface sterilized and stratified at 4 °C for 48 h. Plates were incubated vertically at 22 °C and 70% humidity under long day conditions.

Four biological replicates of around 144 plants each of 5-week-old inflorescence meristem and floral buds, corresponding to floral stages 1 to 13 (SMYTH ET AL., 1990), were collected for Ler-0 plants, and *ap1, ap2, ap3, pi*, and *ag* mutant lines, as done for the initial characterization of spatial gene expression in Arabidopsis flowers (WELLMER ET AL., 2004). These samples were used in the mass spectrometry experiments and for RNA extraction. Two biological replicates were collected for RNA extraction from other tissues: Ler-0 mature flower, young siliques and mature siliques (with seeds) from 5-week-old plants grown in soil (n = 144 plants per replicate), Ler-0 rosette leaves from 19-day-old plants grown in soil (n = 72 per replicate), Ler-0 germinating seeds (2-day-old plants grown in plates, 3 plates per replicate), and Ler-0 seedling roots and shoots (4-day-old plants grown in plates, 2 plates per replicate). Grinded samples of each tissue were preserved at -80 °C until their use.

4.4.2 Peptide extraction

To choose the optimal peptide extraction method, two different techniques were compared, ultrafiltration, and ammonium sulphate precipitation, both followed by reverse phase chromatography. The final extraction method was

chosen based on the results of both techniques for samples of Ler-0 inflorescence tissue and mature flowers. As the number of obtained and identified peaks by mass spectrometry was better for the ultrafiltration with 30K filters followed by reverse phase chromatography, this was the extraction method of choice (ÁLVAREZ-URDIOLA, BORRÀS, ET AL., 2023).

Ultrafiltration. For each sample, ~0.5 g of blended tissue distributed in two 2 mL microcentrifuge tubes were used. A total of 1.2 mL of extraction buffer (phosphate-buffered saline (PBS) 1x, urea 1.5M, DTT 10mM, acetonitrile 2% v/v, trifluoroacetic acid (TFA) 0.5%, MG-132 10µM, Proteinase Inhibitor cocktail cOmplete 1x, and PMSF 1mM) were added to the tissue, mixed by vortexing and incubated the samples with continuous shaking for 1 h at 4 °C. Samples were spined twice for 1 min at 4 °C (max speed) to precipitate cellular debris and solid particles in suspension. All the supernatant of each sample was filtered in 30K- or 100K- Amicon ® Ultra-0.5 Centrifugal Filter devices as indicated by the manufacturer (~500 µL of supernatant at a time were centrifuged at 14,000 x g for 10 min at 4 °C). Filtrates containing the smallest peptides depending on the weight limit of the filter device were kept for reverse phase chromatography.

Ammonium sulphate precipitation. For each sample, ~0.5 g of blended tissue distributed in two 2 mL microcentrifuge tubes were used. A total of 1.2 mL of extraction buffer (PBS 1x, urea 2M, acetonitrile (ACN) 2% v/v, DTT 10mM, acetonitrile 2% v/v, trifluoroethanol (TFE) 5%, Tris pH 7.6 50mM, MG-132 10µM, cOmplete 1x, and PMSF 1mM) were added to the tissue, mixed by vortexing and incubated the samples with continuous shaking for 30 min at 4°C. Samples were spined for 1 min at 4°C (max speed) to precipitate cellular debris and solid particles in suspension. Then, 75% of ammonium sulphate was added to the supernatant to precipitate the proteins in solution. Ammonium sulphate calculator from EnCor Biotechnology Inc. (http://www.encorbio.com/protocols/AM-SO4.htm) was used to calculate the needed amount of ammonium sulphate for each specific sample. The mixes were centrifuged at maximum speed for 25 min at 4 °C, and supernatants were kept for further reverse phase chromatography.

Reverse phase chromatography. Samples obtained with both previous methods were mixed with sample buffer (2% TFA in 20% ACN) in 3:1 sample:sample buffer proportion. Final samples contained 0.5% TFA in 5% ACN. C18 resin columns (Pierce, Thermo Scientific) were prepared as indicated by the manufacturer before loading the samples on top of the resin beds (150 µL at a time). Samples were centrifuged at 1500 x g for 1 min as many times as needed to pass all the sample volume through the resin. Peptides were eluted from the column by adding 21 µL of elution buffer on top (0.1% formic acid in 70% ACN) and centrifuge at 1500 x g for 1 min (this step was repeated twice to increment the final concentration of the samples). The concentration (270 – 830 µg/mL) and amount of total protein (15 – 40 µg) in each sample were quantified using a Qubit Protein Assay Kit.

4.4.3 Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS)

Peptidomics experiments were conducted in collaboration with Dr Eduard Sabidó and Dr Eva Borrás from the proteomics facility at the Center for Genomic Regulation (CRG).

Sample preparation. Samples (10 µg) were reduced with dithiothreitol (30 nmol, 37 °C, 60 min) and alkylated in the dark with iodoacetamide (60 nmol, 25 °C, 30 min). The resulting protein extract was first diluted to 2M urea with 200 mM ammonium bicarbonate for digestion with endoproteinase LysC (1:10 w:w, 37°C, o/n, Wako, cat # 129-02541), and then diluted 2-fold with 200 mM ammonium bicarbonate for trypsin digestion (1:10 w:w, 37°C, 8h, Promega cat # V5113). After digestion, peptide mix was acidified with formic acid and desalted with a MicroSpin C18 column (The Nest Group, Inc) prior to LC-MS/MS analysis.

Chromatographic and mass spectrometric analysis. Samples were analysed using a LTQ-Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EASY-nLC 1000 (Thermo Fisher Scientific (Proxeon), Odense, Denmark). Peptides were loaded onto the 2-cm Nano Trap column with an inner diameter of 100 µm packed with C18 particles of 5 µm particle size (Thermo Fisher Scientific) and were separated by reversed-phase chromatography using a 25-cm column with an inner

diameter of 75 µm, packed with 1.9 µm C18 particles (Nikkyo Technos Co., Ltd. Japan). Chromatographic gradients started at 93% buffer A and 7% buffer B with a flow rate of 250 nl/min for 5 minutes and gradually increased 65% buffer A and 35% buffer B in 60 min. After each analysis, the column was washed for 15 min with 10% buffer A and 90% buffer B. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in acetonitrile. The mass spectrometer was operated in positive ionization mode with nanospray voltage set at 2.1 kV and source temperature at 300°C. Ultramark 1621 for the was used for external calibration of the FT mass analyzer prior the analyses, and an internal calibration was performed using the background polysiloxane ion signal at m/z 445.1200.

The acquisition was performed in data-dependent acquisition (DDA) mode and full MS scans with 1 micro scans at resolution of 60,000 were used over a mass range of m/z 350-2000 with detection in the Orbitrap. Auto gain control (AGC) was set to 1E6, dynamic exclusion (60 seconds) and charge state filtering disqualifying singly charged peptides was activated. In each cycle of DDA analysis, following each survey scan, the top twenty most intense ions with multiple charged ions above a threshold ion count of 5000 were selected for fragmentation. Fragment ion spectra were produced via collision-induced dissociation (CID) at normalized collision energy of 35% and they were acquired in the ion trap mass analyzer. AGC was set to 1E4, isolation window of 2.0 m/z, an activation time of 10 ms and a maximum injection time of 100 ms were used. All data were acquired with Xcalibur software v2.2. Digested bovine serum albumin (New England Biolabs cat # P8108S) was analysed between each sample to avoid sample carryover and to assure stability of the instrument and QCloud (CHIVA ET AL., 2018) has been used to control instrument longitudinal performance during the project.

Data analysis. Acquired spectra were analysed using the Proteome Discoverer software suite (v2.3, Thermo Fisher Scientific) and the Mascot search engine (v2.6, Matrix Science) (PERKINS ET AL., 1999). The data were searched against two different databases (*see* **Reference database**), plus a list common contaminants (BEER ET AL., 2017) and all the corresponding decoy entries. For peptide identification a precursor ion mass tolerance of 7 ppm was used for MS1 level, trypsin was chosen as enzyme, and up to three

missed cleavages were allowed. The fragment ion mass tolerance was set to 0.5 Da for MS2 spectra.

Oxidation of methionine and N-terminal protein acetylation were used as variable modifications whereas carbamidomethylation on cysteines was set as a fixed modification. False discovery rate (FDR) in peptide identification was set to a maximum of 5%. Peptide quantification data were retrieved from the "Precursor ion area detector" node from Proteome Discoverer (v2.0) using 2 ppm mass tolerance for the peptide extracted ion current (XIC). Protein abundance in each condition was estimated using the average of the three most intense peptides per protein group (TOP3). The obtained values were used for subsequent statistical analysis. The raw peptidomics data will be deposited to PRIDE (PEREZ-RIVEROL ET AL., 2022). Median normalisation was performed by subtracting from each logged value the sample median and adding the global dataset median. Biological replicates from the different genotypes clustered together when Principal Component Analysis (PCA) was performed (**Figure 4.25A**). The selection of the candidates was performed using the presence / absence criteria as the expression levels did not provide enough information for the classification of the peptides and proteins (**Figure 4.25B**). The Floral Organ criteria were applied to individual peptidic fragments for each protein and peptide, and to the TOP3 results for each peptide and protein. For a peptide or protein to be considered as specific for an organ, it must be classified for that organ at both levels.

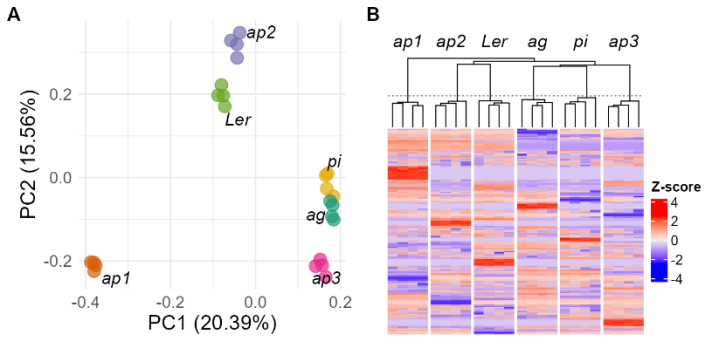


Figure 4.25. Clustering of peptides and proteins quantified in at least one genotype.

A) PCA of the LC-MS/MS results for the different genotypes. **B)** Heatmap of z-scored abundance values for peptides and proteins in the different homeotic mutants.

Reference database. The LC-MS/MS data were searched with a tryptic analysis against a database (DB1) containing 40,798 non-redundant Araport11 (CHENG ET AL., 2017) protein coding genes (downloaded October 2019, Araport11_genes.201606.pep.fasta), 1,684 short Open Reading Frames (sORFs) identified in root and shoot by RiboTaper (HSU ET AL., 2016), and 7,016 putative sORFs identified in intergenic regions (HANADA ET AL., 2007, 2013), plus a list of common contaminants (BEER ET AL., 2017) and all the corresponding decoy entries. For the final peptide identification, a second database (DB2) containing (i) those proteins (with more than 100 aa) and peptides (≤ 100 aa) that had been identified in the LC-MS/MS spectra using DB1 (6,124 proteins and peptides) plus (ii) all peptide sequences (of ≥10 aa) derived from a three-frame translation of lncRNAs (SZCZEŚNIAK ET AL., 2019) (CNTdb2.0) and TUFs (other RNA, lncRNA, antisense lncRNA, antisense RNA, novel transcribed region and uORF genes in Araport11) (TAIR 'nc'). Sequence redundancy at amino acid sequence level was removed by grouping into clusters each subset using CD-HIT (https://www.bioinformatics.org/cd-hit/). The priority order that was used to remove redundancy was: Araport11 > Hsu et al. > Hanada et al. > CNTdb 2.0 > TAIR 'nc'.

MS spectra were matched with the peptides in DB2 in a tryptic and semi-tryptic manner (DB2T, DB2ST). For the final peptide quantification, all the information obtained for the three analyses was kept. The most reliable quantifications were those from the first analysis (DB1T), then the information of those new detections in DB2T and the new peaks in DB2ST were added. For the final dataset, the origin of the information was annotated (DB1T, DB2T, DB2ST or a combination of more than one analysis). Finally, two lists of data were analysed: one with the average abundance of the top three aminoacidic sequences for each accession (ID of peptides or proteins, TOP3) and one with all the aminoacidic partial sequences quantified for each accession (sequences). To each dataset, median normalisation was performed by subtracting from each logged value the sample median and adding the global dataset median.

Comparison with previous data. BLASTp (NCBI, v2.11.0+) was used to compare the amino acid sequences of the peptides detected by LC-MS/MS in this Thesis against the amino acid sequences of novel SEPs detected in (S.

WANG ET AL., 2020). All the sORFs from the MS database were also compared to a list of novel genes described by (R. ZHANG ET AL., 2021), however, I did not find any match between both datasets.

Candidate selection and validation of selection criteria. A gene ontology (GO) enrichment analysis (G. YU ET AL., 2012) of the annotated proteins and peptides classified as organ-specific was performed to check whether the floral organ filter worked properly. Moreover, a correlation network was created using the LC-MS/MS expression levels of the proteins and peptides, followed by a new GO enrichment analysis of the different modules calculated using the Random Matrix Theory.

4.4.4 Re-annotation of Translation Initiation Sites (TIS)

The LC-MS/MS database for peptide identification was created without any limitations for TIS, that is, the CDS of the possible peptides did not have to start necessarily in an AUG codon. As it was possible that there were more suitable TIS for the detected peptides in their corresponding genomic sequences, a set of TIS selection criteria was established for the sequences based on the peptidic fragments detected in the MS analysis for each peptide: i) AUG was selected over near-cognate or other non-AUG codons as TIS, and near-cognate codons were selected over other non-AUG codons. ii) If the detected peptide fragment closer to the annotated TIS of the peptide had a tryptic beginning (i.e., it started with lysine or arginine), the annotated TIS was kept, unless there was a more suitable TIS (according to the first criterium) between the annotated TIS and the codon corresponding to the beginning of the detected fragment. In the latter case, the TIS was re-annotated with the more suitable codon. If there were more than one possibility, the more suitable TIS closer to the annotated start of the peptide was chosen. iii) If the detected peptide fragment closer to the annotated TIS of the peptide had a non-tryptic beginning (i.e., it started with any amino acid but lysine or arginine) and the previous codon was an AUG or a near-cognate codon, it was selected as the new putative TIS. Otherwise, the more suitable TIS closer to the previously annotated start codon was selected (according to the first criteria). iv) If the detected peptide fragment included the annotated TIS, the annotation remained unmodified.

To deepen in the analysis of the sequence of sORFs encoding for the peptide candidates, the online tool SignalP 6.0 was used to find putative secretory signals (https://services.healthtech.dtu.dk/service.php?SignalP).
Candidates were classified as putative precursors of small-secreted peptides (SSPs) using the information from SignalP 6.0 and also the LC-MS/MS data (a peptidic fragment without any tryptic end was found).

4.4.5 Conservation analyses

Analysis of related sequences within the sORF list. A customized database containing exclusively the amino acid sequences of non-redundant sORFs from Hsu et al., 2016, Hanada et al., 2007 and 2013, the CANTATAdb 2.0 and TAIR 'nc' RNA sequences was generated using the makeblastdb program included in the blast+ package (BLAST+, NCBI, v2.10.1+). The database was blasted (protein-protein BLASTp, NCBI, v2.10.1+) against itself. Top ten hits for each query were filtered depending on their bit-score (hit: bit score ≥ self-score*0.6).

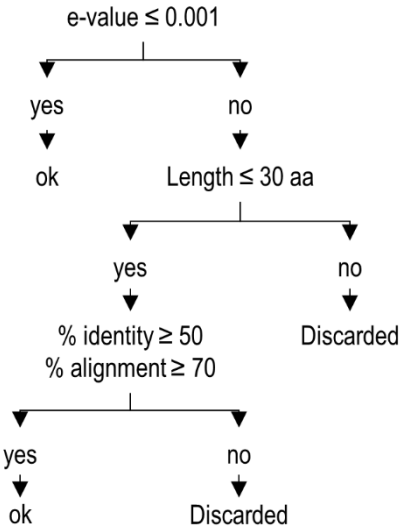


Figure 4.26. Decision tree to select putative homologs among the sequences obtained with BLAST (homology-threshold).

Homology analysis. The amino acidic sequences of the 132 candidates were searched by BLAST (tblastn, NCBI, v2.11.0+) against the genomes of *A. thaliana* (gene family search) and twelve different plant species separately (*A. lyrata, B. oleracea, C. sativa, V. vinifera, C. clementina, C. melo, G. max, M. truncatula, P. trichocarpa, S. lycopersicum, O. sativa* and *Z. mays*) (**Sup Tables 4.6, 4.7**). Sequences with an e-value ≤ 0.001 were classified as putative homologs, as well as sequences with length up to 30 amino acids with more than a 50% of identity and more than a 70% of alignment, independently of their e-value (**Figure 4.26**). Most of the putative homologs passed the e-value threshold (≤ 0.001), and had higher percentages of identity and alignment, independently of their length (**Figure 4.27**). There was a dependency of query coverage on the bit-score of all the tBLASTn (best) hits for all the genomes, shorter peptides presented lower bit-score values, and discarded matches had the lowest bit-score values independently of their coverage (**Figure 4.28**).

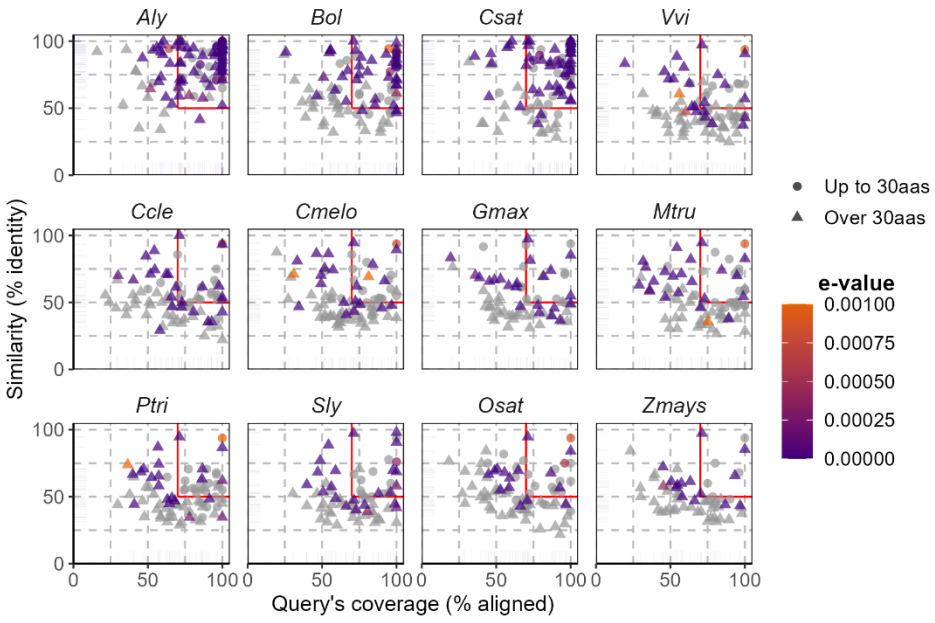


Figure 4.27. Selection criteria for putative homologs.
Scatter plot of alignment vs identity data for each candidate-putative homolog pair coloured by e-value (≤ 0.001). Points represent peptides up to 30 amino acids, and triangles, peptides from 31 to 100 amino acids. Red lines delimit the 70% alignment / 50% identity threshold.

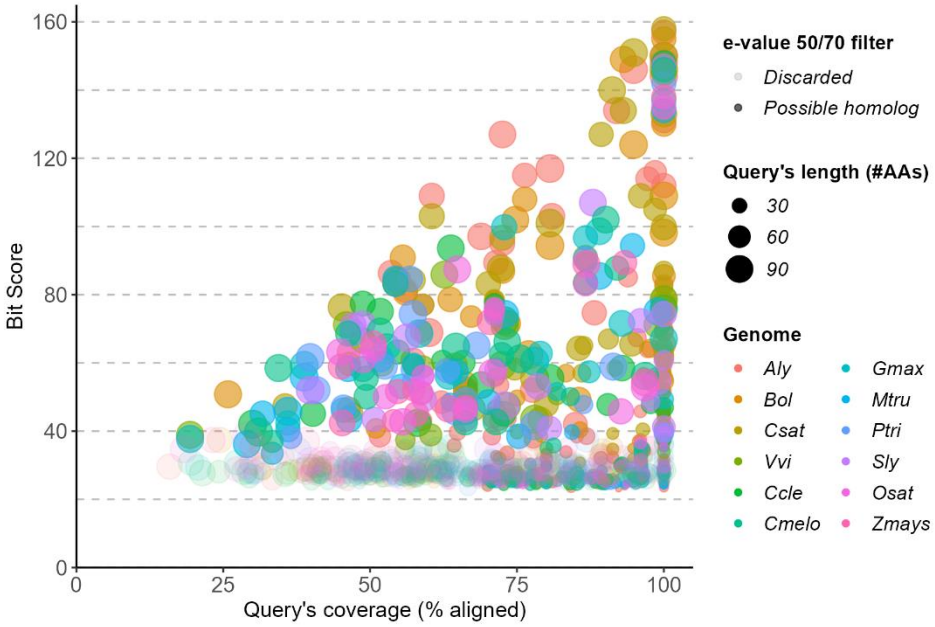


Figure 4.28. Scatterplot showing the dependency of query coverage on the Bit Score of the tBLASTn best hits grouped by genome.

For each *A. thaliana* candidate, the number of matches with tBLASTn for each species fluctuated between 0 and 140, though only 1-36 sequences passed the "homology-threshold" for each candidate (**Sup Table 4.7**). The number of matches per candidate varied depending on the species, although in most of them almost 50% of the candidates had only one match that passed the threshold (46% *A. lyrata*, 25% *B. oleracea*, 8% *C. sativa*, 45% *V. vinifera*, 53% *C. clementina*, 53% *C. melo*, 13% *G. max*, 43% *M. truncatula*, 41% *P. trichocarpa*, 57% *S. lycopersicum,* 57% *O. sativa*, and 45% *Z. mays*) (**Figure 4.29**). For further analyses, only the hit with the lower e-value and the higher percentages of identity and alignment among the putative homologs for each candidate in each species was used.

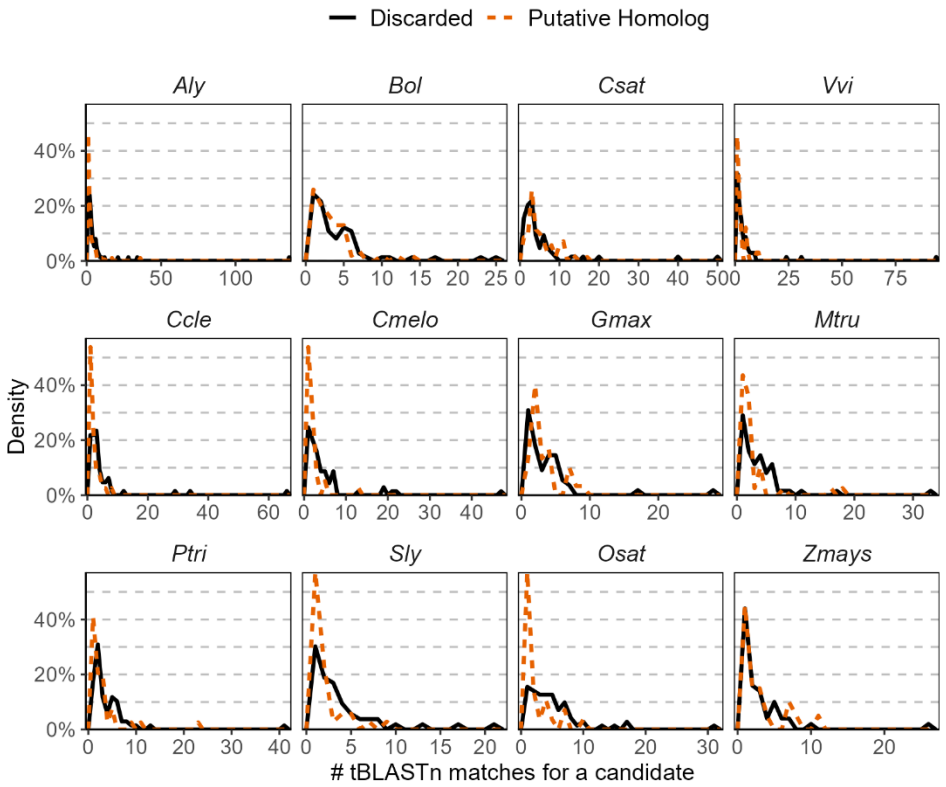


Figure 4.29. Distribution of the number of matches per candidate in each species.
Percentage of matches that did not pass (discarded, black solid line) or did pass (putative homologs, orange dashed line) the homology-threshold.

The nucleotide sequences of the putative homologs were obtained using blasdbcmd (NCBI, v2.11.0+). To check whether the homology can be found in both directions, BLASTx (NCBI, v2.11.0+) was used to compare the resulting homologs with the *A. thaliana* candidates. The CDS of the putative homologs were also blasted against the transcriptome (cDNA and ncRNA databases) and the proteome (peptides and proteins) of their correspondent plant species using blastn and blastx respectively. The same threshold as for the tblastn was used to select sequences in this part. The putative CDS of *A. thaliana* were aligned with those of their putative homologs in the other species using MEGA-X (megacc v10.2.5) (S. Kumar et al., 2012). The alignment was performed for nucleotide sequences and amino acidic sequences, for which CDS were translated using Transeq (EMBOSS online

tool, https://www.ebi.ac.uk/Tools/st/emboss_transeq/), and in-frame STOP codons were removed using *perl*, as the alignment of nucleotides is defective when in-frame STOP codons are present before the 3'-end of the aligned sequences.

Finally, the synonymous and non-synonymous substitution rates and phylogenetic trees were calculated using yn00 (YANG, 2007; YANG & NIELSEN, 2000) or MrBayes (RONQUIST ET AL., 2012) depending on the number of sequences per alignment were available, as MrBayes requires at least 4 taxons to calculate the median synonymous and non-synonymous substitution rate (ω, d_N/d_S). For candidates with one or two homologs, the program yn00 was used to calculate the number of synonymous positions (S), number of non-synonymous positions (N), sequence divergence level (time or distance measured by the expected number of substitutions per codon, t), transition/transversion ratio (κ), synonymous and non-synonymous substitution rate (ω, d_N/d_S), non-synonymous substitution per non-synonymous site (d_N), and synonymous substitutions per synonymous site (d_S). For candidates with homologs in three or more species, MrBayes was the programme of choice. Less than the 25% of the parameters obtained with MrBayes statistical analysis had a total effective sample size (average ESS x 4 runs) lower than 100, thus the analysis could be considered as successful and accurate with the selected parameters (Ngen = 60k, nruns = 4). The program calculates the frequency of site categories (negative, neutral, positive) for each alignment considering the maximum length aligned (longest "length" parameter in tBLASTn results among the different species). However, in the alignments there were gaps because of the different size of the putative homologs in each case. To avoid over-representation of negative sites, the frequency of site categories was re-calculated considering the length of the candidates which was aligned for each pairwise comparison in the alignments. The necessary format modifications (from *.meg to *.nex) were performed using PGDSpider (LISCHER ET AL., 2012).

Genomes, transcriptomes, and proteomes were downloaded from ENSEMBLE (http://ftp.ensemblgenomes.org), except for *V. vinifera* (https://urgi.versailles.inra.fr/files), *C. melo* (https://melonomics.net), and *C. clementina* (https://www.citrusgenomedb.org/analysis/156).

4.4.6 Gene expression of SEPs in different tissues

RNA extraction and cDNA obtention. RNA was extracted from 20-100mg of each sample (*see* **Plant lines, growth conditions, and tissue collection**) using a Maxwell® RSC Plant RNA Kit, and Transcriptor High Fidelity cDNA Synthesis Kit (Roche) was used to obtain cDNA from ~1µg of RNA.

qPCR primer design and testing. The selection of appropriate reference genes for the normalization of qRT-PCR data is a crucial component for successful expression studies (ÁLVAREZ-URDIOLA, BUSTAMANTE, ET AL., 2023). A list of 23 possible combinations of primers was created for classic and novel reference genes, which were previously described (CZECHOWSKI ET AL., 2005) or selected using RefGenes, an online tool based on the Genevestigator database (HRUZ ET AL., 2011) (www.genevestigator.com). qPCR primers for each gene (housekeeping and candidates) were designed using primer-BLAST (primer3 algorithm combined with a BLAST analysis against the *A. thaliana* transcriptome, https://www.ncbi.nlm.nih.gov/tools/primer-blast/) as guide. Amplicons vary from 50 to 178bp length, the melting temperatures (T_m) of the different primers vary from 54 to 61.5°C, their GC content from 29 to 69%, their self-complementarity from 1 to 8 (primer Blast scale), their self-3'-complementarity from 0 to 6, and their length from 16 to 25 bp (**Sup Table 4.8).** Specific amplification of the primers was checked by RT-PCR using a cDNA mix of the 16 tissues of interest as template. The primers of housekeeping genes were checked using as template cDNA of each sample separately and by RT-PCR and qPCR. The primers of 8 housekeeping genes detected at RT-PCR level in all tissues were also tested by qPCR, as well as 2 pairs of primers for candidate genes randomly selected (#006 and #048), to calculate the optimal cDNA concentration for the chip (25 ng of cDNA for each individual reaction). Finally, 5 reference genes were used for the analyses (*ACTIN2, GADPH, UBC, AT4G26410, UBC30*) (**Sup Table 4.8, Figure 4.30**).

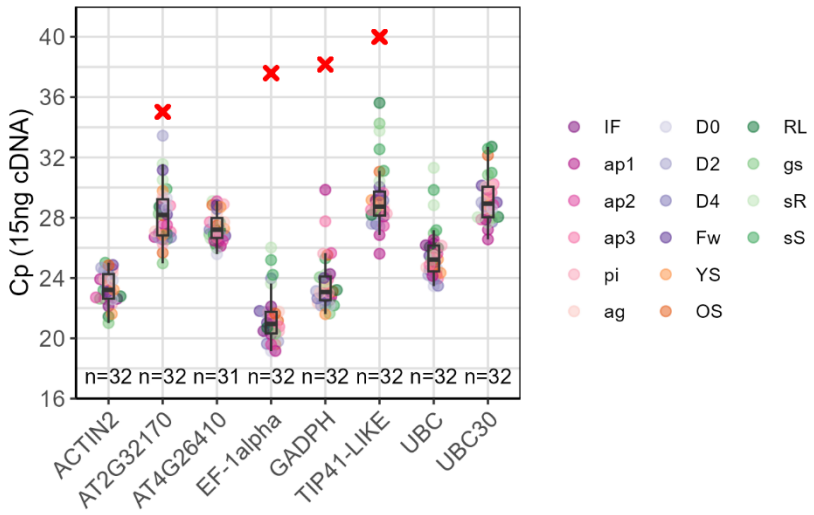


Figure 4.30. Validation of housekeeping genes.
Cp per gene coloured by tissue. Red crosses indicate (if any) the Cp value of negative controls. The number of total samples with a measurement for each gene (n) was 32, except for *AT4G26410* (HK_20), which was 31. Samples: inflorescences of WT plants and the homeotic mutants (IF, *ap1, ap2, ap3, pi, ag*), pAP1:AP1-GR *ap1 cal* inflorescences 0, 2 and 4 days after flower development induction with dexamethasone (D0, D2, D4), WT mature flowers (Fw), young siliques (YS), old siliques (OS), rosette leaves (RL), germinated seeds (gs), seedling roots (sR) and seedling shoots (sS).

Quantitative Real Time PCR – BioMark™ System. The 48x48 array was used following the protocol described in (Álvarez-Urdiola, Bustamante, et al., 2023).

Other data. RNA-seq data were obtained as described in **Chapter 2**.

4.4.7 Generation of mutant reporter lines

Candidate selection. Method 1 considers the database of origin of each candidate, its translation initiation site (TIS), its coordinates within its corresponding mRNA and its homology in different species (41 candidates). As this method is biased for the selection of larger peptides (**Figure 4.31A**), a second selection method (method 2) for smaller peptides (up to 50 amino acids) was established considering exclusively their TIS and coordinates within their corresponding mRNA (26 "novel" candidates) (**Figure 4.31B**).

Method 1 (total n = 41 candidates): i) Araport11 candidates detected with high confidence and medium confidence and less than 20 NAs (n = 10); ii) Hsu et al. candidates classified as discovery candidates and assigned to a floral organ, plus a discovery candidate that lacked gene families in *A. thaliana* (n = 5); iii) CANTATAdb 2.0 and TAIR 'nc' candidates starting with AUG or near-cognate codons and that had homologs in other species and a reasonable start point (near the 5' UTR of the transcript) (n = 22 and 6, respectively).

Method 2: All peptides with up to 50 amino acids, ATG or near-cognate TIS and a start (in the transcript) before the position 800 (n = 44).

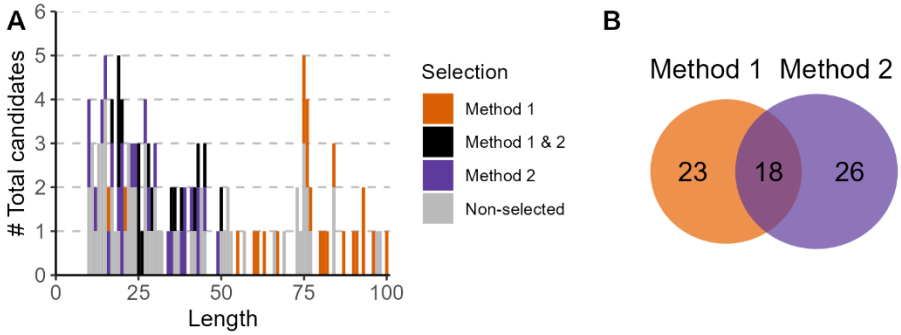


Figure 4.31. Selection of candidates for further analyses.
A) Length distribution of the candidates coloured according to the selection method used (1, 2, both or non-selected). **B)** Venn diagram of the number of candidates selected by each method and the number of candidates in common.

Finally, 37 candidates were selected. Eight peptides selected by Methods 1 and 2 that had less than six putative homologs in *A. thaliana* (without gene families) and/or less than three AUGs upstream their CDS and that were not encoded in antisense lncRNAs (#019, #025, #026, #052, #055, #061, #124, #128). From Method 1, two floral candidates with less than three AUGs upstream (#004, #097) and eight discovery candidates with good qPCR results (#001, #003, #005, #009, #010, #011, #017, #084). From Method 2, six floral candidates with less than three AUGs upstream and that were not encoded in antisense lncRNAs (#032, #044, #077, #116, #121, #131). Besides, 13 extra peptides were selected due to their interesting characteristics: two SSPs (Signal IP6.0) (#014 and #101); a dORF, as there

are no functional peptides encoded in dORFs in plants that have been already characterized according to the literature (#016); the candidate from Hanada et al. (#022) and nine TVGs (qPCR results) with interesting expression patterns (#008, #018, #030, #038, #043, #050, #062, #070, #087).

Promoter amplification. The promoter regions of the peptide candidates (1.5kb upstream the annotated 5' UTR of the transcripts) were amplified using Phusion ® High-Fidelity DNA polymerase (New England Biolabs Inc., #ref:M0530S) and specific primers (**Sup Table 4.9**). PCR products were purified using the NZYGelpure purification kit (nzytech, #ref:MB01101).

Gateway vectors. Purified fragments were cloned in pENTR/D-TOPO entry vectors (Invitrogen; www.thermofisher.com) following the manufacturers' instructions. The resulting plasmids were sequenced to confirm the sequences and the gene cassette transferred into the destiny binary vector pBGWFS7 using the Gateway (Invitrogen) LR-reaction. Final constructs carried the promoter of each candidate (pXXX) fused with GFP and GUS (pXXX:GFP-GUS).

Bacterial strains. Vector cloning was performed in the *Escherichia coli* strain TOP10. Cells were transformed by heat shock and were grown in culture dishes with Luria Bertani medium (LB), agar and the appropriate selection antibiotics. Transformed bacterial colonies were confirmed by colony PCR with M13F and M13R primers for the entry vector and with 5' – CGACCTGCAGGCATGCAAGCTC – 3' and the reverse primer of the promoter of each candidate for the destiny vector. Positive colonies were grown in liquid LB containing the corresponding selection antibiotics. Binary vectors were purified using the QIAprep® Spin Miniprep Kit (QIAGEN, #ref:27106).

Agrobacterium tumefaciens cells strain GV3101 were transformed with the destiny vectors by heat shock (HÖFGEN & WILLMITZER, 1988). *A. tumefaciens* cells were grown 48 h at 28 °C in culture dishes with yeast extract broth (YEB) medium, agar and the appropriate selection antibiotics. Transformed bacterial colonies were confirmed by colony PCR. Positive colonies were grown in liquid YEB containing the corresponding selection antibiotics. These cultures were used for glycerinate-preservation and to be scaled for *A. thaliana* transformation.

Floral dip. Transgenic plants were generated by the floral dip method (CLOUGH & BENT, 1998). Transgenic lines were selected on MS medium supplemented with L-Phosphinothricin (PPT, 16μg/mL).

4.4.8 GUS staining

Inflorescences of at least three independent lines for each pXXX:GFP-GUS construction were placed in 2 mL tubes and kept in acetone 90% for 20 min at -20 °C to remove surface wax. Acetone was removed and samples were washed twice in phosphate buffer pH 7.2 50 mM. After removing the buffer, 1 mL of GUS staining solution (Triton X-100 0.1%, EDTA 1 mM, phosphate buffer 50 mM, potassium ferrocyanide 1 mM, potassium ferricyanide 1 mM, 100 mg of X-Gluc diluted in DMSO) was added to each tube and samples were incubated in vacuum in darkness for at least 30 min. Samples were kept at 37 °C in darkness for 36 h. After removing the GUS staining solution, samples were washed with a series of EtOH dilutions (10-30-50-70-80-96-100% EtOH; 30 min each) at room temperature. Samples were kept on 100% EtOH upon their observation under a stereomicroscope Olympus SZX16.

4.4.9 Generation of knock-out lines

Candidate selection. All candidates with reproducible GUS expression patterns were selected for the generation of mutant lines, except candidates #018 and #019, as both are altORFs (HSU ET AL., 2016) and their modification will alter the main ORF sequence as well. Besides, eight peptides from a previous work in the laboratory were added to the list due to their interesting GUS staining patterns and transcript characterization by RACE PCR (HANADA ET AL., 2007, 2013).

Guide design. Two CRISPR/Cas9 guides were designed for each candidate using CCTop - CRISPR/Cas9 target online predictor (https://cctop.cos.uni-heidelberg.de:8043/index.html) (STEMMER ET AL., 2015) and CRISPR-P 2.0 (http://crispr.hzau.edu.cn/cgi-bin/CRISPR2/CRISPR) (LEI ET AL., 2014). In this step, candidates #025 and #061 were discarded because it was not possible to find suitable guides.

Vector assembly. GoldenGate assembly was used to generate the final constructs using pCBC for guide amplification and pHEE401E as the final vector. For each pair of guides, two forward primers (DT1-BsF, DT1-F0) and two reverse primers (DT2-R0, DT2-BsR) were designed to amplify a fragment of vector pCBC containing a terminator for the first guide, and a promoter for the second one. Forward primers were designed to overlap with each other to add a BsaI site (GGTCTCN) and the first guide, and reverse primers to add the second guide and a BsaI site. $(N)_{20}$ is the guide sequence without the PAM region.

- DT1-BsF: ATATATGGTCTCGATTG$(N)_{20}$GTT
- DT1-F0: TG$(N)_{20}$GTTTTAGAGCTAGAAATAGC
- DT2-R0: AAC$(N)_{20}$CAATCTCTTAGTCGACTCTAC
- DT2-BsR: ATTATTGGTCTCGAAAC$(N)_{20}$CAA

The reaction was conducted as follows:

Component	Volume	Cycling conditions
PrimeStar Buffer	10 µL	One cycle: 95 °C for 2 min 30 cycles: 95 °C for 15 sec; 60°C for 30 sec; 68 °C for 1 min One cycle: 68 °C for 10 min
dNTPs (10mM)	4 µL	
PrimeStar high fidelity polymerase	2 µL	
pCBC	1 µL	
DT1-BsF (20 µM)	1 µL	
DT1-F0 (1 µM)	1 µL	
DT2-R0 (1 µM)	1 µL	
DT2-BsR (20 µM)	1 µL	
ddH20	29 µL	
Total volume	**50 µL**	

PCR amplified fragments were purified from gel and used to conduct the GoldenGate assembly protocol as follows:

Component	Volume	Cycling conditions
PCR purified fragment (~100 ng/µL)	2 µL	5 h at 37 °C 5 min at 50 °C 10 min at 80 °C
pHEE401E (~100 ng/µL)	2 µL	
10x T4 DNA Ligase Buffer	1.5 µL	
10x BSA	1.5 µL	
BsaI	1 µL	
T4 DNA ligase	1 µL	
dd$H_2$0	6 µL	
Total volume	**15 µL**	

Primers and guides used for the generation of *knock-out* lines are included in **Sup Table 4.10**.

Bacterial strains. Vector cloning using 5 µL of GoldenGate reaction mixture was performed in the *E. coli* strain DH5α. Cells were transformed by heat shock and were grown in culture dishes with LB and kanamycin. Transformed bacterial colonies were confirmed by colony PCR with pHEE-seq-Fw (5' – GTCACGACGTTGTAAAACGACG – 3') and pHEE-seq-Rev (5' – CAATGATAAACCAAACGCAAATGC – 3') primers. Positive colonies were grown in liquid LB containing kanamycin. Binary vectors were purified using the NucleoSpin Plasmid, Mini kit for plasmid DNA (Macherey-Nagel, #ref: 740588.50).

A. tumefaciens cells strain GV3101 were transformed with the destiny vectors by electroporation. *A. tumefaciens* cells were grown 48 h at 28 °C in culture dishes with LB medium, agar and the appropriate selection antibiotics (kanamycin, rifampicin, tetracycline and gentamycin). Transformed bacterial colonies were confirmed by colony PCR. Positive colonies were grown in liquid LB containing the corresponding selection antibiotics. These cultures were used for glycerinate-preservation and to be scaled for *A. thaliana* transformation.

Floral dip. Transgenic plants were generated by the floral dip method (CLOUGH & BENT, 1998). Transgenic lines were selected on MS medium supplemented with hygromycin.

Conclusions

Conclusions

Within this work, innovative transcriptomic-proteomic integrative methods and peptide discovery approaches were applied to further the understanding of flower development in the plant model species *Arabidopsis thaliana*. The main conclusions of this Thesis are hereby described in terms of their respective objectives.

Aim 1.- *To establish a chronology of protein expression throughout (early) flower development and correlate these trajectories to unbiased transcript expression data*.

- The customized method used for imputing missing values depending on their nature improved considerably the interpretation of the LC-MS/MS results.
- The size of the transcriptome (i.e., collection of genes) previously known to change its expression during the early stages of flower development was expanded several-fold.
- The correlation between mRNA levels and protein abundance was higher for those gene-protein pairs with significant changes through time for both molecules.
- Around 36% of the quantified gene-protein pairs had a positive correlation between the mRNA levels and protein abundance.
- Gene-protein pairs with opposite patterns for mRNA level and protein abundance were enriched in different hormone responsive pathways, suggesting that there might be regulatory processes (e.g., positive and / or negative feedback loops) affecting mRNA and protein levels differently.
- A total of 230 novel AP1-high confidence targets were identified through the combined analysis of the RNA-seq data and previously published AP1 genome-wide binding data (ChIP-seq).

Aim 2.- *To characterise the flower Arabidopsis peptidome (sORFs and hidden coding sequences in the Arabidopsis genome) and start deciphering its roles in flower development.*

- A total of 1,874 hypothetical peptides were identified in this Thesis using a MS-based method for the identification of novel peptides.
- Sixty hypothetical peptides were classified as possible floral organ-specific peptides.
- A majority of peptide candidates identified as specific or enriched in floral organs were so in stamens, which is in agreement with previously published results for the floral organ differential gene expression of standard, annotated genes.
- The putative or confirmed Translation Initiation Site (TIS) for around 71% of the 1,874 hypothetical peptides was identified as either an AUG or a near-cognate codon (26% and 45%, respectively), and similar percentages were found for the reduced set of 132 candidates (33% and 36% respectively).
- Non-AUG translation initiation is abundant among the identified SEPs. This expands the criteria that should be taken into consideration for protein and peptide predictions from Arabidopsis genomic or transcriptomic sequences.
- Plant SEPs can be conserved across species, but also be species- or family-specific. Sixty-one of the peptide candidates, out of 132, were apparently specific to the Brassicaceae, as they were found exclusively in *A. thaliana, A. lyrata, B. oleracea* and / or *C. sativa*, and 29 of those appeared to be specific to *A. thaliana*.
- There were fourteen candidates with possible homologs in the twelve plant species analysed. The conservation of these sequences could indicate or be related to a conserved function.
- Analysis of gene expression patterns using SEP promoter-GUS reporter fusions revealed distinct and different expression domains, but with most of the analysed SEPs expressed in developing stamens.

References

References

Abraham, P., Gannone, R. J., Adams, R. M., Kalluri, U., Tuskan, G. A., & Hettich, R. L. (2013). Putting the Pieces Together: High-performance LC-MS/MS Provides Network-, Pathway-, and Protein-level Perspectives in Populus. Molecular & Cellular Proteomics : MCP, 12(1), 106. https://doi.org/10.1074/MCP.M112.022996

Akoglu, H. (2018). User's guide to correlation coefficients. Turkish Journal of Emergency Medicine, 18(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Ali, M. S., & Baek, K. H. (2020). Jasmonic acid signaling pathway in response to abiotic stresses in plants. International Journal of Molecular Sciences, 21(2). https://doi.org/10.3390/ijms21020621

Álvarez-Urdiola, R., Borràs, E., Valverde, F., Matus, J. T., Sabidó, E., & Riechmann, J. L. (2023). Peptidomics Methods Applied to the Study of Flower Development. In Methods in molecular biology (Vol. 2686, pp. 509–536). Springer Science+Business Media, LLC, part of Springer Nature 2023. https://doi.org/10.1007/978-1-0716-3299-4_24

Álvarez-Urdiola, R., Bustamante, M., Ribes, J., & Riechmann, J. L. (2023). Gene Expression Analysis by Quantitative Real-Time PCR for Floral Tissues. In Methods in molecular biology (Vol. 2686, pp. 403–428). Springer Science+Business Media, LLC, part of Springer Nature 2023. https://doi.org/10.1007/978-1-0716-3299-4_20

Álvarez-Urdiola, R., Matus, J. T., & Riechmann, J. L. (2023). Multi-Omics Methods Applied to Flower Development. In Methods in Molecular Biology (Vol. 2686, pp. 495–508). Springer Science+Business Media, LLC, part of Springer Nature 2023. https://doi.org/10.1007/978-1-0716-3299-4_23

Amano, Y., Tsubouchi, H., Shinohara, H., Ogawa, M., & Matsubayashi, Y. (2007). Tyrosine-sulfated glycopeptide involved in cellular proliferation and expansion in Arabidopsis. Proc Natl Acad Sci U S A, 104(46), 18333-18338. https://doi.org/10.1073/pnas.0706403104

Anderson, D. M., Anderson, K. M., Chang, C. L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., . . . Olson, E. N. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell, 160(4), 595-606. https://doi.org/10.1016/j.cell.2015.01.009

Anderson, D. M., Makarewich, C. A., Anderson, K. M., Shelton, J. M., Bezprozvannaya, S., Bassel-Duby, R., & Olson, E. N. (2016). Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. Sci Signal, 9(457), ra119. https://doi.org/10.1126/scisignal.aaj1460

Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., & Rasool, M. H. (2017). Proteomics: Technologies and their applications. Journal of Chromatographic Science, 55(2), 182–196. https://doi.org/10.1093/chromsci/bmw167

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., & Couso, J. P. (2014). Extensive translation of small open reading frames revealed by poly-ribo-seq. ELife, 3(August2014), 1–19. https://doi.org/10.7554/eLife.03528

Azpeitia, E., Tichtinsky, G., Le Masson, M., Serrano-Mislata, A., Lucas, J., Gregis, V., Gimenez, C., Prunet, N., Farcot, E., Kater, M. M., Bradley, D., Madueño, F., Godin, C., & Parcy, F. (2021). Cauliflower fractal forms arise from perturbations of floral gene networks. Science, 373(6551), 192–197. https://doi.org/10.1126/SCIENCE.ABG5999/SUPPL_FILE/SCIENCE.ABG5999-SM.PDF

Babina, A. M., Surkov, S., Ye, W., Jerlstrom-Hultqvist, J., Larsson, M., Holmqvist, E., . . . Knopp, M. (2023). Rescue of Escherichia coli auxotrophy by de novo small proteins. elife, 12, e78299. https://doi.org/10.7554/eLife.78299

Bai, B., van der Horst, N., Cordewener, J. H., America, A. H. P., Nijveen, H., & Bentsink, L. (2021). Delayed Protein Changes During Seed Germination. Frontiers in Plant Science, 12(September). https://doi.org/10.3389/fpls.2021.735719

Barashkova, A. S., & Rogozhin, E. A. (2020). Isolation of antimicrobial peptides from different plant sources: Does a general extraction method exist? Plant Methods, 16, 143. https://doi.org/10.1186/S13007-020-00687-1/FIGURES/2

Barghahn, S., Saridis, G., Mantz, M., Meyer, U., Mellüh, J. C., Misas Villamil, J. C., Huesgen, P. F., & Doehlemann, G. (2023). Combination of transcriptomic, proteomic, and degradomic profiling reveals common and distinct patterns of pathogen-induced cell death in maize. Plant Journal, 1–23. https://doi.org/10.1111/tpj.16356

Bartels, S., & Boller, T. (2015). Quo vadis, Pep? Plant elicitor peptides at the crossroads of immunity, stress, and development. Journal of Experimental Botany, 66(17), 5183-5193. https://doi.org/10.1093/jxb/erv180

Bassal, M., Abukhalaf, M., Majovsky, P., Thieme, D., Herr, T., Ayash, M., Tabassum, N., Al Shweiki, M. R., Proksch, C., Hmedat, A., Ziegler, J., Lee, J., Neumann, S., & Hoehenwarter, W. (2020). Reshaping of the Arabidopsis thaliana Proteome Landscape and Co-regulation of Proteins in Development and Immunity. Molecular Plant, 13(12), 1709–1732. https://doi.org/10.1016/j.molp.2020.09.024

Bazin, J., Baerenfaller, K., Gosai, S. J., Gregory, B. D., Crespi, M., & Bailey-Serres, J. (2017). Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. PNAS, E10018–E10027. https://doi.org/10.1073/pnas.1708433114

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., . . . Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. Embo J, 33(9), 981-993. https://doi.org/embj.201488411 [pii]

Beer, L. A., Liu, P., Ky, B., Barnhart, K. T., & Speicher, D. W. (2017). Efficient quantitative comparisons of plasma proteomes using label-free analysis with MaxQuant. Methods in Molecular Biology, 1619, 339–352. https://doi.org/10.1007/978-1-4939-7057-5

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300. https://doi.org/http://www.jstor.org/stable/2346101

Bernhofer, M., Goldberg, T., Wolf, S., Ahmed, M., Zaugg, J., Boden, M., & Rost, B. (2018). NLSdb—major update for database of nuclear localization signals and nuclear export signals. Nucleic Acids Research, 46(D1), D503–D508. https://doi.org/10.1093/NAR/GKX1021

Bessho-Uehara, K., Wang, D. R., Furuta, T., Minami, A., Nagai, K., Gamuyao, R., . . . Ashikari, M. (2016). Loss of function at RAE2, a previously unidentified EPFL, is required for awnlessness in cultivated Asian rice. Proc Natl Acad Sci U S A, 113(32), 8969-8974. https://doi.org/10.1073/pnas.1604849113

Bhati, K. K., Blaakmeer, A., Paredes, E. B., Dolde, U., Eguen, T., Hong, S. Y., . . . Wenkel, S. (2018). Approaches to identify and characterize microProteins and their potential uses in biotechnology. Cellular and Molecular Life Sciences, 75(14), 2529-2536. https://doi.org/10.1007/s00018-018-2818-8

Bhati, K. K., Dolde, U., & Wenkel, S. (2021). MicroProteins: Expanding functions and novel modes of regulation. Mol Plant, 14(5), 705-707. https://doi.org/10.1016/j.molp.2021.01.006

Bhati, K. K., Kruusvee, V., Straub, D., Chandran, A. K. N., Jung, K. H., & Wenkel, S. (2020). Global Analysis of Cereal microProteins Suggests Diverse Roles in Crop Development and Environmental Adaptation. G3 (Bethesda), 10(10), 3709-3717. https://doi.org/10.1534/g3.120.400794

Bi, P., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J. R., Shelton, J. M., . . . Olson, E. N. (2017). Control of muscle formation by the fusogenic micropeptide myomixer. Science, 356(6335), 323-327. https://doi.org/10.1126/science.aam9361

Bishop, D. J., & Hawley, J. A. (2022). Reassessing the relationship between mRNA levels and protein abundance in exercised skeletal muscles. Nature Reviews Molecular Cell Biology, 23(December). https://doi.org/10.1038/s41580-022-00541-3

Blanvillain, R., Young, B., Cai, Y. M., Hecht, V., Varoquaux, F., Delorme, V., . . . Gallois, P. (2011). The Arabidopsis peptide kiss of death is an inducer of programmed cell death. Embo J, 30(6), 1173-1183. https://doi.org/10.1038/emboj.2011.14

Bollier, N., Sicard, A., Leblond, J., Latrasse, D., Gonzalez, N., Gevaudant, F., . . . Delmas, F. (2018). At-MINI ZINC FINGER2 and Sl-INHIBITOR OF MERISTEM ACTIVITY, a Conserved Missing Link in the Regulation of Floral Meristem Termination in Arabidopsis and Tomato. Plant Cell, 30(1), 83-100. https://doi.org/10.1105/tpc.17.00653

Bowman, J. L., Alvarez, J., Weigel, D., Meyerowitz, E. M., & Smyth, D. R. (1993). Control of flower development in Arabidopsis thaliana by APETALA 1 and interacting genes. Development, 119(3), 721–743. https://dev.biologists.org/content/119/3/721

Bowman, J. L., Smyth, D. R., & Meyerowitz, E. M. (1989). Genes directing flower development in arabidopsis. The Plant Cell, 1, 37–52. https://doi.org/10.1105/tpc.19.00276

Bowman, J. L., Smyth, D. R., & Meyerowitz, E. M. (1991). Genetic interactions among floral homeotic genes of Arabidopsis. Development, 112, 1–20.

Breiden, M., & Simon, R. (2016). Q&A: How does peptide signaling direct plant development? BMC Biol, 14, 58. https://doi.org/10.1186/s12915-016-0280-3

Bruce, B. D. (2000). Chloroplast transit peptides: Structure, function and evolution. Trends in Cell Biology, 10(10), 440–447. https://doi.org/10.1016/S0962-8924(00)01833-X

Brunet, M. A., Leblanc, S., & Roucou, X. (2020). Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. Exp Cell Res, 393(1), 112057. https://doi.org/10.1016/j.yexcr.2020.112057

Brunet, M. A., Levesque, S. A., Hunting, D. J., Cohen, A. A., & Roucou, X. (2018). Recognition of the polycistronic nature of human genes is critical to understanding the genotype-

phenotype relationship. Genome Research, 28(5), 609-624. https://doi.org/10.1101/gr.230938.117

Brunet, M. A., Lucier, J. F., Levesque, M., Leblanc, S., Jacques, J. F., Al-Saedi, H. R. H., . . . Roucou, X. (2021). OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. Nucleic Acids Res, 49(D1), D380-D388. https://doi.org/10.1093/nar/gkaa1036

Buhler, E., Fahrbach, E., Schaller, A., & Stuhrwohldt, N. (2023). Sulfopeptide CLEL6 inhibits anthocyanin biosynthesis in Arabidopsis thaliana. Plant Physiol, 193(1), 809-820. https://doi.org/10.1093/plphys/kiad316

Cao, X., & Slavoff, S. A. (2020). Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. Exp Cell Res, 391(1), 111973. https://doi.org/10.1016/j.yexcr.2020.111973

Cao, X., Chen, Y., Khitun, A., & Slavoff, S. A. (2023). BONCAT-based Profiling of Nascent Small and Alternative Open Reading Frame-encoded Proteins. Bio Protoc, 13(1), e4585. https://doi.org/10.21769/BioProtoc.4585

Cardon, T., Fournier, I., & Salzet, M. (2021). Shedding Light on the Ghost Proteome. Trends in Biochemical Sciences, 46(3), 239-250. https://doi.org/10.1016/j.tibs.2020.10.003

Cardon, T., Herve, F., Delcourt, V., Roucou, X., Salzet, M., Franck, J., & Fournier, I. (2020). Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins. Analytical Chemistry, 92(1), 1122-1129. https://doi.org/10.1021/acs.analchem.9b04188

Casson, S. A., Chilley, P. M., Topping, J. F., Evans, I. M., Souter, M. A., & Lindsey, K. (2002). The POLARIS gene of Arabidopsis encodes a predicted peptide required for correct root growth and leaf vascular patterning. Plant Cell, 14(8), 1705-1721. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12172017

Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., & Briggs, S. P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A, 105(52), 21034-21038. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19098097

Causier, B., Schwarz-Sommer, Z., & Davies, B. (2010). Floral organ identity: 20 years of ABCs. Seminars in Cell and Developmental Biology, 21(1), 73–79. https://doi.org/10.1016/j.semcdb.2009.10.005

Chahtane, H., Lai, X., Tichtinsky, G., Rieu, P., Arnoux-Courseaux, M., Cancé, C., Marondedze, C., & Parcy, F. (2023). Flower Development in Arabidopsis. In Methods in Molecular Biology (Vol. 2686, pp. 3–38). Springer Science+Business Media, LLC, part of Springer Nature 2023. https://doi.org/doi.org/10.1007/978-1-0716-3299-4_1

Chen, D., Yan, W., Fu, L. Y., & Kaufmann, K. (2018). Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. Nature Communications, 9(1), 1–13. https://doi.org/10.1038/s41467-018-06772-3

Chen, J., Brunner, A. D., Cogan, J. Z., Nunez, J. K., Fields, A. P., Adamson, B., . . . Weissman, J. S. (2020). Pervasive functional translation of noncanonical human open reading frames. Science, 367(6482), 1140-1146. https://doi.org/10.1126/science.aay0262

Chen, R., Chen, G., & Huang, J. (2017). Shot-gun proteome and transcriptome mapping of the jujube floral organ and identification of a pollen-specific S-locus F-box gene. PeerJ, 2017(7), 1–18. https://doi.org/10.7717/peerj.3588

Chen, R., Xiao, N., Lu, Y., Tao, T., Huang, Q., Wang, S., . . . Yang, Z. (2023). A de novo evolved gene contributes to rice grain shape difference between indica and japonica. Nat Commun, 14(1), 5906. https://doi.org/10.1038/s41467-023-41669-w

Chen, Y. L., Lee, C. Y., Cheng, K. T., Chang, W. H., Huang, R. N., Nam, H. G., & Chen, Y. R. (2014). Quantitative peptidomics study reveals that a wound-induced peptide from PR-1 regulates immune signaling in tomato. Plant Cell, 26(10), 4135-4148. https://doi.org/10.1105/tpc.114.131185

Chen, Y., Li, D., Fan, W., Zheng, X., Zhou, Y., Ye, H., . . . Wang, K. (2020). PsORF: a database of small ORFs in plants. Plant Biotechnol J, 18(11), 2158-2160. https://doi.org/10.1111/pbi.13389

Chen, Y., Su, H., Zhao, J., Na, Z., Jiang, K., Bacchiocchi, A., . . . Slavoff, S. A. (2023). Unannotated microprotein EMBOW regulates the interactome and chromatin and mitotic functions of WDR5. Cell Rep, 42(9), 113145. https://doi.org/10.1016/j.celrep.2023.113145

Chen, Z., Meng, J., Zhao, S., Yin, C., & Luan, Y. (2023). sORFPred: A Method Based on Comprehensive Features and Ensemble Learning to Predict the sORFs in Plant LncRNAs. Interdiscip Sci, 15(2), 189-201. https://doi.org/10.1007/s12539-023-00552-4

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant Journal, 89(4), 789–804. https://doi.org/10.1111/tpj.13415

Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., & Gygi, S. P. (2015). A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol, 33(7), 743-749. https://doi.org/10.1038/nbt.3267

Chien, P. S., Nam, H. G., & Chen, Y. R. (2015). A salt-regulated peptide derived from the CAP superfamily protein negatively regulates salt-stress tolerance in Arabidopsis. Journal of Experimental Botany, 66(17), 5301-5313. https://doi.org/10.1093/jxb/erv263

Chilley, P. M., Casson, S. A., Tarkowski, P., Hawkins, N., Wang, K. L., Hussey, P. J., . . . Lindsey, K. (2006). The POLARIS peptide of Arabidopsis regulates auxin transport and root growth via effects on ethylene signaling. Plant Cell, 18(11), 3058-3072. https://doi.org/tpc.106.040790 [pii]

Chin, S., Kwon, T., Khan, B. R., Sparks, J. A., Mallery, E. L., Szymanski, D. B., & Blancaflor, E. B. (2021). Spatial and temporal localization of SPIRRIG and WAVE/SCAR reveal roles for these proteins in actin-mediated root hair development. Plant Cell, 33(7), 2131-2148. https://doi.org/10.1093/plcell/koab115

Chiva, C., Olivella, R., Borràs, E., Espadas, G., Pastor, O., Solé, A., & Sabidó, E. (2018). QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories. PLoS ONE, 13(1), 1–14. https://doi.org/10.1371/journal.pone.0189209

Chothani, S. P., Adami, E., Widjaja, A. A., Langley, S. R., Viswanathan, S., Pua, C. J., . . . Schafer, S. (2022). A high-resolution map of human RNA translation. Molecular Cell, 82(15), 2885-2899 e2888. https://doi.org/10.1016/j.molcel.2022.06.023

Chu, Q., Ma, J., & Saghatelian, A. (2015). Identification and characterization of sORF-encoded polypeptides. Crit Rev Biochem Mol Biol, 50(2), 134-141. https://doi.org/10.3109/10409238.2015.1016215

Clough, S. J., & Bent, A. F. (1998). Floral dip: A simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. Plant Journal, 16(6), 735–743.

Coen, E. S., & Meyerowitz, E. M. (1991). The war of the whorls: Genetic interactions controlling flower development. Nature, 353(6339), 31–37. https://doi.org/10.1038/353031a0

Colombo, M., Brambilla, V., Marcheselli, R., Caporali, E., Kater, M. M., & Colombo, L. (2010). A new role for the SHATTERPROOF genes during Arabidopsis gynoecium development. Developmental Biology, 337(2), 294–302. https://doi.org/10.1016/J.YDBIO.2009.10.043

Costanzo, E., Trehin, C., & Vandenbussche, M. (2014). The role of WOX genes in flower development. Annals of Botany, 114(7), 1545–1553. https://doi.org/10.1093/aob/mcu123

Couso, J. P., & Patraquim, P. (2017). Classification and function of small open reading frames. Nature Reviews Molecular Cell Biology, 18(9), 575-589. https://doi.org/10.1038/nrm.2017.58

Crappe, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., & Menschaert, G. (2013). Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics, 14, 648. https://doi.org/10.1186/1471-2164-14-648

Crespi, M. D., Jurkevitch, E., Poiret, M., d'Aubenton-Carafa, Y., Petrovics, G., Kondorosi, E., & Kondorosi, A. (1994). enod40, a gene expressed during nodule organogenesis, codes for a non-translatable RNA involved in plant growth. Embo J, 13(21), 5099-5112. https://doi.org/10.1002/j.1460-2075.1994.tb06839.x

Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M., & Drummond, D. A. (2015). Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. PLoS Genetics, 11(5), 1–32. https://doi.org/10.1371/journal.pgen.1005206

Culver, K. D., Allen, J. L., Shaw, L. N., & Hicks, L. M. (2021). Too Hot to Handle: Antibacterial Peptides Identified in Ghost Pepper. J Nat Prod, 84(8), 2200-2208. https://doi.org/10.1021/acs.jnatprod.1c00281

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., & Scheible, W.-R. (2005). Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. Plant Physiology, 139(September), 5–17. https://doi.org/10.1104/pp.105.063743.1

Dai, S., & Chen, S. (2012). Single-cell-type Proteomics: Toward a Holistic Understanding of Plant Function. Molecular & Cellular Proteomics : MCP, 11(12), 1622. https://doi.org/10.1074/MCP.R112.021550

Das, D., Jaiswal, M., Khan, F. N., Ahamad, S., & Kumar, S. (2020). PlantPepDB: A manually curated plant peptide database. Sci Rep, 10(1), 2194. https://doi.org/10.1038/s41598-020-59165-2

De Coninck, B., Carron, D., Tavormina, P., Willem, L., Craik, D. J., Vos, C., … Cammue, B. P. (2013). Mining the genome of Arabidopsis thaliana as a basis for the identification of novel

bioactive peptides involved in oxidative stress tolerance. Journal of Experimental Botany, 64(17), 5297-5307. https://doi.org/10.1093/jxb/ert295

De Giorgi, J., Fuchs, C., Iwasaki, M., Kim, W., Piskurewicz, U., Gully, K., . . . Lopez-Molina, L. (2021). The Arabidopsis mature endosperm promotes seedling cuticle formation via release of sulfated peptides. Developmental Cell, 56(22), 3066-3081 e3065. https://doi.org/10.1016/j.devcel.2021.10.005

Decourcelle, M., Perez-Fons, L., Baulande, S., Steiger, S., Couvelard, L., Hem, S., Zhu, C., Capell, T., Christou, P., Fraser, P., & Sandmann, G. (2015). Combined transcript, proteome, and metabolite analysis of transgenic maize seeds engineered for enhanced carotenoid synthesis reveals pleotropic effects in core metabolism. Journal of Experimental Botany, 66(11), 3141–3150. https://doi.org/10.1093/jxb/erv120

Ding, S., Lv, J., Hu, Z., Wang, J., Wang, P., Yu, J., . . . Shi, K. (2023). Phytosulfokine peptide optimizes plant growth and defense via glutamine synthetase GS2 phosphorylation in tomato. Embo J, 42(6), e111858. https://doi.org/10.15252/embj.2022111858

Ditta, G., Pinyopich, A., Robles, P., Pelaz, S., & Yanofsky, M. F. (2004). The SEP4 Gene of Arabidopsis thaliana Functions in Floral Organ and Meristem Identity. Current Biology, 14(21), 1935–1940. https://doi.org/10.1016/J.CUB.2004.10.028

Dittmar, G., Hernandez, D. P., Kowenz-Leutz, E., Kirchner, M., Kahlert, G., Wesolowski, R., . . . Leutz, A. (2019). PRISMA: Protein Interaction Screen on Peptide Matrix Reveals Interaction Footprints and Modifications- Dependent Interactome of Intrinsically Disordered C/EBPbeta. iScience, 13, 351-370. https://doi.org/10.1016/j.isci.2019.02.026

Djakovic, S., Dyachok, J., Burke, M., Frank, M. J., & Smith, L. G. (2006). BRICK1/HSPC300 functions with SCAR and the ARP2/3 complex to regulate epidermal cell shape in Arabidopsis. Development, 133(6), 1091-1100. https://doi.org/dev.02280 [pii]

D'Lima, N. G., Ma, J., Winkler, L., Chu, Q., Loh, K. H., Corpuz, E. O., . . . Slavoff, S. A. (2017). A human microprotein that interacts with the mRNA decapping complex. Nat Chem Biol, 13(2), 174-180. https://doi.org/10.1038/nchembio.2249

do Amaral, M. N., & Souza, G. M. (2017). The challenge to translate omics data to whole plant physiology: The context matters. Frontiers in Plant Science, 8(December), 8–11. https://doi.org/10.3389/fpls.2017.02146

Doblas, V. G., Smakowska-Luzan, E., Fujita, S., Alassimone, J., Barberon, M., Madalinski, M., . . . Geldner, N. (2017). Root diffusion barrier control by a vasculature-derived peptide binding to the SGN3 receptor. Science, 355(6322), 280-284. https://doi.org/10.1126/science.aaj1562

Dong, X., Wang, D., Liu, P., Li, C., Zhao, Q., Zhu, D., & Yu, J. (2013). Zm908p11, encoded by a short open reading frame (sORF) gene, functions in pollen tube growth as a profilin ligand in maize. Journal of Experimental Botany, 64(8), 2359-2372. https://doi.org/10.1093/jxb/ert093

Du, C., Li, H., Liu, C., & Fan, H. (2021). Understanding of the postgerminative development response to salinity and drought stresses in cucumber seeds by integrated proteomics and transcriptomics analysis. Journal of Proteomics, 232(November 2020), 104062. https://doi.org/10.1016/j.jprot.2020.104062

Duffy, E. E., Finander, B., Choi, G., Carter, A. C., Pritisanac, I., Alam, A., . . . Greenberg, M. E. (2022). Developmental dynamics of RNA translation in the human brain. Nat Neurosci, 25(10), 1353-1365. https://doi.org/10.1038/s41593-022-01164-9

Duncan, O., Trösch, J., Fenske, R., Taylor, N. L., & Millar, A. H. (2017). Resource: Mapping the Triticum aestivum proteome. The Plant Journal, 89(3), 601–616. https://doi.org/10.1111/TPJ.13402

Edfors, F., Danielsson, F., Hallström, B. M., Käll, L., Lundberg, E., Pontén, F., Forsström, B., & Uhlén, M. (2016). Gene-specific correlation of RNA and protein levels in human cells and tissues . Molecular Systems Biology, 12(10), 1–10. https://doi.org/10.15252/msb.20167144

Eguen, T., Straub, D., Graeff, M., & Wenkel, S. (2015). MicroProteins: small size-big impact. Trends Plant Sci, 20(8), 477-482. https://doi.org/10.1016/j.tplants.2015.05.011

Emanuelsson, O., Von Heijne, G., & Schneider, G. (2001). Analysis and prediction of mitochondrial targeting peptides. In Methods in Cell Biology (Vol. 65, Issue 65, pp. 175–187). Academic Press. https://doi.org/10.1016/S0091-679X(01)65011-8

Engelhorn, J., Wellmer, F., & Carles, C. C. (2018). Profiling histone modifications in synchronized floral tissues for quantitative resolution of chromatin and transcriptome dynamics. In Plant Chromatin Dynamics: Methods and protocols, Methods in Molecular Biology (Vol. 1675, pp. 271–296). https://doi.org/10.1007/978-1-4939-7318-7

Fabre, B., Combier, J. P., & Plaza, S. (2021). Recent advances in mass spectrometry-based peptidomics workflows to identify short-open-reading-frame-encoded peptides and explore their functions. Curr Opin Chem Biol, 60, 122-130. https://doi.org/10.1016/j.cbpa.2020.12.002

Favaro, R., Pinyopich, A., Battaglia, R., Kooiker, M., Borghi, L., Ditta, G., Yanofsky, M. F., Kater, M. M., & Colombo, L. (2003). MADS-box protein complexes control carpel and ovule development in Arabidopsis. The Plant Cell, 15(11), 2603–2611. https://doi.org/10.1105/TPC.015123

Feng, Z., Kong, D., Kong, Y., Zhang, B., & Yang, X. (2022). Coordination of root growth with root morphology, physiology and defense functions in response to root pruning in Platycladus orientalis. Journal of Advanced Research, 36, 187–199. https://doi.org/10.1016/j.jare.2021.07.005

Fernandez, A., Hilson, P., & Beeckman, T. (2013). GOLVEN peptides as important regulatory signalling molecules of plant development. Journal of Experimental Botany, 64(17), 5263-5268. https://doi.org/10.1093/jxb/ert248

Ferrándiz, C., Gu, Q., Martienssen, R., & Yanofsky, M. F. (2000). Redundant regulation of meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER. Development (Cambridge, England), 127(4), 725–734. https://doi.org/10.1242/DEV.127.4.725

Fesenko, I., Azarkina, R., Kirov, I., Kniazev, A., Filippova, A., Grafskaia, E., . . . Govorun, V. (2019). Phytohormone treatment induces generation of cryptic peptides with antimicrobial activity in the Moss Physcomitrella patens. BMC Plant Biol, 19(1), 9. https://doi.org/10.1186/s12870-018-1611-z

Fesenko, I., Kirov, I., Kniazev, A., Khazigaleeva, R., Lazarev, V., Kharlampieva, D., . . . Govorun, V. (2019). Distinct types of short open reading frames are translated in plant cells. Genome Research, 29(9), 1464-1477. https://doi.org/10.1101/gr.253302.119

Fesenko, I., Shabalina, S. A., Mamaeva, A., Knyazev, A., Glushkevich, A., Lyapina, I., . . . Koonin, E. V. (2021). A vast pool of lineage-specific microproteins encoded by long non-coding RNAs in plants. Nucleic Acids Res, 49(18), 10328-10346. https://doi.org/10.1093/nar/gkab816

Fields, A. P., Rodriguez, E. H., Jovanovic, M., Stern-Ginossar, N., Haas, B. J., Mertins, P., . . . Weissman, J. S. (2015). A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Molecular Cell, 60(5), 816-827. https://doi.org/10.1016/j.molcel.2015.11.013

Filippova, A., Lyapina, I., Kirov, I., Zgoda, V., Belogurov, A., Kudriaeva, A., . . . Fesenko, I. (2019). Salicylic acid influences the protease activity and posttranslation modifications of the secreted peptides in the moss Physcomitrella patens. J Pept Sci, 25(2), e3138. https://doi.org/10.1002/psc.3138

Fletcher, J. C. (2020). Recent Advances in Arabidopsis CLE Peptide Signaling. Trends Plant Sci, 25(10), 1005-1016. https://doi.org/10.1016/j.tplants.2020.04.014

Fletcher, J. C., Brand, U., Running, M. P., Simon, R., & Meyerowitz, E. M. (1999). Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. Science, 283(5409), 1911-1914. http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=10082464

Frank, M. J., & Smith, L. G. (2002). A small, novel protein highly conserved in plants and animals promotes the polarized growth and division of maize leaf epidermal cells. Current Biology, 12(10), 849-853. https://doi.org/S0960982202008199 [pii]

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., . . . Grimmond, S. M. (2006). The abundance of short proteins in the mammalian proteome. PLoS Genetics, 2(4), e52. https://doi.org/10.1371/journal.pgen.0020052

Fujita, T., Kurihara, Y., & Iwasaki, S. (2019). The Plant Translatome Surveyed by Ribosome Profiling. Plant Cell Physiol, 60(9), 1917-1926. https://doi.org/10.1093/pcp/pcz059

Furumizu, C., & Sawa, S. (2021). The RGF/GLV/CLEL Family of Short Peptides Evolved Through Lineage-Specific Losses and Diversification and Yet Conserves Its Signaling Role Between Vascular Plants and Bryophytes. Front Plant Sci, 12, 703012. https://doi.org/10.3389/fpls.2021.703012

Garcia-Molina, A., Kleine, T., Schneider, K., Mühlhaus, T., Lehmann, M., & Leister, D. (2020). Translational Components Contribute to Acclimation Responses to High Light, Heat, and Cold in Arabidopsis. IScience, 23(7). https://doi.org/10.1016/j.isci.2020.101331

Gautam, H., Sharma, A., & Trivedi, P. K. (2023). Plant microProteins and miPEPs: Small molecules with much bigger roles. Plant Sci, 326, 111519. https://doi.org/10.1016/j.plantsci.2022.111519

Gemperline, E., Keller, C., Jayaraman, D., Maeda, J., Sussman, M. R., Ane, J. M., & Li, L. (2016). Examination of Endogenous Peptides in Medicago truncatula Using Mass Spectrometry Imaging. J Proteome Res, 15(12), 4403-4411. https://doi.org/10.1021/acs.jproteome.6b00471

Ghorbani, S., Lin, Y., Parizot, B., Fernandez, A., Njo, M. F., Peer, Y. Van De, Beeckman, T., & Hilson, P. (2015). Expanding the repertoire of secretory peptides controlling root development with comparative genome analysis and functional assays. 66(17), 5257–5269. https://doi.org/10.1093/jxb/erv346

Goad, D. M., Zhu, C., & Kellogg, E. A. (2017). Comprehensive identification and clustering of CLV3/ESR-related (CLE) genes in plants finds groups with potentially shared function. New Phytologist, 216(2), 605-616. https://doi.org/10.1111/nph.14348

Gong, P., Shen, Q., Zhang, M., Qiao, R., Jiang, J., Su, L., . . . Zhou, X. (2023). Plant and animal positive-sense single-stranded RNA viruses encode small proteins important for viral infection in their negative-sense strand. Mol Plant, 16(11), 1794-1810. https://doi.org/10.1016/j.molp.2023.09.020

Goslin, K., Finocchio, A., & Wellmer, F. (2023). Floral Homeotic Factors: A Question of Specificity. Plants 2023, Vol. 12, Page 1128, 12(5), 1128. https://doi.org/10.3390/PLANTS12051128

Goslin, K., Zheng, B., Serrano-Mislata, A., Rae, L., Ryan, P. T., Kwaśniewska, K., Thomson, B., Ó'Maoiléidigh, D. S., Madueño, F., Wellmer, F., & Graciet, E. (2017). Transcription factor interplay between LEAFY and APETALA1/CAULIFLOWER during floral initiation. Plant Physiology, 174(2), 1097–1109. https://doi.org/10.1104/pp.17.00098

Graeff, M., Straub, D., Eguen, T., Dolde, U., Rodrigues, V., Brandt, R., & Wenkel, S. (2016). MicroProtein-Mediated Recruitment of CONSTANS into a TOPLESS Trimeric Complex Represses Flowering in Arabidopsis. PLoS Genetics, 12(3), e1005959. https://doi.org/10.1371/journal.pgen.1005959

Gramzow, L., Weilandt, L., & Theißen, G. (2014). MADS goes genomic in conifers: towards determining the ancestral set of MADS-box genes in seed plants. Annals of Botany, 114(7), 1407–1429. https://doi.org/10.1093/AOB/MCU066

Gregis, V., Andrés, F., Sessa, A., Guerra, R. F., Simonini, S., Mateos, J. L., Torti, S., Zambelli, F., Prazzoli, G. M., Bjerkan, K. N., Grini, P. E., Pavesi, G., Colombo, L., Coupland, G., & Kater, M. M. (2013). Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in Arabidopsis. Genome Biology, 14(6), 1–26. https://doi.org/10.1186/gb-2013-14-6-r56

Grienenberger, E., & Fletcher, J. C. (2015). Polypeptide signaling molecules in plant development. Current Opinion in Plant Biology, 23, 8-14. https://doi.org/S1369-5266(14)00134-4 [pii]

Griffiths, G. (2020). Jasmonates: biosynthesis, perception and signal transduction. Essays in Biochemistry, 64(3), 501–512. https://doi.org/10.1042/EBC20190085

Grillet, L., Lan, P., Li, W., Mokkapati, G., & Schmidt, W. (2018). IRON MAN is a ubiquitous family of peptides that control iron transport in plants. Nat Plants, 4(11), 953-963. https://doi.org/10.1038/s41477-018-0266-y

Grossmann, J., Roschitzki, B., Panse, C., Fortes, C., Barkow-Oesterreicher, S., Rutishauser, D., & Schlapbach, R. (2010). Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. Journal of Proteomics, 73(9), 1740–1746. https://doi.org/10.1016/j.jprot.2010.05.011

Guillou, M. C., Vergne, E., Aligon, S., Pelletier, S., Simonneau, F., Rolland, A., . . . Renou, J. P. (2022). The peptide SCOOP12 acts on reactive oxygen species homeostasis to modulate cell division and elongation in Arabidopsis primary root. Journal of Experimental Botany, 73(18), 6115-6132. https://doi.org/10.1093/jxb/erac240

Gully, K., Pelletier, S., Guillou, M. C., Ferrand, M., Aligon, S., Pokotylo, I., . . . Aubourg, S. (2019). The SCOOP12 peptide regulates defense response and root elongation in Arabidopsis

thaliana. Journal of Experimental Botany, 70(4), 1349-1365. https://doi.org/10.1093/jxb/ery454

Guo, P., Yoshimura, A., Ishikawa, N., Yamaguchi, T., Guo, Y., & Tsukaya, H. (2015). Comparative analysis of the RTFL peptide family on the control of plant organogenesis. J Plant Res, 128(3), 497-510. https://doi.org/10.1007/s10265-015-0703-1

Guo, Y., Chen, Y., Wang, Y., Wu, X., Zhang, X., Mao, W., . . . Peng, H. (2023). The translational landscape of bread wheat during grain development. Plant Cell, 35(6), 1848-1867. https://doi.org/10.1093/plcell/koad075

Guruceaga, E., Garin-Muga, A., & Segura, V. (2020). MiTPeptideDB: a proteogenomic resource for the discovery of novel peptides. Bioinformatics, 36(1), 205-211. https://doi.org/10.1093/bioinformatics/btz530

Gyuricza, I. G., Chick, J. M., Keele, G. R., Deighan, A. G., Munger, S. C., Korstanje, R., Gygi, S. P., & Churchill, G. A. (2022). Genome-wide transcript and protein analysis highlights the role of protein homeostasis in the aging mouse heart. Genome Research, 32(5), 838–852. https://doi.org/10.1101/gr.275672.121

Han, F., Zhang, X., Yang, L., Zhuang, M., Zhang, Y., Li, Z., Fang, Z., & Lv, H. (2018). iTRAQ-based proteomic analysis of ogura-CMS cabbage and its maintainer line. International Journal of Molecular Sciences, 19(10). https://doi.org/10.3390/ijms19103180

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., & Shiu, S. H. (2010). sORF finder: a program package to identify small open reading frames with high coding potential. Bioinformatics, 26(3), 399-400. https://doi.org/10.1093/bioinformatics/btp688

Hanada, K., Higuchi-Takeuchi, M., Okamoto, M., Yoshizumi, T., Shimizu, M., Nakaminami, K., . . . Matsui, M. (2013). Small open reading frames associated with morphogenesis are hidden in plant genomes. Proc Natl Acad Sci U S A, 110(6), 2395-2400. https://doi.org/1213958110 [pii]

Hanada, K., Zhang, X., Borevitz, J. O., Li, W. H., & Shiu, S. H. (2007). A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. Genome Research, 17(5), 632–640. https://doi.org/10.1101/gr.5836207

Hander, T., Fernandez-Fernandez, A. D., Kumpf, R. P., Willems, P., Schatowitz, H., Rombaut, D., . . . Stael, S. (2019). Damage on plants activates Ca(2+)-dependent metacaspases for release of immunomodulatory peptides. Science, 363(6433). https://doi.org/10.1126/science.aar7486

Hara, K., Kajita, R., Torii, K. U., Bergmann, D. C., & Kakimoto, T. (2007). The secretory peptide gene EPF1 enforces the stomatal one-cell-spacing rule. Genes Dev, 21(14), 1720-1725. https://doi.org/10.1101/gad.1550707

Hartford, C. C. R., & Lal, A. (2020). When Long Noncoding Becomes Protein Coding. Molecular and Cellular Biology, 40(6), e00528-00519. https://doi.org/10.1128/MCB.00528-19

Hazarika, R. R., De Coninck, B., Yamamoto, L. R., Martin, L. R., Cammue, B. P. A., & Van Noort, V. (2017). ARA-PEPs: A repository of putative SORF-encoded peptides in Arabidopsis thaliana. BMC Bioinformatics, 18(1), 1–9. https://doi.org/10.1186/s12859-016-1458-y

He, C., Jia, C., Zhang, Y., & Xu, P. (2018). Enrichment-Based Proteogenomics Identifies Microproteins, Missing Proteins, and Novel smORFs in Saccharomyces cerevisiae. J Proteome Res, 17(7), 2335-2344. https://doi.org/10.1021/acs.jproteome.8b00032

Hellens, R. P., Brown, C. M., Chisnall, M. A. W., Waterhouse, P. M., & Macknight, R. C. (2016). The Emerging World of Small ORFs. Trends Plant Sci, 21(4), 317-328. https://doi.org/10.1016/j.tplants.2015.11.005

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., & Apweiler, R. (2003). IntAct: an open source molecular interaction database. Nucleic Acids Research, 32(Database issue), D452–D455. https://doi.org/10.1093/nar/gkh052

Higuchi-Takeuchi, M., Kondo, T., Shimizu, M., Kim, Y. W., Shinozaki, K., & Hanada, K. (2020). Effect of small coding genes on the circadian rhythms under elevated CO(2) conditions in plants. Plant Mol Biol, 104(1-2), 55-65. https://doi.org/10.1007/s11103-020-01023-w

Hirayama, T., Lei, G. J., Yamaji, N., Nakagawa, N., & Ma, J. F. (2018). The Putative Peptide Gene FEP1 Regulates Iron Deficiency Response in Arabidopsis. Plant Cell Physiol, 59(9), 1739-1752. https://doi.org/10.1093/pcp/pcy145

Höfgen, R., & Willmitzer, L. (1988). Storage of competent cells for Agrobacterium transformation. Nucleic Acids Research, 16(20), 9877.

Hong, S. Y., Sun, B., Straub, D., Blaakmeer, A., Mineri, L., Koch, J., . . . Wenkel, S. (2020). Heterologous microProtein expression identifies LITTLE NINJA, a dominant regulator of jasmonic acid signaling. Proc Natl Acad Sci U S A, 117(42), 26197-26205. https://doi.org/10.1073/pnas.2005198117

Hoogendijk, A. J., Pourfarzad, F., Aarts, C. E. M., Tool, A. T. J., Hiemstra, I. H., Grassi, L., Frontini, M., Meijer, A. B., van den Biggelaar, M., & Kuijpers, T. W. (2019). Dynamic Transcriptome-Proteome Correlation Networks Reveal Human Myeloid Differentiation and Neutrophil-Specific Programming. CellReports, 29(8), 2505-2519.e4. https://doi.org/10.1016/j.celrep.2019.10.082

Horvath, B., Gungor, B., Toth, M., Domonkos, A., Ayaydin, F., Saifi, F., . . . Kalo, P. (2023). The Medicago truncatula nodule-specific cysteine-rich peptides, NCR343 and NCR-new35 are required for the maintenance of rhizobia in nitrogen-fixing nodules. New Phytologist, 239(5), 1974-1988. https://doi.org/10.1111/nph.19097

Hou, S., Liu, D., Huang, S., Luo, D., Liu, Z., Xiang, Q., . . . He, P. (2021). The Arabidopsis MIK2 receptor elicits immunity by sensing a conserved signature from phytocytokines and microbes. Nat Commun, 12(1), 5494. https://doi.org/10.1038/s41467-021-25580-w

Hou, S., Wang, X., Chen, D., Yang, X., Wang, M., Turra, D., . . . Zhang, W. (2014). The secreted peptide PIP1 amplifies immunity through receptor-like kinase 7. PLoS Pathog, 10(9), e1004331. https://doi.org/10.1371/journal.ppat.1004331

Hruz, T., Wyss, M., Docquier, M., Pfaffl, M. W., Masanetz, S., Borghi, L., Verbrugghe, P., Kalaydjieva, L., Bleuler, S., Laule, O., Descombes, P., Gruissem, W., & Zimmermann, P. (2011). RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. https://doi.org/10.1186/1471-2164-12-156

Hsu, P. Y., & Benfey, P. N. (2018). Small but Mighty: Functional Peptides Encoded by Small ORFs in Plants. Proteomics, 18(10), e1700038. https://doi.org/10.1002/pmic.201700038

Hsu, P. Y., Calviello, L., Wu, H. L., Li, F. W., Rothfels, C. J., Ohler, U., & Benfey, P. N. (2016). Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis.

Proc Natl Acad Sci U S A, 113(45), E7126-E7135. https://doi.org/10.1073/pnas.1614788113

Huang, N., Li, F., Zhang, M., Zhou, H., Chen, Z., Ma, X., . . . Zhang, N. (2021). An Upstream Open Reading Frame in Phosphatase and Tensin Homolog Encodes a Circuit Breaker of Lactate Metabolism. Cell Metab, 33(1), 128-144 e129. https://doi.org/10.1016/j.cmet.2020.12.008

Huang, Y., Zhou, L., Hou, C., & Guo, D. (2022). The dynamic proteome in Arabidopsis thaliana early embryogenesis . Development, 149(18). https://doi.org/10.1242/dev.200715

Huffaker, A., Pearce, G., & Ryan, C. A. (2006). An endogenous peptide signal in Arabidopsis activates components of the innate immune response. Proc Natl Acad Sci U S A, 103(26), 10098-10103. https://doi.org/10.1073/pnas.0603727103

Huo, Y., Yang, H., Ding, W., Huang, T., Yuan, Z., & Zhu, Z. (2023). Combined Transcriptome and Proteome Analysis Provides Insights into Petaloidy in Pomegranate. Plants, 12(13). https://doi.org/10.3390/plants12132402

Ikeuchi, M., Yamaguchi, T., Kazama, T., Ito, T., Horiguchi, G., & Tsukaya, H. (2011). ROTUNDIFOLIA4 regulates cell proliferation along the body axis in Arabidopsis shoot. Plant Cell Physiol, 52(1), 59-69. https://doi.org/10.1093/pcp/pcq138

Ingolia, N. T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. Cell, 165(1), 22-33. https://doi.org/10.1016/j.cell.2016.02.066

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J., Jackson, S. E., . . . Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep, 8(5), 1365-1379. https://doi.org/10.1016/j.celrep.2014.07.045

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 324(5924), 218-223. https://doi.org/1168978 [pii]

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell, 147(4), 789-802. https://doi.org/10.1016/j.cell.2011.10.002

Jack, T., Brockman, L. L., & Meyerowitz, E. M. (1992). The homeotic gene APETALA3 of Arabidopsis thaliana encodes a MADS box and is expressed in petals and stamens. Cell, 68(4), 683–697. https://doi.org/10.1016/0092-8674(92)90144-2

Jain, A., Singh, H. B., & Das, S. (2021). Deciphering plant-microbe crosstalk through proteomics studies. Microbiological Research, 242(August 2020), 126590. https://doi.org/10.1016/j.micres.2020.126590

Ji, J., Yang, L., Fang, Z., Zhuang, M., Zhang, Y., Lv, H., Liu, Y., & Li, Z. (2018). Complementary transcriptome and proteome profiling in cabbage buds of a recessive male sterile mutant provides new insights into male reproductive development. Journal of Proteomics, 179(January), 80–91. https://doi.org/10.1016/j.jprot.2018.03.003

Ji, Z., Song, R., Regev, A., & Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. elife, 4, e08890. https://doi.org/10.7554/eLife.08890

Jiang, L., Wang, M., Lin, S., Jian, R., Li, X., Chan, J., Dong, G., Fang, H., Robinson, A. E., Aguet, F., Anand, S., Ardlie, K. G., Gabriel, S., Getz, G., Graubert, A., Hadley, K., Handsaker, R. E.,

Huang, K. H., Kashin, S., … Snyder, M. P. (2020). A Quantitative Proteome Map of the Human Body. Cell, 183(1), 269-283.e19. https://doi.org/10.1016/j.cell.2020.08.036

Jiang, X., Li, H., Wang, T., Peng, C., Wang, H., Wu, H., & Wang, X. (2012). Gibberellin indirectly promotes chloroplast biogenesis as a means to maintain the chloroplast population of expanded cells. Plant Journal, 72(5), 768–780. https://doi.org/10.1111/j.1365-313X.2012.05118.x

Jiao, Y., & Meyerowitz, E. M. (2010). Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. Mol Syst Biol, 6, 419. https://doi.org/msb201076 [pii]

Jin, L., Bi, Y., Hu, C., Qu, J., Shen, S., Wang, X., & Tian, Y. (2021). A comparative study of evaluating missing value imputation methods in label-free proteomics. Scientific Reports, 11(1), 1–11. https://doi.org/10.1038/s41598-021-81279-4

Jing, D., Chen, W., Hu, R., Zhang, Y., Xia, Y., Wang, S., He, Q., Guo, Q., & Liang, G. (2020). An integrative analysis of transcriptome, proteome and hormones reveals key differentially expressed genes and metabolic pathways involved in flower development in loquat. International Journal of Molecular Sciences, 21(14), 1–22. https://doi.org/10.3390/ijms21145107

Jorge, G. L., & Balbuena, T. S. (2021). Identification of novel protein-coding sequences in Eucalyptus grandis plants by high-resolution mass spectrometry. Biochim Biophys Acta Proteins Proteom, 1869(3), 140594. https://doi.org/10.1016/j.bbapap.2020.140594

Jourquin, J., Fernandez, A. I., Wang, Q., Xu, K., Chen, J., Simura, J., . . . Beeckman, T. (2023). GOLVEN peptides regulate lateral root spacing as part of a negative feedback loop on the establishment of auxin maxima. Journal of Experimental Botany, 74(14), 4031-4049. https://doi.org/10.1093/jxb/erad123

Juntawong, P., Girke, T., Bazin, J., & Bailey-Serres, J. (2014). Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. Proc Natl Acad Sci U S A, 111(1), E203-212. https://doi.org/1317811111 [pii]

Kage, U., Powell, J. J., Gardiner, D. M., & Kazan, K. (2020). Ribosome profiling in plants: what is not lost in translation? Journal of Experimental Botany, 71(18), 5323-5332. https://doi.org/10.1093/jxb/eraa227

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. In Nucleic Acids Research (Vol. 28, Issue 1). http://www.genome.ad.jp/kegg/

Käppel, S., Rümpler, F., & Theißen, G. (2023). Cracking the Floral Quartet Code: How Do Multimers of MIKCC-Type MADS-Domain Transcription Factors Recognize Their Target Genes? International Journal of Molecular Sciences 2023, Vol. 24, Page 8253, 24(9), 8253. https://doi.org/10.3390/IJMS24098253

Kaufmann, K., Wellmer, F., Muiño, J. M., Ferrier, T., Wuest, S. E., Kumar, V., Serrano-Mislata, A., Madueño, F., Krajewski, P., Meyerowitz, E. M., Angenent, G. C., & Riechmann, J. L. (2010). Orchestration of floral initiation by APETALA1. Science, 328(85), 85–89. https://doi.org/10.1126/science.1185244

Kawamoto, N., Del Carpio, D. P., Hofmann, A., Mizuta, Y., Kurihara, D., Higashiyama, T., … Simon, R. (2020). A Peptide Pair Coordinates Regular Ovule Initiation Patterns with Seed Number and Fruit Size. Current Biology, 30(22), 4352-4361 e4354. https://doi.org/10.1016/j.cub.2020.08.050

Keller, M., Simm, S., Bokszczanin, K. L., Bostan, H., Bovy, A., Chaturvedi, P., Chen, Y., Chiusano, M. L., Firon, N., Fragkostefanakis, S., Iannacone, R., Jegadeesan, S., Li, H., Mariani, C., Marko, D., Mesihovic, A., Müller, F., Paul, P., Paupiere, M., … Vriezen, W. (2018). The coupling of transcriptome and proteome adaptation during development and heat stress response of tomato pollen. BMC Genomics, 19(1), 1–20. https://doi.org/10.1186/s12864-018-4824-5

Kempin, S. A., Savidge, B., & Yanofsky, M. F. (1995). Molecular basis of the cauliflower phenotype in Arabidopsis. Science, 267(5197), 522–525. https://doi.org/10.1126/SCIENCE.7824951

Kereszt, A., Mergaert, P., Montiel, J., Endre, G., & Kondorosi, E. (2018). Impact of Plant Peptides on Symbiotic Nodule Development and Functioning. Front Plant Sci, 9, 1026. https://doi.org/10.3389/fpls.2018.01026

Khitun, A., & Slavoff, S. A. (2019). Proteomic Detection and Validation of Translated Small Open Reading Frames. Curr Protoc Chem Biol, 11(4), e77. https://doi.org/10.1002/cpch.77

Kim, D., Šimo, L., & Park, Y. (2018). Molecular characterization of neuropeptide elevenin and two elevenin receptors, IsElevR1 and IsElevR2, from the blacklegged tick, Ixodes scapularis. Insect Biochemistry and Molecular Biology, 101, 66–75. https://doi.org/10.1016/j.ibmb.2018.07.005

Kim, Y. S., Kim, S. G., Lee, M., Lee, I., Park, H. Y., Seo, P. J., . . . Park, C. M. (2008). HD-ZIP III activity is modulated by competitive inhibitors via a feedback loop in Arabidopsis shoot apical meristem development. Plant Cell, 20(4), 920-933. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18408069

Kmiec, B., Branca, R. M. M., Berkowitz, O., Li, L., Wang, Y., Murcha, M. W., . . . Teixeira, P. F. (2018). Accumulation of endogenous peptides triggers a pathogen stress response in Arabidopsis thaliana. Plant J, 96(4), 705-715. https://doi.org/10.1111/tpj.14100

Koch, A., Gawron, D., Steyaert, S., Ndah, E., Crappe, J., De Keulenaer, S., . . . Menschaert, G. (2014). A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. Proteomics, 14(23-24), 2688-2698. https://doi.org/10.1002/pmic.201400180

Koehler, G., Rohloff, J., Wilson, R. C., Kopka, J., Erban, A., Winge, P., Bones, A. M., Davik, J., Alsheikh, M. K., & Randall, S. K. (2015). Integrative "omic" analysis reveals distinctive cold responses in leaves and roots of strawberry, fragaria × ananassa 'Korona.' Frontiers in Plant Science, 6(OCTOBER), 1–21. https://doi.org/10.3389/fpls.2015.00826

Koh, M., Ahmad, I., Ko, Y., Zhang, Y., Martinez, T. F., Diedrich, J. K., . . . Bollong, M. J. (2021). A short ORF-encoded transcriptional regulator. Proc Natl Acad Sci U S A, 118(4), e2021943118. https://doi.org/10.1073/pnas.2021943118

Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., & Kageyama, Y. (2007). Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nature Cell Biology, 9(6), 660-665. https://doi.org/10.1038/ncb1595

Kuljanin, M., Dieters-Castator, D. Z., Hess, D. A., Postovit, L.-M., & Lajoie, G. A. (2017). Comparison of sample preparation techniques for large-scale proteomics. Proteomics, 17(1–2), 1600337. https://doi.org/10.1002/pmic.201600337

Kumar, D., Bansal, G., Narang, A., Basak, T., Abbas, T., & Dash, D. (2016). Integrating transcriptome and proteome profiling: Strategies and applications. Proteomics, 16(19), 2533–2544. https://doi.org/10.1002/pmic.201600140

Kumar, M., Carr, P., & Turner, S. R. (2022). An atlas of Arabidopsis protein S-acylation reveals its widespread role in plant cell organization and function. Nature Plants, 8(6), 670–681. https://doi.org/10.1038/s41477-022-01164-4

Kumar, S., Stecher, G., Peterson, D., & Tamura, K. (2012). Sequence analysis MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. 28(20), 2685–2686. https://doi.org/10.1093/bioinformatics/bts507

Kurihara, Y., Makita, Y., Shimohira, H., Fujita, T., Iwasaki, S., & Matsui, M. (2020). Translational landscape of protein-coding and non-protein-coding RNAs upon light exposure in Arabidopsis. Plant and Cell Physiology, 61(3), 536–545. https://doi.org/10.1093/pcp/pcz219

Kurihara, Y., Makita, Y., Shimohira, H., Fujita, T., Iwasaki, S., & Matsui, M. (2020). Translational Landscape of Protein-Coding and Non-Protein-Coding RNAs upon Light Exposure in Arabidopsis. Plant Cell Physiol, 61(3), 536-545. https://doi.org/10.1093/pcp/pcz219

Kute, P. M., Soukarieh, O., Tjeldnes, H., Tregouet, D. A., & Valen, E. (2021). Small Open Reading Frames, How to Find Them and Determine Their Function. Front Genet, 12, 796060. https://doi.org/10.3389/fgene.2021.796060

Ladoukakis, E., Pereira, V., Magny, E., Eyre-Walker, A., & Couso, J. P. (2011). Hundreds of putatively functional small open reading frames in Drosophila. Genome Biol, 12(11), R118. https://doi.org/gb-2011-12-11-r118 [pii]

Lan, Z., Song, Z., Wang, Z., Li, L., Liu, Y., Zhi, S., . . . Qu, L. J. (2023). Antagonistic RALF peptides control an intergeneric hybridization barrier on Brassicaceae stigmas. Cell, 186(22), 4773-4787 e4712. https://doi.org/10.1016/j.cell.2023.09.003

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics, 9. https://doi.org/10.1186/1471-2105-9-559

Lauressergues, D., Couzigou, J. M., Clemente, H. S., Martinez, Y., Dunand, C., Becard, G., & Combier, J. P. (2015). Primary transcripts of microRNAs encode regulatory peptides. Nature, 520(7545), 90-93. https://doi.org/10.1038/nature14346

Lauressergues, D., Ormancey, M., Guillotin, B., San Clemente, H., Camborde, L., Duboe, C., . . . Combier, J. P. (2022). Characterization of plant microRNA-encoded peptides (miPEPs) reveals molecular mechanisms from the translation to activity and specificity. Cell Rep, 38(6), 110339. https://doi.org/10.1016/j.celrep.2022.110339

Lazar, C., Gatto, L., Ferro, M., Bruley, C., & Burger, T. (2016). Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. Journal of Proteome Research, 15, 1116–1125. https://doi.org/10.1021/acs.jproteome.5b00981

Le Signor, C., Aimé, D., Bordat, A., Belghazi, M., Labas, V., Gouzy, J., Young, N. D., Prosperi, J. M., Leprince, O., Thompson, R. D., Buitink, J., Burstin, J., & Gallardo, K. (2017). Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. New Phytologist, 214(4), 1597–1613. https://doi.org/10.1111/nph.14500

Le, J., Mallery, E. L., Zhang, C., Brankle, S., & Szymanski, D. B. (2006). Arabidopsis BRICK1/HSPC300 is an essential WAVE-complex subunit that selectively stabilizes the Arp2/3 activator SCAR2. Current Biology, 16(9), 895-901. https://doi.org/S0960-9822(06)01361-3 [pii]

Leblanc, S., Brunet, M. A., Jacques, J. F., Lekehal, A. M., Duclos, A., Tremblay, A., . . . Roucou, X. (2022). Newfound Coding Potential of Transcripts Unveils Missing Members of Human Protein Communities. Genomics Proteomics Bioinformatics, S1672-0229(1622)00124-00123. https://doi.org/10.1016/j.gpb.2022.09.008

Lee, C. Q. E., Kerouanton, B., Chothani, S., Zhang, S., Chen, Y., Mantri, C. K., . . . Ho, L. (2021). Coding and non-coding roles of MOCCI (C15ORF48) coordinate to regulate host inflammation and immunity. Nat Commun, 12(1), 2130. https://doi.org/10.1038/s41467-021-22397-5

Lee, C., Yen, K., & Cohen, P. (2013). Humanin: a harbinger of mitochondrial-derived peptides? Trends Endocrinol Metab, 24(5), 222-228. https://doi.org/10.1016/j.tem.2013.01.005

Lee, C., Zeng, J., Drew, B. G., Sallam, T., Martin-Montalvo, A., Wan, J., . . . Cohen, P. (2015). The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell Metab, 21(3), 443-454. https://doi.org/10.1016/j.cmet.2015.02.009

Lehmann, B. D., Colaprico, A., Silva, T. C., Chen, J., An, H., Ban, Y., Huang, H., Wang, L., James, J. L., Balko, J. M., Gonzalez-Ericsson, P. I., Sanders, M. E., Zhang, B., Pietenpol, J. A., & Steven Chen, X. (2021). Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes. Nature Communications, 12, 6276. https://doi.org/10.1038/s41467-021-26502-6

Lei, Y., Lu, L., Liu, H. Y., Li, S., Xing, F., & Chen, L. L. (2014). CRISPR-P: A Web Tool for Synthetic Single-Guide RNA Design of CRISPR-System in Plants. Molecular Plant, 7(9), 1494–1496. https://doi.org/10.1093/MP/SSU044

Leon, J., Rojo, E., Titarenko, E., & Sánchez-Serrano, J. J. (1998). Jasmonic acid-dependent and -independent wound signal transduction pathways are differentially regulated by Ca2+/calmodulin in Arabidopsis thaliana. Molecular and General Genetics, 258(4), 412–419. https://doi.org/10.1007/s004380050749

Li, H., Hu, C., Bai, L., Li, H., Li, M., Zhao, X., . . . Shao, Z. (2016). Ultra-deep sequencing of ribosome-associated poly-adenylated RNA in early Drosophila embryos reveals hundreds of conserved translated sORFs. DNA Research, 23(6), 571-580. https://doi.org/10.1093/dnares/dsw040

Li, K., Li, B., Zhang, D., Du, T., Zhou, H., Dai, G., . . . Huang, Z. P. (2023). The translational landscape of human vascular smooth muscle cells identifies novel short ORF-encoded peptide regulators for phenotype alteration. Cardiovasc Res, 119(8), 1763-1779. https://doi.org/10.1093/cvr/cvad044

Li, M., Shao, F., Qian, Q., Yu, W., Zhang, Z., Chen, B., . . . Cao, H. (2021). A putative long noncoding RNA-encoded micropeptide maintains cellular homeostasis in pancreatic beta cells. Mol Ther Nucleic Acids, 26, 307-320. https://doi.org/10.1016/j.omtn.2021.06.027

Li, Y. R., & Liu, M. J. (2020). Prevalence of alternative AUG and non-AUG translation initiators and their regulatory effects across plants. Genome Research, 30(10), 1418–1433. https://doi.org/10.1101/GR.261834.120

Li, Y., Zhou, H., Chen, X., Zheng, Y., Kang, Q., Hao, D., . . . He, S. (2021). SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. Genomics Proteomics Bioinformatics, 19(4), 602-610. https://doi.org/10.1016/j.gpb.2021.09.002

Liang, Y., Zhu, W., Chen, S., Qian, J., & Li, L. (2021). Genome-Wide Identification and Characterization of Small Peptides in Maize. Front Plant Sci, 12, 695439. https://doi.org/10.3389/fpls.2021.695439

Liljegren, S. J., Gustafson-Brown, C., Pinyopich, A., Ditta, G. S., & Yanofsky, M. F. (1999). Interactions among APETALA1, LEAFY, and TERMINAL FLOWER1 Specify Meristem Fate. The Plant Cell, 11(6), 1007. https://doi.org/10.2307/3870794

Lin, H., Niu, L., McHale, N. A., Ohme-Takagi, M., Mysore, K. S., & Tadege, M. (2013). Evolutionarily conserved repressive activity of WOX proteins mediates leaf blade outgrowth and floral organ development in plants. Proceedings of the National Academy of Sciences of the United States of America, 110(1), 366–371. https://doi.org/10.1073/pnas.1215376110

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 27(13), i275-282. https://doi.org/10.1093/bioinformatics/btr209

Lin, N., Chang, K. Y., Li, Z., Gates, K., Rana, Z. A., Dang, J., . . . Rana, T. M. (2014). An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Molecular Cell, 53(6), 1005-1019. https://doi.org/10.1016/j.molcel.2014.01.021

Lin, X., Lin, W., Ku, Y. S., Wong, F. L., Li, M. W., Lam, H. M., Ngai, S. M., & Chan, T. F. (2020). Analysis of Soybean Long Non-Coding RNAs Reveals a Subset of Small Peptide-Coding Transcripts. Plant Physiology, 182(3), 1359–1374. https://doi.org/10.1104/PP.19.01324

Lindeboom, R. G., van Voorthuijsen, L., Oost, K. C., Rodríguez-Colman, M. J., Luna-Velez, M. V, Furlan, C., Baraille, F., Jansen, P. W., Ribeiro, A., Burgering, B. M., Snippert, H. J., & Vermeulen, M. (2018). Integrative multi-omics analysis of intestinal organoid differentiation. Molecular Systems Biology, 14(6), 1–16. https://doi.org/10.15252/msb.20188227

Lischer, H. E. L., Excoffier, L., & Kelso, J. (2012). Data and text mining PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. BIOINFORMATICS APPLICATIONS NOTE, 28(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

Liu, C., Shen, L., Xiao, Y., Vyshedsky, D., Peng, C., Sun, X., . . . Li, C. (2021). Pollen PCP-B peptides unlock a stigma peptide-receptor kinase gating mechanism for pollination. Science, 372(6538), 171-175. https://doi.org/10.1126/science.abc6107

Liu, J., Mehdi, S., Topping, J., Tarkowski, P., & Lindsey, K. (2010). Modelling and experimental analysis of hormonal crosstalk in Arabidopsis. Mol Syst Biol, 6, 373. https://doi.org/10.1038/msb.2010.26

Liu, M. J., Wu, S. H., Wu, J. F., Lin, W. D., Wu, Y. C., Tsai, T. Y., . . . Wu, S. H. (2013). Translational landscape of photomorphogenic Arabidopsis. Plant Cell, 25(10), 3699-3710. https://doi.org/10.1105/tpc.113.114769

Liu, X., Cao, X., Shi, S., Zhao, N., Li, D., Fang, P., Chen, X., Qi, W., & Zhang, Z. (2018). Comparative RNA-Seq analysis reveals a critical role for brassinosteroids in rose (Rosa hybrida) petal defense against Botrytis cinerea infection. BMC Genetics, 19(1), 1–10. https://doi.org/10.1186/s12863-018-0668-x

Liu, Yansheng, Beyer, A., & Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell, 165(3), 535–550. https://doi.org/10.1016/j.cell.2016.03.014

Liu, Yinghao, Fan, W., Cheng, Q., Zhang, L., Cai, T., Shi, Q., Wang, Z., Chang, C., Yin, Q., Jiang, X., & Jin, K. (2022). Multi-omics analyses reveal new insights into nutritional quality changes of alfalfa leaves during the flowering period. Frontiers in Plant Science, 13(November), 1–13. https://doi.org/10.3389/fpls.2022.995031

Liu, Z., Hou, S., Rodrigues, O., Wang, P., Luo, D., Munemasa, S., . . . Shan, L. (2022). Phytocytokine signalling reopens stomata in plant immunity and water loss. Nature, 605(7909), 332-339. https://doi.org/10.1038/s41586-022-04684-3

Lu, S., Zhang, J., Lian, X., Sun, L., Meng, K., Chen, Y., . . . He, Q. Y. (2019). A hidden human proteome encoded by 'non-coding' genes. Nucleic Acids Res, 47(15), 8111-8125. https://doi.org/10.1093/nar/gkz646

Luo, W., Xiao, Y., Liang, Q., Su, Y., & Xiao, L. (2019). Identification of Potential Auxin-Responsive Small Signaling Peptides through a Peptidomics Approach in Arabidopsis thaliana. Molecules, 24(17), 3146. https://doi.org/10.3390/molecules24173146

Luo, X., Cao, D., Zhang, J., Chen, L., Xia, X., Li, H., Zhao, D., Zhang, F., Xue, H., Chen, L., Li, Y., & Cao, S. (2018). Integrated microRNA and mRNA expression profiling reveals a complex network regulating pomegranate (Punica granatum L.) seed hardness. Scientific Reports, 8(1), 1–14. https://doi.org/10.1038/s41598-018-27664-y

Lyapina, I., Filippova, A., & Fesenko, I. (2019). The Role of Peptide Signals Hidden in the Structure of Functional Proteins in Plant Immune Responses. Int J Mol Sci, 20(18), 4343. https://doi.org/10.3390/ijms20184343

Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., . . . Saghatelian, A. (2016). Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. Analytical Chemistry, 88(7), 3967-3975. https://doi.org/10.1021/acs.analchem.6b00191

Ma, J., Saghatelian, A., & Shokhirev, M. N. (2018). The influence of transcript assembly on the proteogenomics discovery of microproteins. PLoS ONE, 13(3), e0194518. https://doi.org/10.1371/journal.pone.0194518

Ma, J., Ward, C. C., Jungreis, I., Slavoff, S. A., Schwaid, A. G., Neveu, J., Budnik, B. A., Kellis, M., & Saghatelian, A. (2014). Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. Journal of Proteome Research, 13(3), 1757–1765. https://doi.org/10.1021/pr401280w

Ma, J., Yan, B., Qu, Y., Qin, F., Yang, Y., Hao, X., . . . Ao, G. (2008). Zm401, a short-open reading-frame mRNA or noncoding RNA, is essential for tapetum and microspore development and can regulate the floret formation in maize. Journal of Cellular Biochemistry, 105(1), 136-146. https://doi.org/10.1002/jcb.21807

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., . . . Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. Genome Biol, 16, 179. https://doi.org/10.1186/s13059-015-0742-x

Magnani, E., de Klein, N., Nam, H. I., Kim, J. G., Pham, K., Fiume, E., . . . Rhee, S. Y. (2014). A comprehensive analysis of microProteins reveals their potentially widespread mechanism of transcriptional regulation. Plant Physiol, 165(1), 149-159. https://doi.org/10.1104/pp.114.235903

Magnani, E., Sjolander, K., & Hake, S. (2004). From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. Plant Cell, 16(9), 2265-2277. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15319480

Magny, E. G., Pueyo, J. I., Pearl, F. M., Cespedes, M. A., Niven, J. E., Bishop, S. A., & Couso, J. P. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science, 341(6150), 1116-1120. https://doi.org/10.1126/science.1238802

Makarewich, C. A., & Olson, E. N. (2017). Mining for Micropeptides. Trends in Cell Biology, 27(9), 685–696. https://doi.org/10.1016/j.tcb.2017.04.006

Mamaeva, A., Taliansky, M., Filippova, A., Love, A. J., Golub, N., & Fesenko, I. (2020). The role of chloroplast protein remodeling in stress responses and shaping of the plant peptidome. New Phytologist, 227(5), 1326-1334. https://doi.org/10.1111/nph.16620

Mandel, M. A., Gustafson-Brown, C., Savidge, B., & Yanofsky, M. F. (1992). Molecular characterization of the Arabidopsis floral homeotic gene APETALA1. Nature 1992 360:6401, 360(6401), 273–277. https://doi.org/10.1038/360273a0

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2018). Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. Briefings in Bioinformatics, 19(2), 286–302. https://doi.org/10.1093/BIB/BBW114

Mardis, E. R. (2017). DNA sequencing technologies: 2006-2016. Nature Protocols, 12(2), 213–218. https://doi.org/10.1038/nprot.2016.182

Martínez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., & Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. Methods Molecular Biology, 16(4), 458–468. https://doi.org/10.1038/s41589-019-0425-0

Martinez, T. F., Lyons-Abbott, S., Bookout, A. L., De Souza, E. V., Donaldson, C., Vaughan, J. M., . . . Barnes, C. A. (2023). Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. Cell Metab, 35(1), 166-183 e111. https://doi.org/10.1016/j.cmet.2022.12.004

Marx, H., Minogue, C. E., Jayaraman, D., Richards, A. L., Kwiecien, N. W., Siahpirani, A. F., Rajasekar, S., Maeda, J., Garcia, K., Del Valle-Echevarria, A. R., Volkening, J. D., Westphall, M. S., Roy, S., Sussman, M. R., Ané, J. M., & Coon, J. J. (2016). A proteomic atlas of the legume, M. truncatula, and its nitrogen fixing endosymbiont, S. meliloti. Nature Biotechnology, 34(11), 1198. https://doi.org/10.1038/NBT.3681

Mateos, J. L., Tilmes, V., Madrigal, P., Severing, E., Richter, R., Rijkenberg, C. W. M., Krajewski, P., & Coupland, G. (2017). Divergence of regulatory networks governed by the orthologous transcription factors FLC and PEP1 in Brassicaceae species. Proceedings of the National Academy of Sciences of the United States of America, 114(51), E11037–E11046. https://doi.org/10.1073/pnas.1618075114

Matsubayashi, Y. (2011). Small post-translationally modified Peptide signals in Arabidopsis. Arabidopsis Book, 9, e0150. https://doi.org/10.1199/tab.0150

Matsubayashi, Y. (2014). Posttranslationally modified small-peptide signals in plants. Annual Review of Plant Biology, 65, 385-413. https://doi.org/10.1146/annurev-arplant-050312-120122

Matsubayashi, Y. (2018). Exploring peptide hormones in plants: identification of four peptide hormone-receptor pairs and two post-translational modification enzymes. Proc Jpn Acad Ser B Phys Biol Sci, 94(2), 59-74. https://doi.org/10.2183/pjab.94.006

Matsubayashi, Y., Ogawa, M., Kihara, H., Niwa, M., & Sakagami, Y. (2006). Disruption and overexpression of Arabidopsis phytosulfokine receptor gene affects cellular longevity and potential for growth. Plant Physiol, 142(1), 45-53. https://doi.org/10.1104/pp.106.081109

Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., . . . Pandolfi, P. P. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature, 541(7636), 228-232. https://doi.org/10.1038/nature21034

Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L. L., . . . Wu, M. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. Nature Reviews Molecular Cell Biology. https://doi.org/10.1038/s41580-022-00566-8

Matus, J. T. (2016). Transcriptomic and metabolomic networks in the grape berry illustrate that it takes more than flavonoids to fight against ultraviolet radiation. Frontiers in Plant Science, 7(AUG2016), 1337. https://doi.org/10.3389/FPLS.2016.01337/BIBTEX

Meleth, S., Deshane, J., & Kim, H. (2005). The case for well-conducted experiments to validate statistical protocols for 2D gels: Different pre-processing = different lists of significant proteins. BMC Biotechnology, 5, 1–15. https://doi.org/10.1186/1472-6750-5-7

Melzer, R., Verelst, W., & Theißen, G. (2009). The class E floral homeotic protein SEPALLATA3 is sufficient to loop DNA in 'floral quartet'-like complexes in vitro. Nucleic Acids Research, 37(1), 144–157. https://doi.org/10.1093/NAR/GKN900

Mendes, M. A., Guerra, R. F., Berns, M. C., Manzo, C., Masiero, S., Finzi, L., Kater, M. M., & Colombo, L. (2013). MADS Domain Transcription Factors Mediate Short-Range DNA Looping That Is Essential for Target Gene Expression in Arabidopsis. The Plant Cell, 25(7), 2560–2572. https://doi.org/10.1105/TPC.112.108688

Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappe, J., Gevaert, K., & Van Damme, P. (2013). Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. Molecular and Cellular Proteomics, 12(7), 1780-1790. https://doi.org/10.1074/mcp.M113.027540

Mergner, J., & Kuster, B. (2022). Plant Proteome Dynamics. Annual Review of Plant Biology, 73, 67–92. https://doi.org/10.1146/annurev-arplant-102620-031308

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., . . . Kuster, B. (2020). Mass-spectrometry-based draft of the Arabidopsis proteome. Nature, 579(7799), 409-414. https://doi.org/10.1038/s41586-020-2094-2

Mise, S., Matsumoto, A., Shimada, K., Hosaka, T., Takahashi, M., Ichihara, K., . . . Nakayama, K. I. (2022). Kastor and Polluks polypeptides encoded by a single gene locus cooperatively

regulate VDAC and spermatogenesis. Nat Commun, 13(1), 1071. https://doi.org/10.1038/s41467-022-28677-y

Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2019). Integrated omics: Tools, advances and future approaches. Journal of Molecular Endocrinology, 62(1), R21–R45. https://doi.org/10.1530/JME-18-0055

Mohsen, J. J., Martel, A. A., & Slavoff, S. A. (2023). Microproteins-Discovery, structure, and function. Proteomics, e2100211. https://doi.org/10.1002/pmic.202100211

Moore, S., Zhang, X., Mudge, A., Rowe, J. H., Topping, J. F., Liu, J., & Lindsey, K. (2015). Spatiotemporal modelling of hormonal crosstalk explains the level and patterning of hormones and gene expression in Arabidopsis thaliana wild-type and mutant roots. New Phytologist, 207(4), 1110-1122. https://doi.org/10.1111/nph.13421

Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., Bader, G. D., & Ferrin, T. E. (2011). ClusterMaker: A multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics, 12, 1–14. https://doi.org/10.1186/1471-2105-12-436

Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I., . . . van Heesch, S. (2022). Standardized annotation of translated open reading frames. Nat Biotechnol, 40(7), 994-999. https://doi.org/10.1038/s41587-022-01369-0

Na, C. H., Barbhuiya, M. A., Kim, M. S., Verbruggen, S., Eacker, S. M., Pletnikova, O., Troncoso, J. C., Halushka, M. K., Menschaert, G., Overall, C. M., & Pandey, A. (2018). Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. Genome Research, 28(1), 25–36. https://doi.org/10.1101/GR.226050.117

Na, Z., Luo, Y., Schofield, J. A., Smelyansky, S., Khitun, A., Muthukumar, S., . . . Slavoff, S. A. (2020). The NBDY Microprotein Regulates Cellular RNA Decapping. Biochemistry, 59(42), 4131-4142. https://doi.org/10.1021/acs.biochem.0c00672

Nagy, G., & Nagy, L. (2020). Motif grammar: The basis of the language of gene expression. Computational and Structural Biotechnology Journal, 18, 2026–2032. https://doi.org/10.1016/J.CSBJ.2020.07.007

Najafi, J., Brembu, T., Vie, A. K., Viste, R., Winge, P., Somssich, I. E., & Bones, A. M. (2020). PAMP-INDUCED SECRETED PEPTIDE 3 modulates immunity in Arabidopsis. Journal of Experimental Botany, 71(3), 850-864. https://doi.org/10.1093/jxb/erz482

Nakaminami, K., Okamoto, M., Higuchi-Takeuchi, M., Yoshizumi, T., Yamaguchi, Y., Fukao, Y., . . . Hanada, K. (2018). AtPep3 is a hormone-like peptide that plays a role in the salinity stress tolerance of plants. Proc Natl Acad Sci U S A, 115(22), 5810-5815. https://doi.org/10.1073/pnas.1719491115

Nakayama, T., Shinohara, H., Tanaka, M., Baba, K., Ogawa-Ohnishi, M., & Matsubayashi, Y. (2017). A peptide hormone required for Casparian strip diffusion barrier formation in Arabidopsis roots. Science, 355(6322), 284-286. https://doi.org/10.1126/science.aai9057

Nanjo, Y., Maruyama, K., Yasue, H., Yamaguchi-Shinozaki, K., Shinozaki, K., & Komatsu, S. (2011). Transcriptional responses to flooding stress in roots including hypocotyl of soybean seedlings. Plant Mol Biol, 77(1-2), 129-144. https://doi.org/10.1007/s11103-011-9799-4

Narita, N. N., Moore, S., Horiguchi, G., Kubo, M., Demura, T., Fukuda, H., . . . Tsukaya, H. (2004). Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation

and alters leaf shape in Arabidopsis thaliana. Plant J, 38(4), 699-713. https://doi.org/10.1111/j.1365-313X.2004.02078.x

Navrot, N., Finnie, C., Svensson, B., & Hägglund, P. (2011). Plant redox proteomics. Journal of Proteomics, 74(8), 1450–1462. https://doi.org/10.1016/j.jprot.2011.03.008

Nelson, B. R., Makarewich, C. A., Anderson, D. M., Winders, B. R., Troupes, C. D., Wu, F., . . . Olson, E. N. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science, 351(6270), 271-275. https://doi.org/10.1126/science.aad4076

Neme, R., Amador, C., Yildirim, B., McConnell, E., & Tautz, D. (2017). Random sequences are an abundant source of bioactive RNAs or peptides. Nat Ecol Evol, 1(6), 0217. https://doi.org/10.1038/s41559-017-0127

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics, 73(11), 2092-2123. https://doi.org/10.1016/j.jprot.2010.08.009

Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. Nat Methods, 11(11), 1114-1125. https://doi.org/10.1038/nmeth.3144

Neville, M. D. C., Kohze, R., Erady, C., Meena, N., Hayden, M., Cooper, D. N., . . . Prabakaran, S. (2021). A platform for curated products from novel open reading frames prompts reinterpretation of disease variants. Genome Research, 31(2), 327-336. https://doi.org/10.1101/gr.263202.120

Ng, M., & Yanofsky, M. F. (2001). Activation of the Arabidopsis B Class Homeotic Genes by APETALA1. 13(April), 739–753.

Niu, Z., Liu, L., Pu, Y., Ma, L., Wu, J., Hu, F., Fang, Y., Li, X., Sun, W., Wang, W., & Bai, C. (2021). iTRAQ-based quantitative proteome analysis insights into cold stress of Winter Rapeseed (Brassica rapa L.) grown in the field. Scientific Reports (Nature), 11, 23434. https://doi.org/10.1038/s41598-021-02707-z

Ó'Maoiléidigh, D. S., Thomson, B., & Wellmer, F. (2023). Floral Induction Systems for the Study of Arabidopsis Flower Development. In Methods in molecular biology (Clifton, N.J.) (Vol. 2686, pp. 285–292). NLM (Medline). https://doi.org/10.1007/978-1-0716-3299-4_12/TABLES/1

Ó'Maoiléidigh, D. S., Wuest, S. E., Rae, L., Raganelli, A., Ryan, P. T., Kwaśniewska, K., Das, P., Lohan, A. J., Loftus, B., Graciet, E., & Wellmer, F. (2013). Control of Reproductive Floral Organ Identity Specification in Arabidopsis by the C Function Regulator AGAMOUS. The Plant Cell, 25(7), 2482–2503. https://doi.org/10.1105/TPC.113.113209

Ogawa-Ohnishi, M., Yamashita, T., Kakita, M., Nakayama, T., Ohkubo, Y., Hayashi, Y., . . . Matsubayashi, Y. (2022). Peptide ligand-mediated trade-off between plant growth and stress response. Science, 378(6616), 175-180. https://doi.org/10.1126/science.abq5735

Ogilvie, H. A., Imin, N., & Djordjevic, M. A. (2014). Diversification of the C-TERMINALLY ENCODED PEPTIDE (CEP) gene family in angiosperms, and evolution of plant-family specific CEP genes. BMC Genomics, 15(1), 870. https://doi.org/10.1186/1471-2164-15-870

Ohyama, K., Ogawa, M., & Matsubayashi, Y. (2008). Identification of a biologically active, small, secreted peptide in Arabidopsis by in silico gene screening, followed by LC-MS-based

structure analysis. Plant J, 55(1), 152-160. https://doi.org/10.1111/j.1365-313X.2008.03464.x

Okamoto, M., Higuchi-Takeuchi, M., Shimizu, M., Shinozaki, K., & Hanada, K. (2014). Substantial expression of novel small open reading frames in Oryza sativa. Plant Signal Behav, 9(2), e27848. https://doi.org/10.4161/psb.27848

Okuda, S., Tsutsui, H., Shiina, K., Sprunck, S., Takeuchi, H., Yui, R., . . . Higashiyama, T. (2009). Defensin-like polypeptide LUREs are pollen tube attractants secreted from synergid cells. Nature, 458(7236), 357-361. https://doi.org/10.1038/nature07882

Olexiouk, V., Van Criekinge, W., & Menschaert, G. (2018). An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res, 46(D1), D497-D502. https://doi.org/10.1093/nar/gkx1130

Oliva-Vilarnau, N., Vorrink, S. U., Ingelman-Sundberg, M., & Lauschke, V. M. (2020). A 3D Cell Culture Model Identifies Wnt/β-Catenin Mediated Inhibition of p53 as a Critical Step during Human Hepatocyte Regeneration. Advanced Science, 7(15), 1–12. https://doi.org/10.1002/advs.202000248

Omidbakhshfard, M. A., Sokolowska, E. M., Di Vittori, V., Perez de Souza, L., Kuhalskaya, A., Brotman, Y., Alseekh, S., Fernie, A. R., & Skirycz, A. (2021). Multi-omics analysis of early leaf development in Arabidopsis thaliana. Patterns, 2(4). https://doi.org/10.1016/J.PATTER.2021.100235

Ormancey, M., Thuleau, P., Combier, J. P., & Plaza, S. (2023). The Essentials on microRNA-Encoded Peptides from Plants to Animals. Biomolecules, 13(2). https://doi.org/10.3390/biom13020206

Orr, M. W., Mao, Y., Storz, G., & Qian, S. B. (2020). Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res, 48(3), 1029-1042. https://doi.org/10.1093/nar/gkz734

Ouspenskaia, T., Law, T., Clauser, K. R., Klaeger, S., Sarkizova, S., Aguet, F., . . . Regev, A. (2022). Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat Biotechnol, 40(2), 209-217. https://doi.org/10.1038/s41587-021-01021-3

Pajoro, A., Biewers, S., Dougali, E., Valentim, F. L., Mendes, M. A., Porri, A., Coupland, G., Van De Peer, Y., Van Dijk, A. D. J., Colombo, L., Davies, B., & Angenent, G. C. (2014). The (r)evolution of gene regulatory networks controlling Arabidopsis plant reproduction: A two-decade history. Journal of Experimental Botany, 65(17), 4731–4745. https://doi.org/10.1093/jxb/eru233

Pajoro, A., Madrigal, P., Muiño, J. M., Matus, J. T., Jin, J., Mecchia, M. A., Debernardi, J. M., Palatnik, J. F., Balazadeh, S., Arif, M., Ó'Maoiléidigh, D. S. Ó., Wellmer, F., Krajewski, P., Riechmann, J., Angenent, G. C., & Kaufmann, K. (2014). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biology, 15(R41). https://doi.org/10.1186/gb-2014-15-3-r41

Palos, K., Dittrich, A. C. N., Ang Yu, L. ', Brock, J. R., Railey, C. E., Wu, H.-Y. L., Sokolowska, E., Skirycz, A., Hsu, P. Y., Gregory, B. D., Lyons, E., Beilstein, M. A., & Nelson, A. D. L. (2022). Identification and functional annotation of long intergenic non-coding RNAs in Brassicaceae. The Plant Cell, 34, 3233–3260. https://doi.org/10.1093/plcell/koac166

Pan, H., & Wang, D. (2017). Nodule cysteine-rich peptides maintain a working balance during nitrogen-fixing symbiosis. Nat Plants, 3(5), 17048. https://doi.org/10.1038/nplants.2017.48

Patel, N., Mohd-Radzman, N. A., Corcilius, L., Crossett, B., Connolly, A., Cordwell, S. J., . . . Djordjevic, M. A. (2018). Diverse Peptide Hormones Affecting Root Growth Identified in the Medicago truncatula Secreted Peptidome. Molecular and Cellular Proteomics, 17(1), 160-174. https://doi.org/10.1074/mcp.RA117.000168

Pathan, M., Samuel, M., Keerthikumar, S., & Mathivanan, S. (2017). Unassigned MS/MS Spectra: Who Am I? In Methods in Molecular Biology (Vol. 1549, pp. 67–74). Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-6740-7_6

Pearce, G., Bhattacharya, R., Chen, Y. C., Barona, G., Yamaguchi, Y., & Ryan, C. A. (2009). Isolation and characterization of hydroxyproline-rich glycopeptide signals in black nightshade leaves. Plant Physiol, 150(3), 1422-1433. https://doi.org/10.1104/pp.109.138669

Pearce, G., Moura, D. S., Stratmann, J., & Ryan, C. A. (2001). Production of multiple plant hormones from a single polyprotein precursor. Nature, 411(6839), 817-820. https://doi.org/10.1038/35081107

Pearce, G., Siems, W. F., Bhattacharya, R., Chen, Y. C., & Ryan, C. A. (2007). Three hydroxyproline-rich glycopeptides derived from a single petunia polyprotein precursor activate defensin I, a pathogen defense response gene. J Biol Chem, 282(24), 17777-17784. https://doi.org/10.1074/jbc.M701543200

Pearce, G., Yamaguchi, Y., Barona, G., & Ryan, C. A. (2010). A subtilisin-like protein from soybean contains an embedded, cryptic signal that activates defense-related genes. Proc Natl Acad Sci U S A, 107(33), 14921-14925. https://doi.org/10.1073/pnas.1007568107

Peeters, M. K. R., & Menschaert, G. (2020). The hunt for sORFs: A multidisciplinary strategy. Exp Cell Res, 391(1), 111923. https://doi.org/10.1016/j.yexcr.2020.111923

Pelaz, S., Ditta, G. S., Baumann, E., Wisman, E., & Yanofsky, M. F. (2000). B and C floral organ identity functions require SEPALLATA MADS-box genes. Nature, 405(6783), 200–203. https://doi.org/10.1038/35012103

Peng, Z., He, S., Gong, W., Xu, F., Pan, Z., Jia, Y., Geng, X., & Du, X. (2018). Integration of proteomic and transcriptomic profiles reveals multiple levels of genetic regulation of salt tolerance in cotton. BMC Plant Biology, 18(128). https://doi.org/10.1186/s12870-018-1350-1

Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D. J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., Walzer, M., Wang, S., Brazma, A., & Vizcaíno, J. A. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Research, 50(D1), D543–D552. https://doi.org/10.1093/NAR/GKAB1038

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probablity-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20, 3551–3557. https://doi.org/10.1097/MAO.0000000000000737

Pinyopich, A., Ditta, G. S., Savidge, B., Liljegren, S. J., Baumann, E., Wisman, E., & Yanofsky, M. F. (2003). Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature, 424(6944), 85–88. https://doi.org/10.1038/NATURE01741

Plaza, S., Menschaert, G., & Payre, F. (2017). In Search of Lost Small Peptides. Annual Review of Cell and Developmental Biology, 33, 391-416. https://doi.org/10.1146/annurev-cellbio-100616-060516

Ponnala, L., Wang, Y., Sun, Q., & Van Wijk, K. J. (2014). Correlation of mRNA and protein abundance in the developing maize leaf. Plant Journal, 78(3), 424–440. https://doi.org/10.1111/TPJ.12482

Prensner, J. R., Abelin, J. G., Kok, L. W., Clauser, K. R., Mudge, J. M., Ruiz-Orera, J., . . . van Heesch, S. (2023). What Can Ribo-Seq, Immunopeptidomics, and Proteomics Tell Us About the Noncanonical Proteome? Molecular and Cellular Proteomics, 22(9), 100631. https://doi.org/10.1016/j.mcpro.2023.100631

Prensner, J. R., Enache, O. M., Luria, V., Krug, K., Clauser, K. R., Dempster, J. M., . . . Golub, T. R. (2021). Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol, 39(6), 697-704. https://doi.org/10.1038/s41587-020-00806-2

Qi, X., Yoshinari, A., Bai, P., Maes, M., Zeng, S. M., & Torii, K. U. (2020). The manifold actions of signaling peptides on subcellular dynamics of a receptor specify stomatal cell fate. elife, 9. https://doi.org/10.7554/eLife.58097

Quinn, M. E., Goh, Q., Kurosaka, M., Gamage, D. G., Petrany, M. J., Prasad, V., & Millay, D. P. (2017). Myomerger induces fusion of non-fusogenic cells and is required for skeletal muscle development. Nat Commun, 8, 15665. https://doi.org/10.1038/ncomms15665

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., & Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biology, 4(10), 1707–1719. https://doi.org/10.1371/journal.pbio.0040309

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., . . . Pritchard, J. K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. elife, 5, e13328. https://doi.org/10.7554/eLife.13328

Rathore, A., Martinez, T. F., Chu, Q., & Saghatelian, A. (2018). Small, but mighty? Searching for human microproteins and their potential for understanding health and disease. Expert Rev Proteomics, 15(12), 963-965. https://doi.org/10.1080/14789450.2018.1547194

Reeves, P. H., Ellis, C. M., Ploense, S. E., Wu, M. F., Yadav, V., Tholl, D., Chételat, A., Haupt, I., Kennerley, B. J., Hodgens, C., Farmer, E. E., Nagpal, P., & Reed, J. W. (2012). A regulatory network for coordinated flower maturation. PLoS Genetics, 8(2). https://doi.org/10.1371/JOURNAL.PGEN.1002506

Rhodes, J., Roman, A. O., Bjornson, M., Brandt, B., Derbyshire, P., Wyler, M., . . . Zipfel, C. (2022). Perception of a conserved family of plant signalling peptides by the receptor kinase HSL3. elife, 11. https://doi.org/10.7554/eLife.74687

Rich-Griffin, C., Stechemesser, A., Finch, J., Lucas, E., Ott, S., & Schäfer, P. (2020). Single-Cell Transcriptomics: A High-Resolution Avenue for Plant Functional Genomics. Trends in Plant Science, 25(2), 186–197. https://doi.org/10.1016/j.tplants.2019.10.008

Roberts, I., Smith, S., De Rybel, B., Van Den Broeke, J., Smet, W., De Cokere, S., . . . Beeckman, T. (2013). The CEP family in land plants: evolutionary analyses, expression studies, and role in Arabidopsis shoot development. Journal of Experimental Botany, 64(17), 5371-5381. https://doi.org/10.1093/jxb/ert331

Rodrigues, V. L., Dolde, U., Sun, B., Blaakmeer, A., Straub, D., Eguen, T., . . . Wenkel, S. (2021). A microProtein repressor complex in the shoot meristem controls the transition to flowering. Plant Physiol, 187(1), 187-202. https://doi.org/10.1093/plphys/kiab235

Rohrig, H., Schmidt, J., Miklashevichs, E., Schell, J., & John, M. (2002). Soybean ENOD40 encodes two peptides that bind to sucrose synthase. Proc Natl Acad Sci U S A, 99(4), 1915-1920. https://doi.org/10.1073/pnas.022664799

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Ohna, S. H. ¨, Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). Software for Systematics and Evolution MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst. Biol, 61(3), 539–542. https://doi.org/10.1093/sysbio/sys029

Rubio-Somoza, I., & Weigel, D. (2013). Coordination of Flower Maturation by a Regulatory Circuit of Three MicroRNAs. PLoS Genetics, 9(3). https://doi.org/10.1371/journal.pgen.1003374

Ruiz Cuevas, M. V., Hardy, M. P., Holly, J., Bonneil, E., Durette, C., Courcelles, M., . . . Yewdell, J. W. (2021). Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep, 34(10), 108815. https://doi.org/10.1016/j.celrep.2021.108815

Ruiz-Orera, J., & Alba, M. M. (2019). Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. Trends in Genetics, 35(3), 186-198. https://doi.org/10.1016/j.tig.2018.12.003

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., & Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. elife, 3, e03523. https://doi.org/10.7554/eLife.03523

Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Canas, J. L., Messeguer, X., & Alba, M. M. (2018). Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat Ecol Evol, 2(5), 890-896. https://doi.org/10.1038/s41559-018-0506-6

Ruiz-Orera, J., Villanueva-Canas, J. L., & Alba, M. M. (2020). Evolution of new proteins from translated sORFs in long non-coding RNAs. Exp Cell Res, 391(1), 111940. https://doi.org/10.1016/j.yexcr.2020.111940

Ryan, C. A., & Pearce, G. (1998). Systemin: a polypeptide signal for plant defensive genes. Annual Review of Cell and Developmental Biology, 14, 1-17. https://doi.org/10.1146/annurev.cellbio.14.1.1

Ryan, C. A., & Pearce, G. (2003). Systemins: a functionally defined family of peptide signals that regulate defensive genes in Solanaceae species. Proc Natl Acad Sci U S A, 100 Suppl 2(Suppl 2), 14577-14580. https://doi.org/10.1073/pnas.1934788100

Saha, G., Park, J. I., Kayum, M. A., & Nou, I. S. (2016). A genome-wide analysis reveals stress and hormone responsive patterns of TIFY family genes in Brassica rapa. Frontiers in Plant Science, 7(June), 1–18. https://doi.org/10.3389/fpls.2016.00936

Samandi, S., Roy, A. V., Delcourt, V., Lucier, J. F., Gagnon, J., Beaudoin, M. C., . . . Roucou, X. (2017). Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. elife, 6, e27860. https://doi.org/10.7554/eLife.27860

Samir, P., & Link, A. J. (2011). Analyzing the cryptome: uncovering secret sequences. AAPS J, 13(2), 152-158. https://doi.org/10.1208/s12248-011-9252-2

Sandmann, C. L., Schulz, J. F., Ruiz-Orera, J., Kirchner, M., Ziehm, M., Adami, E., . . . Hubner, N. (2023). Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Molecular Cell, 83(6), 994-1011 e1018. https://doi.org/10.1016/j.molcel.2023.01.023

Santiago, J., Brandt, B., Wildhagen, M., Hohmann, U., Hothorn, L. A., Butenko, M. A., & Hothorn, M. (2016). Mechanistic insight into a peptide hormone signaling complex mediating floral organ abscission. elife, 5. https://doi.org/10.7554/eLife.15075

Sauter, M. (2015). Phytosulfokine peptide signalling. Journal of Experimental Botany, 66(17), 5161-5169. https://doi.org/10.1093/jxb/erv071

Schlesinger, D., & Elsasser, S. J. (2022). Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. FEBS J, 289(1), 53-74. https://doi.org/10.1111/febs.15769

Schlotterer, C. (2015). Genes from scratch--the evolutionary fate of de novo genes. Trends in Genetics, 31(4), 215-219. https://doi.org/10.1016/j.tig.2015.02.007

Schmelz, E. A., Carroll, M. J., LeClere, S., Phipps, S. M., Meredith, J., Chourey, P. S., . . . Teal, P. E. (2006). Fragments of ATP synthase mediate plant perception of insect attack. Proc Natl Acad Sci U S A, 103(23), 8894-8899. https://doi.org/10.1073/pnas.0602328103

Schommer, C., Palatnik, J. F., Aggarwal, P., Chételat, A., Cubas, P., Farmer, E. E., Nath, U., & Weigel, D. (2008). Control of jasmonate biosynthesis and senescence by miR319 targets. PLoS Biology, 6(9), 1991–2001. https://doi.org/10.1371/journal.pbio.0060230

Seaton, D. D., Graf, A., Baerenfaller, K., Stitt, M., Millar, A. J., & Gruissem, W. (2018). Photoperiodic control of the Arabidopsis proteome reveals a translational coincidence mechanism . Molecular Systems Biology, 14(3), 1–19. https://doi.org/10.15252/msb.20177962

Senis, E., Esgleas, M., Najas, S., Jimenez-Sabado, V., Bertani, C., Gimenez-Alejandre, M., . . . Abad, M. (2021). TUNAR lncRNA Encodes a Microprotein that Regulates Neural Differentiation and Neurite Formation by Modulating Calcium Dynamics. Front Cell Dev Biol, 9, 747667. https://doi.org/10.3389/fcell.2021.747667

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research, 13, 2498–2504. https://doi.org/10.1101/gr.1239303

Sharma, A., Badola, P. K., Bhatia, C., Sharma, D., & Trivedi, P. K. (2020). Primary transcript of miR858 encodes regulatory peptide and controls flavonoid biosynthesis and development in Arabidopsis. Nat Plants, 6(10), 1262-1274. https://doi.org/10.1038/s41477-020-00769-x

Shi, M., Hu, X., Wei, Y., Hou, X., Yuan, X., Liu, J., & Liu, Y. (2017). Genome-wide profiling of small RNAs and degradome revealed conserved regulations of miRNAs on auxin-responsive genes during fruit enlargement in peaches. International Journal of Molecular Sciences, 18(12), 1–14. https://doi.org/10.3390/ijms18122599

Sidhaye, J., Trepte, P., Sepke, N., Novatchkova, M., Schutzbier, M., Dürnberger, G., Mechtler, K., & Knoblich, J. A. (2023). Integrated transcriptome and proteome analysis reveals posttranscriptional regulation of ribosomal genes in human brain organoids. ELife, 12, 1–41. https://doi.org/10.7554/elife.85135

Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. C., & Geromanos, S. J. (2006). Absolute quantification of proteins by LCMSE: A virtue of parallel MS acquisition. Molecular and Cellular Proteomics, 5(1), 144–156. https://doi.org/10.1074/mcp.M500230-MCP200

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., . . . Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol, 9(1), 59-64. https://doi.org/10.1038/nchembio.1120

Smaczniak, C., Immink, R. G. H., Muiño, J. M., Blanvillain, R., Busscher, M., Busscher-Lange, J., Dinh, Q. D., Liu, S., Westphal, A. H., Boeren, S., Parcy, F., Xu, L., Carles, C. C., Angenent, G. C., & Kaufmann, K. (2012). Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. Proceedings of the National Academy of Sciences of the United States of America, 109(5), 1560–1565. https://doi.org/10.1073/pnas.1112871109

Smyth, D. R., Bowman, J. L., & Meyerowitz, E. M. (1990). Early Flower Development in Arabidopsis. The Plant Cell, 2(August), 755–767.

Song, Y. C., Das, D., Zhang, Y., Chen, M. X., Fernie, A. R., Zhu, F. Y., & Han, J. (2023). Proteogenomics-based functional genome research: approaches, applications, and perspectives in plants. Trends Biotechnol. https://doi.org/10.1016/j.tibtech.2023.05.010

Sousa, C., Johansson, C., Charon, C., Manyani, H., Sautter, C., Kondorosi, A., & Crespi, M. (2001). Translational and structural requirements of the early nodulin gene enod40, a short-open reading frame-containing RNA, for elicitation of a cell-specific growth response in the alfalfa root cortex. Molecular and Cellular Biology, 21(1), 354-366. https://doi.org/10.1128/MCB.21.1.354-366.2001

Staudt, A. C., & Wenkel, S. (2011). Regulation of protein function by 'microProteins'. EMBO Rep, 12(1), 35-42. https://doi.org/10.1038/embor.2010.196

Stegmann, M., Zecua-Ramirez, P., Ludwig, C., Lee, H. S., Peterson, B., Nimchuk, Z. L., . . . Huckelhoven, R. (2022). RGI-GOLVEN signaling promotes cell surface immune receptor abundance to regulate plant immunity. EMBO Rep, 23(5), e53281. https://doi.org/10.15252/embr.202153281

Stemmer, M., Thumberger, T., Del, M., Keyer, S., Wittbrodt, J., & Mateo, J. L. (2015). CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. https://doi.org/10.1371/journal.pone.0124633

Straub, D., & Wenkel, S. (2017). Cross-Species Genome-Wide Identification of Evolutionary Conserved MicroProteins. Genome Biol Evol, 9(3), 777-789. https://doi.org/10.1093/gbe/evx041

Stuhrwohldt, N., Dahlke, R. I., Kutschmar, A., Peng, X., Sun, M. X., & Sauter, M. (2015). Phytosulfokine peptide signaling controls pollen tube growth and funicular pollen tube guidance in Arabidopsis thaliana. Physiologia Plantarum, 153(4), 643-653. https://doi.org/10.1111/ppl.12270

Su, G., Kuchinsky, A., Morris, J. H., States, D. J., & Meng, F. (2010). GLay: community structure analysis of biological networks. BIOINFORMATICS APPLICATIONS NOTE, 26(24), 3135–3137. https://doi.org/10.1093/bioinformatics/btq596

Sugano, S. S., Shimada, T., Imai, Y., Okawa, K., Tamai, A., Mori, M., & Hara-Nishimura, I. (2009). Stomagen positively regulates stomatal density in Arabidopsis. Nature. https://doi.org/nature08682 [pii]

Szcześniak, M. W., Bryzghalov, O., Ciomborowska-Basheer, J., & Makałowska, I. (2019). CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. In Methods

in Molecular Biology (Vol. 1933, pp. 415–429). https://doi.org/10.1007/978-1-4939-9045-0_26

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., & Von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Research, 45(D1), D362–D368. https://doi.org/10.1093/nar/gkw937

Szymanski, J., Levin, Y., Savidor, A., Breitel, D., Chappell-Maor, L., Heinig, U., Töpfer, N., & Aharoni, A. (2017). Label-free deep shotgun proteomics reveals protein dynamics during tomato fruit tissues development. The Plant Journal, 90(2), 396–417. https://doi.org/10.1111/TPJ.13490

Tabata, R., Sumida, K., Yoshii, T., Ohyama, K., Shinohara, H., & Matsubayashi, Y. (2014). Perception of root-derived peptides by shoot LRR-RKs mediates systemic N-demand signaling. Science, 346(6207), 343-346. https://doi.org/10.1126/science.1257800

Takáč, T., Šamajová, O., & Šamaj, J. (2017). Integrating cell biology and proteomic approaches in plants. Journal of Proteomics, 169, 165–175. https://doi.org/10.1016/j.jprot.2017.04.020

Takahashi, F., Hanada, K., Kondo, T., & Shinozaki, K. (2019). Hormone-like peptides and small coding genes in plant stress signaling and development. Current Opinion in Plant Biology, 51, 88-95. https://doi.org/10.1016/j.pbi.2019.05.011

Tang, F., Chen, N., Zhao, M., Wang, Y., He, R., Peng, X., & Shen, S. (2017). Identification and functional divergence analysis of WOX gene family in paper mulberry. International Journal of Molecular Sciences, 18(8), 1–18. https://doi.org/10.3390/ijms18081782

Taniguchi, Y., Choi, P. J., Li, G., Chen, H., Babu, M., Hearn, J., Emili, A., & Xie, X. S. (2011). Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Science (New York, N.Y.), 329(5991), 533–539. https://doi.org/10.1126/science.1188308

Tarazona, S., Balzano-Nogueira, L., & Conesa, A. (2018). Multiomics Data Integration in Time Series Experiments. In Comprehensive Analytical Chemistry (1st ed., Vol. 82). Elsevier B.V. https://doi.org/10.1016/bs.coac.2018.06.005

Tavormina, P., De Coninck, B., Nikonorova, N., De Smet, I., & Cammue, B. P. (2015). The Plant Peptidome: An Expanding Repertoire of Structural Features and Biological Functions. Plant Cell, 27(8), 2095-2118. https://doi.org/10.1105/tpc.15.00440

Theißen, G. (2001). Development of floral organ identity: Stories from the MADS house. Current Opinion in Plant Biology, 4(1), 75–85. https://doi.org/10.1016/S1369-5266(00)00139-4

Theißen, G., & Saedler, H. (2001). Floral quartets. Nature, 409(25), 469–471. https://doi.org/10.1111/dom.13526

Theißen, G., Melzer, R., & Ruümpler, F. (2016). MADS-domain transcription factors and the floral quartet model of flower development: Linking plant development and evolution. Development (Cambridge), 143(18), 3259–3271. https://doi.org/10.1242/dev.134080

Thomson, B., & Wellmer, F. (2019). Molecular regulation of flower development. In Current Topics in Developmental Biology (1st ed., Vol. 131, pp. 185–210). Elsevier Inc. https://doi.org/10.1016/bs.ctdb.2018.11.007

Tian, L., Musetti, V., Kim, J., Magallanes-Lundback, M., & DellaPenna, D. (2004). The Arabidopsis LUT1 locus encodes a member of the cytochrome P450 family that is required for carotenoid ε-ring hydroxylation activity. Proceedings of the National Academy of Sciences of the United States of America, 101(1), 402–407. https://doi.org/10.1073/pnas.2237237100

Tong, X., & Liu, S. (2019). CPPred: coding potential prediction based on the global description of RNA sequence. Nucleic Acids Res, 47(8), e43. https://doi.org/10.1093/nar/gkz087

Tong, X., Hong, X., Xie, J., & Liu, S. (2020). CPPred-sORF: Coding Potential Prediction of sORF based on

Toyokura, K., Goh, T., Shinohara, H., Shinoda, A., Kondo, Y., Okamoto, Y., . . . Fukaki, H. (2019). Lateral Inhibition by a Peptide Hormone-Receptor Cascade during Arabidopsis Lateral Root Founder Cell Formation. Developmental Cell, 48(1), 64-75 e65. https://doi.org/10.1016/j.devcel.2018.11.031

Trigg, S. A., Garza, R. M., MacWilliams, A., Nery, J. R., Bartlett, A., Castanon, R., Goubil, A., Feeney, J., O'Malley, R., Huang, S. S. C., Zhang, Z. Z., Galli, M., & Ecker, J. R. (2017). CrY2H-seq: A massively multiplexed assay for deep-coverage interactome mapping. Nature Methods, 14(8), 819–825. https://doi.org/10.1038/nmeth.4343

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Uhrig, R. G., Echevarría-Zomeño, S., Schlapfer, P., Grossmann, J., Roschitzki, B., Koerber, N., Fiorani, F., & Gruissem, W. (2021). Diurnal dynamics of the Arabidopsis rosette proteome and phosphoproteome. Plant Cell and Environment, 44(3), 821–841. https://doi.org/10.1111/pce.13969

Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S. B., Wacholder, A., . . . Carvunis, A. R. (2020). De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat Commun, 11(1), 781. https://doi.org/10.1038/s41467-020-14500-z

Vakirlis, N., Vance, Z., Duggan, K. M., & McLysaght, A. (2022). De novo birth of functional microproteins in the human lineage. Cell Rep, 41(12), 111808. https://doi.org/10.1016/j.celrep.2022.111808

Valdés-López, O., Batek, J., Gomez-Hernandez, N., Nguyen, C. T., Isidra-Arellano, M. C., Zhang, N., Joshi, T., Xu, D., Hixson, K. K., Weitz, K. K., Aldrich, J. T., Paša-Tolí, L., & Stacey, G. (2016). Soybean Roots Grown under Heat Stress Show Global Changes in Their Transcriptional and Proteomic Profiles. Frontiers in Plant Science | Www.Frontiersin.Org, 1, 517. https://doi.org/10.3389/fpls.2016.00517

Valdivia, E. R., Chevalier, D., Sampedro, J., Taylor, I., Niederhuth, C. E., & Walker, J. C. (2012). DVL genes play a role in the coordination of socket cell recruitment and differentiation. Journal of Experimental Botany, 63(3), 1405-1412. https://doi.org/10.1093/jxb/err378

Valentim, F. L., Van Mourik, S., Posé, D., Kim, M. C., Schmid, M., Van Ham, R. C. H. J., Busscher, M., Sanchez- Perez, G. F., Molenaar, J., Angenent, G. C., Immink, R. G. H., & Van Dijk, A. D. J. (2015). A quantitative and dynamic model of the arabidopsis flowering time gene regulatory network. PLoS ONE, 10(2), 1–18. https://doi.org/10.1371/journal.pone.0116973

Van de Velde, W., Zehirov, G., Szatmari, A., Debreczeny, M., Ishihara, H., Kevei, Z., . . . Mergaert, P. (2010). Plant peptides govern terminal differentiation of bacteria in symbiosis. Science, 327(5969), 1122-1126. https://doi.org/10.1126/science.1184057

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends in Genetics, 34(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J. F., Adami, E., Faber, A. B., . . . Hubner, N. (2019). The Translational Landscape of the Human Heart. Cell, 178(1), 242-260 e229. https://doi.org/10.1016/j.cell.2019.05.010

Vanderperre, B., Lucier, J. F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F. M., & Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. PLoS ONE, 8(8). https://doi.org/10.1371/journal.pone.0070698

Vie, A. K., Najafi, J., Liu, B., Winge, P., Butenko, M. A., Hornslien, K. S., . . . Brembu, T. (2015). The IDA/IDA-LIKE and PIP/PIP-LIKE gene families in Arabidopsis: phylogenetic relationship, expression patterns, and transcriptional effect of the PIPL3 peptide. Journal of Experimental Botany, 66(17), 5351-5365. https://doi.org/10.1093/jxb/erv285

Vitorino, R., Guedes, S., Amado, F., Santos, M., & Akimitsu, N. (2021). The role of micropeptides in biology. Cellular and Molecular Life Sciences, 78(7), 3285-3298. https://doi.org/10.1007/s00018-020-03740-3

Vogel, C., & Marcotte, E. M. (2013). Insights into regulation of protein abundance from proteomics and transcriptomis analyses. Nature Reviews Genetics, 13(4), 227–232. https://doi.org/10.1038/nrg3185.Insights

Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., Zecha, J., Asplund, A., Li, L., Meng, C., Frejno, M., Schmidt, T., Schnatbaum, K., Wilhelm, M., Ponten, F., Uhlen, M., Gagneur, J., Hahne, H., & Kuster, B. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Molecular Systems Biology, 15(2), 1–16. https://doi.org/10.15252/msb.20188503

Wang, D., Li, C., Zhao, Q., Zhao, L., Wang, M., Zhu, D., . . . Yu, J. (2009). Zm401p10, encoded by an anther-specific gene with short open reading frames, is essential for tapetum degeneration and anther development in maize. Funct Plant Biol, 36(1), 73-85. https://doi.org/10.1071/FP08154

Wang, J. H., Liu, J., Chen, K., Li, H., He, J., Guan, B., & He, L. (2017). Comparative transcriptome and proteome profiling of two Citrus sinensis cultivars during fruit development and ripening. BMC Genomics, 18(1), 1–13. https://doi.org/10.1186/s12864-017-4366-2

Wang, J., Qiu, Y., Cheng, F., Chen, X., Zhang, X., Wang, H., Song, J., Duan, M., Yang, H., & Li, X. (2017). Genome-wide identification, characterization, and evolutionary analysis of flowering genes in radish (Raphanus sativus L.). BMC Genomics, 18(1), 1–10. https://doi.org/10.1186/s12864-017-4377-z

Wang, L., Clarke, L. A., Eason, R. J., Parker, C. C., Qi, B., Scott, R. J., & Doughty, J. (2017). PCP-B class pollen coat proteins are key regulators of the hydration checkpoint in Arabidopsis thaliana pollen-stigma interactions. New Phytologist, 213(2), 764-777. https://doi.org/10.1111/nph.14162

Wang, P., Wu, T., Jiang, C., Huang, B., & Li, Z. (2023). Brt9SIDA/IDALs as peptide signals mediate diverse biological pathways in plants. Plant Sci, 330, 111642. https://doi.org/10.1016/j.plantsci.2023.111642

Wang, P., Wu, X., Shi, Z., Tao, S., Liu, Z., Qi, K., . . . Zhang, S. (2023). A large-scale proteogenomic atlas of pear. Mol Plant, 16(3), 599-615. https://doi.org/10.1016/j.molp.2023.01.011

Wang, P., Yao, S., Kosami, K. I., Guo, T., Li, J., Zhang, Y., . . . Kawano, Y. (2020). Identification of endogenous small peptides involved in rice immunity through transcriptomics- and proteomics-based screening. Plant Biotechnol J, 18(2), 415-428. https://doi.org/10.1111/pbi.13208

Wang, S., Tian, L., Liu, H., Li, X., Zhang, J., Chen, X., . . . Wu, L. (2020). Large-Scale Discovery of Non-conventional Peptides in Maize and Arabidopsis through an Integrated Peptidogenomic Pipeline. Mol Plant, 13(7), 1078-1093. https://doi.org/10.1016/j.molp.2020.05.012

Wang, X., Wang, Y., Yang, G., Zhao, L., Zhang, X., Li, D., & Guo, Z. (2020). Complementary transcriptome and proteome analyses provide insight into the floral transition in bamboo (Dendrocalamus latiflorus munro). International Journal of Molecular Sciences, 21(22), 1–21. https://doi.org/10.3390/ijms21228430

Wang, Y., Zhang, W. Z., Song, L. F., Zou, J. J., Su, Z., & Wu, W. H. (2008). Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in arabidopsis. Plant Physiology, 148(3), 1201–1211. https://doi.org/10.1104/pp.108.126375

Wellmer, F., & Riechmann, L. (2010). Gene networks controlling the initiation of flower development. Trends in Genetics, 26(12), 519–527. https://doi.org/10.1016/j.tig.2010.09.001

Wellmer, F., Alves-Ferreira, M., Dubois, A., Riechmann, L., & Meyerowitz, E. M. (2006). Genome-Wide Analysis of Gene Expression during Early Arabidopsis Flower Development. PLoS Genetics, 2(7). https://doi.org/10.1371/journal.pgen.0020117

Wellmer, F., Riechmann, L., Alves-Ferreira, M., & Meyerowitz, E. M. (2004). Genome-Wide Analysis of Spatial Gene Expression in Arabidopsis Flowers. The Plant Cell, 16(May), 1314–1326. https://doi.org/10.1105/tpc.021741.termination

Wen, J., Lease, K. A., & Walker, J. C. (2004). DVL, a novel class of small polypeptides: overexpression alters Arabidopsis development. Plant J, 37(5), 668-677. https://doi.org/10.1111/j.1365-313x.2003.01994.x

Whitewoods, C. D. (2021). Evolution of CLE peptide signalling. Seminars in Cell and Developmental Biology, 109, 12-19. https://doi.org/10.1016/j.semcdb.2020.04.022

Whitford, R., Fernandez, A., Tejos, R., Perez, A. C., Kleine-Vehn, J., Vanneste, S., . . . Hilson, P. (2012). GOLVEN secretory peptides regulate auxin carrier turnover during plant gravitropic responses. Developmental Cell, 22(3), 678-685. https://doi.org/10.1016/j.devcel.2012.02.002

Willoughby, A. C., & Nimchuk, Z. L. (2021). WOX going on: CLE peptides in plant development. Current Opinion in Plant Biology, 63, 102056. https://doi.org/10.1016/j.pbi.2021.102056

Wils, C. R., Kaufmann, K., & All, E. B. V. (2017). Gene-regulatory networks controlling in florescence and flower development in Arabidopsis thaliana. BBA - Gene Regulatory Mechanisms, 1860(1), 95–105. https://doi.org/10.1016/j.bbagrm.2016.07.014

Wright, B. W., Yi, Z., Weissman, J. S., & Chen, J. (2022). The dark proteome: translation from noncanonical open reading frames. Trends in Cell Biology, 32(3), 243–258. https://doi.org/10.1016/J.TCB.2021.10.010

Wright, J. C., Mudge, J., Weisser, H., Barzine, M. P., Gonzalez, J. M., Brazma, A., . . . Harrow, J. (2016). Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. Nat Commun, 7, 11778. https://doi.org/10.1038/ncomms11778

Wrzaczek, M., Brosche, M., Kollist, H., & Kangasjarvi, J. (2009). Arabidopsis GRI is involved in the regulation of cell death induced by extracellular ROS. Proc Natl Acad Sci U S A, 106(13), 5412-5417. https://doi.org/10.1073/pnas.0808980106

Wu, F., Shi, X., Lin, X., Liu, Y., Chong, K., Theißen, G., & Meng, Z. (2017). The ABCs of flower development: mutational analysis of AP1/FUL-like genes in rice provides evidence for a homeotic (A)-function in grasses. Plant Journal, 89(2), 310–324. https://doi.org/10.1111/tpj.13386

Wu, H. L., Song, G., Walley, J. W., & Hsu, P. Y. (2019). The Tomato Translational Landscape Revealed by Transcriptome Assembly and Ribosome Profiling. Plant Physiol, 181(1), 367-380. https://doi.org/10.1104/pp.19.00541

Wu, Q., Kuang, K., Lyu, M., Zhao, Y., Li, Y., Li, J., . . . Zhong, S. (2020). Allosteric deactivation of PIFs and EIN3 by microproteins in light control of plant development. Proc Natl Acad Sci U S A, 117(31), 18858-18868. https://doi.org/10.1073/pnas.2002313117

Wu, Y., Tang, Y., Jiang, Y., Zhao, D., Shang, J., & Tao, J. (2018). Combination of transcriptome sequencing and iTRAQ proteome reveals the molecular mechanisms determining petal shape in herbaceous peony (Paeonia lactiflora Pall.). Bioscience Reports, 38(6). https://doi.org/10.1042/BSR20181485

Xanthopoulou, A., Moysiadis, T., Bazakos, C., Karagiannis, E., Karamichali, I., Stamatakis, G., Samiotaki, M., Manioudaki, M., Michailidis, M., Madesis, P., Ganopoulos, I., Molassiotis, A., & Tanou, G. (2022). The perennial fruit tree proteogenomics atlas: a spatial map of the sweet cherry proteome and transcriptome. Plant Journal, 109(5), 1319–1336. https://doi.org/10.1111/tpj.15612

Xing, M., Sun, C., Li, H., Hu, S., Lei, L., & Kang, J. (2018). Integrated analysis of transcriptome and proteome changes related to the ogura cytoplasmic male sterility in cabbage. PLoS ONE, 13(3), 1–22. https://doi.org/10.1371/journal.pone.0193462

Xu, K., Jourquin, J., Xu, X., De Smet, I., Fernandez, A. I., & Beeckman, T. (2023). Dynamic GOLVEN-ROOT GROWTH FACTOR 1 INSENSITIVE signaling in the root cap mediates root gravitropism. Plant Physiol, 192(1), 256-273. https://doi.org/10.1093/plphys/kiad073

Xu, Q., Li, R., Weng, L., Sun, Y., Li, M., & Xiao, H. (2019). Domain-specific expression of meristematic genes is defined by the LITTLE ZIPPER protein DTM in tomato. Commun Biol, 2, 134. https://doi.org/10.1038/s42003-019-0368-8

Xu, R., Li, Y., Sui, Z., Lan, T., Song, W., Zhang, M., . . . Xing, J. (2021). A C-terminal encoded peptide, ZmCEP1, is essential for kernel development in maize. Journal of Experimental Botany, 72(15), 5390-5406. https://doi.org/10.1093/jxb/erab224

Xu, Y., Yu, Z., Zhang, D., Huang, J., Wu, C., Yang, G., . . . Zheng, C. (2018). CYSTM, a Novel Non-Secreted Cysteine-Rich Peptide Family, Involved in Environmental Stresses in

Arabidopsis thaliana. Plant Cell Physiol, 59(2), 423-438. https://doi.org/10.1093/pcp/pcx202

Yadav, A., Bakshi, S., Yadukrishnan, P., Lingwan, M., Dolde, U., Wenkel, S., . . . Datta, S. (2019). The B-Box-Containing MicroProtein miP1a/BBX31 Regulates Photomorphogenesis and UV-B Protection. Plant Physiol, 179(4), 1876-1892. https://doi.org/10.1104/pp.18.01258

Yang, X., Tschaplinski, T. J., Hurst, G. B., Jawdy, S., Abraham, P. E., Lankford, P. K., . . . Tuskan, G. A. (2011). Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Research, 21(4), 634-641. https://doi.org/gr.109280.110 [pii]

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution, 24(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang, Z., & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution, 17(1), 32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236

Yant, L., Mathieu, J., Dinh, T. T., Ott, F., Lanz, C., Wollmann, H., Chen, X., & Schmid, M. (2010). Orchestration of the floral transition and floral development in arabidopsis by the bifunctional transcription factor APETALA2. Plant Cell, 22(7), 2156–2170. https://doi.org/10.1105/tpc.110.075606

Ye, X., Zhao, N., Yu, X., Han, X., Gao, H., & Zhang, X. (2016). Extensive characterization of peptides from Panax ginseng C. A. Meyer using mass spectrometric approach. Proteomics, 16(21), 2788-2791. https://doi.org/10.1002/pmic.201600183

Yeasmin, F., Yada, T., & Akimitsu, N. (2018). Micropeptides encoded in transcripts previously identified as long noncoding RNAs: A new chapter in transcriptomics and proteomics. Frontiers in Genetics, 9(APR), 1–10. https://doi.org/10.3389/fgene.2018.00144

Yin, Y., Adachi, Y., Nakamura, Y., Munemasa, S., Mori, I. C., & Murata, Y. (2016). Involvement of OST1 protein kinase and PYR/PYL/RCAR receptors in methyl jasmonate-induced stomatal closure in arabidopsis guard cells. Plant and Cell Physiology, 57(8), 1779–1790. https://doi.org/10.1093/pcp/pcw102

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. OMICS A Journal of Integrative Biology, 16(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Yu, L., Wang, Y., Liu, Y., Li, N., Yan, J., & Luo, L. (2018). Wound-induced polypeptides improve resistance against Pseudomonas syringae pv. tomato DC3000 in Arabidopsis. Biochemical and Biophysical Research Communications, 504(1), 149-156. https://doi.org/10.1016/j.bbrc.2018.08.147

Yu, T., Tzeng, D. T. W., Li, R., Chen, J., Zhong, S., Fu, D., Zhu, B., Luo, Y., & Zhu, H. (2019). Genome-wide identification of long non-coding RNA targets of the tomato mads box transcription factor rin and function analysis. Annals of Botany, 123(3), 469–482. https://doi.org/10.1093/aob/mcy178

Yu, X., Zhang, Y., Li, T., Ma, Z., Jia, H., Chen, Q., . . . Zhu, D. (2017). Long non-coding RNA Linc-RAM enhances myogenic differentiation by interacting with MyoD. Nat Commun, 8, 14016. https://doi.org/10.1038/ncomms14016

Yu, Y., Guo, S., Ren, Y., Zhang, J., Li, M., Tian, S., Wang, J., Sun, H., Zuo, Y., Chen, Y., Gong, G., Zhang, H., & Xu, Y. (2022). Quantitative Transcriptomic and Proteomic Analysis of Fruit

Development and Ripening in Watermelon (Citrullus lanatus). Frontiers in Plant Science, 13(March), 1–14. https://doi.org/10.3389/fpls.2022.818392

Yuan, N., Dai, C., Ling, X., Zhang, B., & Du, J. (2019). Peptidomics-based study reveals that GAPEP1, a novel small peptide derived from pathogenesis-related (PR) protein of cotton, enhances fungal disease resistance. Molecular Breeding, 39, 156.

Zeng, Y., Tang, Y., Shen, S., Zhang, M., Chen, L., Ye, D., & Zhang, X. (2022). Plant-specific small peptide AtZSP1 interacts with ROCK1 to regulate organ size in Arabidopsis. New Phytologist, 234(5), 1696-1713. https://doi.org/10.1111/nph.18093

Zhai, Q., Zhang, X., Wu, F., Feng, H., Deng, L., Xu, L., Zhang, M., Wang, Q., & Li, C. (2015). Transcriptional Mechanism of Jasmonate Receptor COI1-Mediated Delay of Flowering Time in Arabidopsis. 27(October), 2814–2828. https://doi.org/10.1105/tpc.15.00619

Zhang, C., Wang, J., Wenkel, S., Chandler, J. W., Werr, W., & Jiao, Y. (2018). Spatiotemporal control of axillary meristem formation by interacting transcriptional regulators. Development, 145(24), dev158352. https://doi.org/10.1242/dev.158352

Zhang, H., Zhang, H., & Lin, J. (2020). Systemin-mediated long-distance systemic defense responses. New Phytologist, 226(6), 1573-1582. https://doi.org/10.1111/nph.16495

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A. R., . . . Long, M. (2019). Rapid evolution of protein diversity by de novo origination in Oryza. Nat Ecol Evol, 3(4), 679-690. https://doi.org/10.1038/s41559-019-0822-5

Zhang, M., Zhao, J., Li, C., Ge, F., Wu, J., Jiang, B., . . . Song, X. (2022). csORF-finder: an effective ensemble learning framework for accurate identification of multi-species coding short open reading frames. Brief Bioinform, 23(6). https://doi.org/10.1093/bib/bbac392

Zhang, Q., Vashisht, A. A., O'Rourke, J., Corbel, S. Y., Moran, R., Romero, A., . . . Sampath, S. C. (2017). The microprotein Minion controls cell fusion and muscle formation. Nat Commun, 8, 15664. https://doi.org/10.1038/ncomms15664

Zhang, Q., Wu, E., Tang, Y., Cai, T., Zhang, L., Wang, J., . . . Yang, F. (2021). Deeply Mining a Universe of Peptides Encoded by Long Noncoding RNAs. Molecular and Cellular Proteomics, 20, 100109. https://doi.org/10.1016/j.mcpro.2021.100109

Zhang, R., Kuo, R., Coulter, M., G Calixto, C. P., Carlos Entizne, J., Guo, W., Marquez, Y., Milne, L., Riegler, S., Matsui, A., Tanaka, M., Harvey, S., Gao, Y., Wießner-Kroh, T., Crespi, M., Denby, K., ben Hur, A., Huq, E., Jantsch, M., … Brown, J. W. (2021). A high resolution single molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. https://doi.org/10.1101/2021.09.02.458763

Zhang, Y., Jia, C., Fullwood, M. J., & Kwoh, C. K. (2021). DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. Brief Bioinform, 22(2), 2073-2084. https://doi.org/10.1093/bib/bbaa039

Zhang, Z., Hu, M., Feng, X., Gong, A., Cheng, L., & Yuan, H. (2017). Proteomes and Phosphoproteomes of Anther and Pollen: Availability and Progress. Proteomics, 17(20), 1–12. https://doi.org/10.1002/pmic.201600458

Zhao, C., Zhao, S., Hou, L., Xia, H., Wang, J., Li, C., Li, A., Li, T., Zhang, X., & Wang, X. (2015). Proteomics analysis reveals differentially activated pathways that operate in peanut gynophores at different developmental stages. BMC Plant Biology, 15(1), 1–12. https://doi.org/10.1186/s12870-015-0582-6

Zhao, S., Meng, J., & Luan, Y. (2022). LncRNA-Encoded Short Peptides Identification Using Feature Subset Recombination and Ensemble Learning. Interdiscip Sci, 14(1), 101-112. https://doi.org/10.1007/s12539-021-00464-1

Zhao, S., Meng, J., Kang, Q., & Luan, Y. (2022). Identifying LncRNA-Encoded Short Peptides Using Optimized Hybrid Features and Ensemble Learning. IEEE/ACM Trans Comput Biol Bioinform, 19(5), 2873-2881. https://doi.org/10.1109/TCBB.2021.3104288

Zhao, S., Meng, J., Wekesa, J. S., & Luan, Y. (2023). Identification of small open reading frames in plant lncRNA using class-imbalance learning. Comput Biol Med, 157, 106773. https://doi.org/10.1016/j.compbiomed.2023.106773

Zheng, C., Wei, Y., Zhang, P., Xu, L., Zhang, Z., Lin, K., . . . Chen, Y. (2023). CRISPR/Cas9 screen uncovers functional translation of cryptic lncRNA-encoded open reading frames in human cancer. J Clin Invest, 133(5), e159940. https://doi.org/10.1172/JCI159940

Zheng, E. B., & Zhao, L. (2022). Protein evidence of unannotated ORFs in Drosophila reveals diversity in the evolution and properties of young proteins. elife, 11, e78772. https://doi.org/10.7554/eLife.78772

Zhong, S., Li, L., Wang, Z., Ge, Z., Li, Q., Bleckmann, A., . . . Qu, L. J. (2022). RALF peptide signaling controls the polytubey block in Arabidopsis. Science, 375(6578), 290-296. https://doi.org/10.1126/science.abl4683

Zhong, S., Liu, M., Wang, Z., Huang, Q., Hou, S., Xu, Y. C., . . . Qu, L. J. (2019). Cysteine-rich peptides promote interspecific genetic isolation in Arabidopsis. Science, 364(6443). https://doi.org/10.1126/science.aau9564

Zhou, H., Xiao, F., Zheng, Y., Liu, G., Zhuang, Y., Wang, Z., . . . Lin, H. (2022). PAMP-INDUCED SECRETED PEPTIDE 3 modulates salt tolerance through RECEPTOR-LIKE KINASE 7 in plants. Plant Cell, 34(2), 927-944. https://doi.org/10.1093/plcell/koab292

Zhou, Y., Sarker, U., Neumann, G., & Ludewig, U. (2019). The LaCEP1 peptide modulates cluster root morphology in Lupinus albus. Physiologia Plantarum, 166(2), 525-537. https://doi.org/10.1111/ppl.12799

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C., Cao, X., Han, X., Wang, X., van der Knaap, E., Zhang, Z., Cui, X., Klee, H., Fernie, A. R., Luo, J., & Huang, S. (2018). Rewiring of the Fruit Metabolome in Tomato Breeding. Cell, 172(1–2), 249-261.e12. https://doi.org/10.1016/j.cell.2017.12.019

Zhu, M., & Gribskov, M. (2019). MiPepid: MicroPeptide identification tool using machine learning. BMC Bioinformatics, 20(1), 559. https://doi.org/10.1186/s12859-019-3033-9

Zhu, Q. G., Gong, Z. Y., Wang, M. M., Li, X., Grierson, D., Yin, X. R., & Chen, K. S. (2018). A transcription factor network responsive to high CO 2 /hypoxia is involved in deastringency in persimmon fruit. Journal of Experimental Botany, 69(8), 2061–2070. https://doi.org/10.1093/jxb/ery028

Zhu, Y., Orre, L. M., Johansson, H. J., Huss, M., Boekel, J., Vesterlund, M., . . . Lehtio, J. (2018). Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun, 9(1), 903. https://doi.org/10.1038/s41467-018-03311-y

Zou, L., Pan, C., Wang, M., Cui, L., & Han, B. (2020). Progress on the mechanism of hormones regulating plant flower formation. 42, 739–751. https://doi.org/10.16288/j.yczz.20-014

Acknowledgements

Acknowledgements

How to finish four years (almost five) of work without making a list of "*thank you*" to all the people that have helped me on this journey?

Let's start from the beginning. A special 'thank you' to my PhD supervisors, José Luis, and Tom, for guiding me through this path.

A big thanks to the CRAG services and our collaborators, specially to the Bioinformatics unit, Anna, Víctor and Martí, for their patience and help with all the scripts, especially when I didn't even know what 'Hello world!' meant. Also, thanks to José for helping me during the *early stages of this project development*. Parce, no habría podido recoger todas esas inflorescencias yo sola. Gracias por la acogida cuando llegué a Barcelona.

Thanks to Dr George Coupland for accepting me in his lab, and to all the MPIPZ colleges that I met during my stay in Cologne. It was amazing sharing those months with you, the TATA-bar really helped me to survive to the German experience.

Thanks to the SUMO-lab members for *always* being there. To the current members, Priya, Jaime, Anna, Serena, and Lucas, and to the former, Jordi, Caro, Elisabeth… Gracias, Diana, por ser compi bioinformática en la sombra. Gracias, Sil, por todo, de verdad. Por todas las horas de lab, de chelas, de charlas, de excursiones por pueblitos que no habría visitado si no fuera por Carmen y por ti, por tu cariño y por tus enseñanzas sobre la vida.

A los demás compañeros sufridores del CRAG. Carlos, gracias por estar ahí desde la tercera planta, siempre atento y disfrutando de los dramas de Twitter. Andrés, gracias por enseñarme a usar el Maxwell® y a hacer peso muerto. Mi espalda y mi salud mental lo necesitaban. Toni, gracias por las horas de terapia gratis y por tu ayuda como sabio postdoc; he aprendido mucho gracias a ti. También al resto, Unai, Cris, Ari, Rosa… y voy a parar de poner nombres porque seguro que me dejo a alguien y después cuando lo vea me voy a estresar.

María, una de las mejores sorpresas de Barcelona fue encontrarte a ti aquí. Gracias por ser compañera desde el principio, dentro y fuera del CRAG. Ahora ya está. Podremos darnos ese abrazo de doctoras.

Álex, gracias por el helado y por escucharme casi siempre, porque mis consejos de "*no lo hagas"* te entraron por un oído y te salieron por el otro. Mucha suerte mirando las estrellas.

Patos (y allegados), gracias porque, aunque ahora estemos esparcidos por el mundo, cuando por fin conseguimos vernos siento que estoy en casa. Por cierto, Elenilla, mil gracias por la portada.

Lucía, Sara, Ainhoa, gracias, gracias, GRACIAS por tantas cosas que no tiene sentido hacer una lista. No creo que mi sabiduría como Dra. en Flores os vaya a servir para algo en la vida, pero pienso aprovecharme de vuestros superpoderes jurídico-dentales y de camionera dicharachera hasta el fin de los tiempos.

Gracias a mi familia (a la de garrafa también), por estar apoyándome, aunque no acabéis de entender en qué me estáis apoyando. Y a la pequeña tribu galaico-catalana con toques venezolanos, Inés, Thayron, MariCeli, Arturo, María, só moitas grazas, de corazón.

Mamá, papá, Javi, gracias por ayudarme a poder con todo (aunque primero vaya a llorar). Gracias a los cuatro (Clara también cuenta).

Iago, gracias por ayudarme a mejorar cada día y a intentar superar mi miedo a hacer cosas. Gracias por hacer mi vida más bonita.

Para acabar, quiero dar las gracias a las personas que más me han insistido siempre en que me esforzara y siguiera estudiando. *"Aprovecha, que yo aún me acuerdo de cuando el maestro se quedaba conmigo por las tardes porque no podía ir a clase porque tenía que ayudar en casa"* (Sí, yaya, sí). *"Tú sobre todo sigue con los estudios"* (Que sí). *"Pero entonces, ¿todavía te quedan exámenes?".* Pues, algo así, abuelo(s). Si todo va bien, este será el último "examen"; aunque todavía me queda mucho por aprender.

Appendix

Supplementary material

Full tables of results can be obtained in **Supplementary information** (URL: https://drive.google.com/drive/folders/1I3fjFHOutQn-_JflNtTVg9nEPbmN85vN?usp=drive_link).

The PDF version of this Thesis contains links that can be used to open the specified online document or folder.

Publications

Álvarez-Urdiola, R., Borràs, E., Valverde, F., Matus, J. T., Sabidó, E., & Riechmann, J. L. (2023). Peptidomics Methods Applied to the Study of Flower Development. In Methods in molecular biology (Vol. 2686, pp. 509–536). Springer Science + Business Media, LLC, part of Springer Nature 2023. https://doi.org/10.1007/978-1-0716-3299-4_24

Álvarez-Urdiola, R., Bustamante, M., Ribes, J., & Riechmann, J. L. (2023). Gene Expression Analysis by Quantitative Real-Time PCR for Floral Tissues. In Methods in molecular biology (Vol. 2686, pp. 403–428). Springer Science + Business Media, LLC, part of Springer Nature 2023. https://doi.org/10.1007/978-1-0716-3299-4_20

Álvarez-Urdiola, R., Matus, J. T., & Riechmann, J. L. (2023). Multi-Omics Methods Applied to Flower Development. In Methods in Molecular Biology (Vol. 2686, pp. 495–508). Springer Science + Business Media, LLC, part of Springer Nature 2023. https://doi.org/10.1007/978-1-0716-3299-4_23

Chapter 24

Peptidomics Methods Applied to the Study of Flower Development

Raquel Álvarez-Urdiola, Eva Borràs, Federico Valverde, José Tomás Matus, Eduard Sabidó, and José Luis Riechmann

Abstract

Understanding the global and dynamic nature of plant developmental processes requires not only the study of the transcriptome, but also of the proteome, including its largely uncharacterized peptidome fraction. Recent advances in proteomics and high-throughput analyses of translating RNAs (ribosome profiling) have begun to address this issue, evidencing the existence of novel, uncharacterized, and possibly functional peptides. To validate the accumulation in tissues of sORF-encoded polypeptides (SEPs), the basic setup of proteomic analyses (i.e., LC-MS/MS) can be followed. However, the detection of peptides that are small (up to ~100 aa, 6–7 kDa) and novel (i.e., not annotated in reference databases) presents specific challenges that need to be addressed both experimentally and with computational biology resources. Several methods have been developed in recent years to isolate and identify peptides from plant tissues. In this chapter, we outline two different peptide extraction protocols and the subsequent peptide identification by mass spectrometry using the database search or the de novo identification methods.

Key words Peptidome, Ultrafiltration, Ammonium sulphate, Reverse-phase chromatography, C-18, Arabidopsis, Mass spectrometry, Database

1 Introduction

Although a variety of peptides have been well documented in both animal and plant genomes, until recently the coding potential of eukaryotic short open reading frames (sORFs) at the genome-wide level had mostly been overlooked. One of the reasons behind this gap is the computational and experimental difficulties for their identification and functional characterization, and particularly for determining whether these sequences are in fact translated. However, it has become clear over the past few years that small peptides (usually defined as shorter than 100 amino acids in length) constitute an important part, largely still uncharacterized, of the eukaryotic proteome [1–13]. Moreover, the massive and widespread

José Luis Riechmann and Cristina Ferrándiz (eds.), *Flower Development: Methods and Protocols*, Methods in Molecular Biology, vol. 2686, https://doi.org/10.1007/978-1-0716-3299-4_24,

transcription of the eukaryotic genome and the pervasive translation of long noncoding RNAs (lncRNAs) habilitate sORFs and the resulting small peptides as raw materials for de novo gene origin and evolution [14–19].

In plants, several peptides have been functionally characterized as key players in diverse signalling pathways of plant development, including flower formation and maturation, in Arabidopsis and other plant species (i.e., [20–22]). Moreover, the presence of novel, uncharacterized Arabidopsis small peptides has been inferred from transcriptome data, in particular ribosome profiling (Poly-Ribo-Seq) experiments [23–25], leaving the door open for their identification through proteomics and peptidomics approaches. In fact, in studies with human cells and for selected SEPs identified from lncRNAs, primarily by Poly-Ribo-Seq, it was experimentally estimated that SEPs can be present in the cell at concentrations that are within the range of typical cellular proteins [26], that SEPs can exhibit different and specific subcellular localizations [27, 28], and that they can carry out important biological functions (e.g., [29–33]). Furthermore, in addition to transcriptomics, computational tools have also been used as a source of knowledge on new potentially coding sORFs, in plants as well as in other eukaryotic organisms and bacteria (e.g., [34–36]).

The sources of peptides that altogether would constitute the peptidome of a plant are several and include the following: (1) processing from larger functional or nonfunctional precursors; (2) additional short open reading frames (sORFs) in known protein-coding genes (up- or downstream the main ORF, in introns, as short splice variants or in a different reading frame from that of the main ORF); and (3) sORFs in long noncoding RNAs (lncRNAs), transcripts of unknown function (TUFs), intergenic regions, junctions, and microRNA precursors [37–40]. For instance, computational analyses suggested that several thousands of novel, potentially coding sORFs could exist in the intergenic regions of the *Arabidopsis* genome [35]. In fact, it was found that when overexpressed, some of those novel sORFs could induce developmental alterations in plant size, leaf number and shape, fertility, or cause lethality, raising the possibility that (many) sORFs with coding potential but that are still uncharacterized in plant genomes might be associated with morphogenesis [37] and other developmental and physiological processes.

RNA-based methods are a very powerful tool to detect potentially translating sORFs, and the analysis of ribosome profiling data obtained from a variety of eukaryotic organisms provided strong support to the idea that lncRNAs are an important source of new peptides [41, 42]. Ribosome profiling has also been used to demonstrate extensive translation of open reading frames, including novel sORFs, in plant species such as *Arabidopsis* [23, 25, 43], maize [44] and tomato [45]. The evaluation of the coding

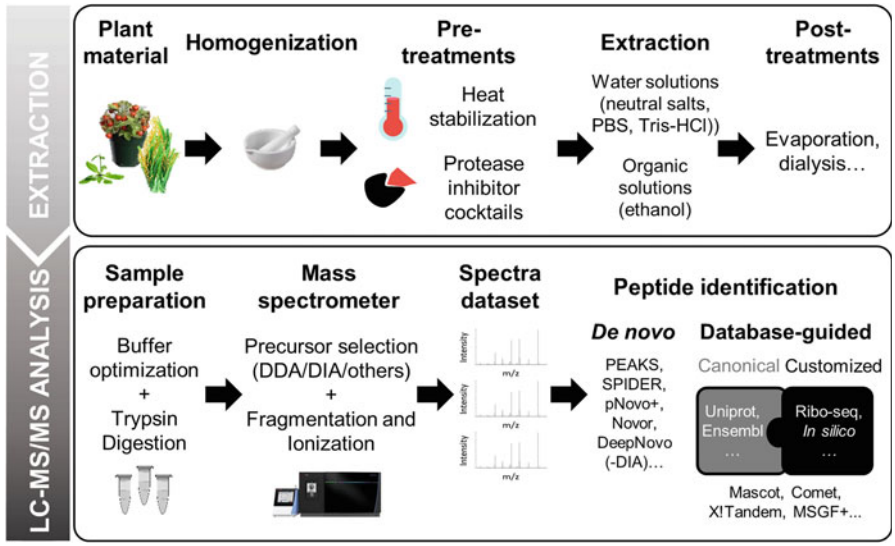


Fig. 1 Workflow for peptide discovery and characterization based on mass spectrometry. Extraction method and MS data analysis

potential of the sequences identified through ribosome profiling is mostly computational but there are mass spectrometry (MS)-based methods able to detect peptides that are translated from novel sORFs, thereby directly validating the protein-coding potential of the transcripts [27, 38, 44, 46–53].

In parallel, the improvement of mass spectrometry and data interpretation bioinformatic algorithms have facilitated the analysis of complex protein mixtures. However, the detection of novel plant peptides derived from small ORFs that are not annotated in reference databases presents specific challenges that need to be addressed, both experimentally and with computational resources (Fig. 1).

The first requirement is an efficient and high-quality extraction from abundant starting material, for which several methods have been developed and optimized. Most basic protocols used for protein extraction from plant tissue are trichloroacetic acid (TCA)-acetone and phenol-based methods. The optimal composition of the extraction buffer depends on the species and tissue of interest [54, 55], but other aspects must be considered, such as heat treatment of the sample to diminish nonspecific protease digestions [38, 56, 57] or the addition of protease inhibitors to avoid protein degradation [38, 44, 46, 49, 58, 59] (Table 1). Besides, the processing and degradation of cellular proteins can generate peptidic fragments that increase the complexity of the peptidome sample, deteriorating the signal-to-noise ratio in the experiments. Therefore, strategies to separate larger proteins from peptides prior to LC-MS/MS analyses are crucial to improve the

Table 1
Overview of peptide extraction methods applied in LC-MS/MS studies in different plant species in recent years

Species	Sample type	Highlights of the extraction method characteristics	Identification workflow	Identified peptides	References
A. thaliana	Col-0 leaves (four-leaf stage)	Heat treatment of the sample to diminish nonspecific protease digestions Trichloroacetic acid (TCA) precipitation 10 kDa MWCO filter	1. Database: six-frame translation of the complete *A. thaliana* genome 2. LC-MS/MS (DDA, Mascot)	1860 novel SEPs	[38]
A. thaliana	Leaf tissue and leaf protoplast	SDS-PAGE (10% gel) 10 kDa MWCO filter	1. Canonical database (Araport11) 2. LC-MS/MS (DDA, ProteinLynx, ProteinPilot)	127 protein-derived auxin-responsive peptides	[54]
Medicago truncatula	Root cultures, xylem sap	*o*-chlorophenol/acetone precipitation Size exclusion chromatography	Nano-LC-ESI-MS/MS (DDA, Proteome Discoverer, SEQUEST) matching spectra against three databases: (1) 225 sequences from known members of different peptide families in *M. truncatula* identified using BLAST; (2) *M. truncatula* whole protein database; (3) ~940 C-terminally encoded peptide (CEP) sequences from different species	12 peptide hormones	[122]
Z. mays	Inbred line B73 seeds	Phenol extraction and ammonium sulphate precipitation	1. Identification of sORFs based on Ribo-seq data 2. LC-MS/MS (DIA, MaxQuant)	2695 small peptides (up to 100aa)	[44]

Z. mays	Maize inbred line B73 leaves (three-leaf stage)	Heat treatment of the sample to diminish non-specific protease digestions Trichloroacetic acid (TCA) precipitation 10 kDa MWCO filter	1. Database: six-frame translation of the complete *Z. mays* genome 2. LC-MS/MS (DDA, Mascot)	1993 novel SEPs	[38]
Oryza sativa	Leaves of *O. sativa* L. ssp. *japonica* cv. Nipponbare and suspension cells derived from Nipponbare calli	Anion-exchange chromatography and acetone precipitation SDS-PAGE using (16.5% gel)	1. Database: rice genome (MSU7) considering and considering three different possible PTMs 2. LC-MS/MS (DDA, Mascot)	236 annotated and 52 unannotated novel small secreted proteins (SSPs)	[47]
Solanum lycopersicum	Tomato leaves	Recovery of analytes (peptides) using Sep-Pak C18 Cartridges	1. Tomato hypothetical peptide database (TomHT database) and randomized databases (Ran Databases) 2. LC-MS/MS (DDA, Mascot)	46 unique peptides derived from 25 pre-proteins	[123]
Capsicum chinense x frutescens	Aerial tissue (~12-week-old plants)	30 kDa MWCO filter	1. Database generated using antimicrobial peptides (AMPs) prediction algorithms + canonical *C-chinense* reference proteome 2. LC-MS/MS (DDA, Mascot)	14 AMPs	[59]
Panax ginseng	Ginseng radix (dry root)	Methanol based extraction.	1. Nano-LC-MS/MS (DIA, PEAKS) with canonical database search (Swissprot) 2. De novo nano-LC-MS/MS (DIA, SPIDER) peptide identification	308 peptides	[90]

(continued)

Table 1
(continued)

Species	Sample type	Highlights of the extraction method characteristics	Identification workflow	Identified peptides	References
Eucalyptus grandis	Plant stem material	Organic solvent precipitation SDS-PAGE	1. Database: three-frame translation (Virtual Ribosome v.2.0) of mRNA, ncRNA, and transcribed RNA sequences publicly available combined with *E. grandis* protein database 2. Database-guided LC-MS/MS (DDA, PEAKS) + novel peptide mapping and genomic classification (BLAST)	41 novel peptides	[48]
Physcomitrella patens	Protonemata	Gel filtration	1. sORF database generated using sORFfinder at genome and transcriptome level 2. LC-MS/MS (DDA, MaxQuant)	828 peptide sequences	[49]

identification and sequence coverage of low-abundance peptides [55, 60]. Peptides can be separated and purified using different methods such as electrophoresis gels [27, 47, 48, 61] or molecular weight cut-off (MWCO) filters [38, 52, 54, 58, 59, 62] (Table 1). Moreover, the optimal polypeptide size for detection by LC-MS/MS is approximately 10–20 amino acids, suggesting that trypsin (or trypsin + Lys-C) cleavage is crucial for high-sensitivity SEP detection. Nevertheless, smaller peptides that may be amenable to protease cleavage should be detectable as well [26].

An additional difficulty lies in undersampling (i.e., identification of only a subset of the peptides) by conventional data acquisition methods [63]. According to a study to optimize a SEP discovery MS workflow using human samples [52], SEP detection is stochastic due to their size and expression characteristics. Therefore, to avoid undersampling and thus identify more SEPs, it is often more efficient to perform multiple technical and/or biological replicates (multiple runs on the MS platform) than, for example, introduce extensive fractionation methods before LC-MS/MS analyses (as in [26]).

For peptide identification from tandem mass spectra, there are two approaches that could be used: database search and de novo sequencing. In database search, all potential peptide sequences included in a specified database are retrieved for each spectrum, and each peptide-spectrum match is scored via a scoring function calculated by database search engines (such as SEQUEST [64], Mascot [65], Phenyx [66], X! Tandem [67], OMSSA [68], pFind [69], InsPecT [70], ByOnic [71], Comet [72], MS-GF+ [73], MaxQuant [74], or MSTracer [75]). This guided approach is widely used for peptidomics and proteomics, and can be based on canonical (well-annotated) protein databases (e.g., UniProt) or customized databases containing putative SEPs identified by bioinformatic (e.g., sORFinder) [76] or transcriptomic analyses (i.e., RNA-sequencing or ribosome profiling).

The annotation of the genome of the organism under study is the first source for preparing the database for MS database search (i.e.,, all proteins and peptides that are already known or identified from that genome). However, for the identification of novel SEPs in MS data, it is necessary to design more specific, expanded databases that should also include the potential novel coding sORFs. Current integrated peptidomics pipelines include different database creation strategies, from the use of ribosome profiling data to the six-frame or three-frame identification of sORFs at the genome or transcriptome level, respectively (Tables 2 and 3; *see* also the **Notes** section). For instance, a recent MS-based study identified over 1000 novel human proteins derived from alternative ORFs identified by RNA-seq (mostly corresponding to SEPs, 57aa median length) [27]. In plants, approximately 70,000 transcribed sORFs were detected in *Physcomitrella patents* (moss) using "sORF

Table 2
Repositories for SEP-database generation

Database	Description	Collected data	Organism	References
ARA-PEP	Putative peptides encoded by sORFs in the *A. thaliana* genome	Tiling arrays, RNA-seq data, and other publicly available datasets	*A. thaliana*	[39]
PsORF	sORFs across different plant species	Genomic, transcriptomic, ribo-seq, and MS data	35 plant species	[104]
PlantPepDB	Manually curated database of plant-derived peptides	Experimentally validated peptides, peptides with evidence at transcript level, based on computational predictions or inferred by homology	Several plant species including algae, bryophyte, angiosperms, and gymnosperms	[124]
RPFdb v2.0	Genome-wide information of translated mRNA	Ribo-seq samples	Plants: *A. thaliana* Others: 28 different species	[103]
CANTATAdb 2.0	lncRNA data from plant and algae	lncRNA identified computationally using publicly available RNA-seq data	39 plant species (including three algae)	[125]
AlnC	Angiosperm lncRNA Catalogue	lncRNA in angiosperms (1KP transcriptome data)	682 angiosperm plant species (809 tissues)	[126]
GWIPS-viz	Online visualization tool for ribo-seq data	Ribo-seq samples	Plants: *A.thaliana, Z. mays* Others: bacteria, animals, etc.	[102]
uORFlight	Database for the evaluation of uORF frequency among different accessions	uORF identified in genome and transcriptome annotations	Plants: *A. thaliana, O. sativa, B. napus, G. max, G. raimondii, M. truncatula, S. lycopersicum, S. tuberosum, T. aestivum, Z. mays* Others: fungus, metazoan, and vertebrate	[127]
uORFdb	Comprehensive literature database on eukaryotic uORFs	uORF-related references; manually curated from all uORF-related literature listed at the PubMed database	Plants: *A. thaliana* Others: human, mouse, rat, virus, yeast, etc.	[105]

(continued)

Table 2
(continued)

Database	Description	Collected data	Organism	References
C-PAmP	Computationally predicted plant antimicrobial peptides	Selection of peptides included in the Antimicrobial Peptide Database (APD) and the Collection of Anti-Microbial Peptides (CAMP)	2112 plant species in UniProtKB/Swiss-Prot	[128]
StraPep	Structure database of bioactive peptides	Structural data collected from UniProtKB and PDB	452 different species including bacteria, yeast, animals, humans, and plants	[129]
DRAMP 3.0	Manually curated data repository of antimicrobial peptides	Peptides retrieved from Pubmed, Swiss-prot, and Lens	Variety of organisms, including bacteria, archaea, protists, fungi, animals, and plants	[130]

Finder" [76], from which 828 distinct peptide sequences were identified by LC-MS/MS [49]. Customized peptide databases can also be derived from the six-frame translation of genomic sequences, an approach that has been successfully used in micro-organisms [62, 77], and recently also in both monocot and dicot plants, where a total of 1993 and 1860 SEPs were identified in maize and Arabidopsis, respectively [38]. Altogether, these and other studies illustrate the existence of a substantial, uncharted fraction of the eukaryotic proteome that is mainly composed of small proteins (peptidome) (Table 1).

In contrast to database search, for de novo peptide sequencing, peptide sequences are extracted directly from tandem mass spectra using specific algorithms such as PEAKS [78], SPIDER [79], UniNovo [80], pNovo+ [81], Novor [82], DeepNovo [83], or DeepNovo-DIA [84]. The de novo method is less powerful than database search, as many spectra cannot be unambiguously sequenced due to incomplete fragmentation. In addition, the de novo method is relatively slow when compared with the database-search engines, and the large search space of all possible amino acid sequences for each spectrum often leads to higher false discovery rates. Moreover, the complexity of tandem mass spectra can be significantly increased when posttranslational modifications (PTMs) are considered as well [85]. Some algorithms have been used for solving the de novo identification problems involving dynamic programming, integer linear programming, machine learning or other methods, and advances in mass spectrometry

Table 3
Tools for database design

Tool	Description	Method	Example	References
SPADA	Small peptide alignment discovery application. Free software tool that identifies and predicts the gene structure for short peptides with one or two exons	Sequence similarity	Creation of an *M. truncatula* small secreted peptide database (MtSSPdb) using SPADA and sORF Finder [131]	[112]
sORF Finder	Program package for the identification of sORFs with high coding potential	Codon pattern, codon substitution and cross-species conservation	51 new sORFs identified using sORF finder and the ARA-PEP repository (LC-MS/MS results) [46]	[35, 76]
PhyloCSF	Phylogenetic Codon Substitution Frequencies: method to determine whether a multispecies nucleotide sequence alignment is likely to represent a protein-coding region	Codon pattern, codon substitution and cross-species conservation	Identification of small peptide-coding "long noncoding" RNAs in soybean [132]	[34]
MiPepid	RNA-seq sORF annotation in mammalian species	Machine learning	Identification of 82 novel species-specific translated sORFs (LC-MS/MS) from lncRNA (database generated using MiPepid) [19]	[113]
lncPepid	RNA-seq sORF annotation in plants. A discovery tool trained using maize and Arabidopsis data that considers sequence composition and physicochemical properties	Machine learning		[115]
CPPred-sORF	Predicts the coding potential of sORFs based on non-AUG initiation of translation	Machine learning	sORF finder, miPepid, CPPred, and DeepCPP used as control groups [115]	[114]
DeepCPP	Optimization of CPPred	Deep learning	sORF finder, miPepid, CPPred, and DeepCPP used as control groups [115]	[116]

(continued)

Table 3
(continued)

Tool	Description	Method	Example	References
RiboTaper	Statistical approach that identifies translated regions based on the characteristic three-nucleotide periodicity of ribo-seq data	Ribo-seq	Identification of uORFs, dORFs, and altORFs in *A. thaliana* [23]	[106]
PRICE	Computational method that models experimental noise to resolve overlapping sORFs and noncanonical translation initiation in an accurate manner	Ribo-seq	Validation of the method using major histocompatibility complex class I (MHC I) peptidomics [107]	[107]
RiboCode	Unbiased method to recover the signal of active translation from the ribo-seq data	Ribo-seq	Identification of 9388 sORF encoding peptides (2-100aa) in maize, from which 2695 SEPs were verified by MS data [44]	[108]
RiboStreamR	Quality control platform for Ribo-seq data in the form of an R shine web application	Ribo-seq		[109]
RiboPlotR	Visualization package written in R. Representation of RNA-seq coverage and Ribo-seq reads in genomic coordinates for all annotated transcript isoforms of a gene	Ribo-seq	RiboPlotR combines transcriptome annotation files, standard RNA-seq bam files, and Ribo-seq P-site position/count files to plot RNA-seq and Ribo-seq data with genomic coordinates for each isoform. Tested in Arabidopsis and tomato [110]	[110]
RiboNT	Noise-tolerant sORF predictor that can use RPFs with poor periodicity	Ribo-seq	Identification of sORFs in Arabidopsis seedlings that are evolutionary conserved in diverse plant species [111]	[111]

instruments have improved de novo sequencing results [86]. However, further optimizations of algorithms, particularly with respect to data confidence, are still necessary to turn the technique into an actual alternative to commonly used database search peptide identification methods. Despite the difficulties, de novo identification has been successfully implemented for SEP detection in several

plant species [48, 59, 87–89]. When combined with classic database search strategies, de novo approaches can help to provide more comprehensive results [59, 90, 91] (Table 1). In fact, several research groups have developed software that directly combines both, database search and de novo sequencing, for peptide identification from mass spectra [71, 92].

Despite the advances in mass spectrometry and data interpretation, however, a problem still not fully addressed is the (high) number of unassigned spectra. New mass spectrometry sampling methods, such as data-independent acquisition (DIA [93]), together with the development of new machine learning tools to predict peptide fragmentation [94–97] promise a bright and very exciting future in the peptidome field, in which a significant amount of information will be confidently recovered from the acquired data.

In this chapter, we provide two plant peptide extraction methods based on different extraction buffers and precipitation techniques, and describe an example of an LC-MS/MS pipeline, also introducing some suggestions for database design.

2 Materials

2.1 General

1. Protein low-binding microcentrifuge tubes (1.5 or 2 mL).
2. Mortar and pestle.
3. Liquid nitrogen.

2.2 Ultrafiltration

1. Extraction buffer: 1× phosphate-buffered saline (PBS), 1.5 M urea, 10 mM dithiothreitol (DTT), 2% v/v acetonitrile (ACN), 0.5% v/v trifluoroacetic acid (TFA), 10 μM MG-132 proteasome inhibitor, 1 tablet of Proteinase Inhibitor cocktail cOmplete (Roche) per each 50 mL of buffer, and 1 mM phenylmethylsulfonyl fluoride (PMSF) (*see* **Note 1**). Prepare fresh for each experiment (*see* **Note 2**).
2. 10× phosphate buffer saline (PBS): 1.37 M NaCl, 0.027 M KCl, 80 mM Na_2HPO_4, and 20 mM KH_2PO_4 pH 7 (NaOH). Prepare 1 L and autoclave it.
3. Ultra-0.5 mL 30-K centrifugal filter devices (Amicon®) (*see* **Note 3**).

2.3 Ammonium Sulphate Precipitation

1. Extraction buffer: 1× PBS, 2 M urea, 2% v/v acetonitrile, 10 mM DTT, 5% v/v trifluoroethanol (TFE), 50 mM Tris–HCl pH 7.6, 10 μM MG-132, 1 tablet of Proteinase Inhibitor cocktail cOmplete (Roche) per each 50 mL of buffer, and 1 mM PMSF. Prepare fresh for each experiment (*see* **Note 2**).
2. Ammonium sulphate (salt, EM/HPLC grade).

2.4 Reverse-Phase Chromatography Peptide Extraction

1. C18 spin columns, containing 8 mg of resin each (Pierce, Thermo Scientific).
2. Activation solution: 50% v/v ACN in distilled water (400 μL per sample).
3. Equilibration solution: 0.5% v/v TFA in 5% v/v ACN (400 μL per sample).
4. Sample buffer: 2% v/v TFA in 20% v/v ACN (1 μL for every 3 μL of sample) (*see* **Note 4**).
5. Wash solution: 0.5% v/v TFA in 5% v/v ACN (400–800 μL per sample) (*see* **Note 5**).
6. Elution buffer: 0.1% v/v formic acid in 70% v/v ACN (42 μL per sample) (*see* **Note 6**).
7. Qubit protein assay kit.
8. Qubit fluorometer.

2.5 LC-MS/MS

1. DL-dithiothreitol (DTT) (*see* **Note 7**).
2. Iodoacetamide.
3. Urea.
4. Ammonium bicarbonate.
5. Lysyl endopeptidase.
6. Trypsin.
7. Formic acid.
8. MicroSpin C18 columns (The Nest Group, Inc).
9. Nano Trap C18 columns with an inner diameter of 100 μm packed with C18 particles of 5 μm particle size (Thermo Fisher Scientific) (optional, depending on the setup of each laboratory).
10. Reverse-phase chromatography columns (C18, 2 μm, 15–50 cm length) (*see* **Note 8**).
11. Buffer A: 0.1% v/v formic acid in water.
12. Buffer B: 0.1% v/v formic acid in acetonitrile.
13. Bovine serum albumin (New England Biolabs cat # P8108S).
14. Orbitrap Eclipse mass spectrometer (Thermo Fisher Scientific) (*see* **Note 9**).
15. EASY-nLC 1000 (Thermo Fisher Scientific).

3 Methods

Below we provide two peptide extraction methods based on different extraction buffers and precipitation techniques (*see* Subheadings 3.1 and 3.2), both of which are to be followed by a reverse-phase chromatography (*see* Subheading 3.3) (Fig. 2) and describe an example of an LC-MS/MS pipeline (*see* Subheading 3.4).

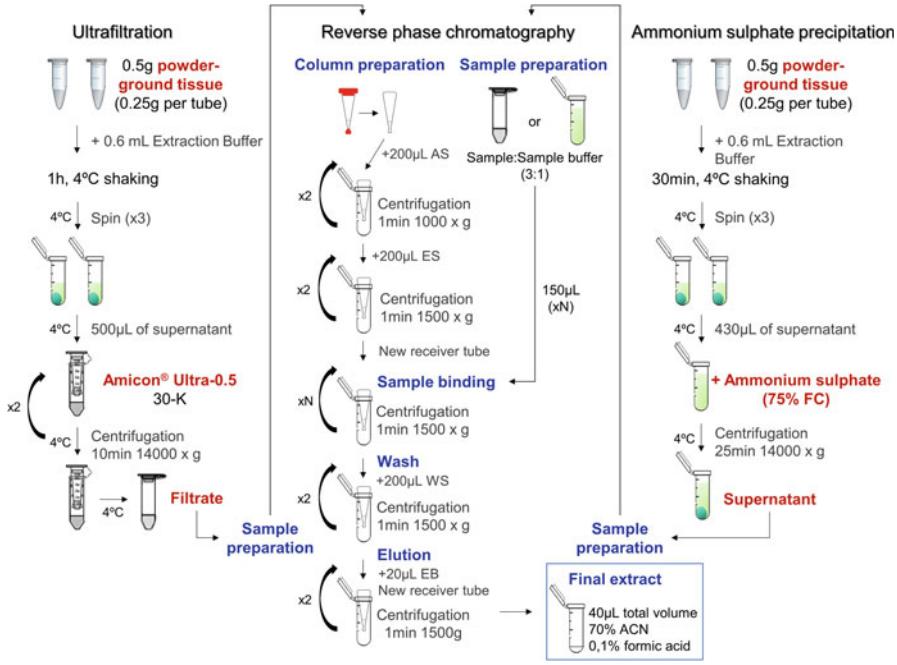


Fig. 2 Schematic representation of the two peptide extraction methods described in this chapter. AS, activation solution; ES, equilibration solution; WS, wash solution; EB, elution buffer

3.1 Ultrafiltration

1. Collect tissue of interest with clean material (*see* **Note 10**) and freeze directly in liquid nitrogen. Keep at −80 °C until required.
2. Using a different mortar and pestle for each sample, grind the tissue with liquid nitrogen until obtaining a whitish fine powder (*see* **Note 11**).
3. Collect 0.5 g of blended tissue distributed in two 2 mL Eppendorf tubes (*see* **Note 12**).
4. Add a total of 1.2 mL of extraction buffer to 0.5 g of tissue, vortex immediately, and transfer to ice while preparing the rest of the samples (*see* **Note 13**).
5. Incubate the samples with continuous shaking for 1 h at 4 °C.
6. Spin the samples for 1 min at 4 °C in a microcentrifuge (max speed, ≥14,000 × g) to precipitate cellular debris and solid particles in suspension. Repeat as many times as necessary (*see* **Note 14**).
7. Insert each Amicon filter device in one of the provided microcentrifuge tubes.
8. Add up to 500 µL of the clean supernatant in the Amicon filter device and centrifuge at 14,000 × g for 10 min at 4 °C as indicated by manufacturer (*see* **Notes 15** and **16**). Repeat until all sample has passed through the same Amicon filter.

9. Keep the filtrate in the provided microcentrifuge tubes (flow-through) (*see* **Note 17**). Keep the samples on ice to immediately continue with the reverse-phase chromatography or store them at −80 °C until use.

3.2 Ammonium Sulphate Precipitation

1. Collect tissue of interest with clean material (*see* **Note 10**) and freeze directly in liquid nitrogen. Keep at −80 ° C until required.
2. Using a different mortar and pestle for each sample, grind the tissue with liquid nitrogen until obtaining a whitish fine powder (*see* **Note 11**).
3. Collect 0.5 g of blended tissue distributed in two 2 mL Eppendorf tubes (*see* **Note 12**).
4. Add a total of 1.2 mL of extraction buffer to 0.5 g of tissue, vortex immediately, and transfer to ice while preparing the rest of the samples (*see* **Note 13**).
5. Incubate the samples shaking for 30 min at 4 °C.
6. Spin the samples for 1 min at 4 °C in a microcentrifuge (max speed, ≥14,000 × g) to precipitate cellular debris and solid particles in suspension (*see* **Note 14**).
7. Add 75% (w/v) of ammonium sulphate to the supernatant to precipitate the proteins in solution at 4 °C. The salt must be added little by little pipetting slowly each time until proteins precipitate (*see* **Note 18**).
8. Centrifuge at maximum speed (≥14,000 − g) for 25 min at 4 °C.
9. Place the supernatant in a new low-binding protein tube (smaller peptides will remain in the supernatant, whereas larger proteins precipitate). Keep the samples on ice to immediately continue with the reverse-phase chromatography or store them at −80 °C until use.

3.3 Reverse-Phase Chromatography Peptide Extraction

Prepare the reverse phase chromatography C18 columns as indicated by the manufacturer protocol. In brief:

Sample preparation:

1. Mix 3:1 parts of sample:sample buffer. The final sample mix will contain approximately 0.5% TFA in 5% ACN (*see* **Note 19**).

Column preparation:

2. Tap the column to settle the resin on the bottom of each column. Remove top and bottom caps (in that order). Place the column into a 2 mL receiver tube.
3. Add 200 μL of activation solution to wet the resin. Make sure to rinse the walls of the spin column (*see* **Note 20**).

4. Centrifuge at 1000 × g for 1 min. Discard the flow-through and repeat **steps 3** and **4**.
5. Add 200 μL of equilibration solution, centrifuge at 1500 × g for 1 min, and discard the flow-through. Repeat this step once.

Sample binding:

6. Place the column into a receiver tube and load up to 150 μL of sample on top of resin bed (*see* **Note 21**).
7. Centrifuge at 1500 × g for 1 min. Repeat **steps 6** and **7** as many times as needed to load all the sample in the same column (*see* **Notes 22** and **23**).

Column wash:

8. Add 200 μL of wash solution to the column and centrifuge at 1500 × g for 1 min. Repeat this step once (*see* **Note 5**).

Elution:

9. Place the column in a new protein low-binding receiver tube and add 21 μL of elution buffer to the top of the resin bed.
10. Centrifuge at 1500 × g for 1 min and repeat **steps 9** and **10** with the same receiver tube.
11. Quantify the concentration and amount of total protein in each sample using a Qubit protein assay kit: Mix 199 μL of Qubit buffer with 1 μL of Qubit reagent for each sample. Add 2 μL of sample to 198 μL of the reaction mixture, vortex, and spin the tube. Incubate at room temperature for 15 min before measuring.
12. Store the samples at −80 °C until further analysis.

3.4 LC-MS/MS

3.4.1 Sample Preparation

1. Prepare or dissolve samples in 6 M urea, 200 mM ammonium bicarbonate.
2. Reduce the samples (10 μg of protein) with 30 nmols of dithiothreitol at 37 °C for 1 h.
3. Alkylate the samples (10 μg of protein) in the dark with 60 nmols of iodoacetamide at 25 °C for 30 min.
4. Dilute the sample extract to 2 M urea with 200 mM ammonium bicarbonate for digestion with endoproteinase LysC (1:10 w:v), and incubate at 37 °C overnight.
5. Dilute twofold with 200 mM ammonium bicarbonate for trypsin digestion (1:10 w:w), and incubate at 37 °C for 8 h.
6. After digestion, add formic acid (10% v/v of the final volume) to acidify the peptide mix.
7. Desalt the samples with MicroSpin C18 columns prior to LC-MS/MS analysis, following manufacturer's instructions.

3.4.2 Chromatographic and Mass Spectrometric Analysis

1. Load the peptides onto the analytical column (C18, 2 μm, 15–50 cm length).
2. Separation of the peptides by reverse-phase chromatography with the corresponding columns.
3. Chromatographic gradients start at 93% buffer A and 7% buffer B with a flow rate of 250 nL/min for 5 min and gradually increase 65% buffer A and 35% buffer B in 60 min.
4. After each analysis, wash the column for 15 min with 10% buffer A and 90% buffer B.
5. Peptide eluates are dried in a vacuum centrifuge, and resuspended with buffer A at a final concentration of 1 μg/μL prior to analysis by LC-MS/MS.
6. Operate the mass spectrometer to acquire peptide spectra (*see* **Note 24)**.

3.4.3 Data Analysis for Database-Search Peptide Identification

1. Search the acquired spectra against the desired peptide database (*see* **Note 25**), plus a list of common contaminants (suggested: [98]), and all the corresponding decoy entries.
2. Set the parameters accordingly to the experimental and mass spectrometric settings and, if appropriate, select variable post-translational modifications to be detected (*see* **Notes 26** and**27**).
3. Determine the peptide abundance estimation [99, 100].
4. Add the information to the appropriate repositories (*see* **Note 28**).

4 Notes

1. Octyl-glucoside, a detergent, could be added (0.1% v/v) to the extraction buffer. The use of detergents is only necessary for the extraction and solubilization of hydrophobic peptides and proteins. However, the presence of detergents in peptide samples decreases chromatographic resolution in LC-MS/MS. Thus, they must be removed prior to MS analysis [101]. As a general rule for MS/MS experiments, keep laboratory wear and high-quality chemicals separated from the rest of the laboratory materials, always use gloves and, if possible, disposable plastic material of the highest quality.
2. Prepare a new extraction buffer on every extraction day as protease inhibitors could not work properly otherwise. MG-132 is available from several suppliers (we have routinely used MG-132 from Sigma-Aldrich). Proteinase Inhibitor cocktail cOmplete is from Roche. Different extraction buffers have been proposed in the recent years, and their final composition

needs to be selected considering the final objective of the study and the type of analytes of interest (e.g., phosphopeptides), because its formulation may affect the final state of the peptides and proteins in the samples (Table 1).

3. The Amicon® Ultra-0.5 product line includes five different cut-offs depending on its nominal molecular weight limit (NMWL); 30-K (30 kDa filter) devices are recommended, as peptides would normally be below the 30 kDa cut-off.
4. ACN can be substituted for methanol in all sample preparation buffers, depending on the desired composition of the final elution buffer.
5. The required washing volume will be dependent upon amount and type of contaminants present in the samples. Samples already containing large amounts of urea or >100 mM ammonium bicarbonate derived from the extraction buffer (Table 1) need to be washed one or two additional times.
6. The elution buffer used can be tailored to the downstream application. Acceptable buffers include 50–70% (v/v) ACN or methanol with or without 0.1% (v/v) TFA. For best results in LC-MS/MS analysis, TFA is replaced with 0.1% (v/v) formic acid.
7. Reagents for LC-MS/MS can be obtained from several suppliers. As an example, we list here the specific products we use: urea (GE Healthcare; Sigma-Aldrich, P/N 17-1319-01), ammonium bicarbonate (BioUltra, ≥99.5% (T); Sigma-Aldrich, P/N 09830), iodoacetamide (BioUltra; Sigma-Aldrich, P/N I1149), DL-dithiothreitol (for electrophoresis, ≥99%; Sigma-Aldrich, P/N D9163), formic acid for analysis EMSURE® (ACS Reag. Merck, P/N 1.00264.0100), sequencing grade modified trypsin (Promega, P/N V5111), and lysyl endopeptidase (Wako Chemicals GmbH, P/N 129-02541).
8. Suitable reverse-phase chromatography columns that we have used are, for instance, 25 cm columns with an inner diameter of 75 μm, packed with 1.9 μm C18 particles (Nikkyo Technos Co.); and 50 cm columns with an inner diameter of 75 μm, packed with 2 μm C18 particles (EASY-Column, Thermo Fisher Scientific, ES903).
9. This is just a concrete example of a "modern high-resolution mass spectrometer"; other instruments could be used.
10. To reduce sample contamination with human proteins (i.e., keratins and collagen) during sample collection, the use of nitrile gloves and laboratory coats is recommended. Take precaution to avoid hair contamination. If flower organs or tissues are going to be dissected, cool tweezers and any other sampling instrument with liquid nitrogen.

11. Keep samples (before and after grinding) always frozen by pouring liquid nitrogen in the mortar sporadically. Cool collection spatulas before using them to collect homogenized tissue.
12. The extraction yields are around 1 mg of total protein for each 0.5 g of tissue. Peptides might represent about 1% of the total protein, and therefore the expected yield for these extraction methods would be 10–15 μg of peptides. For Arabidopsis inflorescences, a volume of 1 mL of blended tissue in a 2 mL Eppendorf tube is equivalent to approximately 0.5 g of tissue. Dividing the sample in different tubes facilitates its dissolution in the extraction buffer, that is, using tubes with only 0.25 g (equivalent to 0.5 mL of volume) of blended tissue. After finishing the entire extraction protocol (including the reverse-phase chromatography with C18 columns), 10–30 μg of total peptides are obtained when using 30-K filters. The efficiency of the ammonium sulphate precipitation method may be lower (~6 μg of total peptides) (Fig. 2).
13. If total sample has been divided in two tubes, add approximately 0.6 mL of extraction buffer to each tube (with 0.25 g of blended tissue).
14. After each 1 min spin, transfer the supernatant to a new tube. Be careful to avoid both the pellet and remaining particles in suspension. Repeat the spin in a new tube as many times as needed until supernatant is clear (2 or 3 times should be enough).
15. When the sample has been divided in two tubes, the efficiency of using one single Amicon filter for all subsamples and the same collection tube is sufficient to achieve a suitable yield.
16. The required centrifugation time may vary according to the NMWL of the columns used. This protocol is defined for 30-K (or upper) devices, yet a higher centrifugation time is necessary for 10-K or 3-K devices (15 and 30 min, respectively).
17. The filtrate contains the smallest peptides depending on the weight limit of the filter device. However, if needed, it is possible to recover the concentrated solute by placing the filter device upside down in a clean microcentrifuge tube and centrifuging at 1000 × g for 2 min at 4 °C. For optimal recovery, it is important to perform the reverse spin immediately after filtrating. Besides, desalting, buffer exchange or diafiltration of this concentrated solute can be accomplished before eluting it by reconstituting the concentrate retained in the column to the original sample volume with the desired solvent and repeating the ultrafiltration process from the beginning to the concentrated solute elution.

18. Ammonium sulphate calculator from EnCor Biotechnology Inc. (http://www.encorbio.com/protocols/AM-SO4.htm) (selecting 4 °C temperature) can be used to calculate the needed amount of ammonium sulphate for each specific sample. The salt addition will increase the sample volume, which should be considered for the reverse-phase chromatography. For smallest peptides, ammonium sulphate could be added up to 80–85%.
19. The final exact concentrations of TFA and ACN will vary according to the extraction buffer, that is, ultrafiltration or ammonium sulphate precipitation protocol. In these examples, the concentration of the sample:sample buffer mix prior to reverse-phase chromatography would be 6.5% (v/v) of ACN for both extraction methods, 0.875% (v/v) TFA for ultrafiltration and 0.5% (v/v) TFA for ammonium sulphate precipitation. Nevertheless, these slight variations do not appear to result in significant differences in the efficiency of the reverse-phase chromatography process.
20. Add solutions carefully, especially in the activation step. Pour the solution through the walls of the column to avoid producing irregularities in the resin.
21. Each column can bind up to 30 μg of total peptide from 10 to 150 μL sample volumes.
22. In some cases, the extraction yield can be increased by recovering the flow-through and recentrifuging it after each step.
23. Flow-through may be retained to confirm sample binding.
24. 1–2 μg of peptides are loaded onto an analytical column (25 cm C18 2 μm particle size) using an autosampler device (e.g., EASY nLC 1000 and Thermo Fisher Scientific) and the peptides are then separated by reverse-phase chromatography using a water-acetonitrile chromatographic gradient. Modern high-resolution mass spectrometers are recommended for data acquisition (e.g., Orbitrap or qTOF). The mass spectrometer is operated in data-dependent acquisition (DDA) mode, in which a full MS scan is recorded in each cycle, followed by the fragmentation of the 10–30 most intense precursor ions to obtain the fragment ion spectra.
25. Obtained raw data are analyzed using a database search strategy. However, the results are susceptible to the characteristics of the reference database used for peptide identification. It is advisable to add the lists of putative SEPs to a database containing the canonical peptides and proteins of each organism (available in ENSEMBL, Uniprot, or other databases). The total number of sequences included in the database is also important, as an excessively large database (e.g., over 100,000 sequences) may lead to a higher false discovery rate in the

identifications. There are several different approaches for the identification of novel potential SEP sequences to be included in the reference database.

One approach that has often been used is to make use of Ribo-seq data. There are multiple repositories that contain sORFs identified by Ribo-seq for plant species such as GWIPS-viz (*Arabidopsis thaliana* and *Zea mays*) [102], RPFdb v2.0 (*A. thaliana*) [103], PsORF (35 plant species) [104], and uORFdb (*A. thaliana* and others) [105] (Table 2); as well as several tools for the analysis of Ribo-seq data and sORF identification such as RiboTaper [106], PRICE [107], RiboCode [108], RiboStreamR [109], RiboPlotR [110], or RiboNT [111] (Table 3).

An alternative (and complementary) approach for the identification of putative novel SEPs is to make use of the genome or lncRNA transcriptome sequences through sORF-prediction tools such as SPADA [112], sORF Finder [35, 76], PhyloCSF (Phylogenetic Codon Substitution Frequencies) [34], MiPepid [113], CPPred [114], lncPepid [115], or DeepCPP [116] (Table 3).

In addition, the putative peptide databases can also be derived from the six-frame translation of the corresponding genome sequence or from the three-frame translation of transcriptomic datasets (e.g., RNA-seq data and lncRNA), an approach referred to as peptidogenomics [117]. It is a strategy that has been successfully implemented in microorganisms [62, 77], and plants [38].

An additional consideration for the generation of the putative SEP database is whether the presence of translation initiation codons in the ORFs (the standard ATG or noncanonical codons such as CTG or ACG; *see* [118–120]) is a requirement or not, as both approaches have been used (e.g., [35, 38]).

26. Once the database has been constructed, the raw LC-MS/MS data needs to be interpreted using a database search engine (such as SEQUEST [64], Mascot [65], Phenyx [66], X! Tandem [67], OMSSA [68], pFind [69], InsPecT [70], ByOnic [71], Comet [72], MS-GF+ [73], MaxQuant [74], or MSTracer [75]). As example, the Mascot search engine (v2.6) can be used, using the search parameters accordingly to the experimental and mass spectrometry settings. For peptide identification a precursor ion mass tolerance below 10–20 ppm is recommended, whereas the fragment ion mass tolerance can go from 10 to 20 ppm for high-resolution mass analyzers (Orbitrap and TOF) to 0.5 Da if a linear ion trap is used for the analysis of the tandem mass spectra. Common peptide modifications such as oxidation of methionine and N-terminal protein acetylation are used as variable modifications. False discovery rate (FDR) in peptide identification is set to a maximum of 1%.

27. Validation of (selected) identified peptides is highly recommended due to the intrinsic limitations of FDR estimation when working with large databases, although this cannot be done yet in a high-throughput manner. Peptide identifications that pass the FDR threshold can be further validated with the purchase and full LC-MS/MS characterization of synthetic peptides with the same identified sequence (e.g., [38]), and/or by comparison with the fragmentation patterns and retention time predicted by the new machine learning algorithms (e.g., Prosit and MS^{2}PIP) [94, 95, 97].
28. Share data and results in a public repository. Data sharing in the public domain is the standard for omics research and a requirement for publication. For proteomics, the Proteomics IDEntifications (PRIDE) database (https://www.ebi.ac.uk/pride/) at the European Bioinformatics Institute (EMBL-EBI, Hinxton, Cambridge, UK) has enabled public data deposition of MS data since 2004, and its archival component has become the largest repository for proteomics data sharing worldwide [121]. The PRIDE database provides access to most of the experimental proteomics data described in MS-related scientific publications. Moreover, several repositories for sORFs and SEPs in plants have been developed with different purposes and using information from multiple in silico and experimental approaches (Table 2).

Acknowledgments

Our work on peptidomics was supported by grant BFU2014-58289-P (funded by MICIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe") and by grant 2017SGR718 (from the Agencia de Gestió d'Ajuts Universitaris I de Recerca) to JLR, and by institutional grant SEV-2015-0533 (funded by MCIN/AEI/10.13039/501100011033) and by the CERCA Programme/Generalitat de Catalunya. R.A. is supported by fellowship PRE2018-084278 funded by MCIN/AEI/ 10.13039/501100011033 and by "ESF Investing in your future." The CRG/UPF Proteomics Unit is part of the Spanish Infrastructure for Omics Technologies (ICTS OmicsTech). We also acknowledge "Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya" (2017SGR595) and support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa, and the CERCA Programme/Generalitat de Catalunya.

References

1. Tavormina P, De Coninck B, Nikonorova N, De Smet I, Cammuea BPA (2015) The plant peptidome: an expanding repertoire of structural features and biological functions. Plant Cell 27(8):2095–2118
2. Hsu PY, Benfey PN (2018) Small but mighty: functional peptides encoded by small ORFs in plants. Proteomics 18:1700038
3. Brunet MA, Leblanc S, Roucou X (2020) Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. Exp Cell Res 393(1):112057
4. Brunet MA, Levesque SA, Hunting DJ, Cohen AA, Roucou X (2018) Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. Genome Res 28(5): 609–624
5. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I et al (2022) Standardized annotation of translated open reading frames. Nat Biotechnol 40(7): 994–999
6. Lyapina I, Ivanov V, Fesenko I (2021) Peptidome: chaos or inevitability. Int J Mol Sci 22: 13128
7. Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, Macknight RC (2016) The emerging world of small ORFs. Trends Plant Sci 21(4):317–328
8. Takahashi F, Hanada K, Kondo T, Shinozaki K (2019) Hormone-like peptides and small coding genes in plant stress signaling and development. Curr Opin Plant Biol 51:88–95
9. Andrews SJ, Rothnagel JA (2014) Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet 15(3):193–204
10. Couso JP, Patraquim P (2017) Classification and function of small open reading frames. Nat Rev Mol Cell Biol 18(9):575–589
11. Plaza S, Menschaert G, Payre F (2017) In search of lost small peptides. Annu Rev Cell Dev Biol 33:391–416
12. Wright BW, Yi Z, Weissman JS, Chen J (2022) The dark proteome: translation from noncanonical open reading frames. Trends Cell Biol 32(3):243–258
13. Orr MW, Mao Y, Storz G, Qian SB (2021) Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res 48(3):1029–1042
14. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R et al (2015) Origins of de novo genes in human and chimpanzee. PLoS Genet 11(12):e1005721
15. Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà MM (2018) Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat Ecol Evol 2(5):890–896
16. Ruiz-Orera J, Albà MM (2019) Translation of small open reading frames: roles in regulation and evolutionary innovation. Trends Genet 35(3):186–198
17. Ruiz-Orera J, Villanueva-Cañas JL, Albà MM (2020) Evolution of new proteins from translated sORFs in long non-coding RNAs. Exp Cell Res 391(1):111940
18. Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L et al (2021) Uncovering de novo gene birth in yeast using deep transcriptomics. Nat Commun 12(1):604
19. Fesenko I, Shabalina SA, Mamaeva A, Knyazev A, Glushkevich A, Lyapina I et al (2021) A vast pool of lineage-specific microproteins encoded by long non-coding RNAs in plants. Nucleic Acids Res 49(18): 10328–10346
20. Goto H, Okuda S, Mizukami A, Mori H, Sasaki N, Kurihara D et al (2011) Chemical visualization of an attractant peptide, LURE. Plant Cell Physiol 52(1):49–58
21. Santiago J, Brandt B, Wildhagen M, Hohmann U, Hothorn LA, Butenko MA et al (2016) Mechanistic insight into a peptide hormone signaling complex mediating floral organ abscission. eLife 5:e15075
22. Covey PA, Subbaiah CC, Parsons RL, Pearce G, Lay FT, Anderson MA et al (2019) A pollen-specific RALF from tomato that regulates pollen tube elongation. Plant Physiol 153:703–715
23. Hsu PY, Calviello L, Wu HYL, Li FW, Rothfels CJ, Ohler U et al (2016) Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. Proc Natl Acad Sci U S A 113(45):E7126–E7135
24. Juntawong P, Girke T, Bazin J, Bailey-Serres J (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. Proc Natl Acad Sci U S A 111(1):E203–E212
25. Bazin J, Baerenfaller K, Gosai SJ, Gregory BD, Crespi M, Bailey-Serres J (2017) Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational

regulation. Proc Natl Acad Sci U S A 114(46): E10018–E10027

26. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ et al (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 9(1):59–64
27. Vanderperre B, Lucier JF, Bissonnette C, Motard J, Tremblay G, Vanderperre S et al (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS One 8(8):e70698
28. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M et al (2014) Extensive translation of small open reading frames revealed by poly-ribo-seq. eLife 3:e03528
29. Huang JZ, Chen M, Chen D, Gao XC, Zhu S, Huang H et al (2017) A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. Mol Cell 68(1): 171–184
30. Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F et al (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science 351:271–275
31. Makarewich CA, Baskin KK, Munir AZ, Bezprozvannaya S, Sharma G, Khemtong C et al (2018) MOXI is a mitochondrial micropeptide that enhances fatty acid β-oxidation. Cell Rep 23(13):3701–3709
32. Prensner JR, Enache OM, Luria V, Krug K, Clauser KR, Dempster JM et al (2021) Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol 39(6):697–704
33. Boix O, Martinez M, Vidal S, Giménez-Alejandre M, Palenzuela L, Lorenzo-Sanz L et al (2022) pTINCR microprotein promotes epithelial differentiation and suppresses tumor growth through CDC42 SUMOylation and activation. Nat Commun 13(1): 6840
34. Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27(13):i275–i282
35. Hanada K, Zhang X, Borevitz JO, Li WH, Shiu SH (2007) A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. Genome Res 17(5):632–640
36. Miravet-Verde S, Ferrar T, Espadas-García G, Mazzolini R, Gharrab A, Sabido E et al (2019) Unraveling the hidden universe of small proteins in bacterial genomes. Mol Syst Biol 15(2):e8290
37. Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K et al (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes. Proc Natl Acad Sci U S A 110(6):2395–2400
38. Wang S, Tian L, Liu H, Li X, Zhang J, Chen X et al (2020) Large-scale discovery of non-conventional peptides in maize and Arabidopsis through an integrated peptidogenomic pipeline. Mol Plant 13(7):1078–1093
39. Hazarika RR, De Coninck B, Yamamoto LR, Martin LR, Cammue BPA, Van Noort V (2017) ARA-PEPs: a repository of putative SORF-encoded peptides in Arabidopsis thaliana. BMC Bioinformatics 18(1):37
40. Couzigou J-M, Lauressergues D, Bécard G, Combier J-P, Ecard GB (2015) miRNA-encoded peptides (miPEPs): a new tool to analyze the roles of miRNAs in plant biology. RNA Biol 12:1178–1180
41. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM (2014) Long non-coding RNAs as a source of new peptides. eLife 3:e03523
42. Hartford CCR, Lal A (2020) When long noncoding becomes protein coding. Mol Cell Biol 40(6):e00528–e00519
43. Kurihara Y, Makita Y, Shimohira H, Fujita T, Iwasaki S, Matsui M (2020) Translational landscape of protein-coding and non-protein-coding RNAs upon light exposure in Arabidopsis. Plant Cell Physiol 61(3):536–545
44. Liang Y, Zhu W, Chen S, Qian J, Li L (2021) Genome-wide identification and characterization of small peptides in maize. Front Plant Sci 12:695439
45. Wu HYL, Song G, Walley JW, Hsu PY (2019) The tomato translational landscape revealed by transcriptome assembly and ribosome profiling. Plant Physiol 181(1):367–380
46. Mergner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A et al (2020) Mass-spectrometry-based draft of the Arabidopsis proteome. Nature 579:409–414
47. Wang P, Yao S, Kosami K-I, Guo T, Li J, Zhang Y et al (2020) Identification of endogenous small peptides involved in rice immunity through transcriptomics-and proteomics-based screening. Plant Biotechnol J 18:415–428
48. Jorge GL, Balbuena TS (2021) Identification of novel protein-coding sequences in Eucalyptus grandis plants by high-resolution mass

spectrometry. Biochim Biophys Acta Proteins Proteom 1869:140594

49. Fesenko I, Kirov I, Kniazev A, Khazigaleeva R, Lazarev V, Kharlampieva D et al (2019) Distinct types of short open reading frames are translated in plant cells. Genome Res 29(9):1464–1477
50. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, Aguet F et al (2021) Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat Biotechnol 40:209–217
51. Chen J, Brunner AD, Cogan JZ, Nuñez JK, Fields AP, Adamson B et al (2020) Pervasive functional translation of noncanonical human open reading frames. Science 367:140–146
52. Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J et al (2014) Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J Proteome Res 13(3):1757–1765
53. Flower CT, Chen L, Jung HJ, Raghuram V, Knepper MA, Yang CR (2020) Genetic and genomics investigation of structure and function of the kidney: an integrative proteogenomics approach reveals peptides encoded by annotated lincRNA in the mouse kidney inner medulla. Physiol Genomics 52(10):485
54. Luo W, Xiao Y, Liang Q, Su Y, Xiao L (2019) Identification of potential auxin-responsive small signaling peptides through a peptidomics approach in arabidopsis thaliana. Molecules 24:3146
55. Barashkova AS, Rogozhin EA (2020) Isolation of antimicrobial peptides from different plant sources: does a general extraction method exist? Plant Methods 16:143
56. Damerval C, De Vienne D, Zivy M, Thiellement H (1986) Technical improvements in two-dimensional electrophoresis increase the level of genetic variation detected in wheat-seedling proteins. Electrophoresis 7(1):52–54
57. Chatterjee M, Gupta S, Bhar A, Das S (2012) Optimization of an efficient protein extraction protocol compatible with two-dimensional electrophoresis and mass spectrometry from recalcitrant phenolic rich roots of chickpea (Cicer arietinum L.). Int J Proteomics 2012: 536963
58. Shi Y, Li J, Li L, Lin G, Bilal AM, Smagghe G et al (2021) Genomics, transcriptomics, and peptidomics of Spodoptera frugiperda (Lepidoptera, Noctuidae) neuropeptides. Arch Insect Biochem Physiol 106:e21740
59. Culver KD, Allen JL, Shaw LN, Hicks LM (2021) Too hot to handle: antibacterial peptides identified in ghost pepper. J Nat Prod 84:2200–2208
60. Kuljanin M, Dieters-Castator DZ, Hess DA, Postovit L-M, Lajoie GA (2017) Comparison of sample preparation techniques for large-scale proteomics. Proteomics 17(1–2): 1600337
61. Flower CT, Chen L, Jung HJ, Raghuram V, Knepper MA, Yang C-R (2020) An integrative proteogenomics approach reveals peptides encoded by annotated lincRNA in the mouse kidney inner medulla. Physiol Genomics 52:485–491
62. Cao S, Liu X, Huang Y, Yan Y, Zhou C, Shao C et al (2021) Proteogenomic discovery of sORF-encoded peptides associated with bacterial virulence in Yersinia pestis. Commun Biol 4:1248
63. Grossmann J, Roschitzki B, Panse C, Fortes C, Barkow-Oesterreicher S, Rutishauser D et al (2010) Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. J Proteome 73(9):1740–1746
64. Eng JK, Mccormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5:977–989
65. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probablity-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551–3557
66. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 3(8):1454–1463
67. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20(9):1466–1467
68. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM et al (2004) Open mass spectrometry search algorithm. J Proteome Res 3:958–964
69. Fu Y, Yang Q, Sun R, Li D, Zeng R, Ling CX et al (2004) Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. Bioinformatics 20(12):1948–1954
70. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M et al (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 77(14):4626–4639

71. Bern M, Cai Y, Goldberg D (2007) Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. Anal Chem 79(4):1393–1400
72. Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. Proteomics 13(1):22–24
73. Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5(1):5277
74. Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc 11(12):2301–2319
75. Zeng X, Ma B (2021) MSTracer: a machine learning software tool for peptide feature detection from liquid chromatography-mass spectrometry data. J Proteome Res 20(7): 3455–3462
76. Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K, Shiu S-H (2010) sORF finder: a program package to identify small open reading frames with high coding potential. Bioinformatics 26(3):399–400
77. Yang X, Jensen SI, Wulff T, Harrison SJ, Long KS (2016) Identification and validation of novel small proteins in Pseudomonas putida. Environ Microbiol Rep 8(6):966–674
78. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A et al (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17:2337–2342
79. Han Y, Ma B, Zhang K (2005) Spider: software for protein identification from sequence tags with de novo sequencing error. J Bioinforma Comput Biol 3(3):697–716
80. Jeong K, Kim S, Pevzner PA (2013) UniNovo: a universal tool for de novo peptide sequencing. Bioinformatics 29(16): 1953–1962
81. Chi H, Chen H, He K, Wu L, Yang B, Sun R-X et al (2013) pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. J Proteome Res 12:615–625
82. Ma B (2015) Novor: real-time peptide de novo sequencing software. J Am Soc Mass Spectrom 26:1885–1894
83. Tran NH, Zhang X, Xin L, Shan B, Li M (2017) De novo peptide sequencing by deep learning. Proc Natl Acad Sci U S A 114(31): 8247–8252
84. Tran NH, Qiao R, Xin L, Chen X, Liu C, Zhang X et al (2019) Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nat Methods 16(1):63–66
85. Pathan M, Samuel M, Keerthikumar S, Mathivanan S (2017) Unassigned MS/MS spectra: who am I? In: Keerthikumar S, Mathivanan S (eds) Proteome bioinformatics. Methods in molecular biology, vol 1549. Humana Press, New York, pp 67–74
86. Muth T, Renard BY (2018) Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Brief Bioinform 19(5): 954–970
87. Wu H, Johnson MC, Lu CH, Fritsche KL, Thomas AL, Lai Y et al (2015) Peptidomics study of anthocyanin-rich juice of elderberry. Talanta 131:640–644
88. Gemperline E, Keller C, Jayaraman D, Maeda J, Sussman MR, Ané J-MA et al (2016) Examination of endogenous peptides in Medicago truncatula using mass spectrometry imaging. J Proteome Res 15:4403–4411
89. Gemperline E, Keller C, Li L (2016) Mass spectrometry in plant-omics. Anal Chem 88(7):3422–3434
90. Ye X, Zhao N, Yu X, Han X, Gao H, Zhang X (2016) Extensive characterization of peptides from Panax ginseng C. A. Meyer using mass spectrometric approach. Proteomics 16: 2788–2791
91. Zhang K, Mckinlay C, Hocart CH, Djordjevic MA (2006) The Medicago truncatula small protein proteome and peptidome. J Proteome Res 12:3355–3367
92. Wang X, Li Y, Wu Z, Wang H, Tan H, Peng J (2014) JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. Mol Cell Proteomics 13(12):3663–3673
93. Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, Schubert OT et al (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol 32:219–223
94. Wilhelm M, Zolg DP, Graber M, Gessulat S, Schmidt T, Schnatbaum K et al (2021) Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. Nat Commun 12:3346
95. Gessulat S, Schmidt T, Zolg DP, Samaras P, Schnatbaum K, Zerweck J et al (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods 16:509–518
96. Ekvall M, Truong P, Gabriel W, Wilhelm M, Käll L (2022) Prosit transformer: a transformer for prediction of MS2 spectrum intensities. J Proteome Res 21(5):1359–1364

97. Gabriels R, Martens L, Degroeve S (2019) Updated MS^2PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. Nucleic Acids Res 47(W1):W295–W299
98. Beer LA, Liu P, Ky B, Barnhart KT, Speicher DW (2017) Efficient quantitative comparisons of plasma proteomes using label-free analysis with MaxQuant. Methods Mol Biol 1619:339–352
99. Gerster S, Kwon T, Ludwig C, Matondo M, Vogel C, Marcotte EM et al (2014) Statistical approach to protein quantification. Mol Cell Proteomics 13(2):666–677
100. Fabre B, Lambour T, Bouyssié D, Menneteau T, Monsarrat B, Burlet-Schiltz O et al (2014) Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. EuPA Open Proteom 4:82–86
101. Yeung YG, Stanley ER (2010) Rapid detergent removal from peptide samples with ethyl acetate for mass spectrometry analysis. Curr Protoc Protein Sci 16(16):12
102. Michel AM, Fox G, Kiran A M, De Bo C, O'Connor PBF, Heaphy SM et al (2014) GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Res 42: D859–D864
103. Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z (2019) RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res 47:D230–D234
104. Chen Y, Li D, Fan W, Zheng X, Zhou Y, Ye H et al (2020) PsORF: a database of small ORFs in plants. Plant Biotechnol J 18:2158–2160
105. Wethmar K, Barbosa-Silva A, Andrade-Navarro MA, Leutz A (2014) uORFdb--a comprehensive literature database on eukaryotic uORF biology. Nucleic Acids Res 42: D60–D67
106. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M et al (2016) Detecting actively translated open reading frames in ribosome profiling data. Nat Methods 13(2):165–170
107. Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP et al (2018) Improved Ribo-seq enables identification of cryptic translation events. Nat Methods 15(5):363–366
108. Xiao Z, Huang R, Xing X, Chen Y, Deng H, Yang X (2018) De novo annotation and characterization of the translatome with ribosome profiling data. Nucleic Acids Res 46(10):e61
109. Perkins P, Mazzoni-Putman S, Stepanova A, Alonso J, Heber S (2019) RiboStreamR: a web application for quality control, analysis, and visualization of Ribo-seq data. BMC Genomics 20:422
110. Larry Wu H-Y, Yingshan Hsu P (2021) RiboPlotR: a visualization tool for periodic Ribo-seq reads. Plant Methods 17:124
111. Song B, Jiang M, Gao L (2021) RiboNT: a noise-tolerant predictor of open reading frames from ribosome-protected footprints. Life (Basel) 11(7):701
112. Zhou P, Silverstein KAT, Gao L, Walton JD, Nallu S, Guhlin J et al (2013) Detecting small plant peptides using SPADA (small peptide alignment discovery application). BMC Bioinformatics 14(1):335
113. Zhu M, Gribskov M (2019) MiPepid: MicroPeptide identification tool using machine learning. BMC Bioinformatics 20(1):559
114. Tong X, Hong X, Xie J, Liu S (2020) CPPred-sORF: coding potential prediction of sORF based on non-AUG. bioRxiv. https://doi.org/10.1101/2020.03.31.017525
115. Zhao S, Meng J, Luan Y (2022) LncRNA-encoded short peptides identification using feature subset recombination and ensemble learning. Interdiscip Sci 14(1):101–112
116. Zhang Y, Jia C, Fullwood MJ, Kwoh CK (2021) DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. Brief Bioinform 22(2):2073–2084
117. Kersten RD, Yang Y, Xu Y, Cimermancic P, Nam S-J, Fenical W et al (2011) A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat Chem Biol 7(11):794–802
118. Cao X, Slavoff SA (2020) Non-AUG start codons: expanding and regulating the small and alternative ORFeome. Exp Cell Res 391(1):111973
119. Na CH, Barbhuiya MA, Kim MS, Verbruggen S, Eacker SM, Pletnikova O et al (2018) Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. Genome Res 28(1):25–36
120. Li YR, Liu MJ (2020) Prevalence of alternative AUG and non-AUG translation initiators and their regulatory effects across plants. Genome Res 30(10):1418–1433
121. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S et al (2022) The PRIDE database resources in 2022: a hub for mass

spectrometry-based proteomics evidences. Nucleic Acids Res 50(D1):D543–D552

122. Patel N, Mohd-Radzman NA, Corcilius L, Crossett B, Connolly A, Cordwell SJ et al (2018) Diverse peptide hormones affecting root growth identified in the Medicago truncatula secreted peptidome. Mol Cell Proteomics 17(1):160–174
123. Chen YL, Lee CY, Cheng KT, Chang WH, Huang RN, Nam HG et al (2014) Quantitative peptidomics study reveals that a wound-induced peptide from PR-1 regulates immune signaling in tomato. Plant Cell 26(10): 4135–4148
124. Das D, Jaiswal M, Khan FN, Ahamad S, Kumar S (2020) PlantPepDB: a manually curated plant peptide database. Sci Rep 10(1):2194
125. Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Makałowska I (2019) CANTATAdb 2.0: expanding the collection of plant long noncoding RNAs. In: Chekanova JA, Wang HLV (eds) Plant long non-coding RNAs, Methods in molecular biology, vol 1933. Humana Press, New York, pp 415–429
126. Singh A, Vivek AT, Kumar S (2021) AlnC: an extensive database of long non-coding RNAs in angiosperms. PLoS One 16(4):e0247215
127. Niu R, Zhou Y, Zhang Y, Mou R, Tang Z, Wang Z et al (2020) uORFlight: a vehicle toward uORF-mediated translational regulation mechanisms in eukaryotes. Database 2020:baaa007
128. Niarchou A, Alexandridou A, Athanasiadis E, Spyrou G (2013) C-PAmP: large scale analysis and database construction containing high scoring computationally predicted antimicrobial peptides for all the available plant species. PLoS One 8(11):e79728
129. Wang J, Yin T, Xiao X, He D, Xue Z, Jiang X et al (2018) StraPep: a structure database of bioactive peptides. Database 2018:bay038
130. Shi G, Kang X, Dong F, Liu Y, Zhu N, Hu Y et al (2022) DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. Nucleic Acids Res 50(D1):D488–D496
131. Boschiero C, Dai X, Lundquist PK, Roy S, de Bang TC, Zhang S et al (2020) MtSSPdb: the Medicago truncatula small secreted peptide database. Plant Physiol 183(1):399–413
132. Lin X, Lin W, Ku YS, Wong FL, Li MW, Lam HM et al (2020) Analysis of soybean long non-coding RNAs reveals a subset of small peptide-coding transcripts. Plant Physiol 182(3):1359–1374

Chapter 20

Gene Expression Analysis by Quantitative Real-Time PCR for Floral Tissues

Raquel Álvarez-Urdiola, Mariana Bustamante, Joana Ribes, and José Luis Riechmann

Abstract

Real-time, or quantitative, reverse transcription polymerase chain reaction (qRT-PCR) is a powerful method for rapid and reliable quantification of mRNA abundance. Although it has not featured prominently in flower development research in the past, the availability of novel techniques for the synchronized induction of flower development, or for the isolation of cell-specific mRNA populations, suggests that detailed quantitative analyses of gene expression over time and in specific tissues and cell types by qRT-PCR will become more widely used. In this chapter, we discuss specific considerations for studying gene expression by using qRT-PCR, such as the identification of suitable reference genes for the experimental set-up used. In addition, we provide protocols for performing qRT-PCR experiments in a multiwell plate format (with the LightCycler® 480 system, Roche) and with nanofluidic arrays (BioMark™ system, Fluidigm), which allow the automatic combination of sets of samples with sets of assays, and significantly reduce reaction volume and the number of liquid-handling steps performed during the experiment.

Key words Real-time PCR, qRT-PCR, Quantitative PCR (qPCR), SYBR Green I dye

1 Introduction

Differential gene expression, over time or among different cell and tissue types, is central to the developmental processes of all organisms. In flower development studies, this aspect of gene function has usually been approached by using methods to characterize spatial patterns or domains of gene expression, such as in situ hybridization and promoter-reporter gene fusions. Several groups have also progressed in the characterization of flower development in different plant species using quantitative real-time reverse transcription polymerase chain reaction (qRT-PCR) analyses [1–6], although this technique has not traditionally featured prominently in flower development research. Nevertheless, as a result of the development of techniques for the synchronized induction of

José Luis Riechmann and Cristina Ferrándiz (eds.), *Flower Development: Methods and Protocols*, Methods in Molecular Biology, vol. 2686, https://doi.org/10.1007/978-1-0716-3299-4_20,

flower development and for the isolation of cell-specific mRNA populations, detailed quantitative analyses of gene expression over time and in specific tissues and cells are becoming more broadly used. QRT-PCR is a powerful method for rapid and reliable quantification of mRNA abundance, which involves three processes: (i) the conversion of mRNA into cDNA via reverse-transcription; (ii) the amplification of the resulting cDNA by PCR; and (iii) the detection and quantification in real time of the synthesized PCR amplification products [7–9]. The reliability of the data obtained in qRT-PCR experiments can be affected by several factors that impact those processes, including template quality (RNA integrity [9, 10]), purity [9, 11] and quantity, efficiency of the RT reaction, PCR primer design, and efficiency of the PCR amplification [9]. To compensate for between-sample variations in the amount of starting material and in the efficiency of the qRT-PCR process, expression levels of the genes of interest are reported relative to one or more reference genes that are presumed to be uniformly and stably expressed across the tissues or conditions tested in the experiment, and whose abundance reflects the total amount of mRNA present in each sample. Thus, the reliability of qRT-PCR analyses is largely affected by the suitability of the gene (or genes) that is selected as a reference, that is, by whether or not such a gene really fulfils the requirements of a normalization control [12, 13].

Housekeeping genes, which function in basic cellular processes and are expressed in all cells of an organism, have often been used as reference genes to normalize the data in qRT-PCR experiments (e.g., genes such as glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), elongation factor-1α (*EF-1α*), actin (*ACT*), or tubulin (*TUB*)). Although the initial evidence indicating that housekeeping genes are stably expressed was obtained using methods that are mostly qualitative (for instance, RNA gel-blots and end-point RT-PCR), subsequent studies demonstrated that in some circumstances their expression may be regulated or be unstable, showing changes in transcript levels throughout development or among different conditions or tissues. Besides, housekeeping genes are usually expressed at higher levels than the typical genes of interest. For these reasons, using them as reference genes may introduce biases in the results obtained by qRT-PCR [12, 13]. For example, in a series of experiments designed to assess traditional *Arabidopsis* reference genes (including *ACT2, ACT7, ACT8, ADENINE PH OSPHORIBOSYLTRANSFERASE 1* (*APT1*), *EF1α, EUKARYOTIC TRANSLATION INITIATION FACTOR 4A1* (*eIF4A*), *TUB2, TUB6, TUB9, UBIQUITIN 4* (*UBQ4*), *UBQ5, UBQ10,* and *UBQ11*), it was found that *eIF4A* would appear to be stably expressed over the course of silique development when *APT1, UBQ5,* or *eF1α* were used to normalize the data, whereas its expression would appear quite variable when *TUB6* was used as reference gene [13].

In summary, the validity of "housekeeping" reference genes is not universal, and is highly dependent on the experimental conditions [12]. Thus, the selection of appropriate reference genes for the normalization of qRT-PCR data has emerged as a crucial component for successful expression studies carried out with this technology, and statistical algorithms like *geNorm* [14] or *BestKeeper* [15] have been developed for that purpose (*see* **Note 1**).

Concomitantly, the use of genome-wide technologies (i.e., initially DNA microarrays and subsequently RNA-Seq) to characterize gene expression changes across many different tissues and developmental stages, environmental conditions, or in response to biotic and abiotic stresses or perturbations has resulted in very rich datasets (e.g., [16]) that can be mined to identify novel, better suited reference genes for the desired experimental set-up. For instance, Czechowski et al. analyzed a very large set of Arabidopsis data obtained with Affymetrix ATH1 GeneChip arrays to identify several hundred genes that outperform traditional reference genes in terms of expression stability throughout development and under a range of environmental conditions [17]. Subsequent qRT-PCR experiments performed with a subset of those novel reference genes confirmed that they showed superior expression stability and lower absolute expression levels [17] (*see* **Note 2**). The results obtained in *Arabidopsis* have informed the selection of reference genes in other plant species, as the corresponding orthologous genes may also show stable expression (e.g., in Leafy spurge, *see* [18]). If candidate reference genes are selected based on orthology, however, their suitability needs to be confirmed experimentally, as such character is not always maintained across all experimental conditions in all organisms [9] (for instance, *see* [19]).

Candidate reference gene selections for various species, such as maize [20–23], rice [24–27], wheat [28–30], or strawberry [31, 32] and for specific conditions, tissues, or developmental stages (e.g., rice anther development, wheat meiosis, or strawberry fruits) have also been published. In addition, a literature review by Joseph et al. compiled a collection of reference genes for Arabidopsis and other plant species [33] (*see* Table 1).

The approach of using genome-wide data to select reference genes has been further expanded and refined with *RefGenes*, an online tool that allows easy identification of condition-specific reference genes [34]. *RefGenes* is based on the Genevestigator database of normalized and well-annotated microarray and RNA-Seq experiments and is accessible through the Genevestigator web page (www.genevestigator.com). The appropriateness of using condition-specific reference genes is based on the observation that for each biological context a subset of stable genes exists that has a smaller variance than either commonly used reference genes or genes that were selected for their stability across all conditions

Table 1
Arabidopsis general reference genes according with their expression stability under different conditions [33]

Accession number	Gene	Primers (5′–3′)	Conditions	
At1g50010	α-Tubulin	GATGTACCGTGGTGATGTC GAGCCTCTGAAAATTCTCC	Abiotic stress	Sulfate starvation, salt, drought, ABA
AT3G18780	Actin 2	CTTGCACCAAGCAGCATGAA CCGATCCAGACACTGTAC TTCCTT	Abiotic stress	Dehydration, cold, salt, oxidative, exposure to high light intensity
		TATGTGGCTATTCAGGCTGT TGGCGGTGCTTCTTCTCTG	Abiotic stress	Salt, mannitol, drought, and cold
		ATGCCATCCTCCGTCTTGAC CGCTCTGCTGTTGTGGTGAA	Biotic stress	*A. tumefaciens, H. schachtii, B. cinerea, P. syringae* pv. *maculicola, P. syringae* pv. *tomato*
At3g53750	Actin 3	GAGGCTCCTCTTAACCCAA TACAATTTCCCGCTCTGC	Abiotic stress	Salt stress, drought stress, ABA
At1g49240	Actin 8	TATGTGGCTATTCAGGCTGT TGGCGGTGCTTCTTCTCTG	Abiotic stress	Salt, mannitol, drought, and cold
		GGTGATGGTGTGTCT ACTGAGCACAATGTTAC	Biotic stress	*A. tumefaciens*
At1g13440	GAPDH	TTGGTGACAACAGG TCAAGCA AAACTTGTCGCTCAATGCAA TC	Abiotic stress	Salt, mannitol, drought, and cold
At2g41540	GAPDH	GAAGCAAGGCAAAGAAAT GAAGCAAGGCAAAGAAAT	Biotic stress	*A. tumefaciens*
At5g25760	UBC21	TTCAAATGGACCGCTCTTA TCA AAACACCGCCTTCGTAAGGA	Biotic stress	*A. tumefaciens*
At1g64230	UBC28	TCCAGAAGGATCCTCCAAC TTCCTGCAGT ATGG TTACGAGAAAGACACCGCC TGAATA	Abiotic stress	Salt, osmotic, temperature, radiomimetic, oxidative, UV, Zebularine, Trichostatin A, Sodium butyrate
At3g62250	UBQ5	GTAAACGTAGGTGAGTCC GACGCTTCATCTCGTCC	Abiotic stress	Drought, mannitol, and salt
		GACGCTTCATCTCGTCC GTAAACGTAGGTGAGTCC	Biotic stress	*B. cinerea; P. syringae pv. maculicola, P. syringae pv. Tomato*
At5g62690	Tubulin 2	CTCTGACCTCCGAAAGC TTGC	Abiotic stress	

(continued)

Table 1
(continued)

Accession number	Gene	Primers (5′–3′)	Conditions	
		TCACCTTCTTCATCCGCAG TT		Sucrose, NaCl, mannitol, paclobutrazol, hormonal
		AGCAATACCAAGATGCAAC TGCG	Biotic stress	*B. cinerea; P. syringae pv. maculicola*, *P. syringae pv. Tomato*
		TAACTAAATTATTCTCAGTAC TCTTCC		
At5g15710	F-BOX	TTTCGGCTGAGAGGTTCGAG T	Abiotic stress	Metal stress
		GATTCCAAGACG TAAAGCAGATCAA		
At5g08290	YLS8	TTACTGTTTCGGTTGTTC TCCATTT	Abiotic stress	Metal stress
		CACTGAATCATG TTCGAAGCAAGT		
At2g28390	SAND	AACTCTATGCAGCATTTGA TCCACT	Abiotic stress	Metal stress
		TGATTGCATATCTTTA TCGCCATC		
		AACTCTATGCAGCATTTGA TCCACT	Biotic stress	*P. infestans, A. laibachii*
		TGATTGCATATCTTTA TCGCCATC		
At5g60390	EF1-α	TGAGCACGCTCTTCTTGC TTTCA	Abiotic stress	Metal stress
		GGTGGTGGCATCCATCTTG TTACA		
AT5G46630	AP2M (CACS)	TCGATTGCTTGG TTTGGAAGAT	Development	Different tissues, organs, developmental stages, and genotypes
		GCACTTAGCGTGGACTCTG TTTGATC		
At1g58050	Helicase	CCATTCTACTTTTTGGCGGC T	Development	Different tissues, organs, developmental stages, and genotypes
		TCAATGGTAACTGATCCAC TCTGATG		
AT4G26410	Expressed	GAGCTGAAGTGGCTTCCA TGAC	Development	Different tissues, organs, developmental stages, and genotypes
		GGTCCGACATACCCATGA TCC		
AT4G34270	TIP41-like	GTGAAAACTG TTGGAGAGAAGCAA	Development	Different tissues, organs, developmental stages, and genotypes
		TCAACTGGATACCC TTTCGCA		

Primer sequences indicated in the table correspond to those used in the original experiment, as referenced in [33]

[34]. In other words, there is no gene that is universally stable, and the most appropriate set of reference genes for each biological context and specific experimental condition does vary.

Through *RefGenes*, users can select the transcriptomic experiments that are most similar to their chosen experimental conditions (including tissue, developmental stage, treatment, etc.). Afterward, the user indicates the set of target genes of interest (up to ten genes can be tested at once). A search is then triggered to identify those genes that have the lowest variance within the selected set of experiments and a range of expression that is similar to that of the target gene set. The result of the search is graphically displayed, showing the top 20–25 best candidate reference genes for the selected conditions. The behavior of these candidate genes in the chosen (or in additional) tissues or experimental conditions can then be explored using the *Conditions* tool of Genevestigator [35].

It is worth noting that the novel candidate reference genes that are identified using *RefGenes* and the aforementioned algorithms (*geNorm* or *Bestkeeper*) should be validated for the specific biological conditions of the experiments to be performed, for example, tissue type [36], growth conditions [24, 37], stresses [22, 38], treatments [39], etc. The evaluation of reference genes should be done by comparing the results with those obtained for other algorithms, experimentally, and preferably together with commonly used reference genes.

The use of *RefGenes* to select reference genes for flower development studies is illustrated in Figs. 1, 2, and 3, and in Table 2. Ten genes that participate in and/or are expressed at early stages of Arabidopsis flower development were used as target set to search for reference genes using the genome-wide expression profiling data available in Genevestigator (*SUPERMAN -SUP*, At3g23130-, *LEAFY -LFY*, At5g61850-, *AGL24* -At4g24540-, *YABBY3 -YAB3*, At4g00180-, *APETALA2 -AP2*, AT4g36920-, *AGL42* -At5g62165-, *SHATTERPROOF2 -SHP2*, At2g42830-, *AGAMOUS -AG*, AT4g18960-, *SEPALLATA3 -SEP3*, At1g24260-, and *APETALA3 -AP3*, At3g54340-, *see* [40]). *RefGenes* returns a list of candidate novel reference genes (Fig. 1, Table 2), which in this chapter are then compared to traditional reference genes (list of genes from [17]) and to reference genes for developmental processes (genes from [33] included in Table 1). The novel reference genes and the reference genes specifically selected for studying plant development are more stably expressed throughout all plant stages of development, and their mean expression level is generally lower than that of traditional reference genes, and thus closer to that of the typical genes of interest (*see* Fig. 2). Besides, novel reference genes selected for flower development studies are more stably expressed in floral tissues than traditional reference genes and the reference genes selected for studying other developmental processes (Fig. 3).

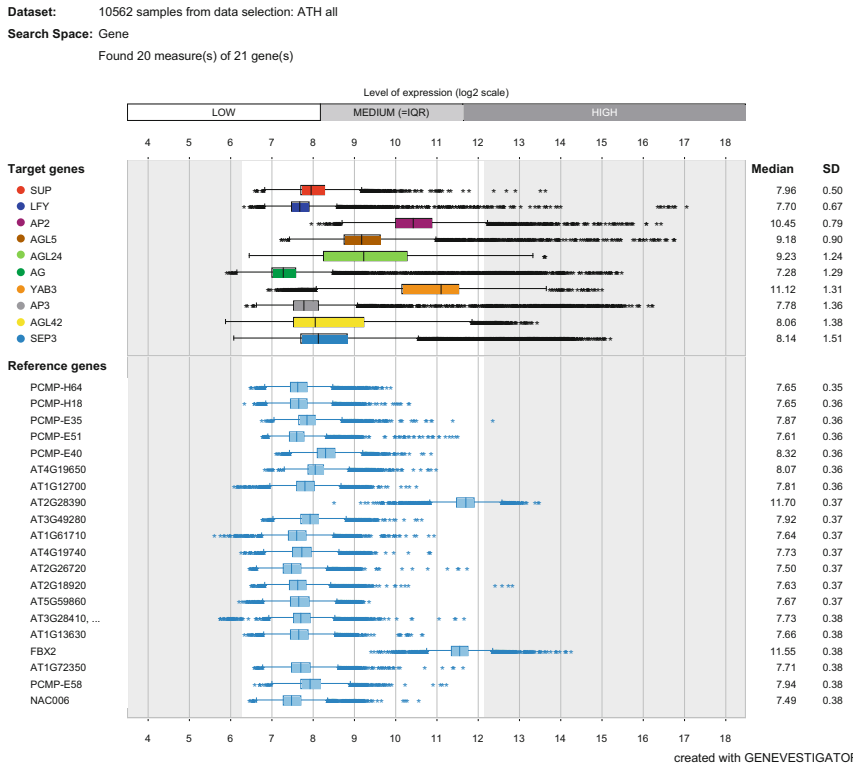


Fig. 1 Example of output results obtained when using the *RefGenes* tool (Anatomy-Inflorescence category in Genevestigator) with a set of floral regulatory genes (*SUP* (AT3G23130), *LFY* (AT5G61850), *AGL24* (AT4G24540), *YAB3* (AT4G00180), *AP2* (AT4G36920), *AGL42* (AT5G62165), *SHP2* (AT2G42830), *AG* (AT4G18960), *SEP3* (AT1G24260), and *AP3* (AT3G54340))

The detection of product formation in real-time during the amplification reaction of qRT-PCR experiments is carried out by measuring the emission signal from either fluorescent double-stranded DNA-binding dyes (such as SYBR® Green I and EvaGreen®, see below), or template-specific fluorescent probes (such as the TaqMan® probe technology). A general protocol for using SYBR Green I dye in a qRT-PCR experiment performed in a LightCycler® 480 Real-Time PCR system (Roche) is provided in this chapter (equally suited real-time PCR machines are available from various manufacturers). In addition to standard real-time PCR systems, in which reactions are performed either in thin-wall PCR tubes or in multiwell plates, newer systems based on nanofluidic arrays (such as the BioMark™ system, Fluidigm) have been developed for high-throughput analyses. These arrays contain nanofluidic networks that allow the automatic combination of sets of samples with sets of assays, significantly reducing reaction volume (and thus the amount of material needed to perform an assay) and the number of liquid-handling steps performed during the experiment. A protocol for a qRT-PCR experiment using EvaGreen® and the BioMark™ system is also provided.

2 Materials

2.1 Tissue Collection and RNA Extraction

1. RNase-free microcentrifuge tubes (1.5 mL).
2. Plastic pellet pestles for 1.5 mL microcentrifuge tubes (optional: a mixer motor or an electric drill).
3. Forceps (e.g., Dupont size #5).
4. Liquid nitrogen.
5. Vortex.
6. Microcentrifuge.

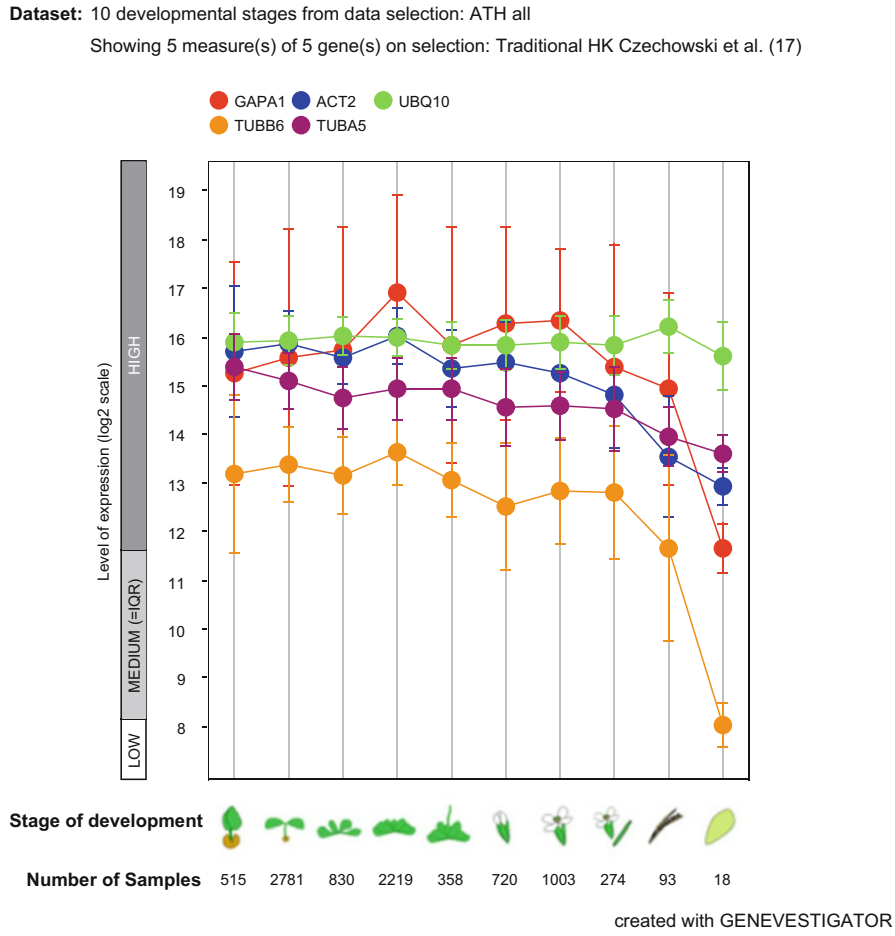


Fig. 2 Expression characteristics during plant development of some commonly used and novel reference genes in Arabidopsis inflorescences. (**a**) Traditional reference genes: *GAPDH* (AT3G26650, GAPA1), *ACT2* (AT3G18780), *UBQ10* (AT4G05320), *TUBB6* (AT5G12250), *TUBA5* (AT5G19780) [17]. (**b**) Reference genes for developmental processes: *AP2M* (*AT5G46630*), *AT1G58050*, *AT4G26410*, *AT4G34270* [33]. (**c**) Novel reference genes based on the expression of floral regulatory genes: *AT2G28390*, *AT5G15710*, *VPS45 (AT1G77140)*, *AT5G10700*, and *CLT2* (*AT4G24460*)

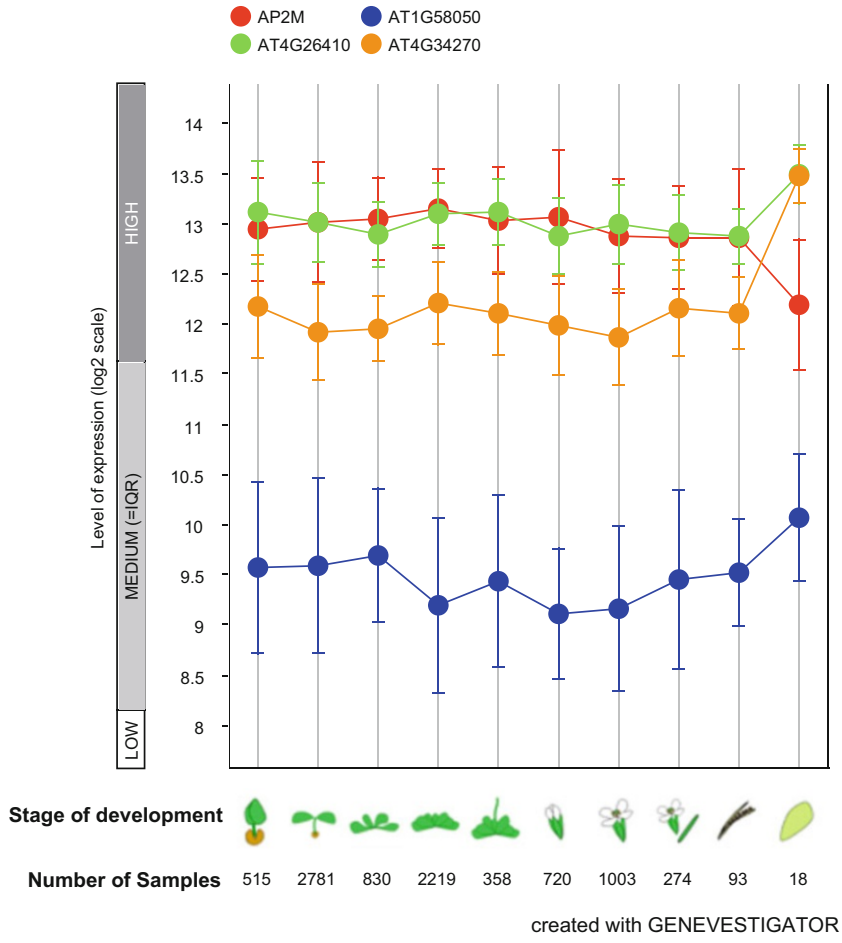


Fig. 2 (continued)

7. Spectrum Plant Total RNA Kit (Sigma-Aldrich) or an equivalent total RNA isolation kit or reagents (*see* **Note 3**).
8. Spectrophotometer (such as a Nanodrop).
9. Agilent Bioanalyzer and associated reagents (Agilent RNA 6000 Nano kit).

2.2 Reverse Transcription Reaction

1. High-Capacity cDNA Reverse Transcription Kit (e.g., Applied Biosystems; other commercial kits are available, but the protocols provided below are based on this kit) containing dNTPs (100 mM), MultiScribe reverse transcriptase (50 U/mL), reverse transcription Random Primers, reverse transcription buffer (10×), RNase inhibitor (20 U/mL).

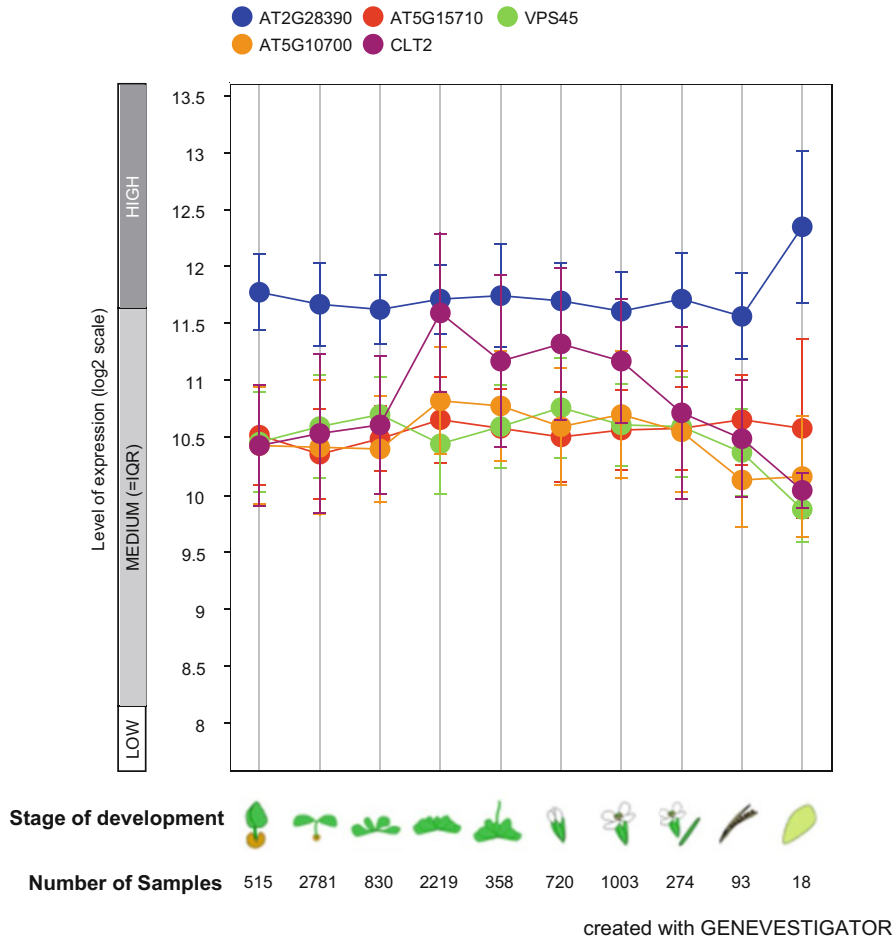


Fig. 2 (continued)

2. RNase-free PCR-tubes.
3. Nuclease-free water.

2.3 Quantitative Real Time PCR—LightCycler® 480 System

1. LightCycler® 480 SYBR Green I Master (Roche Diagnostics; other commercial kits are available, but the protocols provided in the following text are based on this kit): ready-to-use hot-start PCR mix containing FastStart Taq DNA Polymerase, reaction buffer, dNTP mix (with dUTP, instead of dTTP), SYBR Green I dye, and $MgCl_2$.
2. LC 480 Multiwell Plate 96 (Roche Diagnostics) (*see* **Note 4**).
3. Forward and reverse PCR primers at 100 μM each.
4. Nuclease-free water.

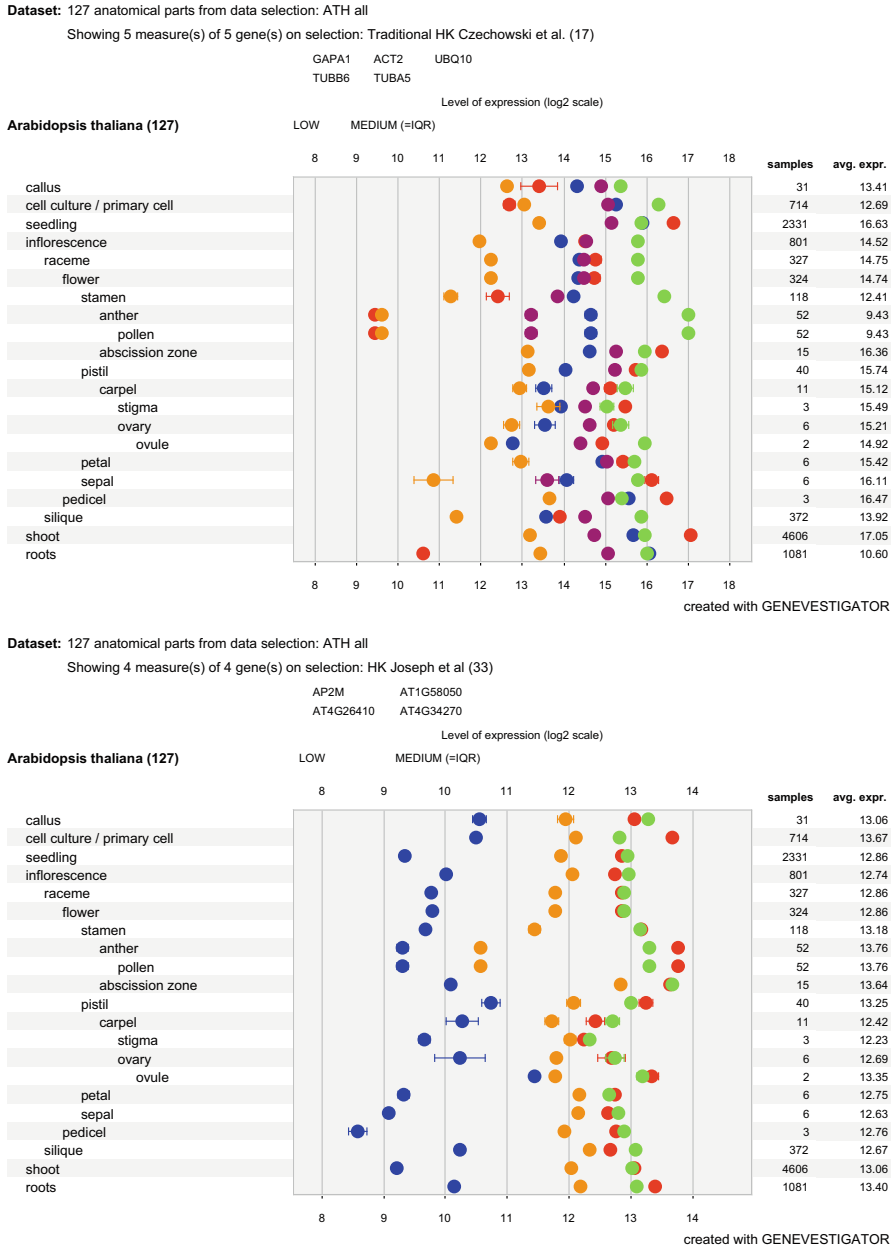


Fig. 3 Expression characteristics in different floral tissues of some commonly used and novel reference genes in Arabidopsis inflorescences. (**a**) Traditional reference genes. (**b**) Reference genes for developmental processes. (**c**) Novel reference genes based on the expression of floral regulatory genes (as in Fig. 2)

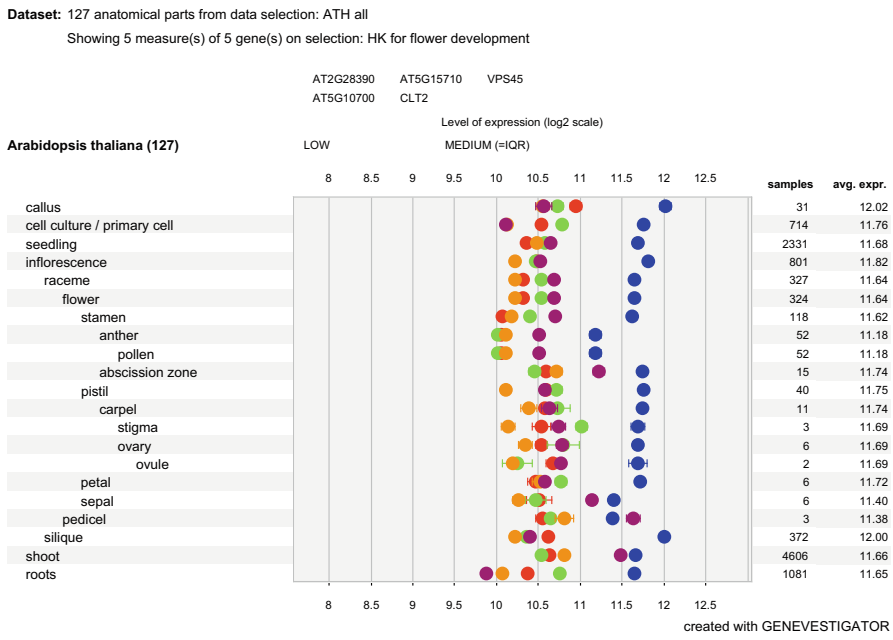


Fig. 3 (continued)

2.4 Quantitative Real Time PCR—BioMark™ System

1. TaqMan PreAmp Master Mix 2× (Applied Biosystems).
2. SsoFast EvaGreen SuperMix with Low ROX (Biorad): 2× real-time PCR mix, containing dNTPs, Sso7d fusion polymerase, $MgCl_2$, ROX passive reference dye and stabilizers.
3. 2× Assay Loading Reagent (Fluidigm).
4. 20× DNA Binding Dye Sample Loading Reagent (Fluidigm).
5. Exonuclease I (*E. coli*) (20,000 U/mL; New England Biolabs).
6. Exonuclease I Reaction Buffer 10× (New England Biolabs).
7. Forward and reverse PCR primers at 100 μM each.
8. Nuclease-free water.
9. TE Buffer: 10 mM Tris-HCl, pH 8.0, 1.0 mM EDTA (TEKnova).
10. DNA Suspension Buffer; 10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA (TEKnova).
11. 48.48 Dynamic Array IFC (Fluidigm).

3 Methods

The performance of the primers for a qRT-PCR experiment is crucial for obtaining high-quality results, and several aspects must be considered for successful primer design (*see* **Note 5**). There are

Table 2
Candidate novel reference genes for Arabidopsis proteins and peptides expressed in floral tissues identified using *RefGenes*. These genes were selected using as search set a list of floral regulatory genes

Gene	Annotation	Search set
AT2G28390	SAND family protein (MON1)	Floral regulatory genes
AT5G15710	Galactose oxidase/kelch repeat superfamily protein	Floral regulatory genes
AT1G77140	Vacuolar protein sorting 45 (VPS45)	Floral regulatory genes
AT5G10700	Peptidyl-tRNA hydrolase II (PTH2) family protein	Floral regulatory genes
AT4G24460	CRT (chloroquine-resistance transporter)-like transporter 2 (CLT2)	Floral regulatory genes
AT5G22760	PHD finger family protein (DDP2)	Floral regulatory genes
AT5G11380	1-deoxy-D-xylulose 5-phosphate synthase 3 (DXPS3)	Floral regulatory genes
AT5G04270	DHHC-type zinc finger family protein (PAT15)	Floral regulatory genes
AT1G50170	Sirohydrochlorin ferrochelatase B (SIRB)	Floral regulatory genes
AT3G59000	F-box/RNI-like superfamily protein	Floral regulatory genes
AT2G36480	ENTH/VHS family protein	Floral regulatory genes
AT5G52880	F-box family protein	Floral regulatory genes
AT5G65620	Zincin-like metalloproteases family protein (TOP1)	Floral regulatory genes
AT5G60750	CAAX amino terminal protease family protein. Encodes a chloroplast endoproteinase required for photosynthetic acclimation to higher light intensities (SCO4)	Floral regulatory genes
AT5G64970	Mitochondrial substrate carrier family protein	Floral regulatory genes
AT3G61180	RING/U-box superfamily protein	Floral regulatory genes
AT2G41790	Insulinase (Peptidase family M16) family protein	Floral regulatory genes
AT5G13050	5-formyltetrahydrofolate cycloligase (5FCL)	Floral regulatory genes
AT5G04920	EAP30/Vps36 family protein (VPS36)	Floral regulatory genes
AT3G59770	SacI homology domain-containing protein / WW domain-containing protein (SAC9)	Floral regulatory genes

many online resources for primer design, some of which also provide access to a consultative design service, such as:

- Oligoarchitect: http://www.oligoarchitect.com/LoginServlet
- RealTimeDesign: https://www.biosearchtech.com/support/tools/design-software/realtimedesign-software
- QuantPrime: http://www.quantprime.de/
- IDT-qPCR: http://eu.idtdna.com/scitools/Applications/RealTimePCR/

- Primer3: http://primer3.sourceforge.net/
- Primer-BLAST: http://www.ncbi.nlm.nih.gov/tools/primer-blast/

3.1 Tissue Collection and RNA Extraction

RNA quality (integrity and purity) is a critical factor for qRT-PCR experiments.

1. Harvest at least 100 mg of the desired plant tissue (e.g., inflorescences), into a 1.5 mL RNase-free microcentrifuge tube containing liquid nitrogen.
2. Grind the tissue to a fine powder with the pellet pestles (and a mixer motor), keeping the bottom of the tube immersed in liquid nitrogen throughout the grinding process to avoid RNA degradation (*see* **Notes 6** and 7).
3. Follow the manufacturer's instructions for the RNA extraction kit.
4. Analyze the integrity of the isolated RNA using a Bioanalyzer (or by using the 3′/5′ integrity assay, *see* [9]) and determine the concentration by absorption at 260 nm (e.g., with a Nanodrop spectrophotometer).

3.2 Reverse Transcription Reaction

The reverse transcription reaction to synthesize cDNA from the starting RNA material can be performed with various priming strategies, enzymes, and experimental conditions [8, 9]. However, to compare gene expression data across different experiments or laboratories, these variables should be kept constant, particularly ensuring that the same amount of RNA is added to each reaction (or that the enzyme/protocol used results in a proportional cDNA yield).

1. Prepare an RT master mix in a 1.5 mL tube:

Component	Volume (per reaction) (μL)
Water	4.2
10× RT Buffer (1×)	2
25× dNTP Mix (100 mM)	0.8
10× RT Random Primers	2
MultiScribe Reverse Transcriptase	1

2. Add 10 μL of Master Mix to each individual PCR-tube. Then add 100–1000 ng of each RNA sample, in a volume of 10 μL. The final reaction volume is 20 μL. No-RT control reaction (s) should be included in the experiment.

3. Briefly centrifuge the tubes to collect the contents and to eliminate any air bubbles.
4. Place the tubes in a thermal cycler using the following conditions:

	Step 1	Step 2	Step 3	Step 4
Temperature (°C)	25	37	85	4
Time	10 min	120 min	5 min	∞

5. Store cDNA samples at 4 °C (short term) or at −20 °C (for up to 6 months).

3.3 Quantitative Real Time PCR: LightCycler® 480 System

1. Set up the samples:
   1.1. Every gene/primer-pair combination used in a qPCR should be tested to calculate primer efficiency (*see* **Note 8**).
   1.2. The cDNA samples resulting from the RT reaction can be diluted in water, to obtain a final estimated concentration between 5 and 10 ng/μL (estimation based on the initial amount of RNA used in the RT reaction). This concentration range is ideal for the qRT-PCR. All amplification reactions should have a similar concentration of cDNA.
2. Before loading the PCR plate, and in order to minimize pipetting errors, it is important to prepare master mixes for each primer pair used. The accuracy of qPCR is highly dependent on accurate pipetting and thorough mixing of solutions. The protocol provided here uses SYBR® Green I chemistry, but other PCR-product detection chemistries could be used (*see* **Note 9**). To prepare the qPCR Master Mix, add components in the following order:

Component	Volume (per reaction) for 96-well plate (μL)
LC480 SYBR® Green I Master (2×) (Roche Diagnostics)	10
Water	6.4
Primer Forward (10 μM)	0.8
Primer Reverse (10 μM)	0.8

3. Loading the plate: Once all master mixes for each pair of primers are prepared, start loading the plate by adding first the Master Mix (18 μL) and then the cDNA samples (2 μL).

Avoid producing bubbles. The final reaction volume in each well is 20 μL. Then add the No Template Control (NTC) and no-Reverse Transcription control (no-RT, or RT) reactions (*see* **Note 10**). Seal the plate with LightCycler® 480 Sealing Foil by pressing it firmly to the plate surface, using your hand or a scraper. Sealing the plate properly is crucial to eliminate evaporation at high temperatures.

4. Place the multiwell plate in a standard swing-bucket centrifuge equipped with a rotor for multiwell plates with suitable adaptors. Balance it with a suitable counterweight (e.g., another multiwell plate). Centrifuge the plate at 1500 × *g* for 2 min.
5. Load the multiwell plate into the LightCycler® 480 Instrument and set-up the qPCR program (annealing temperature in the PCR is primer-dependent):

	Temperature (°C)	Time	Acquisition
Activation	95	10 min	None
PCR (45 Cycles)	95	10 s	None
	60	30 s	None
	72	30 s	Single
Melting	95	2 s	None
	65	15 s	None
	95	–	Continuous
Cooling	40	30 s	None

3.4 Quantitative Real Time PCR: BioMark™ System

BioMark System arrays allow for the automatic combination of sets of samples with sets of assays, significantly reducing reaction volume and the number of liquid-handling steps performed during the experiment. For instance, using the 48 × 48 array (as described in this protocol), 48 different samples (e.g., time-points in a time-course experiment) can be tested with up to 48 different assays (e.g., genes).

1. Specific Target Amplification (STA): This step is recommended to increase the number of copies of target DNA.
   1.1. STA Primer Mix (500 nM):
      1.1.1. Pool together 1 μL aliquots of all 100 μM primer sets to be included in the STA reaction (up to 100 different assays).
      1.1.2. Add DNA Suspension Buffer to make the final volume 200 μL.
      1.1.3. Vortex to mix and briefly spin reaction tube.

1.2. STA Pre-Mix:

1.2.1. In a DNA-free hood, prepare a Pre-Mix for the STA reaction:

Component	Volume (per reaction) (μL)
TaqMan PreAmp Master Mix	2.5
500 nM pooled STA Primer Mix	0.5
Water	0.75

1.2.2. Add 3.75 μL of STA Pre-Mix for each sample in a 96-well plate.

1.2.3. Add 1.25 μL of cDNA (at 10–20 ng/μL) to each reaction well, making a final volume of 5 μL. Include a no-PreAmplification control: add water instead of cDNA.

1.2.4. Seal the plate properly. Then, vortex and briefly spin the plate.

1.3. STA thermal cycle reaction:

1.3.1. Place the plate into the thermal cycler and run the following program (annealing temperature in the PCR is primer-dependent):

	Activation	16 cycles		Hold
Temperature (°C)	95	95	60	4
Time	10 min	15 s	4 min	∞

1.3.2. Eliminate the unincorporated primers from the STA amplification reaction. Prepare Exonuclease Mix as follows:

Component	Per 5 μL Sample
Water	1.4 μL
Exonuclease I Reaction Buffer	0.2 μL
Exonuclease I (20 units/μL)	0.4 μL

1.3.3. Add 2 μL of Exonuclease Mix to each 5 μL STA reaction. Vortex, centrifuge, and place in a thermal cycler.

	Digest	Inactivate	Hold
Temperature (°C)	37	80	4
Time	30 min	15 min	∞

1.3.4. Dilute the STA reaction to an appropriate final product concentration, as shown in the following text. A minimum dilution of five-fold should be used.

	Volume of water or TE Buffer		
Volume of STA Rx	5-fold dilution	10-fold dilution	20-fold dilution
7 μL	18 μL	43 μL	93 μL

Store diluted STA products at −20 °C or use immediately for on-chip PCR.

2. Sample and Assay Mix preparation:

2.1. Prepare Sample mix as shown in the following text:

Component	Volume per inlet with overage (μL)
2× SsoFast EvaGreen Supermix with Low ROX	3.0
20× DNA Binding Dye Sample Loading Reagent	0.3

2.2. In a new 96-well plate aliquot 3.3 μL of Sample mix and add 2.7 μL of each STA and Exo I-treated and diluted sample.

2.3. Seal the plate properly. Then, vortex and spin plate. Keep on ice.

2.4. Prior to preparing the Assay mix, combine the two primers of each primer pair making a final concentration of 20 μM.

2.5. Prepare Assay mix as shown in the following text:

Component	Volume per inlet with overage (μL)
2× Assay Loading Reagent	3.0
1× DNA Suspension Buffer	2.4

2.6. In a new 96-well plate, aliquot 5.4 μL of Assay mix and add in 1 μL of the 100 μM combined forward and reverse primers primer pair mix. The final concentration of each primer pair is 5 μM in the inlet and 500 nM in the final reaction.

2.7. Seal the plate properly. Then, vortex and spin the plate. Keep on ice.

3. Priming the 48 × 48 Dynamic Array™ IFC.

3.1. Inject control line fluid into each accumulator on the chip. Load the chip within 60 min of priming (refer to instrument manufacturer's instructions for details).

3.2. Remove and discard the blue protective film from the bottom of the chip.

3.3. Place the chip into the IFC controller for the 48 × 48 Dynamic Array IFC.

3.4. Run the Prime script for the 48 × 48 Dynamic Array IFC.

3.5. Pipette 5 μL of each assay and 5 μL of each sample into their respective inlets on the chip. Avoid creating bubbles while vortexing and when transferring reagents to the IFC, failure to do so may result in a decrease in data quality.

3.6. Place the chip to the IFC controller and run the Load Mix program.

3.7. After the program has run, take out the chip from the IFC controller and remove any dust particle from the chip surface.

3.8. Place the chip in the Biomark System and run the following program (annealing temperature in the PCR is primer-dependent):

	Activation	30 Cycles		Melting	
Temperature (°C)	95	96	60	60	95
Time	60 s	5 s	20 s	3 s	1 °C/3 s

3.5 Data Analysis

Different methodologies can be used for determination of the Quantification Cycle, Cq [41] (previously referred to as Ct/Cp/take off point):

- The threshold cycle method measures the Cq at a constant fluorescence level. These constant threshold methods assume that all samples have the same amplicon DNA concentration at the threshold fluorescence. The strength of this method is that it is extremely robust, but the threshold value needs to be adjusted for each experiment.
- The second derivative method calculates Cq as the second derivative maximum of the amplification curve. It is not user-dependent and is widely used.

Before performing the actual analysis, it is important to validate the data according to a variety of criteria (preferably following the Minimum Information for Publication of Real Time PCR Experiments: MIQE guidelines) (*see* **Note 11**, [41]). In particular:

- Check amplification curves. A normal amplification plot has three distinct phases: linear baseline, exponential, and plateau.
- Check controls (RT-, NTC).

- Check that the slope of the standard curve is between −3.2 and −3.5.
- Check technical replicates. They should be within 0.5 Cq of each other.
- Check melting peaks (when using a binding dye, or probes such as Molecular Beacons or Scorpions that are not hydrolyzed during the reaction) to verify that single, specific amplification products have been synthesized in the reaction.

3.5.1 Absolute Quantification

Absolute quantification relies on measurement to a standards curve constructed using the real-time PCR data obtained from amplification of these standards of known concentrations of template. Commonly, standards are derived from purified dsDNA plasmid, in vitro-transcribed RNA or in vitro-synthesized ssDNA. A standard curve (plot of Cq value against log of amount of standard) is generated using different dilutions of the standard. The Cq value of the target is compared with the standard curve, allowing calculation of the initial amount of the target. It is important to select an appropriate standard for the type of nucleic acid to be quantified. This method requires having the same efficiency of amplification in all reactions (reactions with experimental samples and reactions with the external standards). When using absolute quantification for determination of mRNA concentration, it is usual to correct absolute copy number of the specific target relative to absolute copy number of one or more reference genes.

3.5.2 Relative Quantification

Relative quantification relies on comparing the expression level of a target gene relative to a reference gene between a control sample and the test samples. Normalization to reference genes is the most common method for controlling for variation in qRT-PCR experiments. It is used to measure the relative change in mRNA expression levels. Many mathematical models are available. Most common relative quantification methods are:

(a) Pfaffl model [42]: combines gene quantification and normalization into a single calculation (Eq. 1). This model adjusts the amplification efficiencies (E) from target and reference genes in order to correct differences between the two assays.

$$\text{Ratio} = \frac{\left(E_{\text{target}}\right)^{\Delta \text{Cq target (control}-\text{sample)}}}{\left(E_{\text{reference}}\right)^{\Delta \text{Cq reference (control}-\text{sample)}}} \quad (1)$$

(b) $2^{-\Delta\Delta Cq}$ method [43]: This is a simpler version of the first model. Target and control amplification efficiency (E_{target} and $E_{\text{reference}}$) are assumed to be maximum (100%, i.e., a

value of 2, indicating amplicon doubling during each cycle) (Eq. 2). In addition, the relative expression of the target in all test samples is compared to that in a control or calibrator sample.

$$\text{Ratio} = 2^{-[\Delta Cq\ \ \text{Sample}-\Delta Cq\ \ \text{control}]} \quad (2)$$

4 Notes

1. *geNorm* is a widely used algorithm to determine the most stable reference from a given set of candidate genes on the basis of the *M* value (the *M* value is the internal control gene-stability measure, defined as the average pair-wise variation of a particular gene with all other control genes; genes with the lowest M values have the most stable expression) [18]. *geNorm* calculates and compares the *M* value of each pair of genes, and eliminates the gene with the highest *M* value, and then repeats this process with the remaining genes until the pair of genes with the lowest *M* value is identified. Thus, the genes forming this pair are considered as optimal reference genes among the initial candidate set.
2. The genome-wide analyses performed by Czechowski et al. led to the identification of many novel reference gene candidates, with purportedly better expression characteristics than traditional reference genes [17]. In these analyses the SD/MV ratio (SD/mean expression value, i.e., the coefficient of variation, or CV) for each gene in all the given experimental conditions (developmental series, abiotic stress series, hormone series, nutrient starvation and re-addition series, diurnal series, light series, and biotic stress series) is calculated. The gene that has the lowest CV value is considered as the gene with the most stable expression, and therefore a potential reference gene. Through these analyses, 25 reference genes, including 20 novel and 5 traditional ones, were recommended [17]. These genes were then validated by qRT-PCR and their expression stability ranked using the *geNorm* algorithm.
3. There are specific plates and films for the LC480 system that have been designed to ensure the best heat transfer from the thermal block and minimal autofluorescence, which is important to achieve a good signal-to-noise ratio in the detection of amplification products. In this protocol, we suggest using the LC 480 Multiwell Plate 96 from Roche.
4. The RNA preparation should be free of contaminating genomic DNA, so we recommend using a previously tested commercial kit for RNA isolation (*see* **Note 10**).

5. For primer design, it is important to consider the following points: (1) PCR products should be short (ideal length is from 70 to 250 bp). (2) The gene-specific forward and reverse primers should have similar melting temperatures (Tm) and length. (3) Primers should be between 15 and 25 nucleotides long and with a G/C content of around 50%. (4) Primers should have low or no self-complementarity to avoid the formation of primer dimers. (5) For the same reason, avoid pairs of primers that show sequence complementarity at their 3′ ends. (6) Primers that span introns or cross intron/exon boundaries are advantageous because they allow to distinguish amplification from cDNA or from contaminant genomic DNA. Primers should be ordered with desalt purification. Primer stock solutions should be prepared with DNase/RNase-free water. Make aliquots to avoid contamination and repeated freezing/thawing. Original stock of PCR primers should be stored at −20 °C and working dilutions at 4 °C for up to 2 weeks.
6. The presence of liquid nitrogen inside the microcentrifuge tubes during tissue grinding should be avoided, to prevent potential loss of tissue by nitrogen spill, or by the popping of the tube if closed with liquid nitrogen inside. Tubes can be pre-chilled in liquid nitrogen. As an alternative for grinding the tissue, mortar and pestle could be used instead of pellet pestles and an electric drill.
7. Both fresh and frozen (−80 °C) tissue can be used as starting material, and ground plant material can be stored at −80 °C before RNA purification. However, do not allow the frozen material to thaw before grinding or before the first solution of the RNA purification procedure is added.
8. Make a 4-step dilution series (1:4 dilutions) from cDNA samples. To evaluate the efficiency of the PCR reaction, it is important to generate at least one standard curve for each primer pair. A standard curve graph is made by plotting the Ct/Cp values on the *y*-axis and the logarithm of the input amounts on the *x*-axis. The slope of the line of this plot will give the efficiency of the reaction according to the equation $E = [10\text{^}(-1/\text{slope})] - 1$; slope should be between −3.2 and −3.5 and $R^2 > 0.98$.
9. SYBR® Green I and EvaGreen® are the most used dye chemistries, due to cost and simple optimization process. However, these dyes bind to any double-stranded DNA formed in the reaction, including primer-dimers and other non-specific reaction products, which may result in an overestimation of the target concentration. Other methods, such as hydrolysis probes, may also be used. Probe-based qRT-PCR relies on

the sequence-specific detection of a desired PCR product. It utilizes a fluorescently labelled target-specific probe, which results in increased specificity and sensitivity.

10. No template controls (NTC) should be included for each pair of primers tested to ensure that there is no reagent contamination. In these control reactions, water is added instead of sample, so no amplification is expected. In case the NTC reaction shows the synthesis of amplification products (i.e., the presence of a contaminant), measures such as pipette decontamination, using new primers aliquots, or thorough bench cleaning might be necessary. No reverse transcription controls (no-RT, or RT-) are used to detect the presence of contaminant genomic DNA in the RNA samples. If the RT-reaction shows the synthesis of amplification products, the corresponding RNA samples should be treated with DNase prior to their use in the reverse transcription reaction. If the primers were designed to span an intron or an intron/exon boundary, it is not necessary to perform a no-RT control.
11. MIQE Guidelines [41]. The MIQE guidelines were published in response to the recognition that several publications contain little information that describes the qPCR or that gives the reader the opportunity to determine the quality of the experiment. The result of these omissions is that several publications contain misleading conclusions based on inadequate quality control of the technical process. The MIQE guidelines contain a step-by-step guide and checklist, which leads the experimenter through the process of experiment validation. This has the additional function of providing a framework for publication analysis by peer reviewers and journal editors. Several publishing houses are now requiring that MIQE guidelines are followed for papers containing qPCR data.

Acknowledgments

Work in the authors' laboratory was supported by grant BFU2014-58289-P (funded by MICIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe") and by grant 2017SGR718 (from the Agencia de Gestió d'Ajuts Universitaris I de Recerca) to JLR, and by institutional grant SEV-2015-0533 (funded by MCIN/AEI/10.13039/501100011033) and by the CERCA Programme / Generalitat de Catalunya. R.A. is supported by fellowship PRE2018-084278 funded by MCIN/AEI/ 10.13039/501100011033 and by "ESF Investing in your future." We acknowledge the contributions of Jian Jin, Oriol Casagran, and Tania Nolan to the previous, first edition of this chapter.

References

1. Aamir M, Karmakar P, Singh VK et al (2021) A novel insight into transcriptional and epigenetic regulation underlying sex expression and flower development in melon (Cucumis melo L.). Physiol Plant 173(4):1729–1764. https://doi.org/10.1111/ppl.13357
2. Haider S, Bashir MA, Habib U et al (2021) Phenotypic characterization and RT-qPCR analysis of flower development in F1 transgenics of Chrysanthemum × grandiflorum. Plan Theory 10(8):1681. https://doi.org/10.3390/plants10081681
3. de Moura SM, Rossi ML, Artico S et al (2020) Characterization of floral morphoanatomy and identification of marker genes preferentially expressed during specific stages of cotton flower development. Planta 252(4):71. https://doi.org/10.1007/s00425-020-03477-0
4. Deng MH, Zhao K, Lv JH et al (2020) Flower transcriptome dynamics during nectary development in pepper (Capsicum annuum L.). Genet Mol Biol 43(2):e20180267. https://doi.org/10.1590/1678-4685-GMB-2018-0267
5. Moschin S, Nigris S, Ezquer I et al (2021) Expression and functional analyses of Nymphaea caerulea MADS-box genes contribute to clarify the complex flower patterning of water lilies. Front Plant Sci 12:730270. https://doi.org/10.3389/fpls.2021.730270
6. Serra-Picó M, Hecht V, Weller JL et al (2022) Identification and characterization of putative targets of VEGETATIVE1/FULc, a key regulator of development of the compound inflorescence in pea and related legumes. Front Plant Sci 13:765095. https://doi.org/10.3389/fpls.2022.765095
7. Bustin SA, Benes V, Nolan T, Pfaffl MW (2005) Quantitative real-time RT-PCR – a perspective. J Mol Endocrinol 34(3):597–601. https://doi.org/10.1677/jme.1.01755
8. Kubista M, Andrade JM, Bengtsson M et al (2006) The real-time polymerase chain reaction. Mol Asp Med 27(2–3):95–125. https://doi.org/10.1016/j.mam.2005.12.007
9. Nolan T, Hands RE, Bustin SA (2006) Quantification of mRNA using real-time RT-PCR. Nat Protoc 1(3):1559–1582. https://doi.org/10.1038/nprot.2006.236
10. Fleige S, Pfaffl MW (2006) RNA integrity and the effect on the real-time qRT-PCR performance. Mol Aspects Med 27(2–3):126–139. https://doi.org/10.1016/j.mam.2005.12.003
11. Nolan T, Hands RE, Ogunkolade W, Bustin SA (2006) SPUD: a quantitative PCR assay for the detection of inhibitors in nucleic acid preparations. Anal Biochem 351(2):308–310. https://doi.org/10.1016/j.ab.2006.01.051
12. Gutierrez L, Mauriat M, Pelloux J, Bellini C, Van Wuytswinkel O (2008) Towards a systematic validation of references in real-time RT-PCR. Plant Cell 20(7):1734–1735. https://doi.org/10.1105/tpc.108.059774
13. Gutierrez L, Mauriat M, Guénin S et al (2008) The lack of a systematic validation of reference genes: a serious pitfall undervalued in reverse transcription-polymerase chain reaction (RT-PCR) analysis in plants. Plant Biotechnol J 6(6):609–618. https://doi.org/10.1111/j.1467-7652.2008.00346.x
14. Vandesompele J, De Preter K, Pattyn F et al (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3(7):RESEARCH0034. https://doi.org/10.1186/gb-2002-3-7-research0034
15. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper-Excel-based tool using pair-wise correlations. Biotechnol Lett 26(6):509–515. https://doi.org/10.1023/b:bile.0000019559.84305.47
16. Schmid M, Davison TS, Henz SR et al (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet 37(5):501–506. Available from: https://www.nature.com/articles/ng1543
17. Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible W-R (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. Plant Physiol [Internet] 139 (September):5–17. https://doi.org/10.1038/ng1543
18. Chao WS, Doğramaci M, Foley ME et al (2012) Selection and validation of endogenous reference genes for qRT-PCR analysis in leafy spurge (Euphorbia esula). PLoS One 7(8):e42839. https://doi.org/10.1371/journal.pone.0042839
19. Caldana C, Scheible WR, Mueller-Roeber B, Ruzicic S (2007) A quantitative RT-PCR platform for high-throughput expression profiling of 2500 rice transcription factors. Plant Methods 3:7. https://doi.org/10.1186/1746-4811-3-7

20. Alves Oliveira D, Tang JD, Warburton ML (2021) Reference gene selection for RT-qPCR analysis in maize kernels inoculated with Aspergillus flavus. Toxins 13(6):386. https://doi.org/10.3390/toxins13060386
21. Galli V, da Silva Messias R, dos Anjos e Silva SD, Rombaldi CV (2013) Selection of reliable reference genes for quantitative real-time polymerase chain reaction studies in maize grains. Plant Cell Rep 32(12):1869–1877. https://doi.org/10.1007/s00299-013-1499-x
22. Lin Y, Zhang C, Lan H, Gao S et al (2014) Validation of potential reference genes for qPCR in maize across abiotic stresses, hormone treatments, and tissue types. PLoS One 9(5): e95445. https://doi.org/10.1371/journal.pone.0095445
23. Manoli A, Sturaro A, Trevisan S et al (2012) Evaluation of candidate reference genes for qPCR in maize. J Plant Physiol 169(8): 807–815. https://doi.org/10.1016/j.jplph.2012.01.019
24. Auler PA, Benitez LC, do Amaral MN et al (2017) Selection of candidate reference genes and validation for real-time PCR studies in rice plants exposed to low temperatures. Genet Mol Res 16(2):16029695. https://doi.org/10.4238/gmr16029695
25. Auler PA, Benitez LC, do Amaral MN et al (2017) Evaluation of stability and validation of reference genes for RT-qPCR expression studies in rice plants under water deficit. J Appl Genet 58(2):163–177. https://doi.org/10.1007/s13353-016-0374-1
26. Ji Y, Tu P, Wang K et al (2014) Defining reference genes for quantitative real-time PCR analysis of anther development in rice. Acta Biochim Biophys Sin Shanghai 46(4): 305–312. https://doi.org/10.1093/abbs/gmu002
27. Bevitori R, Oliveira MB, Grossi-de-Sá MF et al (2014) Selection of optimized candidate reference genes for qRT-PCR normalization in rice (Oryza sativa L.) during Magnaporthe oryzae infection and drought. Genet Mol Res 13(4): 9795–9805. https://doi.org/10.4238/2014
28. Garrido J, Aguilar M, Prieto P (2020) Identification and validation of reference genes for RT-qPCR normalization in wheat meiosis. Sci Rep 10(1):2726. https://doi.org/10.1038/s41598-020-59580-5
29. Tenea GN, Peres Bota A, Cordeiro Raposo F, Maquet A (2011) Reference genes for gene expression studies in wheat flag leaves grown under different farming conditions. BMC Res Notes 4:373. https://doi.org/10.1186/1756-0500-4-373
30. Wu D, Dong J, Yao YJ et al (2015) Identification and evaluation of endogenous control genes for use in quantitative RT-PCR during wheat (Triticum aestivum L.) grain filling. Genet Mol Res 14(3):10530–10542. https://doi.org/10.4238/2015.September.8.15
31. Zhang Y, Peng X, Liu Y et al (2018) Evaluation of suitable reference genes for qRT-PCR normalization in strawberry (Fragaria × ananassa) under different experimental conditions. BMC Mol Biol 19(1):8. https://doi.org/10.1186/s12867-018-0109-4
32. Galli V, Borowski JM, Perin EC et al (2015) Validation of reference genes for accurate normalization of gene expression for real time-quantitative PCR in strawberry fruits using different cultivars and osmotic stresses. Gene 554(2):205–214. https://doi.org/10.1016/j.gene.2014.10.049
33. Joseph JT, Poolakkalody NJ, Shah JM (2018) Plant reference genes for development and stress response studies. J Biosci 43(1): 173–187. https://doi.org/10.1007/s12038-017-9728-z
34. Hruz T, Wyss M, Docquier M et al (2011) RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. BMC Genomics 12:156. https://doi.org/10.1186/1471-2164-12-156
35. Wang M, Bhullar NK (2021) Selection of suitable reference genes for qRT-PCR gene expression studies in rice. In: Bandyopadhyay A, Thilmony R (eds) Rice genome engineering and gene editing. Methods in molecular biology, vol 2238. Humana, New York, pp 293–312. https://doi.org/10.1007/978-1-0716-1068-8_20
36. Zheng T, Chen Z, Ju Y et al (2018) Reference gene selection for qRT-PCR analysis of flower development in Lagerstroemia indica and L. speciosa. PLoS One 13(3):e0195004. https://doi.org/10.1371/journal.pone.0195004
37. Škiljaica A, Jagić M, Vuk T et al (2022) Evaluation of reference genes for RT-qPCR gene expression analysis in Arabidopsis thaliana exposed to elevated temperatures. Plant Biol 24(2):367–379. https://doi.org/10.1111/plb.13382
38. Xu W, Dong Y, Yu Y et al (2020) Identification and evaluation of reliable reference genes for quantitative real-time PCR analysis in tea plants under differential biotic stresses. Sci Rep 10(1): 2429. https://doi.org/10.1038/s41598-020-59168-z

39. Joseph JT, Poolakkalody NJ, Shah JM (2019) Screening internal controls for expression analyses involving numerous treatments by combining statistical methods with reference gene selection tools. Physiol Mol Biol Plants 25(1): 289–301. https://doi.org/10.1007/s12298-018-0608-2
40. Wellmer F, Alves-Ferreira M, Dubois A et al (2006) Genome-wide analysis of gene expression during early Arabidopsis flower development. PLoS Genet 2(7):e117. https://doi.org/10.1371/journal.pgen.0020117.eor
41. Bustin SA, Benes V, Garson JA et al (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55(4):611–622. https://doi.org/10.1373/clinchem.2008.112797
42. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT–PCR. Nucleic Acids Res 29(9):e45. https://doi.org/10.1093/nar/29.9.e45
43. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2−ΔΔCT method. Methods 25(4):402–408. https://doi.org/10.1006/meth.2001.1262

Chapter 23

Multi-Omics Methods Applied to Flower Development

Raquel Álvarez-Urdiola, José Tomás Matus, and José Luis Riechmann

Abstract

Developmental processes in multicellular organisms depend on the proficiency of cells to orchestrate different gene expression programs. Over the past years, several studies of reproductive organ development have considered genomic analyses of transcription factors and global gene expression changes, modeling complex gene regulatory networks. Nevertheless, the dynamic view of developmental processes requires, as well, the study of the proteome in its expression, complexity, and relationship with the transcriptome. In this chapter, we describe a dual extraction method—for protein and RNA—for the characterization of genome expression at proteome level and its correlation to transcript expression data. We also present a shotgun proteomic procedure (LC-MS/MS) followed by a pipeline for the imputation of missing values in mass spectrometry results.

Key words Protein extraction, RNA extraction, Proteomics, Transcriptomics, Flower development, LC-MS/MS, Arabidopsis

1 Introduction

The capacity of cells to orchestrate different gene expression programs is crucial for developmental processes in multicellular organisms, and it is hardwired and encoded in the genome in the form of *cis*-regulatory sequences that interact with transcription factors, co-regulators, and other types of regulatory proteins or RNAs, as well as of epigenetic marks, altogether determining when, where, and how genes are expressed. For the past 20 years, the exponential advances in technologies and informatics tools for generating and processing large biological datasets (omics) have added new approaches to development studies in plants. Through the use of genomics and transcriptomics (in particular, RNA-Seq, ChIP-Seq, and other high-throughput sequencing-derived methods), the hierarchical levels of plant genetic and molecular organization are being described in detail. In particular, several studies of reproductive organ development have considered genome-wide analyses of transcription factor DNA-binding and global gene expression

José Luis Riechmann and Cristina Ferrándiz (eds.), *Flower Development: Methods and Protocols*, Methods in Molecular Biology, vol. 2686, https://doi.org/10.1007/978-1-0716-3299-4_23,

changes (e.g., [1–5]) and modeled complex gene regulatory networks (reviewed in [6–9]). Even so, a global and comprehensive view of developmental processes would also benefit from the characterization of the corresponding proteome.

The analysis of the proteome of eukaryotic cells is challenging due to the substantial diversity in the properties of the individual proteins that compose it (e.g., abundance, stability, molecular weight, structure, hydrophobicity, hydrophilicity, posttranslational modifications (PTMs), and so on). Nevertheless, along with an enhancement of throughput, sensitivity, and resolution of analytical technologies in MS, computational methods have been developed focusing on the identification and quantification of proteins in complex samples [10–13]. In plants, MS-based proteomics approaches have been applied for the measurement of differential protein expression or the detection of PTMs (e.g., [14, 15]) in different tissues and biological processes (reviewed in [13]). Deep proteome studies have led to the development of proteome atlases of the major plant organs for different plant species [16–21]. Besides, cell type-specific proteome studies are crucial for a better understanding of the unique biological functions and properties of individual cell types in a tissue [22], as well as subcellular plant proteomics and predictions [23–25]. As the proteome is in constant flux, several proteome studies are based on temporal series during developmental processes or stress responses [26–29].

Furthermore, results from more than one type of omics can be matched in order to obtain deeper insights into biological processes [16, 30–33]. These integration studies are usually referred as multi-omics, trans-omics, or integrated omics in current literature. Quantitative proteomics allows to study at a genome-wide level the correlation between mRNA expression levels and the abundance of the corresponding proteins (reviewed in [34, 35]), an issue that has been extensively studied in different species and processes during the past few years. For instance, in plants combined transcriptome-proteome analyses have already been used to study petal shape [36], carotenoid synthesis [37], photoperiodic control of the proteome [38], or leaf development [39], as well as reproductive development; in particular, embryogenesis [40], male reproductive development [41–43], and flower development either in general [44, 45] or focusing on the functions of specific proteins [46].

In these combined studies, the interpretation of the existence, or lack thereof, of correlation between the changes in transcript dynamics and protein abundance, and its biological meaning, is still a lingering issue: numerous studies conclude that there is not a strong correlation between the levels of these macromolecules [41, 43, 47–51], whereas in others such correlation is more apparent [38–40, 45]. The lack of correlation could be in part derived from the difficulties to obtain truly comparable datasets at the

transcript and protein levels, and because the sensitivity of extraction and quantification techniques for mRNAs and proteins highly differ. However, the observed differences might also be caused by posttranslational regulation of protein levels [47], or by their different expression and degradation kinetics, as longer protein half-lives buffer changes in mRNA levels [48–51]. Time-lapse studies could be an approach for addressing this gap, as successive analyses at different time points could allow the discovery of correlative behaviors of protein and mRNA levels through time [52, 53].

In addition, a major concern in label-free quantitative proteomics that hinders the subsequent data analysis and its comparison with other omics data is the high rate of missing values. Three types of missing values can be defined, depending on the nature of the missingness: (1) missing completely at random (MCAR) and (2) missing at random (MAR) values, which are due to minor errors or stochastic fluctuations and to conditional dependencies, respectively; and (3) missing not at random (MNAR) values, which have a targeted effect [54]. Depending on the nature of these "not assigned values" (NAs), different methods can be used to impute them. As there are many types of NAs that coexist in most quantitative datasets, hybrid strategies of imputation could be a better approach [54, 55].

In this chapter, we describe a protocol for common extraction of total proteins and RNA from the same Arabidopsis inflorescence samples to maximize comparability between the proteomic and transcriptomic data. We also present a shotgun proteomic procedure by liquid chromatography-tandem mass spectrometry (LC-MS/MS), and a pipeline for the imputation of missing values in the mass spectrometry results to distinguish the nature of the missingness and to treat NAs accordingly.

2 Materials

1. Mortar and pestle.
2. Liquid nitrogen.
3. Microcentrifuge tubes.

2.1 Protein Extraction

1. Protein low-binding tubes (2 mL).
2. Isopropanol.
3. 0.3 M guanidine in 95% ethanol.
4. 90% ethanol.
5. SDS-PAGE 5× buffer.
6. E buffer: 125 mM Tris–HCl pH 8.8, 1% (w/v) SDS, 10% (v/v) glycerol, 50 mM $Na_2S_2O_5$ [56].

2.2 RNA Extraction

1. RNase free tubes (1.5 mL).
2. Trizol.
3. Chloroform.
4. Phenol:chloroform:isoamyl alcohol (25:24:1).
5. LiCl 3 M.
6. 85% and 100% (v/v) ethanol.
7. DEPC water.

2.3 LC-MS/MS

1. DL-dithiothreitol (DTT) (*see* **Note 1**).
2. Iodoacetamide.
3. Urea.
4. Ammonium bicarbonate.
5. Endoproteinase LysC.
6. Trypsin.
7. Formic acid.
8. MicroSpin C18 columns (The Nest Group, Inc).
9. Nano Trap C18 columns with an inner diameter of 100 μm packed with C18 particles of 5 μm particle size (Thermo Fisher Scientific) (optional, depending on the setup of each laboratory).
10. Reverse-phase chromatography columns (C18, 2 μm, 15–50 cm length) (*see* **Note 2**).
11. Buffer A: 0.1% formic acid in water.
12. Buffer B: 0.1% formic acid in acetonitrile.
13. Bovine serum albumin (New England Biolabs cat # P8108S).
14. Orbitrap Eclipse mass spectrometer (Thermo Fisher Scientific) (*see* **Note 3**).
15. EASY-nLC 1000 (Thermo Fisher Scientific).

3 Methods

3.1 Protein Extraction

1. With a different mortar and pestle for each sample, grind the tissue (i.e., inflorescences) with liquid nitrogen until obtaining a whitish fine powder (*see* **Notes 4** and **5**).
2. Place the powder in a microcentrifuge tube (~250 mg per sample).
3. Add 1 mL of Trizol, vortex for at least 15 s until it is completely homogenized, and incubate on ice for 5 min. This step must be done in an extraction hood.

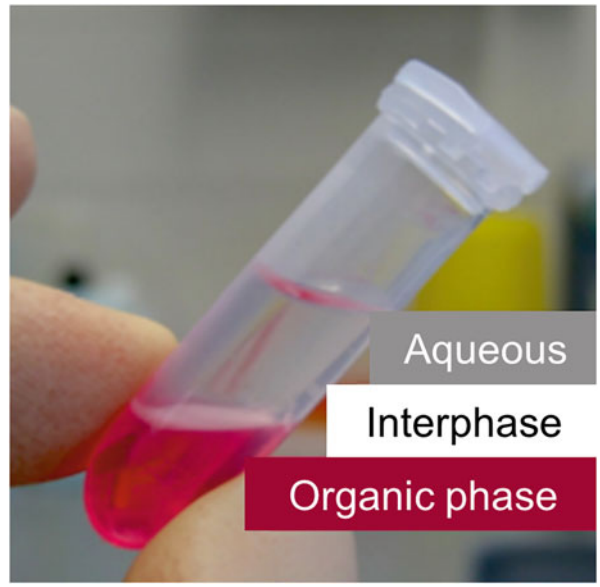


Fig. 1 Picture of the three phases formed in **step 4** of the protein extraction method (*see* Subheading 3.1)

4. Add 200 μL of chloroform, vortex for 15 s, incubate on ice for 5 min, and centrifuge at 4 °C for 15 min at maximum speed (*see* **Note 6**) (Fig. 1).

5.a. Transfer 500–600 μL of the top, aqueous phase into a clean microcentrifuge tube (RNase free) and add the same volume of phenol:chloroform:isoamyl alcohol, vortex for 10 s, incubate on ice for 5 min, and centrifuge at 4 °C for 15 min at maximum speed (to continue with RNA extraction from the sample, *see* Subheading 3.2).

5.b. Add 300 μL of ethanol 100% to the organic phase in the original microcentrifuge tube to continue with protein extraction. Incubate on ice.

6. Centrifuge for 10 min at 2000 g. Place the supernatant in a clean 2 mL microcentrifuge tube (protein low bind).

7. Add 1 mL of isopropanol and incubate at room temperature for 10 min (*see* **Note 7**).

8. Centrifuge at 4 °C for 10 min at 12,000 g. Discard supernatant, which contains phenol, into a container adequate for its controlled elimination.

9. Wash by resuspending the pellet in 2 mL of a solution of 0.3 M guanidine in 95% ethanol (*see* **Note 8**).

10. Sonicate in a sonication bath for 5 min and centrifuge at 4 °C for 5 min at 8000 g.

11. Repeat the washing procedure (**steps 9** and **10**) twice. The obtained pellet can be stored at −20 °C for months.
12. Wash again by the same procedure (**steps 9–11**) with 90% ethanol.
13. Let the pellet dry for a few minutes and resuspend in an appropriate buffer (*see* **Note 9**).
14. Quantify by Bradford with 1 and 2 μL of sample. Add SDS-PAGE 5× buffer to obtain a final 1× concentration when loading the gel.

3.2 RNA Extraction

1. Transfer approximately 500 μL of the top, aqueous phase after the centrifugation in protein extraction **step 5.a** to a clean microcentrifuge tube (RNase free) and add 1 volume (500 μL) of pure isopropanol. Shake and mix.
2. Incubate on ice for 15 min, centrifuge at 4 °C for 10 min at maximum speed, and discard supernatant.
3. Resuspend the pellet in 750 μL of LiCl 3 M, incubate on ice for 10 min, and centrifuge at 4 °C for 10 min at maximum speed.
4. Discard supernatant and wash the pellet with 500 μL of ethanol 85% (v/v), vortexing gently for 10 s.
5. Centrifuge at 4 °C for 10 min at maximum speed and discard supernatant.
6. Let the pellet dry and resuspend in 21 μL of diethylpyrocarbonate (DEPC)-treated water (*see* **Note 10**).
7. Sample quantification with NanoDrop spectrophotometer.

3.3 LC-MS/MS

3.3.1 Sample Preparation

1. Prepare or dissolve protein samples (*see* Subheading 3.1, **step 13**) in 6 M urea 200 nM ammonium bicarbonate.
2. Reduce the samples (10 μg of protein) with 30 nmol DTT at 37 °C for 1 h.
3. Alkylate the samples (10 μg of protein) in the dark with 60 nmol of iodoacetamide at 25 °C for 30 min.
4. Dilute the protein extract to 2 M urea with 200 mM ammonium bicarbonate for digestion with endoproteinase LysC (1:10 w:w), and incubate 37 °C overnight.
5. Dilute twofold with 200 mM ammonium bicarbonate for trypsin digestion (1:10 w:w), and incubate at 37 °C for 8 h.
6. After digestion, add formic acid (10% of the final volume) to acidify the peptide mix.
7. Desalt the samples with MicroSpin C18 columns prior to LC-MS/MS analysis, following manufacturer's instructions.

3.3.2 Chromatographic and Mass Spectrometric Analysis

1. Load the peptides onto the analytical column (C18, 2 μm, 15–50 cm length).
2. Separation of the peptides by reverse-phase chromatography with the corresponding columns.
3. Chromatographic gradients start at 93% buffer A and 7% buffer B with a flow rate of 250 nL/min for 5 min and gradually increase 65% buffer A and 35% buffer B in 60 min.
4. After each analysis, wash the column for 15 min with 10% buffer A and 90% buffer B.
5. Peptide eluates are dried in a vacuum centrifuge, and resuspended with buffer A at a final concentration of 1 μg/μL prior to analysis by LC-MS/MS.
6. Operate the mass spectrometer to acquire peptide spectra (*see* **Note 11**).

3.3.3 Data Analysis

1. Search the acquired spectra against the desired peptide database (*see* **Note 12**), plus a list of common contaminants (suggested: [57]), and all the corresponding decoy entries.
2. Set the parameters accordingly to the experimental and mass spectrometric settings and, if appropriate, select variable post-translational modifications to be detected (*see* **Note 13**).
3. Determine the protein abundance estimation [58, 59].
4. Add the information to the appropriate repositories (*see* **Note 14**).

3.3.4 Treatment of Missing Values and Data Imputation

1. Missing values should first be classified as M(C)AR or MNAR depending on their nature. For instance, for a given protein, if the data from all replicates of the same condition or time point show NAs, probably they are MNAR missing values, whereas if there is only one missing value out of four replicates, it is probably a MAR. Other cases may be more difficult to classify as M(C)AR or MNAR, for instance if there are two NAs out of four replicates. In those instances, other parameters can be considered, for example, the presence or absence of NAs in the adjacent time points (in a time-course experiment) or in the most similar samples in the experiment.
2. Discard all proteins with MNARs or MARs in every sample.
3. Replace MNARs by the minimum of detection of the dataset (deterministic minimum imputation method [60]).
4. Estimate the remaining MARs and MCARs by other imputation method (e.g., k-nearest neighbor (kNN) imputation [61]).

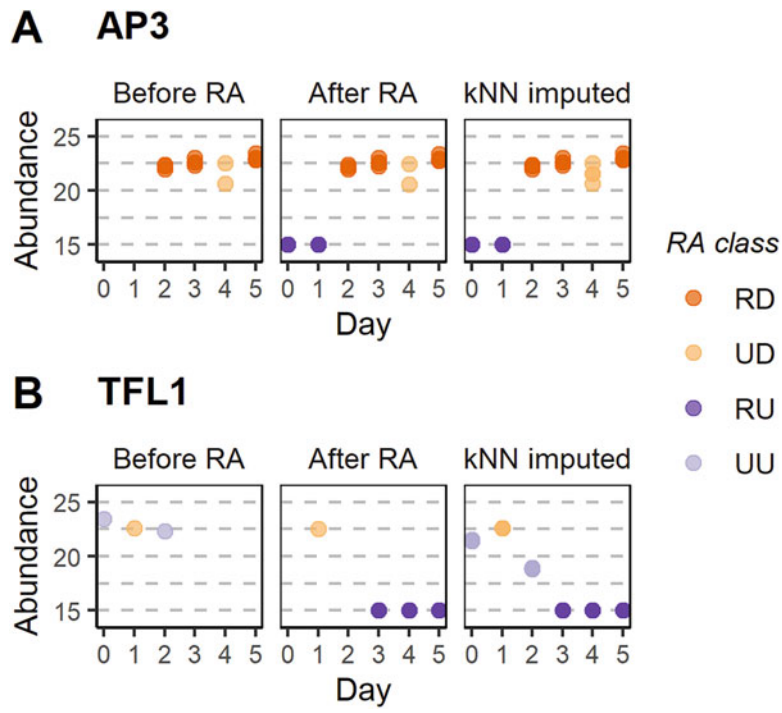


Fig. 2 Stringent analysis to identify reliably undetected and detected fraction of a proteome. The analysis allows to impute values for MAR and MNAR considering their biological meaning. The figure illustrates results from a time-course experiment using the Arabidopsis floral induction system pAP1:AP1-GR *ap1cal* [1], in which samples were collected at 1-day intervals after floral induction (day 0), up to day 5. Log2 TOP3 abundances through time of two flower development regulators, APETALA 3 (AP3) (**a**) and TERMINAL FLOWER 1 (TFL1) (**b**), before and after the "reliability analysis" (RA), and after kNN imputation (from left to right) ($n = 4$ biological replicates)

3.3.5 Example: Treatment of Missing Values in a Time Series Experiment

This missing value classification and data imputation approach can be readily used in, for instance, time-course developmental studies [1, 62], as illustrated in Fig. 2 as an example. In this case, the data processing pipeline consisted on:

1. Classification of each time point (day) for each protein depending on its number of NAs (number of replicates with missing values at a certain time point) and the number of NAs of its immediately adjacent days (neighbors).
   (a) Neighbors are considered as:
      - Unreliable neighbor: Over 50% NAs.
      - Reliable neighbor: Up to 50% NAs (included).
   (b) Initial and final time points are considered as:
      - Reliably undetected: 100% NAs (MNARs).
      - Unreliably undetected: Over 50% NAs (included) (unclear MNARs) + unreliable neighbor.

- Unreliably detected: Over 50% NAs (included) (unclear MARs) + reliable neighbor.
- Reliably detected: Up to 35% NAs (MARs).

(c) Intermediate time points are considered as:

- Reliably undetected: 100% NAs + unreliable neighbors (MNARs).
- Unreliably undetected: Over 50% NAs (included) + unreliable neighbors (probably MNARs).
- Unreliably detected: Over 50% NAs (included) + reliable neighbors (probably MARs).
- Reliably detected: Up to 35% NAs (MARs).

2. Replace reliably undetected time points by the minimum of detection of the dataset (deterministic minimum imputation method [60]).
3. Replace unreliably undetected time points by NAs in all replicates.
4. Discard all proteins which are reliably or unreliably undetected in every time point.
5. Estimate the remaining NAs by k-nearest neighbor (kNN) imputation ($k = 10$) [61].

4 Notes

1. Reagents for LC-MS/MS can be obtained from several suppliers. As an example, we list here the specific products we use: urea (GE Healthcare; Sigma-Aldrich, P/N 17-1319-01), ammonium bicarbonate (BioUltra, ≥99.5% (T); Sigma-Aldrich, P/N 09830), iodoacetamide (BioUltra; Sigma-Aldrich, P/N I1149), DL-dithiothreitol (for electrophoresis, ≥99%; Sigma-Aldrich, P/N D9163), formic acid for analysis EMSURE® (ACS Reag. Merck, P/N 1.00264.0100), sequencing grade modified trypsin (Promega, P/N V5111), and lysyl endopeptidase (Wako Chemicals GmbH, P/N 129-02541).
2. Suitable reverse-phase chromatography columns are, for instance, 25 cm columns with an inner diameter of 75 μm, packed with 1.9 μm C18 particles (Nikkyo Technos Co.); and 50 cm columns with an inner diameter of 75 μm, packed with 2 μm C18 particles (EASY-Column, Thermo Fisher Scientific, ES903).
3. This is just a concrete example of a "modern high-resolution mass spectrometer"; other instruments could be used.
4. For sample collection, to reduce sample contamination with human proteins (i.e., keratins and collagen), make sure to

always use nitrile gloves (instead of latex) and laboratory coats. Pipets, materials, and solutions exclusively used for proteomics. Take precaution to avoid hair contamination. If flower organs or tissues are going to be dissected, cool tweezers and any other sampling instrument with liquid nitrogen.

5. If samples are grown in petri dishes (e.g., Arabidopsis seedlings), discard white clots which correspond to agar.
6. Three phases are formed, the aqueous phase contains RNA (~550 μL, transparent), the interphase, DNA (white), and the organic phase, proteins and lipids (~450 μL, pink) (Fig. 1).
7. It is possible to stop the protocol here and store the samples at −20 °C for a few days.
8. Use a pipette crushing against the bottom of the tube and leave in a colloidal suspension as thin as possible.
9. Resuspend final proteins in acetonitrile, acetic, or formic acid, depending on the analysis protocol. For Western Blot, use E buffer [56]. The buffer volume should be chosen depending on the desired protein concentrations, varying from 20 to 50 μL.
10. Use high pure water, reagents, and products.
11. 1–2 μg of peptides are loaded onto an analytical column (25 cm, C18 2 μm particle size) using an autosampler device (e.g., EASY nLC 1000, Thermo Fisher Scientific) and the peptides are then separated by reverse-phase chromatography using a water-acetonitril chromatographic gradient. Modern high-resolution mass spectrometers are recommended for data acquisition (e.g., Orbitrap or qTOF). The mass spectrometer is operated in data-dependent acquisition (DDA) mode, in which a full MS scan is recorded in each cycle, followed by the fragmentation of the 10–30 most intense precursor ions to obtain the fragment ion spectra.
12. The results may vary significantly depending on the characteristics of the reference database for peptide identification. It is possible to use public repositories of proteins for the different organisms or to design a specific database.
13. Once the database has been constructed, the raw LC-MS/MS data needs to be interpreted using a database search engine (such as SEQUEST [63], Mascot [64], Phenyx [65], X! Tandem [66], OMSSA [67], pFind [68], InsPecT [69], ByOnic [70], Comet [71], MS-GF+ [72], MaxQuant [73], or MSTracer [74]). As example, the Mascot search engine (v2.6) can be used, using the search parameters accordingly to the experimental and mass spectrometry settings. For peptide identification a precursor ion mass tolerance below 10–20 ppm is recommended, whereas the fragment ion mass tolerance can go from 10 to 20 ppm for high-resolution mass analyzers

(Orbitrap, TOF) to 0.5 Da if a linear ion trap is used for the analysis of the tandem mass spectra. Common peptide modifications such as oxidation of methionine and N-terminal protein acetylation are used as variable modifications. False discovery rate (FDR) in peptide identification is set to a maximum of 1%.

14. Share data and results in a public repository. Data sharing in the public domain is the standard for omics research and a requirement for publication. For proteomics, the Proteomics IDEntifications (PRIDE) database (https://www.ebi.ac.uk/pride/) at the European Bioinformatics Institute (EMBL-EBI, Hinxton, Cambridge, UK) has enabled public data deposition of MS data since 2004, and its archival component has become the largest repository for proteomics data sharing worldwide [75]. The PRIDE database provides access to most of the experimental proteomics data described in MS-related scientific publications.

Acknowledgments

Work in the authors' laboratory was supported by grant BFU2014-58289-P (funded by MICIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe") and by grant 2017SGR718 (from the Agencia de Gestió d'Ajuts Universitaris I de Recerca) to JLR, and by institutional grant SEV-2015-0533 (funded by MCIN/AEI/10.13039/501100011033) and by the CERCA Programme/Generalitat de Catalunya. R.A. is supported by fellowship PRE2018-084278 funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future." We are grateful to Eva Borràs and Eduard Sabidó from the CRG/UPF Proteomics Unit for their advice and help in proteomics research.

References

1. Kaufmann K, Wellmer F, Muiño JM, Ferrier T, Wuest SE, Kumar V et al (2010) Orchestration of floral initiation by APETALA1. Science 328(85):85–89
2. Ó'Maoiléidigh DS, Thomson B, Raganelli A, Wuest SE, Ryan PT, Kwašniewska K et al (2015) Gene network analysis of Arabidopsis thaliana flower development through dynamic gene perturbations. Plant J 83(2):344–358
3. Chen D, Yan W, Fu LY, Kaufmann K (2018) Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. Nat Commun 9:4534
4. Pajoro A, Madrigal P, Muiño JM, Matus JT, Jin J, Mecchia MA et al (2014) Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol 15:R41
5. Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, Hanczaryk K et al (2012) Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. Proc Natl Acad Sci U S A 109(33):13452–13457

6. Wellmer F, Riechmann JL (2010) Gene networks controlling the initiation of flower development. Trends Genet 26(12):519–527
7. Wils CR, Kaufmann K (2017) Gene-regulatory networks controlling inflorescence and flower development in Arabidopsis thaliana. Biochim Biophys Acta Gene Regul Mech 1860(1): 95–105
8. Pajoro A, Biewers S, Dougali E, Valentim FL, Mendes MA, Porri A et al (2014) The (r)-evolution of gene regulatory networks controlling Arabidopsis plant reproduction: a two-decade history. J Exp Bot 65(17): 4731–4745
9. Heisler MG, Jönsson H, Wenkel S, Kaufmann K (2022) Context-specific functions of transcription factors controlling plant development: from leaves to flowers. Curr Opin Plant Biol 69:102262
10. Takáč T, Šamajová O, Šamaj J (2017) Integrating cell biology and proteomic approaches in plants. J Proteome 169:165–175
11. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH (2017) Proteomics: technologies and their applications. J Chromatogr Sci 55(2): 182–196
12. Grossmann J, Roschitzki B, Panse C, Fortes C, Barkow-Oesterreicher S, Rutishauser D et al (2010) Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. J Proteome 73(9):1740–1746
13. Mergner J, Kuster B (2022) Plant proteome dynamics. Annu Rev Plant Biol 73:67–92
14. Zhang Z, Hu M, Feng X, Gong A, Cheng L, Yuan H (2017) Proteomes and phosphoproteomes of anther and pollen: availability and progress. Proteomics 17(20). https://doi.org/10.1002/pmic.201600458
15. Navrot N, Finnie C, Svensson B, Hägglund P (2011) Plant redox proteomics. J Proteome 74(8):1450–1462
16. Mergner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A et al (2020) Mass-spectrometry-based draft of the Arabidopsis proteome. Nature 579(7799):409–414
17. Kumar M, Carr P, Turner SR (2022) An atlas of Arabidopsis protein S-acylation reveals its widespread role in plant cell organization and function. Nat Plants 8(6):670–681
18. Abraham P, Gannone RJ, Adams RM, Kalluri U, Tuskan GA, Hettich RL (2013) Putting the pieces together: high-performance LC-MS/MS provides network-, pathway-, and protein-level perspectives in Populus. Mol Cell Proteomics 12(1):106–119
19. Szymanski J, Levin Y, Savidor A, Breitel D, Chappell-Maor L, Heinig U et al (2017) Label-free deep shotgun proteomics reveals protein dynamics during tomato fruit tissues development. Plant J 90(2):396–417
20. Duncan O, Trösch J, Fenske R, Taylor NL, Millar AH (2017) Resource: mapping the Triticum aestivum proteome. Plant J 89(3): 601–616
21. Marx H, Minogue CE, Jayaraman D, Richards AL, Kwiecien NW, Siahpirani AF et al (2016) A proteomic atlas of the legume, M. truncatula, and its nitrogen fixing endosymbiont, S. meliloti. Nat Biotechnol 34(11):1198
22. Dai S, Chen S (2012) Single-cell-type proteomics: toward a holistic understanding of plant function. Mol Cell Proteomics 11(12): 1622–1630
23. Emanuelsson O, Von Heijne G, Schneider G (2001) Analysis and prediction of mitochondrial targeting peptides. Methods Cell Biol 65:175–187
24. Bruce BD (2000) Chloroplast transit peptides: structure, function and evolution. Trends Cell Biol 10(10):440–447
25. Bernhofer M, Goldberg T, Wolf S, Ahmed M, Zaugg J, Boden M et al (2018) NLSdb—major update for database of nuclear localization signals and nuclear export signals. Nucleic Acids Res 46(D1):D503–D508
26. Bassal M, Abukhalaf M, Majovsky P, Thieme D, Herr T, Ayash M et al (2020) Reshaping of the Arabidopsis thaliana proteome landscape and co-regulation of proteins in development and immunity. Mol Plant 13(12):1709–1732
27. Feng Z, Kong D, Kong Y, Zhang B, Yang X (2022) Coordination of root growth with root morphology, physiology and defense functions in response to root pruning in Platycladus orientalis. J Adv Res 36:187–199
28. Jain A, Singh HB, Das S (2021) Deciphering plant-microbe crosstalk through proteomics studies. Microbiol Res 242:126590
29. Niu Z, Liu L, Pu Y, Ma L, Wu J, Hu F et al (2021) iTRAQ-based quantitative proteome analysis insights into cold stress of winter rapeseed (Brassica rapa L.) grown in the field. Sci Rep 11:23434
30. Koehler G, Rohloff J, Wilson RC, Kopka J, Erban A, Winge P et al (2015) Integrative "omic" analysis reveals distinctive cold responses in leaves and roots of strawberry, fragaria × ananassa 'Korona'. Front Plant Sci 6:826

31. Le Signor C, Aimé D, Bordat A, Belghazi M, Labas V, Gouzy J et al (2017) Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. New Phytol 214(4):1597–1613
32. Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C et al (2018) Rewiring of the fruit metabolome in tomato breeding. Cell 172(1–2):249–261.e12
33. Lehmann BD, Colaprico A, Silva TC, Chen J, An H, Ban Y et al (2021) Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes. Nat Commun 12:6276
34. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA et al (2018) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform 19(2):286–302
35. Kumar D, Bansal G, Narang A, Basak T, Abbas T, Dash D (2016) Integrating transcriptome and proteome profiling: strategies and applications. Proteomics 16(19):2533–2544
36. Wu Y, Tang Y, Jiang Y, Zhao D, Shang J, Tao J (2018) Combination of transcriptome sequencing and iTRAQ proteome reveals the molecular mechanisms determining petal shape in herbaceous peony (Paeonia lactiflora Pall.). Biosci Rep 38(6):BSR20181485
37. Decourcelle M, Perez-Fons L, Baulande S, Steiger S, Couvelard L, Hem S et al (2015) Combined transcript, proteome, and metabolite analysis of transgenic maize seeds engineered for enhanced carotenoid synthesis reveals pleotropic effects in core metabolism. J Exp Bot 66(11):3141–3150
38. Seaton DD, Graf A, Baerenfaller K, Stitt M, Millar AJ, Gruissem W (2018) Photoperiodic control of the Arabidopsis proteome reveals a translational coincidence mechanism. Mol Syst Biol 14(3):e7962
39. Omidbakhshfard MA, Sokolowska EM, Di Vittori V, Perez de Souza L, Kuhalskaya A, Brotman Y et al (2021) Multi-omics analysis of early leaf development in Arabidopsis thaliana. Patterns 2(4):100235
40. Huang Y, Zhou L, Hou C, Guo D (2022) The dynamic proteome in Arabidopsis thaliana early embryogenesis. Development 149(18): dev200715
41. Keller M, Simm S, Bokszczanin KL, Bostan H, Bovy A, Chaturvedi P et al (2018) The coupling of transcriptome and proteome adaptation during development and heat stress response of tomato pollen. BMC Genomics 19(1):447
42. Ji J, Yang L, Fang Z, Zhuang M, Zhang Y, Lv H et al (2018) Complementary transcriptome and proteome profiling in cabbage buds of a recessive male sterile mutant provides new insights into male reproductive development. J Proteome 179:80–91
43. Xing M, Sun C, Li H, Hu S, Lei L, Kang J (2018) Integrated analysis of transcriptome and proteome changes related to the ogura cytoplasmic male sterility in cabbage. PLoS One 13(3):e0193462
44. Jing D, Chen W, Hu R, Zhang Y, Xia Y, Wang S et al (2020) An integrative analysis of transcriptome, proteome and hormones reveals key differentially expressed genes and metabolic pathways involved in flower development in loquat. Int J Mol Sci 21(14):5107
45. Chen R, Chen G, Huang J (2017) Shot-gun proteome and transcriptome mapping of the jujube floral organ and identification of a pollen-specific S-locus F-box gene. PeerJ 5: e3588
46. Lu D, Ni W, Stanley BA, Ma H (2016) Proteomics and transcriptomics analyses of Arabidopsis floral buds uncover important functions of ARABIDOPSIS SKP1-LIKE1. BMC Plant Biol 16:61
47. Vogel C, Marcotte EM (2012) Insights into regulation of protein abundance from proteomics and transcriptomis analyses. Nat Rev Genet 13(4):227–232
48. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4(10):1707–1719
49. Taniguchi Y, Choi PJ, Li G, Chen H, Babu M, Hearn J et al (2011) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science 329(5991): 533–539
50. Csárdi G, Franks A, Choi DS, Airoldi EM, Drummond DA (2015) Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. PLoS Genet 11(5):1005206
51. Oliva-Vilarnau N, Vorrink SU, Ingelman-Sundberg M, Lauschke VM (2020) A 3D cell culture model identifies Wnt/β-catenin mediated inhibition of p53 as a critical step during human hepatocyte regeneration. Adv Sci 7(15):2000248
52. Simões T, Novais SC, Natal-da-Luz T, Devreese B, de Boer T, Roelofs D et al (2019) Using time-lapse omics correlations to integrate toxicological pathways of a formulated fungicide in a soil invertebrate. Environ Pollut 246:845–854

53. Tarazona S, Balzano-Nogueira L, Conesa A (2018) Multiomics data integration in time series experiments. Compr Anal Chem 82: 505–532
54. Lazar C, Gatto L, Ferro M, Bruley C, Burger T (2016) Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. J Proteome Res 15:1116–1125
55. Jin L, Bi Y, Hu C, Qu J, Shen S, Wang X et al (2021) A comparative study of evaluating missing value imputation methods in label-free proteomics. Sci Rep 11:1720
56. Martínez-García JF, Monte E, Quail PH (1999) A simple, rapid and quantitative method for preparing Arabidopsis protein extracts for immunoblot analysis. Plant J 20: 251–257
57. Beer LA, Liu P, Ky B, Barnhart KT, Speicher DW (2017) Efficient quantitative comparisons of plasma proteomes using label-free analysis with MaxQuant. Methods Mol Biol 1619: 339–352
58. Gerster S, Kwon T, Ludwig C, Matondo M, Vogel C, Marcotte EM et al (2014) Statistical approach to protein quantification. Mol Cell Proteomics 13(2):666–677
59. Fabre B, Lambour T, Bouyssié D, Menneteau T, Monsarrat B, Burlet-Schiltz O et al (2014) Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. EuPA Open Proteom 4:82–86
60. Meleth S, Deshane J, Kim H (2005) The case for well-conducted experiments to validate statistical protocols for 2D gels: different pre-processing = different lists of significant proteins. BMC Biotechnol 5:7
61. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R et al (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17(6):520–525
62. Wellmer F, Alves-Ferreira M, Dubois A, Riechmann L, Meyerowitz EM (2006) Genome-wide analysis of gene expression during early Arabidopsis flower development. PLoS Genet 2(7):e117
63. Eng JK, Mccormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5: 977–989
64. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probablity-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551–3557
65. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 3(8):1454–1463
66. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20(9):1466–1467
67. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM et al (2004) Open mass spectrometry search algorithm. J Proteome Res 3:958–964
68. Fu Y, Yang Q, Sun R, Li D, Zeng R, Ling CX et al (2004) Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. Bioinformatics 20(12):1948–1954
69. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M et al (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 77(14): 4626–4639
70. Bern M, Cai Y, Goldberg D (2007) Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. Anal Chem 79(4): 1393–1400
71. Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. Proteomics 13(1):22–24
72. Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5(1):5277
73. Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc 11(12):2301–2319
74. Zeng X, Ma B (2021) MSTracer: a machine learning software tool for peptide feature detection from liquid chromatography-mass spectrometry data. J Proteome Res 20(7): 3455–3462
75. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S et al (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res 50(D1):D543–D552