

Origin, evolution and role of ORFs in the birth of new genes and regulatory elements

Origine, évolution et rôle des ORFs dans les mécanismes d'apparition de nouveaux gènes et éléments régulateurs

Thèse préparée et soutenue publiquement par Claire Patiou le 27/03/2025, pour obtenir le
grade de Docteur en biologie des populations et écologie,

devant le jury composé de :

Rapporteurs:

Karine Alix (Professeure, AgroParisTech)
Li Zhao (Associate Professor, The Rockefeller University)

Examineurs:

Juliette De Meaux (Professeure, University of Cologne)
Jacinthe Azevedo-Favory (Chargée de recherche CNRS, LGDP - UMR5096 CNRS/UPVD)
Nikolaos Vakirlis (Junior Investigator, IFBR, BSRC “Alexander Fleming”)

Directeur de thèse:

Vincent Castric (Directeur de recherche CNRS, EEP - UMR 8198 CNRS/U. Lille)

Encadrants:

Eléonore Durand (Maîtresse de conférences, EEP - UMR 8198 CNRS/U. Lille)
Sylvain Legrand (Maître de conférences, EEP - UMR 8198 CNRS/U. Lille)

Président du Jury:

Xavier Vekemans (Professeur des universités, EEP - UMR 8198 CNRS/U. Lille)



Table of contents

General introduction.....	5
I. Short Open Reading Frames as a source of <i>de novo</i> genes.....	7
1. A short history of <i>de novo</i> gene emergence.....	7
2. Defining <i>de novo</i> genes.....	9
3. Factors necessary for gene emergence.....	10
4. Models of <i>de novo</i> gene emergence.....	11
5. First steps of <i>de novo</i> gene birth , within a life cycle.....	14
II. Open Reading Frames in microRNA transcripts: the case of miORFs.....	15
1. microRNAs : a key type of regulatory element.....	15
2. The coding potential of miRNA primary transcripts.....	17
3. How do miRpeptides evolve?.....	20
Combining long-read transcripts sequencing with Ribosome Profiling : scientific strategy and methods optimization.....	22
Scientific questions and thesis plan.....	28
1. Chapter 1: Characterization of a repertoire of <i>de novo</i> ORFs: origins, conservation, functional potential.....	28
2. Chapter 2: Repertoire and conservation of miORFs - origins and evolution of the miRNA-miPEP system.....	28
References.....	30
CHAPTER I.....	37
Abstract.....	39
Introduction.....	40
Results.....	43
Ribosome Profiling sequencing yields evidence of ongoing translation.....	43
Identifying individual ORFs with translation signals.....	46
Synteny analysis identifies many <i>de novo</i> ORFs being translated.....	48
<i>De novo</i> ORFs of varying conservation levels display a continuum of features.....	49
<i>De novo</i> ORF - encoded peptides are highly disordered.....	52
A select few <i>de novo</i> ORFs have the potential to be proto-genes.....	53
Discussion.....	56
Pervasive translation with a rapid evolutionary turnover.....	56
The first steps of <i>de novo</i> gene emergence.....	57
From random intergenic ORF to canonical gene.....	59
Material and methods.....	61
Plant Materials and Growth Conditions.....	61
Lysates preparation.....	61
Ribosome footprints.....	62
Ribosomal RNAs depletion.....	63

Library preparation and sequencing.....	63
Ribo-seq Data Analysis.....	64
Syntenic region annotation and alignment.....	65
Intrinsic disorder prediction.....	66
Acknowledgements.....	66
Bibliography.....	67
Supplementary material.....	75
CHAPTER II.....	83
Abstract.....	85
Introduction.....	86
Results.....	89
Long-read Nanopore sequencing yields miRNA primary transcripts.....	89
ORFs are abundant in primiRNAs, which are scanned by ribosomes.....	91
Conserved miRNAs have conserved coding potential, through both conserved and divergent miORFs.....	94
The primary transcripts of young and old miRNAs showcase coding potential across species.....	95
The miRNA-miPEP system can emerge from inverted gene duplications.....	99
Discussion.....	100
miRNA loci produce a wide variety of transcripts - some of which can be coding.....	100
Regulating regulators: the high prevalence and functional importance of miORFs.....	101
Material and Methods.....	104
Plant Materials and Growth Conditions.....	104
Ribosome Profiling data preparation and analysis.....	104
Long-read Nanopore sequencing.....	105
miRNA precursor annotation.....	106
miRNA primary transcript annotation.....	106
ORF conservation analysis.....	107
Bibliography.....	108
Supplementary material.....	113
Perspectives.....	115
Exploitation of the full long-read RNA sequencing data.....	115
More precise and exhaustive characterization of alternative transcription.....	115
Extension of conservation analyzes: more analyzes and more species.....	116
Genome-wide investigation of the miRNA-miPEP system's origins.....	116
Conclusions and perspectives.....	117
Conclusion.....	117
1. <i>de novo</i> gene emergence from translated short ORFs.....	117
2. The importance of alternative miRNA primary transcripts and their translation in the regulation of mature miRNA levels.....	118
Perspectives.....	123

1. <i>de novo</i> ORF-encoded peptides: population dynamics and functional potential.....	123
2. The origins, conservation and functional potential of the miRNA-miPEP system.....	124
References.....	126
Acknowledgements.....	129
To my thesis jury and CSI committee members.....	129
To my financers.....	129
À mes encadrants.....	129
To my collaborators.....	130
To my fellow lab members.....	130
À mes proches.....	131
Abstract (english).....	132
Résumé (français).....	133

General introduction

Understanding the genetic basis for the origin and evolution of novel phenotypes, and ultimately biological diversity, is one of the most fundamental questions in evolutionary biology. In the quantitative genetics field, phenotypes have classically been described as the combination of both genetic and environmental elements, possibly interacting with one another. Hence, phenotypic novelty may be attributed to changes in either the genetics, the environment, or the interaction between the genetic and environmental components of a given (group of) individual(s). Evolutionary genomics is specifically interested in the genetic component of phenotypes i.e. genotypes, based on the assumption that changes in genotypes can be reflected in changes in phenotypes. In this context, phenotypic novelty might be derived from a wide range of genetic modifications, which fall into two main categories: changes in the sequence of the genes themselves, or changes in their expression. Both categories are impacted by evolutionary forces such as selection, drift, migration and mutation. Pre-existing genes may acquire beneficial, deleterious or neutral mutations, which could result in their loss by pseudogenization, a gain, loss or change of function, and an emergence or a modification of new and pre-existing alleles, among others. New genes can also emerge with either redundant or novel functions, by a plethora of mutational mechanisms that include duplication followed by divergence, horizontal gene transfer or exon shuffling (McLysaght and Hurst 2016). On the other hand, changes in the expression of pre-existing genes can greatly impact the resulting phenotypes. The actors of gene expression regulation are known as “regulatory elements”, which were first described in 1961 through the discovery of “transcription factors” (Jacob and Monod 1961), that bind to genes’ promoters and either inhibit or enhance gene expression (Yu and Gerstein 2006). Initially thought of as having a relatively minor role in phenotypic determination when compared to non-regulatory protein-coding genes, their importance was notably highlighted in 1975 by a seminal paper by Mary-Claire King and Allan Charles Wilson (King and Wilson 1975). Comparing human and chimpanzees protein-coding genes, these researchers concluded that “their macromolecules are so alike that regulatory mutations may account for their biological differences”. Hence, to fully explain the major phenotypic differences between the two species, it was necessary to take into account regulatory elements. Throughout the past 60 years, many more kinds of regulatory “non-coding” RNAs were discovered and identified as key actors in biology, across all Life kingdoms. They play crucial roles in allowing for a time and space - specific expression of protein-coding genes, being

involved in essential biological processes like development or environmentally-induced responses.

In this context, we are studying the dynamics of emergence of new genes and regulatory elements in the *Arabidopsis* genus. To do so, we choose to focus on a type of genetic element known as Open Reading Frames (ORFs). ORFs are defined as sequences which begin with a “START” codon and end with a “STOP” codon, both located within the same triplet-based reading frame. A sub-category of ORFs, the CDSs (Coding DNA Sequence), are present within genes, and translated into polypeptides with a known or presumed function (Andrews and Rothnagel 2014). CDSs are generally characterized by a set of properties: length greater than 300 base pairs (Hanada et al. 2013), higher sequence conservation compared to intergenic regions, presence of upstream promoters, termination and splice sites, canonical initiation codons (ATGs), among others (Andrews and Rothnagel 2014). Yet, ORFs that do not meet these criteria are frequently found in non-genic regions. Specifically, the coding potential (i.e. the ability to encode a polypeptide, which may or may not be functional) of "short" ORFs (Hanada et al. 2013), referred to as "sORFs", has been particularly neglected due to the use of size thresholds when characterizing functional regions of the genome (Yeasmin et al. 2018), below which the sequences were considered by default as "non-coding", and therefore not functional. Yet, the role of some sORFs as regulatory elements has already been demonstrated in many species and genomic contexts (Andrews and Rothnagel 2014; Orr et al. 2021), including in *Medicago truncatula* and *Arabidopsis thaliana*, influencing traits such as flowering or root development thanks to their peptide-encoding abilities (Hanada et al. 2007; Hanada et al. 2013; Lauressergues et al. 2015; Fesenko et al. 2019; Fesenko et al. 2021; Lauressergues et al. 2022). Furthermore, sORFs can also play a role in the emergence of new genes, through a plethora of evolutionary scenarios, for example: they may emerge *de novo* from an intergenic region, arise from an ancient gene undergoing pseudogenization or by duplication and subsequent divergence (McLysaght and Hurst 2016; Xia et al. 2025).

In this thesis, I wish to characterize the origins, evolution and functional potential of sORFs in the emergence of both regulatory elements, using miRNA-encoded peptides and new genes as examples, with a focus on those that arise *de novo*.

I. Short Open Reading Frames as a source of *de novo* genes

1. A short history of *de novo* gene emergence

Genes are considered as highly complex elements, that require an array of characteristics to achieve a “function”, i.e. to induce a phenotype (according to certain philosophical theories of function, others would define biological function as the result of selection, and this is still a heavily debated topic in philosophy of Science (Doolittle et al. 2014; Doolittle 2018)). Such characteristics are numerous, and include the presence of promoters inducing transcription, the presence of sites of polyadenylation, or the ability to encode functional protein domains. The complexity level of the molecular organization of genes very naturally induced researchers to believe that their apparition by purely random mutations would be extremely unlikely (Xia et al. 2025). Indeed, François Jacob is famously quoted as saying that “the probability of a functional protein appearing *de novo* by random association of amino acids is almost zero” (Jacob 1977), and this quote embodies the paradigm that historically dominated evolutionary genetics until the early 21st century. In this paradigm, evolution is a “tinkerer”, simply reusing pre-existing coding material to generate new genes. These tinkering processes include, among others, exon shuffling, gene duplication, mobile genetic elements, horizontal gene transfer, gene fusion/fission and retroposition (Long et al. 2003; Zhou et al. 2008). Among them, gene duplication was long considered the major source of new genes, a view exemplified by Ohno's 1970 book “Evolution by Gene Duplication” (Ohno 1970). Canonical, functional genes that were well integrated into regulation and cellular networks would be likely to become duplicated across their lifetime. Being functionally redundant, the duplicates would experience relaxed selective pressure, allowing them to diverge and progressively acquire new functions, while still retaining the necessary sequence characteristics of genes, ultimately giving rise to new, functional genes (Luchetta et al. 2005). Otherwise, duplicates might diverge to the point of function loss, and progressively become pseudogenized.

Yet, this paradigm was challenged by the progressive discovery of genes that lacked homology to other known coding sequences in a diversity of organisms, and would come to be characterized as *de novo*. In contrast to gene duplication, *de novo* gene emergence implies a mutational origin from a historically non-genic sequence. As François Jacob described it, although *de novo* gene emergence was not considered to be impossible, it was seen as a very minor contributor to genetic novelty (McLysaght and Hurst 2016; Van Oss and Carvunis 2019).

The first examples of *de novo* genes were discovered in 2006, when two papers were published identifying *de novo* genes in *Drosophila*, expressed in male reproductive tissues. The first characterized the ancestral non-coding origins of the *Acp* gene, expressed in accessory glands and found only in *D. erecta* and *D. yakuba* (Begun et al. 2006). The second demonstrated the existence of five *de novo* genes in *D. melanogaster*, four of which are X-linked, and which are expressed mainly in the testes (Levine et al. 2006). Historically, the first *de novo* genes were identified experimentally in *Drosophila*, leading to the “testes first” hypothesis, according to which the rapid evolution of male reproductive traits in certain organisms (*drosophila*, mice, etc.) is primarily due to the frequent appearance of *de novo* genes (Kaessmann 2010). Following the publication of those seminal papers, many *de novo* genes would be uncovered in a wide variety of clades. In the yeast *Saccharomyces cerevisiae*, the *BSC4* gene was found to lack any identifiable homolog in the *Saccharomyces* genus (Cai et al. 2008). In plants, the QQS (Quaqueine Starch) gene, involved in the regulation of starch synthesis, and lacking homologs, was discovered in *A. thaliana* (Li et al. 2009) and in *Oryza sativa*, the *OsDR10* gene, involved in the defense response to pathogenic infections, had identifiable homologs only in the closest genus, *Leersia* (Xiao et al. 2018). In 2009, four *de novo* genes were characterized for the first time in *Homo sapiens*, which were derived from ancestral non-coding sequences found in primates (Knowles and McLysaght 2009). Further studies were able to uncover *de novo* genes on a wider, genomic scale, initially in organisms like *S. cerevisiae* (Carvunis et al. 2012) or *D. melanogaster* (Zhao et al. 2014)

Furthermore, it was demonstrated in mice that, taking into account the entire evolutionary time of a given genome, every position could have been subject to transcription (during a short evolutionary period) at least once: this is referred to as “ubiquitous” transcription (Neme and Tautz 2016). As such, any non-coding portion of the genome could thus have been subject to *de novo* gene evolution, and this view was quickly extended to other organisms (Neme and Tautz 2016). This ubiquity of transcriptional events was incorporated into the formulation of *de novo* gene emergence models, as it would make previously unexpressed regions of the genome “visible” to selection, and explain how non-genic regions could be expressed without promoters, transcription factors or regulatory elements. While it is possible for initially non-genic regions of the genome to be transcribed, it has yet to be proven that they can be translated in the context of protein-coding genes *de novo* origination.

2. Defining *de novo* genes

“New” genes designate genes that have appeared in a limited number of taxa, and are therefore taxonomically restricted. However, the restriction of orthologs to a few taxa can be explained by several processes: a strong divergence of the sequence (loss of homology), a total disappearance of the sequence in certain taxa, or a *de novo* appearance of the sequence (Vakirlis et al. 2020), as represented on **Figure 1**. *De novo* genes have the property of having arisen from a historically non-genic sequence. They include, but are not limited to, certain taxonomically restricted genes: some highly conserved genes may have originally arisen *de novo*. Despite this, Taxonomically Restricted Genes, also known as “orphans”, are often used as proxies to characterize evolutionarily “young” genes. It has been suggested that almost a third of all genes, all genomes included, were “orphaned” and had thus potentially emerged *de novo* (Tautz and Domazet-Lošo 2011). Compared with older, conserved and established genes, these younger genes possess a number of traits that set them apart. For example, they are shorter (both in total length and in exon length); expressed less prominently, more often specifically to a tissue and developmental stage; possess fewer exons; evolve more rapidly and are more similar to non-genic sequences (Neme and Tautz 2013; Van Oss and Carvunis 2019).

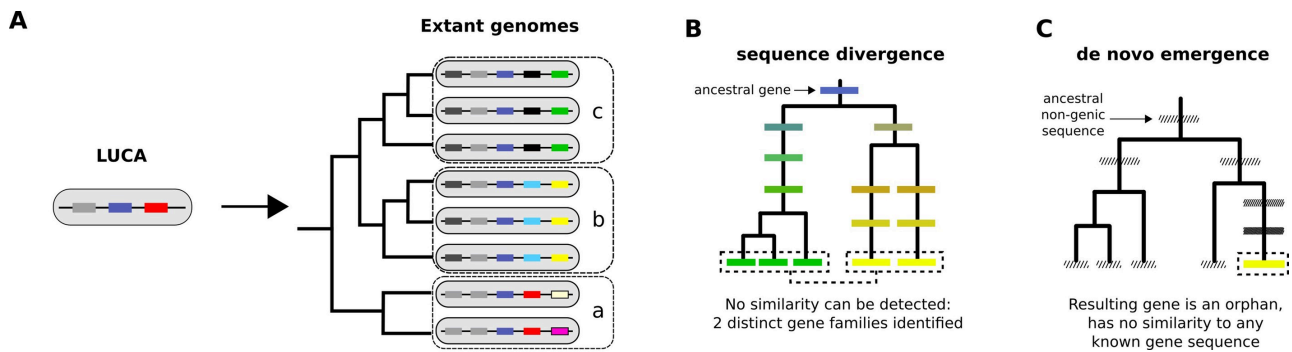


Figure 1: Three mechanisms by which genes can be annotated as TRGs (Taxonomically Restricted Genes): (A) disappearance of an ancestral sequence in certain taxa, (B) divergence of an ancestral sequence after duplication, until loss of homology, or (C) appearance of *de novo* gene (Vakirlis et al. 2020).

When it comes to annotating novel genes, we seek to retain only those gene sequences for which we can find no homolog beyond a taxon of interest. It is therefore necessary to have access to the genomes of several species, more or less distant from each other, in an exhaustive and complete manner. The age of a potential new gene will then be determined by

the phylogeny of the species in which we detect the presence or absence of homologs. A suitable tool for such a search is BLAST (Best Local Alignment Search Tool, (Altschul et al. 1990)), which takes as input the sequence of a gene of interest (or its polypeptide product) in a focal species, and searches all available databases for homologs in distant species. This method is known as phylostratigraphy (Neme and Tautz 2013). Criticisms have been leveled against phylostratigraphy, notably by Moyers and Zhang (Moyers and Zhang 2015; Moyers and Zhang 2016; Moyers and Zhang 2017; Moyers and Zhang 2018), who refer to it as “phylostratigraphic bias”. Indeed, a high number of false negatives appear to be produced by phylostratigraphy (up to 30% of genes identified as orphans would not be so), due to BLAST's inability to identify homologous sequences that have diverged significantly. Worse still, the score assigned to matches by BLAST is directly influenced by many factors, including the length and divergence rate of the gene of interest. The characteristics of young genes mentioned above could in fact be explained, at least in part, by this methodological artifact (Weisman et al. 2020).

To overcome this problem synteny information can be used, which is based on the detection of non-genic regions that are orthologous to potential new genes by sequence alignment between closely related species. Synteny uses genes flanking a region of interest as “anchors” for an alignment. It is based on the prerequisite that the sequences used as anchors for the alignment are conserved, and that the order of these sequences is also conserved. As chromosomal rearrangements and genome duplications are common in Angiosperms (Tang et al. 2008), this method is conservative as well as more difficult to implement for distantly related species.

3. Factors necessary for gene emergence

Although the appearance of *de novo* genes occurs randomly through mutations, certain factors nevertheless favor the emergence of the characteristics required to constitute a gene. The transition from a non-coding to a coding sequence thus requires the passage through a number of steps, namely (Wilson and Masel 2011):

- 1) Transcription of the DNA sequence into RNA (Vakirlis et al. 2018)
- 2) Stable maintenance of this RNA (non-degradation by the exosome)
- 3) The association of this stable RNA with the ribosomal machinery
- 4) Translation of this RNA by ribosomes
- 5) Avoidance of peptide degradation by the proteasome

6) The stable, non-toxic conformation of the peptide product

The key starting point for this transition is transcription of a non-coding region. To be transcribed, a gene requires regulatory regions, which must contain binding sites, transcriptional enhancers, transcription initiation sites and an open chromatin state (Tautz and Domazet-Lošo 2011). Several factors have been suggested as favoring this transcription, first and foremost: proximity to ancestral gene regions. Indeed, the presence of nearby genes facilitates transcription of non-genic flanking regions, through transcriptional “overflow”, bidirectional promoter activity, or even spurious transcription (McLysaght and Hurst 2016). Transcription enhancers, which “recruit” transcription factors and promote an open chromatin state, are effective over long distances, and a new gene located in the intron of a pre-existing gene could share the activity of its enhancer. Transcription enhancers themselves could also be the subject of mutations, thereby evolving the gene regulatory network they underpin, facilitating the birth of *de novo* genes and their integration into regulatory networks (Majic and Payne 2020).

Even if transcribed, a genomic region must contain at least one ORF to be considered as potentially coding, with the probability of its emergence depending on mutation biases and GC content. These mutation biases are numerous, and include the nature of the mutation itself, with a bias favoring the transition from A/T to C/G (Ossowski et al. 2010; Zhu et al. 2014); the speed of DNA replication, with slower replication associated with a higher mutation rate (Lang and Murray 2011); or the general (epi)genetic context (Weng et al. 2019). GC content also strongly influences the presence of *de novo* genes; on the one hand by favoring the presence of long ORFs (lower probability of a STOP codon appearing), and on the other by containing more transcription factor binding sites (in some species, including mammals) (McLysaght and Hurst 2016). Among other things, recombination “hot spots” also appear to be conducive to the appearance of *de novo* genes, thanks to the GC-biased gene conversion repair of meiotic recombination (Vakirlis et al. 2018). Such regions tend to be nucleosome-free, which promotes transcriptional activity, in a GC-rich region that is already prone to *de novo* gene emergence.

4. Models of *de novo* gene emergence

While there is a general consensus on the conditions necessary for the emergence of *de novo* genes, the processes by which these conditions are brought about and lead to this

emergence are debated. Two models are often opposed: that of the continuous spectrum (Carvunis et al. 2012), and that of pre-adaptation (Wilson et al. 2017). According to this first model, represented on **Figure 2**, intergenic ORFs would appear randomly by mutation. Some of them would be expressed (transcribed and translated) at very low levels, temporarily, in particular by the mechanism of ubiquitous transcription. These expressed regions would be “protogenes”, intermediaries between the non-gene and the gene. These protogenes would initially be short, simple and poorly expressed, and thus have no large phenotypic effect. They would serve as a reservoir for selection to act: the majority would rapidly lose all translational activity, but a few could be retained because of the selective advantage they confer to the organism. They would progressively acquire characteristics specific to conserved genes: with a gradual increase in size (for instance due to the loss of STOP codon, fusion with other ORFs or indels conferring a change in the ORF reading frame) and expression (recruitment of existing regulatory elements, emergence of new regulatory elements, etc.). There would thus be a continuum of properties between non-genic and genic sequences.

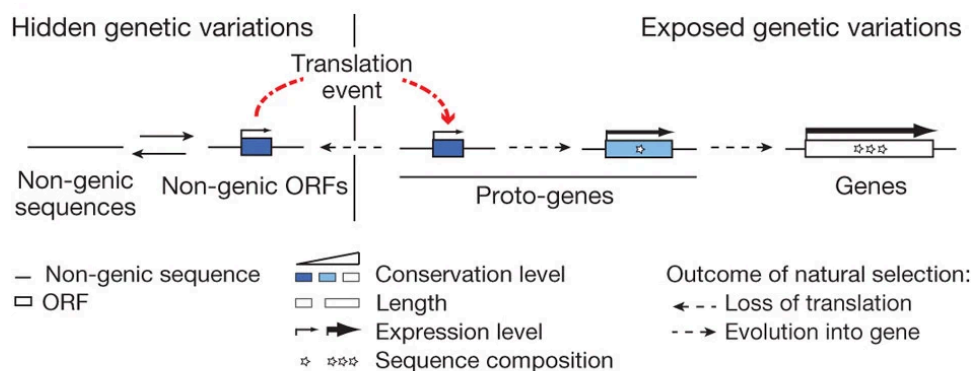


Figure 2: Continuum model suggested by Carvunis et al., 2012. *De novo* genes would appear by “protogenization” of non-genic sequences, with a progressive increase in sequence length, conservation and expression (Carvunis et al. 2012).

On the contrary, according to the pre-adaptation model (see **Figure 3**), young genes carry traits of ancestral genes that are “exaggerated”, so that they only appear in regions that have been particularly suitable and pre-adapted (Wilson et al. 2017). Among such traits, “intrinsic structural disorder” (ISD) has been used to defend this model, qualifying the degree of stability of a potential polypeptide produced by a given sequence on the basis of its amino acid composition. In mice, young genes have a very high ISD, established genes an intermediate ISD, and non-genic sequences a low ISD (Wilson et al. 2017). However, studies in yeast seem to indicate

the opposite trend: younger genes have a lower ISD than older ones (Carvunis et al. 2012; Wilson et al. 2017). It would seem, however, that the annotation procedure determining the “young” character of a gene is inconsistent between species, and excluding some of these annotations completely reverses the model's conclusions (Wilson et al. 2017). Indeed, the very definition of a gene plays a key role in the interpretation of the data, and it is crucial to note that (Wilson et al. 2017) have totally excluded species-specific genes from their analyses, even though these would include the most recently emerged *de novo* genes, which are the ones investigated in this thesis.

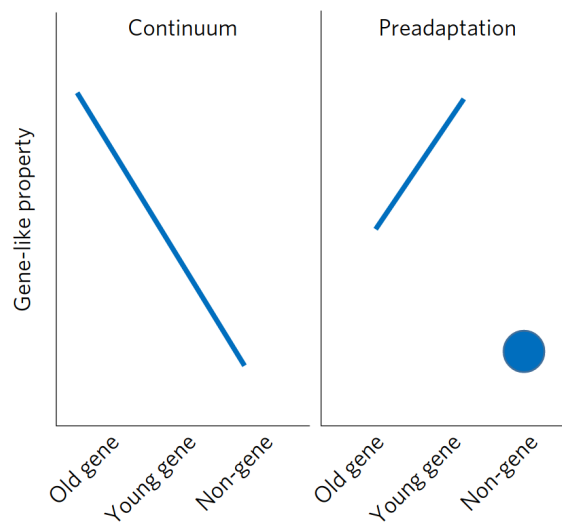


Figure 3: Comparison between the continuum model, and the preadaptation model for *de novo* gene emergence (Wilson et al. 2017). These two models make incompatible predictions regarding the properties of emerging genes, relative to the properties of old genes and non-genic regions. In the case of pre-adaptation, only “pre-adapted” regions can be subject to *de novo* gene emergence.

Indeed, evidence from an array of organisms has since shed light on the progressive, spectrum-like properties of *de novo* genes as they become older (Ruiz-Orera et al. 2018; Van Oss and Carvunis 2019; Bornberg-Bauer et al. 2021; Papadopoulos et al. 2021; Peng and Zhao 2024). These studies revealed that genes that have recently arisen *de novo* have characteristics that sets them apart from older, more conserved genes. They tend to be shorter (both in total gene and exon length), less expressed or expressed specifically to restricted

tissues/developmental stages; they possess fewer exons, evolve more rapidly, and are more similar to non-genic sequences (Neme and Tautz 2013; Van Oss and Carvunis 2019).

5. First steps of *de novo* gene birth , within a life cycle

Before we can speak of a *de novo* gene, it is necessary for any sequence to pass through several stages: transcription, translation and functionalization. This entire process is referred to as “proto-genization” by Carvunis et al. 2012. Here, we are particularly interested in characterizing the first steps in *de novo* gene emergence (see **Figure 4**), which would thus be integrated into a broader cycle of births and deaths of *de novo* genes, with periods of evolutionary radiation characterized by greater retention of *de novo* genes (Neme and Tautz 2016). However, it would appear that a significant portion of these expressed ORFs evolve neutrally, and their expression is thus not sufficient to qualify them as *de novo* genes: rather, they would serve as a reservoir for the appearance of new genes (Ruiz-Orera et al. 2018; Durand et al. 2019). Ultimately, a tiny fraction of ORFs transcribed (ubiquitously), then translated (at low levels) would lead to the emergence of *de novo* genes, and the relative importance of the *de novo* mechanism compared to other mechanisms generating new genes thus remains to be elucidated.

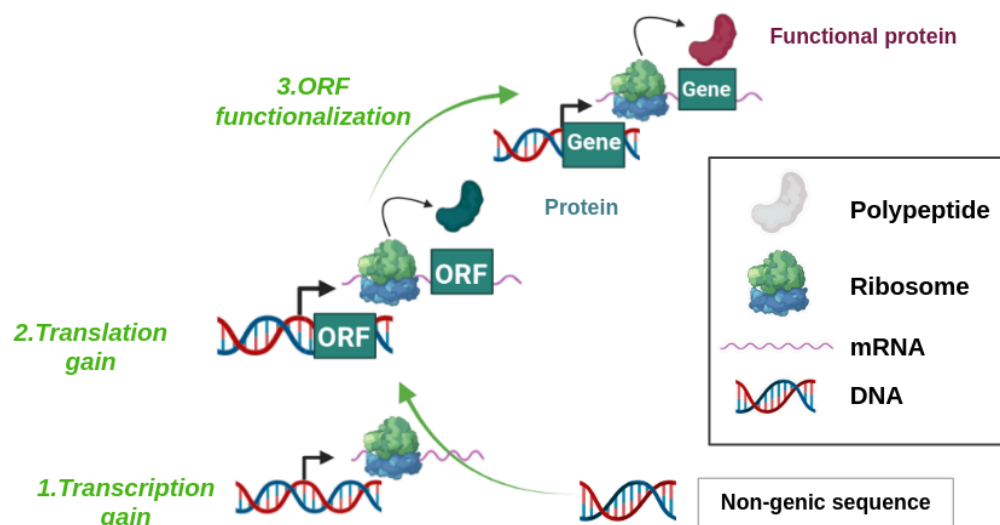


Figure 4: Model of *de novo* gene emergence, from a non-genic region, by progressive acquisition of transcription, translation and function.

Similarly, the population aspects of *de novo* gene emergence and their spread in populations remain largely unknown. A study in *Drosophila* (Zhao et al. 2014) suggests that while many

polymorphic *de novo* genes are generated in male tissues, most are lost through genetic drift, but a few may be maintained through selection. Thus, whereas infrequent *de novo* genes are rather deleterious, short, not very complex and not very expressed; frequent *de novo* genes are rather neutral, long, more complex and more expressed. Selection would thus favor greater length, expression rate and complexity (number of exons). More recently, a study in *A. thaliana* populations uncovered 782 *de novo* genes, which might have spread and become fixed within a population by remaining expressed at low levels, thanks to high degrees of methylation (Li et al. 2016). Hence, they suggest a “neutralization by DNA methylation model”, according to which Transposable Elements (TEs), siRNAs or even antisense transcripts could enhance DNA methylation within *de novo* genes, allowing for their population spread without inducing deleterious effects (Li et al. 2016).

Overall, while it is now clear that *de novo* genes are important contributors to evolutionary novelty, their prevalence and general properties have been studied in a few model species only. These species are often distantly related to one another, and studies tend to focus on previously annotated genes, resulting in the lack of a clear picture of the origins and short-term evolutionary dynamics of *de novo* genes. This is especially true in plants, in which *de novo* gene birth has been investigated by a handful of species only (Xiao et al. 2009; Li et al. 2016).

II. Open Reading Frames in microRNA transcripts: the case of miORFs

1. microRNAs : a key type of regulatory element

In plants, a mature miRNA is in most cases a 21-22 base-pair long (more rarely 24 bp) one-stranded RNA molecule capable, by almost perfect sequence complementarity, of degrading the messenger RNA of a target gene, or inhibiting its translation. In plants, most (67%) miRNA genes are independent transcription units with exons, introns, and their own promoters, which are transcribed by RNAPol II, capped and poly-adenylated (Zhang et al. 2015; Stepien et al. 2017). Initially transcribed as primary transcripts (pri-miRNAs), such transcripts are then recruited into Dicer bodies (D-bodies) within the nucleus, by what we may call the “microprocessor”: an assembly of Dicer-Like 1 (DCL1), Hyponastic Leaves 1 (HYL1) and Serrate (SE), as well as other auxiliary proteins (Zhang et al. 2015; Ding and Zhang 2023). This longer pri-miRNA (from a few hundred to over a thousand bases) is processed by the microprocessor into a precursor (pre-miRNA) exhibiting a stem-loop structure (of variable length), itself processed into a miRNA duplex, with one of the strands becoming the mature

miRNA (Wang et al. 2019). This duplex is exported to the cytoplasm, where an Argonaute protein (AGO1) selects the mature miRNA strand by recognizing its 5' terminal nucleotide (Mi et al. 2008), and then integrates it into the RNA-induced Silencing Complex (RISC). This RISC complex is then able to target specific gene messenger RNAs (mRNAs), thanks to their sequence complementarity to the miRNA, and inhibit the expression of the gene by mRNA degradation or mRNA translation inhibition. Other biogenesis pathways exist, and one can briefly mention the possibility of pri-miRNA cleavage by other DCL proteins, such as DCL2 and DCL3, which lead to mature RNAs 22 and 24 nucleotides long respectively; and other Argonaute proteins (AGO4, AGO6, AGO9) selecting mature RNAs of a different sizes (Rogers and Chen 2013). The simplified genesis pathway is represented below in **Figure 5**:

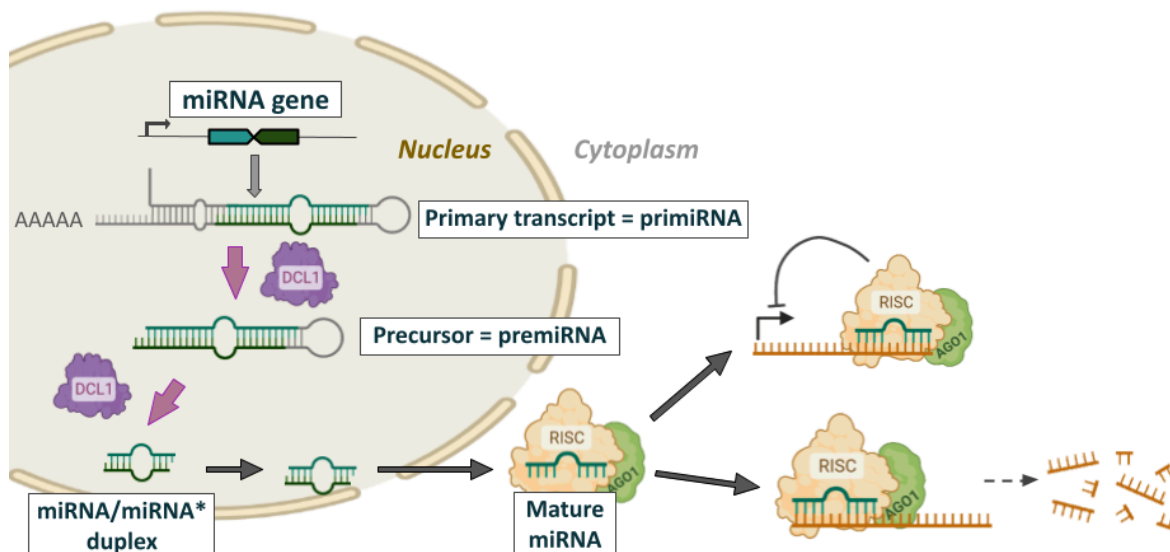


Figure 5: Simplified diagram of the genesis and possible mechanisms of action of miRNAs in plants. The primary transcript (pri-miRNA) is cleaved by an endoribonuclease, such as Dicer-Like 1 (DCL1), into a precursor (pre-miRNA). This precursor is then itself cleaved by DCL1 to release the miRNA/miRNA* duplex. The latter is then exported to the cytoplasm, where it is taken over by an Argonaute protein (e.g. AGO1), selecting just one of the two strands and bringing it together with other proteins to form the RISC complex (RNA-Induced Silencing Complex). This complex is then able to recognize a target gene by sequence complementarity with the mature miRNA, and inhibit expression of this gene by degradation of its mRNA, or inhibition of its translation.

Furthermore, a single miRNA locus can lead to the production of various miRNA primary transcripts, both in the form of transcripts of varying length (attributable to alternative

transcription start sites and/or alternative polyadenylation sites (Szarzynska et al. 2009; Bielewicz et al. 2013)) and of splicing isoforms. Indeed, the literature on miRNA primary transcript splicing reveals that most plant miRNAs contain at least one intron, and could be subject to splicing (Zhang et al. 2015; Stepien et al. 2017). A number of studies revealed that the presence of introns in miRNA primary transcripts is crucial, and may even be necessary for the accumulation of the mature miRNA, such as in the cases of miR163, miR161, and miR400 (Yan et al. 2012; Bielewicz et al. 2013), with environmentally-induced splicing exemplified by miR842/846 (Jia and Rock 2013).

Functionally, miRNAs are involved in a wide range of physiological processes, including leaf morphogenesis, stress response and root development (Millar 2020), through the regulation of key transcription factors. In allowing for precise spatial and time specific expression of genes, they are crucial for development and responding to environmental changes, including abiotic ones (Song et al. 2019). Mutations in key processing components such as DCL1 lead to complete lethality in embryos (when expression is abolished), or severe developmental defects (when weakly expressed) (Schauer et al. 2002).

2. The coding potential of miRNA primary transcripts

MiRNAs have long been considered non-coding, and even less likely to lead to the production of functional polypeptides. However, Lauressergues and his team revealed in 2015 the presence and functional character of peptides encoded by short ORFs located within the 5' region of miRNA primary transcripts (Lauressergues et al. 2015), in *A. thaliana* and *M. truncatula*, which we will from then on call “miORFs”. They were able to demonstrate that the miRpeptides (miPEPs) thus produced were then able to positively regulate the expression of the same primary transcript from which they originated. They suggested the following model to elucidate the possible molecular mechanisms of this regulation (**Figure 6**):

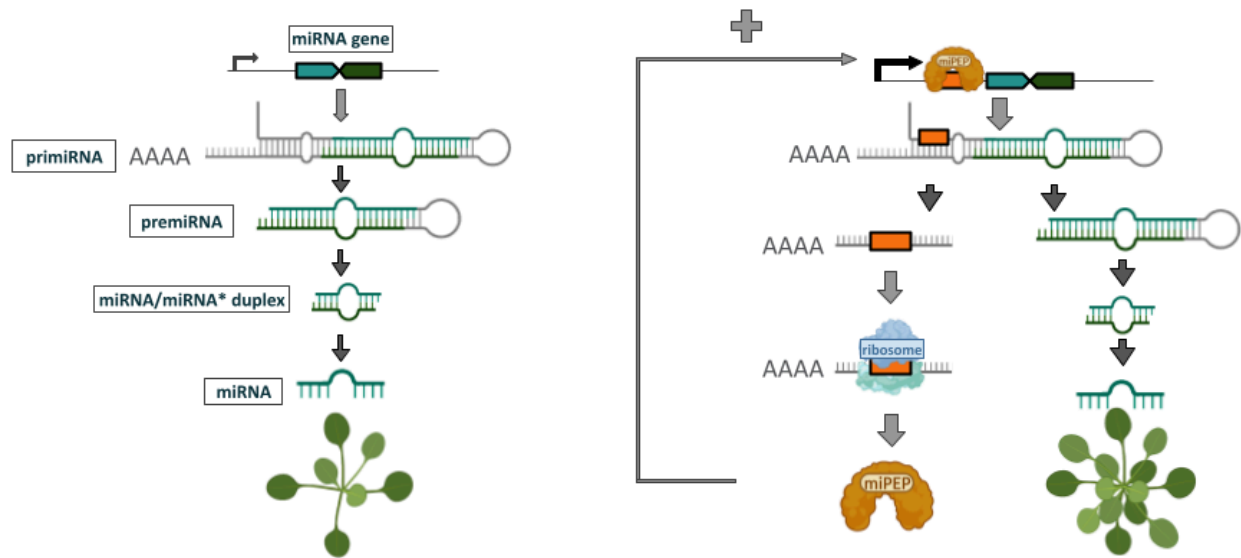


Figure 6: Left: Molecular processing leading to the production of a mature miRNA, from a first primary transcript (pri-miRNA) cleaved into a second precursor (pre-miRNA), further cleaved into a miRNA/miRNA* duplex. Right: Mode of action of miPEPs: translated from a miORF in the 5' region, they are able to bind to their miORF and induce an overexpression of the corresponding miRNA, leading to visual phenotypic effects.

Inducing the overexpression of the miPEP (or applying the synthetic miPEP to plants) could then lead to visible phenotypic effects, such as a reduction in lateral root production, due to the quantitatively greater presence of mature miRNA (Lauressergues et al. 2015).

Between 2015 and 2022, several studies were conducted to further characterize the functional potential of these miPEPs in other species, and within other tissues. These have identified peptides such as AtmiPEP164b, capable of inhibiting root growth in *Barbarea vulgaris* (Ormancey et al. 2021); miPEP165a, capable of promoting root growth and influencing flowering time in *A. thaliana* (Ormancey et al. 2020); or miPEP172c, capable of stimulating nodulation in *Glycine max* (Couzigou et al. 2016). However, the majority of candidate miPEPs have not yet been the target of any functional validation, and a more exhaustive repertoire of candidates has yet to be characterized. Major questions remain: are miORFs common within miRNAs, do they have a regulatory role, and by what molecular mechanisms does the genesis and regulatory activity of miPEPs take place?

To answer the latter question, more recently, the same team (Lauressergues et al. 2022) proposed a more detailed version of the previously presented model. According to this, miRNA

genes could in fact be alternatively transcribed into two types of transcripts. The first one, short, contains only the miORF. It would be exported outside the nucleus to be translated into a miPEP, which would then be reintegrated into the nucleus and positively regulate the expression of the miRNA from which it originated. The second, longer transcript, in which only the stem-loop is excised and conserved, leads to the mature miRNA (see **Figure 7** for the updated model).

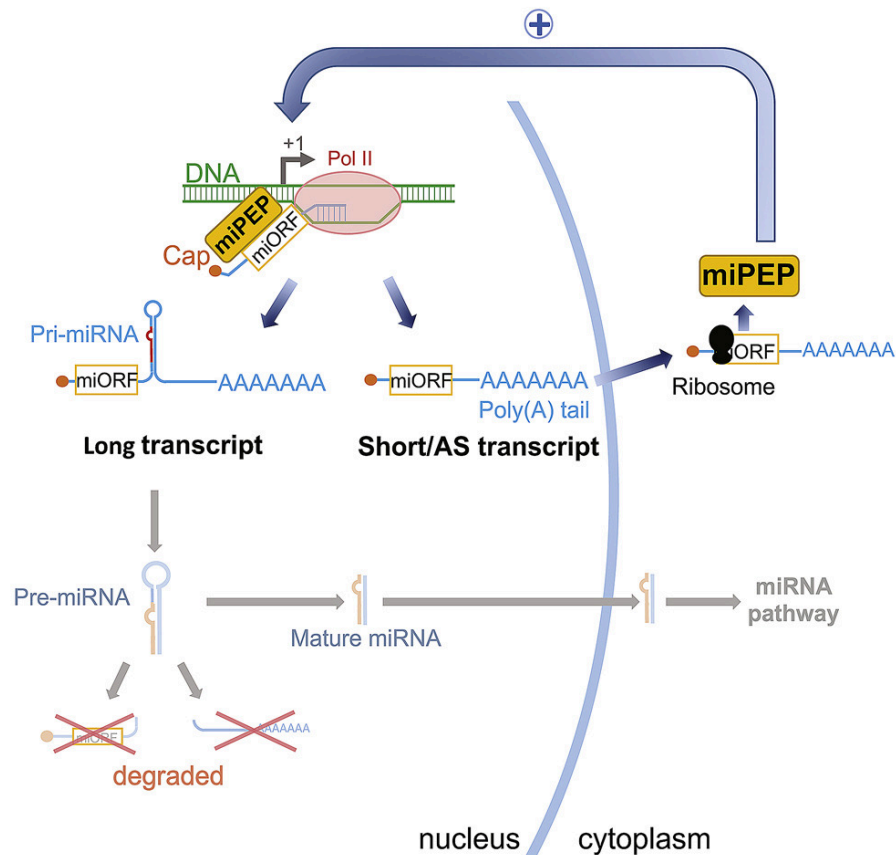


Figure 7: Molecular mechanism of the miRNA-miPEP system suggested and refined upon by Laressergues and his team in 2015 and 2022. From a single miRNA locus, two types of transcripts are generated: a first, short one, which contains the miORF but not the precursor, and which is exported to the cytoplasm where it is translated into a miPEP. The second, longer one, which contains both the miORF and the precursor, and which is canonically processed into a functional miRNA/miRNA* duplex, without miORF translation. The internalization into the nucleus of a miPEP translated from and, binding to its corresponding miORF, enhances the transcription of the host miRNA, leading to visible phenotypic effect (Laressergues et al. 2022).

Furthermore, it has recently been discovered that miPEPs also exist within the animal kingdom: miPEP133 in humans, a tumor suppressor localized to mitochondria (Kang et al. 2020), and MSAmiP in *Drosophila*, localized within the nucleus of male accessory gland secondary cells, involved in male reproductive success (Immarigeon et al. 2021). It has even been suggested that the mechanism of pri-miRNA retroamplification by its corresponding miPEP is not necessarily common to all miPEPs identified in humans (Prel et al. 2021).

Yet, questions remain as to the exact mechanism by which miPEPs operate: how is the primary transcript exported to the cytoplasm? Is miPEP translation possible within the nucleus? How is the miPEP reintegrated into the nucleus? Is this mechanism of action universal or species-specific? (Prasad et al. 2021). How can a miPEP be specific to the pri-miRNA from which it originates? (Lauressergues et al. 2022). Recently, protein to corresponding DNA specificity was also reported for protein coding genes, with “complementary peptides” being able to positively regulate the translation of their corresponding CDS’ upstream/downstream ORF (Ormancey et al. 2023). However, the generality of such results have yet to be confirmed by further studies.

3. How do miRpeptides evolve?

One of the key unanswered questions about miPEPs is the evolution of their repertoires. Approximately 50% of miRNA genes are shared among flowering plants, but many have been gained later on, and plant miRNA repertoires are considered to be highly lineage-specific (Nozawa et al. 2012). They seem to be subject to a rapid cycle of death and birth, giving rise to a labile repertoire between species (Fahlgren et al. 2007; Fahlgren et al. 2010; Pavan et al. 2024). Moreover, these repertoires appear to be highly dynamic. “Young” miRNA genes are less conserved, weakly expressed and have few or no targets. Thus, new miRNAs are continually generated within lineages, but are also continually lost. Although miRNAs were previously thought to primarily evolve from reverse duplication of protein-coding genes (Fahlgren et al. 2010), recent literature tends to show that transposable elements are actually the main contributors to miRNA gene birth in *A. halleri* and *A. lyrata*, with Miniature inverted transposable elements (MITE) from the Harbringer and Mariner superfamilies being the most prolific (Pavan et al. 2025).

Intergenic ORFs, meanwhile, also have their own evolutionary dynamics. In mice, for example, most peptides derived from the expression of species-specific sORFs appear to evolve neutrally (Ruiz-Orera et al. 2018). Fixation of mutations at the sORF level, enabling a progressive gain in translation, could thus be achieved by simple genetic drift, as long as the polypeptide product is not toxic (Ruiz-Orera and Albà 2019). Furthermore, while the presence of upstream ORFs (uORFs) is often conserved between species, their sequence as such does not appear to be evolutionarily constrained, and high rates of divergence have been observed in *Drosophila* (Zhang et al. 2018). Repertoires of short ORFs are thus considered highly labile, and evolve rapidly (Durand et al. 2019).

Hence, while the evolution of miRNA genes on the one hand, and sORFs on the other, has already been characterized, that of miORFs in themselves has not been the subject of any study to date. In this second chapter, we wish to further explore the evolutionary dynamics of the miRNA-miPEP system, and in particular its genomic origins. If several mechanisms have already been suggested for the emergence of miRNA genes, such as inverted protein coding gene duplications, miRNA gene duplications, emergence from a MITE (Pavan et al. 2025) or spontaneous genomic hairpin formation (Voinnet 2009; Nozawa et al. 2012) nothing is known about the emergence of miPEPs. One particular point of contention is whether miPEPs emerge before, after, or simultaneously to their corresponding miRNA; especially since they act as positive regulators of miRNA expression. Therefore, we have suggested several models of emergence for the miRNA-miPEP system, which are non-exclusive and non-exhaustive (**Figure 8**).

The first model implies the acquisition of a spontaneous sORF from random point mutations, a process of *de novo* origination now considered to be relatively common for new genes (McLysaght and Hurst 2016; Van Oss and Carvunis 2019). Here, the sORF could immediately benefit from the high transcription level of the pre-existing miRNA gene if it is advantageous, or be eliminated if toxic. The second model involves the apparition of a miRNA gene from an inverted duplication of a protein-coding gene, located downstream of an uORF. The sORF, derived from the uORF, could have already been filtered by natural selection and is supposedly non-toxic. Finally, the third model implies the acquisition of both a sORF and a miRNA precursor, via the partial inverted duplication of a protein-coding gene. Here, the sORF derived

from a fraction of the ancestral CDS and some known protein domains may be conserved, which may provide hints as to the possible mode of regulatory action of the sORF. Other models could be proposed and involve TEs, known to also play a role in the gain of new miRNA genes (Pavan et al. 2025).

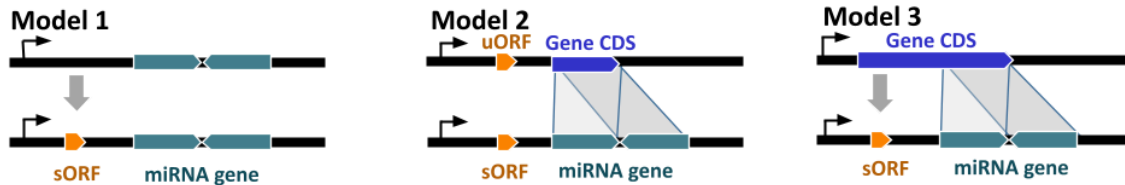


Figure 8: Some of the molecular models suggested for the origins of the miPEP-miRNA system, which are not exhaustive.

Combining long-read transcripts sequencing with Ribosome Profiling : scientific strategy and methods optimization

We hypothesize that the functional potential of short ORFs has so far been largely underestimated, due to the technological and statistical limitations facing their characterization. In particular, a key element to evaluate the functional and evolutionary importance of ORFs is the ability to characterize their translation, both in *de novo* ORFs, as well as within miRNA primary transcripts. In this respect, the recent emergence of Ribosome Profiling (RPF) technology now enables real-time, genome-wide characterization of translation events (Ingolia et al. 2009). Its principle is to sequence RNA fragments that have been protected from RNase digestion by the presence of ribosomes. In this way, only sequences associated with ribosomes are preserved, and it is possible to infer the translational status of any region of the genome. A feature of RPF data is its periodic nature during translation: indeed, when a coding region is scanned by ribosomes, these will tend to accumulate at the first base of each codon. Thus, the fragments of RPF protected by these ribosomes will follow a trinucleotide periodicity along the coding sequences (cf. **Figure 9**).

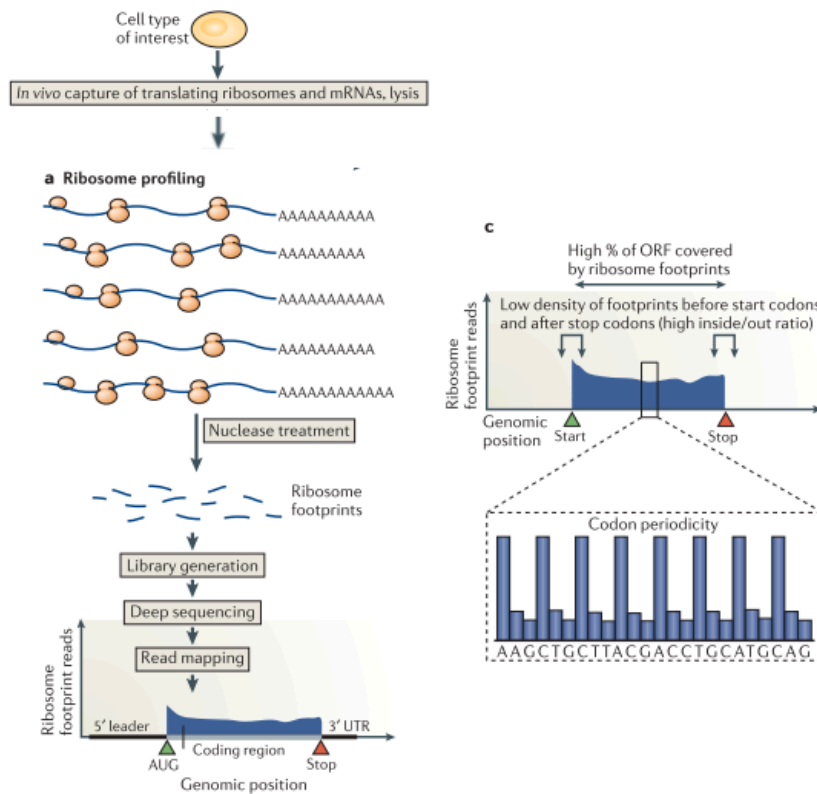


Figure 9: Adapted from Brar, 2015: General Ribosome Profiling experimental protocol, and periodicity of associated reads (Brar and Weissman 2015). Ribosome-associated mRNAs (known as “polysomes”) are digested by a nuclease, where the ribosomes act as protectors, leaving behind protected RNA fragments called “ribosome footprints”. Those footprints are further prepared to generate a library, which is then sequenced and mapped to the genome. In protein coding regions, the footprints will tend to accumulate on the first base of each codon, because that is where ribosomes tend to stay the longest when actively translating.

Hence, the presence of triplet periodicity at a given location within a sequence is considered to be the “gold standard” of translation evidence, and can be detected and analyzed by a wide array of tools (Brar and Weissman 2015).

As for miRNAs, they are annotated on the basis of their precursors. Databases such as miRBase (Griffiths-Jones et al. 2006) list potential miRNA precursors annotated solely on the basis of their sequence, and currently number 326 in *A. thaliana*. However, as mentioned earlier, it is not the precursor that is directly transcribed from the gene, but the primary transcript. As primary transcripts have no distinguishing features, apart from containing one or more future precursors, their *in silico* annotation is complex, and is usually limited to the position of the precursors. However, it has been shown that sORFs encoding miPEPs are located upstream of miRNA precursors, on the pri-miRNA. This is why the use of long-read RNA sequencing data is crucial to the annotation of whole primary transcripts, and therefore to the discovery of miPEPs.

For this purpose, we have chosen to use Nanopore long-reads technology (Oxford Nanopore Technologies).

The global scientific plan of the thesis is represented below in **Figure 10**, showcasing the use of RPF data as well as long-read RNAseq data:

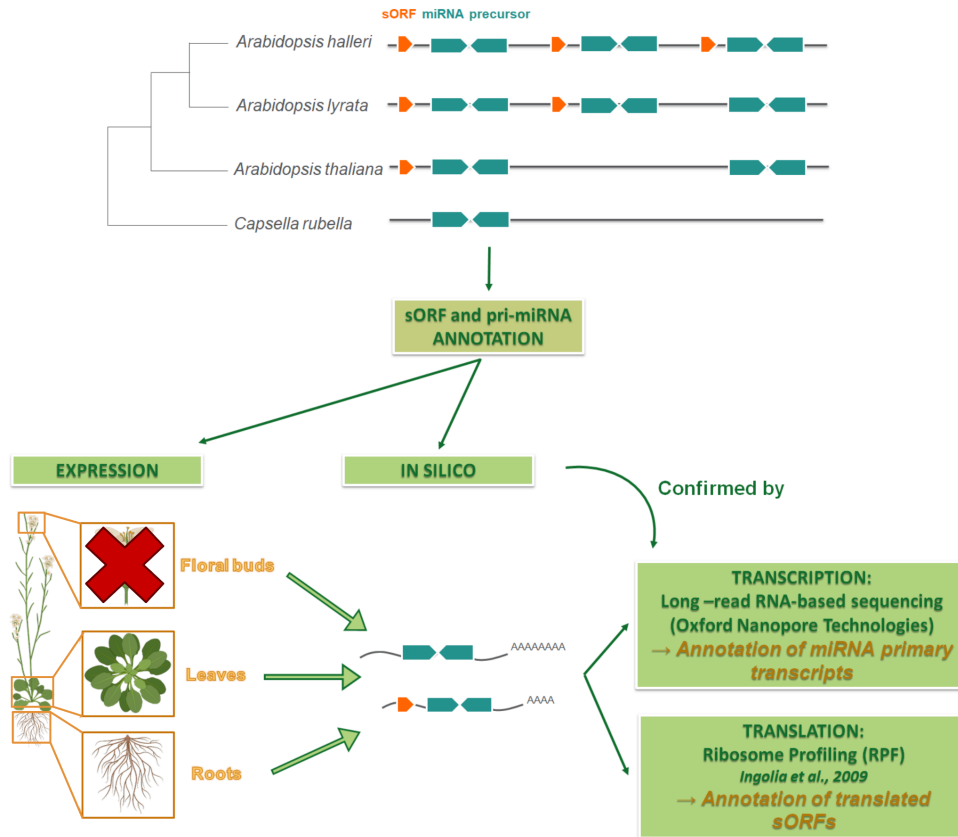


Figure 10: Schematic representation of the methodological workflow for annotating pri-miRNAs and sORFs. On the left, the tissues from which the RNAs were extracted and purified, including whole primary transcripts and mRNAs, within the species represented in the phylogeny. The original plan called for the use of flower buds, leaves and roots, but was later updated due to the lack of flowering in *A. halleri* and *A. lyrata* individuals. On the right, the 2 main branches of analysis: transcription and translation, using Nanopore long-read RNA sequencing and Ribosome Profiling sequencing respectively.

The original experimental setup called for the production of floral buds, roots, and leaves for each of our four species: *A. thaliana*, *A. halleri*, *A. lyrata* and *C. rubella*. While we were able to obtain significant amounts of each tissue type for *A. thaliana* and *C. rubella*, only one *A. halleri* individual flowered over the course of the growing season, and none of the *A. lyrata* individuals did. This meant we had to revisit the experimental setup by removing the floral buds tissue type. In total, ten *A. thaliana*, eight *A. halleri*, eight *A. lyrata* and six *C. rubella* individuals were

sampled. Flash-frozen samples were stored by tissue type, individual, and sampling date, at -80°C. We chose to keep three biological replicates (three distinct individuals) per species.

These frozen samples were to be used for the production of RPF and long-read RNAseq data in every species. However, the scientific plan was further revised after two main steps.

The first one in producing RPF libraries, is related to the production of polysome profiles, which are evidence that the RPF lysates contain mRNAs undergoing translation, and are used as a quality control step in RPF library preparation (see **Figure 11**).

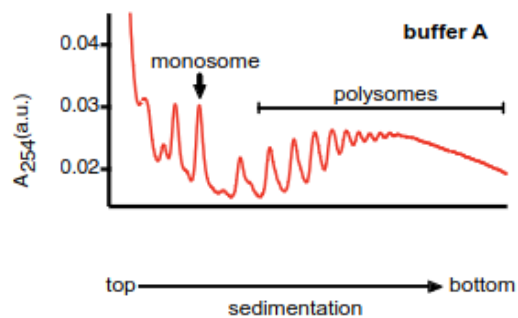


Figure 11: To monitor the active translation of mRNAs, we can perform “polysome profiling”, a method which allows us to assess the association of mRNAs with either one ribosome (monosomes) or several ribosomes (polysomes). To do so; polysomes and monosomes from the same RPF lysate are separated on a sucrose density gradient, and the gradient’s absorbance at 260nm is progressively read by a spectrophotometer (Lecampion et al. 2016). This absorbance is then plotted against the sedimentation gradient, which is called a “polysome profile”. In high quality lysates, the profiles display a first, high peak corresponding to monosomes, followed by a series of wavelets each corresponding to a polysome size (with an increasing number of ribosomes associated to mRNAs), as is represented in this figure, adapted from (Hsu et al. 2016).

Within our own RPF lysates, we took one aliquot of *A. thaliana* leaves (replicate 1) and one aliquot of *C. rubella* leaves (replicate 2), which were separately loaded on a sucrose gradient; and then analysed to control the quality of the polysome extraction (Baudin-Baillieu et al. 2016). We expected to observe a first main peak corresponding to monosomes; and then several; much smaller wavelets corresponding to polysomes of a varying number of attached ribosomes; as had been observed previously for *A. thaliana* flower buds lysates of similar concentration (**Figure 11**). However, we were able to observe neither the main monosome peak nor the succeeding wavelets (**Figure 12**). We then repeated the experiment, but with two aliquots from the same *A. halleri* leaves lysate (replicate 1). This time, we were able to observe part of the main monosome peak in one of the aliquots, but still none of the polysome wavelets (**Figure 13**). This inability reflects the fact that there was an insufficient quantity of polysomal material in

our samples, despite a sufficient concentration measured in absorption units. Our hypothesis is that, compared to flower buds, seedlings or young tissues; mature leaf tissues are expressed at a much lower rate, leading to lower number of polysomes in samples. This would correspond to what has previously been reported for plant polysome profiles in the literature: the shape of the polysome profile depends on growth conditions, plant age and tissue; and the overall amount of both monosomes and polysomes is lower in older tissues where cells do not divide as frequently (Lecampion et al. 2016).

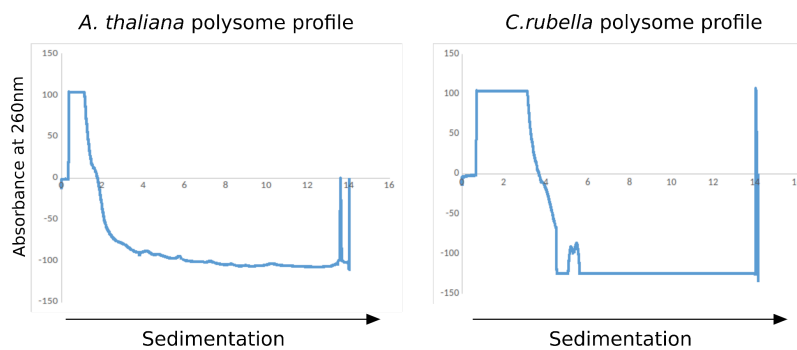


Figure 12: Polysome profiles of *A. thaliana* and *C. rubella* leaves RPF lysates, where the absorption is measured at 260nm along the sedimentation gradient.

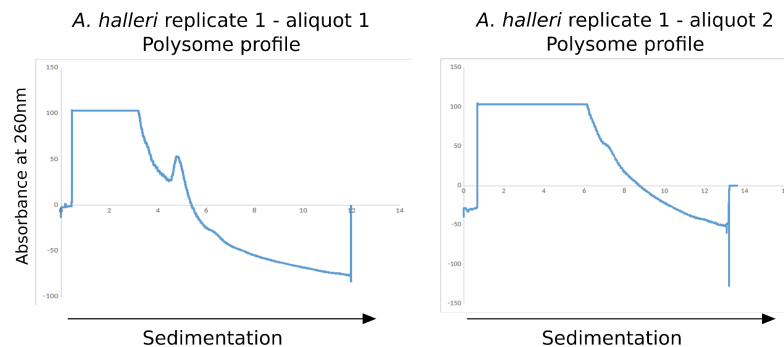


Figure 13: Polysome profiles of 2 aliquots of *A. halleri* leaves RPF lysates, where the absorption is measured at 260nm along the sedimentation gradient.

Nevertheless, we went on with the RPF library preparation, since the quality of the dataset can only ultimately be judged by the absence or presence of periodicity among protein coding genes. After sequencing libraries, we evaluated their quality, notably by producing metagene plots (Chapter 1). We chose to exclude our *A. thaliana* and *C. rubella* libraries from further downstream analyses due to their low quality. For *A. lyrata* and *A. halleri*, we selected one library per species (AHA_rep-2 and ALY_rep-1) with the highest quality for deeper sequencing. For *A. thaliana*, we used the public, high-quality RPF datasets from (Hsu et al. 2016; Wu and Hsu 2022).

This ribosome profiling data allowed us to detect translation events across the genomes of *A. halleri*, *A. lyrata* and *A. thaliana*; but not *C. rubella* and our own samples of *A. thaliana* as

originally planned. Still, enough data was available to be used for the core analyzes of both chapter 1 and 2.

The second step involved the preparation of long-read RNA sequencing data. Nanopore long-read RNA-seq data had seldom been used to annotate miRNA primary transcripts, and our first challenge was to provide a proof of concept of this method. Hence, preliminary data were produced from the sequencing of both *A. thaliana* and *A. halleri* floral buds for which different types of libraries were tested. The aim of this pilot study was two-fold: to identify the best library type for miRNA annotation, and to find out the sequencing depth at which a plateau was reached in terms of genome coverage, and beyond which further sequencing would bring little supplementary information. Libraries were prepared according to three protocols: the first extracts total RNA, which is reverse-transcribed into cDNA, amplified by PCR and then sequenced: PCRtot. A second, in which total RNAs are extracted, purified and then followed by the same steps as above: retro-transcription, amplification and sequencing: PCRp. The final method differs from the first in only one respect: there is no PCR amplification step, and it is referred to as “cDNA direct”. For *A. halleri*, only PCRtot and PCRp libraries could be produced, while for *A. thaliana*, each type of library could be produced and were complemented by public “RNA direct” data (no PCR or retro-transcription steps) from (Parker et al. 2020). Comprehensive analysis of libraries revealed that our cDNA direct library enabled the annotation of the longest and most numerous miRNA primary transcripts, while requiring little starting material and providing a high sequence number yield (contrary to RNA direct, which performed similarly in miRNA annotation, but which has a much poorer yield). Furthermore, a genome coverage plateau was reached at around four to five millions reads.

Based on the results of this pilot study, we then produced new long-read RNAseq data in the remaining species (i.e. all except *A. thaliana*), this time for leaves and roots (as well as plantules for *A. lyrata*). Library preparation kits proved less efficient than expected, with a significantly lower level of coverage compared to preliminary libraries (most likely due to the change in Nanopore chemistry and kit protocols). To obtain a coverage similar to what had been obtained for *A. thaliana*, several individuals from the same species were sampled. In the end, we prepared libraries for five individuals in *A. lyrata*, three individuals in *A. halleri* and two individuals in *C. rubella*.

This long-read Nanopore data allowed us to annotate miRNA primary transcripts in *A. halleri*, *A. lyrata*, *A. thaliana* and *C. rubella*, but not with the same tissue types as had initially been

planned, and with a lesser coverage than initially expected. Yet, the sequencing depths that were reached were still enough to annotate numerous primary miRNA transcripts across species.

Scientific questions and thesis plan

1. Chapter 1: Characterization of a repertoire of *de novo* ORFs: origins, conservation, functional potential

Recent scientific literature has highlighted the possibility of genes appearing *de novo*, from a non-coding region. Several models have been suggested to explain the mechanisms of *de novo* emergence, but the first steps in the conversion of a non-gene region into a gene region are still poorly understood. Few studies have investigated *de novo* gene birth in plant species, and were limited to distantly related model species and previously annotated genes. None have investigated the earlier steps of *de novo* gene birth, and only a handful of high-quality ribosome profiling datasets are available in plants. In this first chapter, I will seek to identify a repertoire of translated *de novo* ORFs by synteny, and to characterize the “starting material” for *de novo* gene emergence in the *Arabidopsis* genus. Combining ribosome profiling datasets from closely related species with synteny analyzes will allow us to obtain high-level evidence of the translation and *de novo* origin of ORFs, while circumventing “phylostratigraphic bias”, all within a short evolutionary timescale.

2. Chapter 2: Repertoire and conservation of miORFs - origins and evolution of the miRNA-miPEP system

Previous work has demonstrated the existence of functional miPEPs, translated from primary miRNA transcripts in *Arabidopsis thaliana*, *Arabidopsis lyrata* and *Medicago truncatula*. However, the scope of these studies was limited by their focus on just a few miRNA families, and they do not allow us to determine either the frequency or the conservation of the miPEP-miRNA system. Furthermore, while the evolutionary origins of miRNAs have been the subject of extensive study, those of the miPEP-miRNA system remain totally unknown to this day. Neither the molecular mechanisms by which this system emerged, nor its evolutionary dynamics after its emergence, are known. In order to elucidate these issues, I will assess the

prevalence of translated sORFs within the primary transcripts of all miRNA families, as well as their conservation within the *Arabidopsis* genus. Finally, I will explore whether the data are consistent with contrasted putative scenarios for the emergence of these molecular systems.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
- Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15:193–204.
- Baudin-Baillieu A, Hatin I, Legendre R, Namy O. 2016. Translation Analysis at the Genome Scale by Ribosome Profiling. In: Devaux F, editor. *Yeast Functional Genomics*. Vol. 1361. *Methods in Molecular Biology*. New York, NY: Springer New York. p. 105–124. Available from: https://link.springer.com/10.1007/978-1-4939-3079-1_7
- Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* 172:1675–1681.
- Bielewicz D, Kalak M, Kalyna M, Windels D, Barta A, Vazquez F, Szweykowska-Kulinska Z, Jarmolowski A. 2013. Introns of plant pri-miRNAs enhance miRNA biogenesis. *EMBO Rep.* 14:622–628.
- Bornberg-Bauer E, Hlouchova K, Lange A. 2021. Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* 68:175–183.
- Brar GA, Weissman JS. 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* 16:651–664.
- Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* 179:487–496.
- Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charleat B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487:370–374.
- Couzigou JM, André O, Guillotin B, Alexandre M, Combier JP. 2016. Use of microRNA-encoded peptide miPEP172c to stimulate nodulation in soybean. *New Phytol.* 211:379–381.
- Ding N, Zhang B. 2023. microRNA production in *Arabidopsis*. *Front. Plant Sci.* 14:1–11.
- Doolittle WF. 2018. We simply cannot go on being so vague about “function.” *Genome Biol.* 19:18–20.
- Doolittle WF, Brunet TDP, Linquist S, Gregory TR. 2014. Distinguishing between “Function” and “Effect” in Genome Biology. :1234–1237.
- Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. 2019. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* 29:932–943.
- Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law

- TF, Grant SR, Dangl JL, et al. 2007. High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA genes. *PLoS ONE* 2.
- Fahlgren N, Jogdeo S, Kasschau KD, Sullivan CM, Chapman EJ, Laubinger S, Smith LM, Dasenko M, Givan SA, Weigel D, et al. 2010. MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana. *Plant Cell* 22:1074–1089.
- Fesenko I, Kirov I, Kniazhev A, Khazigaleeva R, Lazarev V, Kharlampieva D, Grafkskaia E, Zgoda V, Butenko I, Arapidi G, et al. 2019. Distinct types of short open reading frames are translated in plant cells. *Genome Res.* 29:1464–1477.
- Fesenko I, Shabalina SA, Mamaeva A, Knyazev A, Glushkevich A, Lyapina I, Ziganshin R, Kovalchuk S, Kharlampieva D, Lazarev V, et al. 2021. A vast pool of lineage-specific microproteins encoded by long non-coding RNAs in plants. *Nucleic Acids Res.* 49:10328–10346.
- Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 34:140–144.
- Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, Nishi R, Ohashi C, Iida K, Tanaka M, et al. 2013. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc. Natl. Acad. Sci. U. S. A.* 110:2395–2400.
- 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.* 17:632–640.
- Hsu PY, Calviello L, Wu HYL, Li FW, Rothfels CJ, Ohler U, Benfey PN. 2016. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 113:E7126–E7135.
- Immarigeon C, Frei Y, Delbare SYN, Gligorov D, Almeida PM, Grey J, Fabbro L, Nagoshi E, Billeter JC, Wolfner MF, et al. 2021. Identification of a micropeptide and multiple secondary cell genes that modulate Drosophila male reproductive success. *Proc. Natl. Acad. Sci. U. S. A.* 118.
- Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Resolution Using Ribosome Profiling. *Science* 324:218–223.
- Jacob F. 1977. Evolution and Tinkering. *Science* 196:1161–1166.
- Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3:318–356.
- Jia F, Rock CD. 2013. MIR846 and MIR842 comprise a cisgenic MIRNA pair that is regulated by abscisic acid by alternative splicing in roots of Arabidopsis. *Plant Mol. Biol.* 81:447–460.
- Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–1326.

- Kang M, Tang B, Li J, Zhou Z, Liu K, Wang R, Jiang Z, Bi F, Patrick D, Kim D, et al. 2020. Identification of miPEP133 as a novel tumor-suppressor microprotein encoded by miR-34a pri-miRNA. *Mol. Cancer* 19.
- King M-C, Wilson AC. 1975. Evolution at Two Levels in Humans and Chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. *Science* 188:107–116.
- Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res.* 19:1752–1759.
- Lang GI, Murray AW. 2011. Mutation Rates across Budding Yeast Chromosome VI Are Correlated with Replication Timing. *Genome Biol. Evol.* 3:799–811.
- Lauressergues D, Couzigou JM, San Clemente H, Martinez Y, Dunand C, Bécard G, Combier JP. 2015. Primary transcripts of microRNAs encode regulatory peptides. *Nature* 520:90–93.
- Lauressergues D, Ormancey M, Guillotin B, San Clemente H, Camborde L, Duboé C, Tourneur S, Charpentier P, Barozet A, Jauneau A, et al. 2022. Characterization of plant microRNA-encoded peptides (miPEPs) reveals molecular mechanisms from the translation to activity and specificity. *Cell Rep.* 38.
- Lecampion C, Floris M, Fantino JR, Robaglia C, Laloi C. 2016. An Easy Method for Plant Polysome Profiling. *J. Vis. Exp.*:54231.
- Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* 103:9935–9939.
- Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009. Identification of the novel protein QQS as a component of the starch metabolic network in *Arabidopsis* leaves. *Plant J.* 58:485–498.
- Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, Ge S, Guo YL. 2016. On the origin of de novo genes in *Arabidopsis thaliana* populations. *Genome Biol. Evol.* 8:2190–2202.
- Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* 4:865–875.
- Luchetta P, Maurel M-C, Higuët D. 2005. Evolution moléculaire.
- Majic P, Payne JL. 2020. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Mol. Biol. Evol.* 37:1165–1178.
- McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: What, how and why. *Nat. Rev. Genet.* 17:567–578.
- Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C, et al. 2008. Sorting of Small RNAs into *Arabidopsis* Argonaute Complexes Is Directed by the 5' Terminal Nucleotide. *Cell* 133:116–127.

- Millar AA. 2020. The function of miRNAs in plants. *Plants* 9.
- Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.* 32:258–267.
- Moyers BA, Zhang J. 2016. Evaluating Phylostratigraphic Evidence for Widespread de Novo Gene Birth in Genome Evolution. *Mol. Biol. Evol.* 33:1245–1256.
- Moyers BA, Zhang J. 2017. Further simulations and analyses demonstrate open problems of phylostratigraphy. *Genome Biol. Evol.* 9:1519–1527.
- Moyers BA, Zhang J. 2018. Toward reducing phylostratigraphic errors and biases. *Genome Biol. Evol.* 10:2037–2048.
- Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* 14.
- Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife* 5:1–20.
- Nozawa M, Miura S, Nei M. 2012. Origins and evolution of microRNA genes in plant species. *Genome Biol. Evol.* 4:230–239.
- Ohno S. 1970. Evolution by Gene Duplication. Berlin, Heidelberg: Springer Berlin Heidelberg
Available from: <http://link.springer.com/10.1007/978-3-642-86659-3>
- Ormancey M, Guillotin B, Merret R, Camborde L, Duboé C, Fabre B, Pouzet C, Impens F, Van Haver D, Carpentier MC, et al. 2023. Complementary peptides represent a credible alternative to agrochemicals by activating translation of targeted proteins. *Nat. Commun.* 14:1–13.
- Ormancey M, Guillotin B, San Clemente H, Thuleau P, Plaza S, Combier JP. 2021. Use of microRNA-encoded peptides to improve agronomic traits. *Plant Biotechnol. J.* 19:1687–1689.
- Ormancey M, Le Ru A, Duboé C, Jin H, Thuleau P, Plaza S, Combier JP. 2020. Internalization of miPEP165a into arabidopsis roots depends on both passive diffusion and endocytosis-associated processes. *Int. J. Mol. Sci.* 21:1–16.
- Orr MW, Mao Y, Storz G, Qian SB. 2021. Alternative ORFs and small ORFs: Shedding light on the dark proteome. *Nucleic Acids Res.* 48:1029–1042.
- Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. *Science* 327:92–94.
- Papadopoulos C, Callebaut I, Gelly J-C, Hatin I, Namy O, Renard M, Lespinet O, Lopes A. 2021. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Res.* 31:2303–2315.
- Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, Hall AJW, Barton GJ,

- Simpson GG. 2020. Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6A modification. *eLife* 9:1–35.
- Pavan F, Azevedo Favory J, Lacoste E, Beaumont C, Louis F, Blassiau C, Cruaud C, Labadie K, Gallina S, Genete M, et al. 2024. The evolutionary history and functional specialization of microRNA genes in *Arabidopsis halleri* and *A. lyrata*. Available from: <http://biorxiv.org/lookup/doi/10.1101/2024.05.03.592357>
- Pavan F, Lacoste E, Castric V, Legrand S. 2025. Scenarios for the emergence of new miRNA genes in the plant *Arabidopsis halleri*. *Plant Cell Physiol.*:pcaf004.
- Peng J, Zhao L. 2024. The origin and structural evolution of de novo genes in *Drosophila*. *Nat. Commun.* 15:810.
- Prasad A, Sharma N, Prasad M. 2021. Noncoding but Coding: Pri-miRNA into the Action. *Trends Plant Sci.* 26:204–206.
- Prel A, Dozier C, Combiér JP, Plaza S, Besson A. 2021. Evidence That Regulation of Pri-miRNA/miRNA Expression Is Not a General Rule of miPEPs Function in Humans. *Int. J. Mol. Sci.* 2021 Vol 22 Page 3432 22:3432.
- Rogers K, Chen X. 2013. Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell* 25:2383–2399.
- Ruiz-Orera J, Albà MM. 2019. Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. *Trends Genet.* 35:186–198.
- Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messegueur X, Albà MM. 2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.* 2:890–896.
- Schauer SE, Jacobsen SE, Meinke DW, Ray A. 2002. DICER-LIKE1: blind men and elephants in *Arabidopsis* development. *Trends Plant Sci.* 7:487–491.
- Song X, Li Y, Cao X, Qi Y. 2019. MicroRNAs and Their Regulatory Roles in Plant–Environment Interactions. *Annu. Rev. Plant Biol.* 70:489–525.
- Stepien A, Knop K, Dolata J, Taube M, Bajczyk M, Barciszewska-Pacak M, Pacak A, Jarmolowski A, Szweykowska-Kulinska Z. 2017. Posttranscriptional coordination of splicing and miRNA biogenesis in plants. *Wiley Interdiscip. Rev. RNA* 8.
- Szarzynska B, Sobkowiak L, Pant BD, Balazadeh S, Scheible WR, Mueller-Roeber B, Jarmolowski A, Szweykowska-Kulinska Z. 2009. Gene structures and processing of *Arabidopsis thaliana* HYL1-dependent pri-miRNAs. *Nucleic Acids Res.* 37:3083–3093.
- Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and Collinearity in Plant Genomes. *Science*:486–489.
- Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12:692–702.

- Vakirlis N, Carvunis AR, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* 9:1–23.
- Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2018. A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* 35:631–645.
- Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet.* 15:1–23.
- Voinnet O. 2009. Origin, Biogenesis, and Activity of Plant MicroRNAs. *Cell* 136:669–687.
- Wang J, Mei J, Ren G. 2019. Plant microRNAs: Biogenesis, homeostasis, and degradation. *Front. Plant Sci.* 10:1–12.
- Weisman CM, Murray AW, Eddy SR. 2020. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* 18:1–24.
- Weng M-L, Becker C, Hildebrandt J, Neumann M, Rutter MT, Shaw RG, Weigel D, Fenster CB. 2019. Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in *Arabidopsis thaliana*. *Genetics* 211:703–714.
- Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* 1.
- Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* 3:1245–1252.
- Wu HYL, Hsu PY. 2022. A custom library construction method for super-resolution ribosome profiling in *Arabidopsis*. *Plant Methods* 18:1–8.
- Xia S, Chen J, Arsala D, Emerson JJ, Long M. 2025. Functional innovation through new genes as a general evolutionary process. *Nat. Genet.* [Internet]. Available from: <https://www.nature.com/articles/s41588-024-02059-0>
- Xiao W, Liu H, Li Y, Li X, Xu C, Long M, Wang S. 2009. A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS ONE* 4:1–12.
- Xiao Z, Huang R, Xing X, Chen Y, Deng H, Yang X. 2018. De novo annotation and characterization of the translome with ribosome profiling data. *Nucleic Acids Res.* 46:E61.
- Yan K, Liu P, Wu CA, Yang GD, Xu R, Guo QH, Huang JG, Zheng CC. 2012. Stress-Induced Alternative Splicing Provides a Mechanism for the Regulation of MicroRNA Processing in *Arabidopsis thaliana*. *Mol. Cell* 48:521–531.
- Yeasmin F, Yada T, Akimitsu N. 2018. Micropeptides encoded in transcripts previously identified as long noncoding RNAs: A new chapter in transcriptomics and proteomics. *Front. Genet.* 9:1–10.
- Yu H, Gerstein M. 2006. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci.* 103:14724–14731.
- Zhang H, Dou S, He F, Luo J, Wei L, Lu J. 2018. Genome-wide maps of ribosomal occupancy

- provide insights into adaptive evolution and regulatory roles of uORFs during *Drosophila* development. Chen X, editor. *PLOS Biol.* 16:e2003903.
- Zhang S, Liu Y, Yu B. 2015. New insights into pri-miRNA processing and accumulation in plants. *Wiley Interdiscip. Rev. RNA* 6:533–545.
- Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* 343:769–772.
- Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Res.* 18:1446–1455.
- Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci.* [Internet] 111. Available from: <https://pnas.org/doi/full/10.1073/pnas.1323011111>

CHAPTER I

Pervasive translation of short open reading frames and *de novo* gene emergence in *Arabidopsis*

Authors

Claire Patiou¹, Christelle Blassiau¹, Isabelle Hatin², Lars Eicholt³, Chloé Ponitzki¹, Laurence Debacker¹, Olivier Namy², Sylvain Legrand¹, Vincent Castric¹, Eléonore Durand¹

¹ Univ. Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France

² Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France

³ Institute for Evolution and Biodiversity, University of Muenster, 48149 Muenster, Germany

Author for correspondence : eleonore.durand@univ-lille.fr

Abstract

Increasing evidence has unveiled ancestrally non-genic sequences as prolific reservoirs for the de novo emergence of new genes . Across clades, some of these de novo genes were proven to have substantial phenotypic effects, and have helped improve our understanding of the emergence of biological function. Yet, very little is known about the starting material from which de novo proto-genes emerge. In plants specifically, few studies have characterized the mechanisms of de novo gene birth, and were limited to model species. Focusing on non-model species from the *Arabidopsis* genus, *A. halleri* and *A. lyrata*, we produced Ribosome Profiling (RPF) data to annotate genome-wide events of translation. With the help of synteny analysis, we identified translated Open Reading Frames (ORFs) that emerged de novo within the *Arabidopsis* genus. Doing so, we annotated a total of 738 de novo ORFs, most of which were found within intergenic regions. When compared to conserved Coding DNA Sequences (CDS), translated de novo ORFs were significantly shorter and less expressed. Strikingly, we observed a continuum of GC content within de novo conservation classes. Most de novo ORFs were species-specific, transiently existing and being expressed. Across conservation classes, 31 previously annotated CDS seemed to be promising proto-genes candidates. Our results highlight the abundance of translation events outside of conserved CDS, and the role they play in providing the starting material for the emergence of novel genes.

Introduction

According to François Jacob, "The probability of a functional protein appearing *de novo* by random association of amino acids is almost zero" (Jacob 1977). This quote embodies the paradigm that historically dominated evolutionary genetics until the early 21st century. According to this paradigm, evolution is a "tinkerer", simply reusing pre-existing coding material to generate new genes. These tinkering processes include gene and genome duplication, horizontal gene transfer, exon shuffling, gene fusion/fission, retroposition or the activity of mobile genetic elements, among others (Long et al. 2003; Zhou et al. 2008). In contrast, *de novo* gene emergence implies a mutational origin from historically non-coding nucleotide sequences. This mechanism was seen as highly unlikely, as the molecular characteristics of genes were considered far too complex to be the result of chance: if not impossible, it was seen as a very minor contributor to genetic novelty (McLysaght and Hurst 2016; Van Oss and Carvunis 2019). Instead, gene duplication was long considered the major source of new genes, a view exemplified by Ohno's 1970 book "Evolution by Gene Duplication" (Ohno 1970). Yet, this paradigm was challenged by the progressive discovery, in a diversity of organisms, of genes that lacked homology to any other known protein-coding sequences, some of which would come to be characterized as *de novo* (Begun et al. 2006; Levine et al. 2006; Cai et al. 2008; Knowles and McLysaght 2009; Xiao et al. 2009; Tautz and Domazet-Lošo 2011). The first *de novo* genes were discovered in *Drosophila* (Begun et al. 2006; Levine et al. 2006), yeast (Cai et al. 2008) and primates (Knowles and McLysaght 2009). Some were extensively characterized at the functional level and have key fitness consequences, including the iconic antifreeze glycoprotein in cod, which emerged as a vital component of survival in polar environments (Baalsrud et al. 2018). Overall, it is now clear that *de novo* genes are important drivers of the evolutionary process (Blevins et al. 2021; Montañés et al. 2023; Peng and Zhao 2024).

The mechanisms by which *de novo* genes emerge have been debated, and two main visions have been proposed: that of a pre-adaptation of non-genic sequences to the appearance of genes, according to which *de novo* genes would have exaggerated "gene-like" properties (Wilson et al. 2017), and that of a continuum between the properties of non-genic and genic regions (Carvunis et al. 2012) in which "proto-genes" progressively acquire "gene-like" properties. "Gene-like" properties would include characteristics such as a greater length, expression or conservation level, as well as the use of canonical START codons, lower disorder, and distinctive GC levels. While the former theory received some evidential support (Jin et al.

2025), and the latter theory received support by evidence from an array of organisms (Ruiz-Orera et al. 2018; Van Oss and Carvunis 2019; Bornberg-Bauer et al. 2021; Papadopoulos et al. 2021; Peng and Zhao 2024), it remains unclear how *de novo* genes acquire their molecular properties as they become older. Starting from untranscribed and untranslated non-genic nucleotide sequences, either progressive lengthening, or selection of the longest transcripts would give rise to the starting material for *de novo* gene emergence before or after the Open Reading Frame (ORF) sequence has even appeared (Iyengar and Bornberg-Bauer 2023). Subsequent steps include the gain of translation, the avoidance of proteasome degradation and the stable and non-toxic conformation of the polypeptide (Wilson and Masel 2011). Most studies have focused on novel protein-coding sequences that had already been annotated as genes, and hence harbor clearly established “gene-like” characteristics (Ruiz-Orera et al. 2015; Vakirlis et al. 2018; Bornberg-Bauer et al. 2021; Peng and Zhao 2024). However, little is known about the very first steps of *de novo* gene emergence, especially the properties of the starting non-coding nucleotide sequences and how they eventually acquire “gene-like” characteristics according to the continuum or the pre-adaptation model.

Studying the early steps of this process is challenging for at least two reasons. First, depending on the set of thresholds chosen to distinguish non-genic from genic elements (Dinger et al. 2008), millions of potential peptide-encoding sequences can be annotated in the intergenic regions of any genome. Selecting the most promising gene candidates among them requires translational evidence, which is hard to obtain, particularly outside of canonical protein-coding sequences, whose expression tends to be low. Obtaining translational evidence is nonetheless possible by using Ribosome Profiling (RPF), a powerful sequencing technology to provide a detailed snapshot of translational activity at the genome-wide scale (Ingolia et al. 2019; Kiniry et al. 2020). Yet, few RPF datasets have been produced outside of a few well-studied model organisms (Gilmore et al. 2015; Kim et al. 2020; Zhao et al. 2024). Second, confidently asserting that a given ORF has emerged *de novo* requires the identification of a clear homologous non-coding region. However, detecting homology lacks power beyond a certain degree of divergence, resulting in a potentially serious “phylostratigraphic bias” (Moyers and Zhang 2015; Moyers and Zhang 2016; Moyers and Zhang 2017; Moyers and Zhang 2018). The best way to circumvent this problem is to use synteny information and only compare intergenic regions flanked by anchors of genes that have remained syntenic. Because genomes can be rapidly and drastically rearranged over the course of evolution, this involves comparing species that are sufficiently closely related that synteny relationships remain preserved. Previous studies

have tackled either of those two problems, on one hand by producing RPF data in either one or a handful of distantly related species (Ruiz-Orera et al. 2018; Xiao et al. 2018), or on the other hand by performing thorough syntenic analyzes between closely related species, without RPF data from each of them (Vakirlis et al. 2020; Roginski et al. 2024). To our knowledge, a single paper has combined synteny analyzes with the production of several RPF datasets from closely related species, allowing for the characterization of a high translational turnover of emerging *de novo* ORFs in yeast (Durand et al. 2019). Hence, combining genomic and translation data in a comparative framework has shown great potential in deciphering the first steps of *de novo* gene birth within a short evolutionary timescale, but has so far been under-exploited.

The genome-wide study of *de novo* genes, as well as the functional characterization of a few candidates, focused mostly on a handful of model organisms, including *Drosophila* (Peng and Zhao 2024), yeast (Vakirlis et al. 2018) and primates (Ruiz-Orera et al. 2015). In comparison, little is currently known about how *de novo* genes arise in plants, and very few have been functionally verified. One early example of a plant *de novo* gene is the qua-quine-starch (QQS) gene in *A. thaliana*, which was found to encode a 59 amino-acid polypeptide involved in the regulation of starch synthesis in response to external and internal perturbations (Li et al. 2009). A more recent example, also in *A. thaliana*, is the Sun Wu-Kong (SWK) *de novo* gene, which was found to encode a 8kD protein involved in seed germination under osmotic stress (Jin et al. 2025). However, beyond such specific examples, only few studies have investigated the global emergence of *de novo* genes in plants, and most of them focused on model species like *A. thaliana* and *Oryza sativa* (Silveira et al. 2013; Cui et al. 2015; Li et al. 2016; Zhang et al. 2019). In *A. thaliana*, they revealed the existence of 782 *de novo* genes with a series of interesting properties, such as the fact that *de novo* genes tend to be shorter, more highly methylated and more likely to be expressed in pollen. Yet, a strong limitation is that these studies focused on genes that had already been characterized as such, and none has investigated the earlier steps of *de novo* gene birth, such as the role short, translated ORFs might play. Translation itself is under-characterized, with only a handful of high-quality RPF datasets available in plants, all from distantly related species such as *O. sativa* (Zhu et al. 2023), *Zea mays* (Chotewutmontri et al. 2018) or *A. thaliana* (Hsu et al. 2016; Wu and Hsu 2022). Again however, these studies focused mostly on genes that had already been annotated. The fact that pervasive translation activity can be detected in intergenic regions has been demonstrated in *A. thaliana* (Hanada et al. 2007; Hanada et al. 2013), but the role this may play in the context of *de novo* gene

emergence has not been explored. Hence, it has so far not been possible to assess the importance of short, translated intergenic ORFs in *de novo* gene birth in plants.

Here, we aimed to characterize the first steps of *de novo* gene emergence in the *Arabidopsis* genus by specifically investigating the extent to which short ORFs could serve as a reservoir for the appearance of new genes (Ruiz-Orera et al. 2018; Durand et al. 2019). To obtain high-level evidence for the discovery of novel protein-coding ORFs (Prensner et al. 2023), we produced RPF data to specifically select ORFs that exhibit evidence of translational activity. We focused on three species of the *Arabidopsis* genus: *A. halleri*, *A. lyrata* and *A. thaliana*. *A. halleri* and *A. lyrata* diverged less than one million years ago (Roux et al. 2011), and they diverged about five million years ago from *A. thaliana*. The recent divergence between these species provides high power to the synteny analysis, resulting in a unique opportunity to reliably identify the non-coding sequences that are homologous to the putative *de novo* ORFs. Overall, our analysis reveals hundreds of translated *de novo* ORFs within intergenic syntenic regions, shedding new light on the very first steps of the emergence of *de novo* genes in plants.

Results

Ribosome Profiling sequencing yields evidence of ongoing translation

Our overall strategy was based on four closely related Brassicaceae species for which high-quality reference genome assemblies are available (*A. thaliana*; (Lamesch et al. 2012), *A. lyrata*; (Hu et al. 2011), *A. halleri*; (Pavan et al. 2024) and *Capsella rubella*; (Slotte et al. 2013)). First we broadly annotated all possible ORFs across these four genomes, using a relaxed sequence-based bioinformatic annotation of ORFs, allowing for alternative START codons (ATG, TTG, CTG, GTG, ACG), a minimal length of 9 nucleotides, regardless of overlap with gene annotations. This yielded a total of 19, 23 and 3 million ORF annotations in *A. halleri*, *A. lyrata* and *A. thaliana* respectively (**Table S3**). Second, we sought to obtain direct experimental evidence for translation activity. To that end, we grew plants of the four species under hydroponic conditions and extracted total RNAs from a controlled mixture of leaf and root tissue (with three biological replicates per species; *i.e.* a total of 12 samples). We prepared RPF libraries and obtained a total of 90 to 140 million raw sequencing reads per replicate (**Table S1**). Reads were filtered based on their length (between 25 and 35 nucleotides) and quality (minimum quality score of 20), followed by a bioinformatic step to remove sequences with

similarity to rRNAs. Substantial rRNA contamination remained in our libraries (50 to 70% of reads after quality control, **Table S1**), despite the ribo-depletion included in the libraries preparation protocol. The mapping rates of sequencing reads to their respective reference genomes varied across species (with lower mapping success in *A. halleri*, possibly due to contamination of the hydroponic culture medium by *Acidovorax* bacteria, **Table S1**). In the end, we could use between 0.8 and 16 million mapped reads per replicate for the detection of translation signals.

We first evaluated the quality of RPF libraries by examining metagene plots produced by RibORF (Ji 2018). A high quality RPF dataset would exhibit a noticeable nucleotide triplet periodicity, a stark increase in coverage at the start codon position, and a pronounced decrease at the stop codon position, reflecting the dynamics of ribosome occupancy along transcripts. Our libraries displayed strong heterogeneity in terms of coverage depth, as well as overall quality (**Figure S1**). We chose to exclude our *A. thaliana* and *C. rubella* libraries from further downstream analyses due to their suboptimal overall quality. For *A. lyrata* and *A. halleri*, we chose to select one replicate library with the highest quality (AHA_rep-2 and ALY_rep-1) for deeper sequencing. Further sequencing yielded between 18 and 30 million mapped reads available for the detection of translation signals. For *A. thaliana*, we used the high-quality RPF datasets from (Hsu et al. 2016; Wu and Hsu 2022), comprising a total of 40 to 110 million mapped reads (**Table S2**). Ribosomal footprints are expected to be around 28 nucleotides long (*i.e.* the length of the mRNA segment protected by the ribosome (Ingolia et al. 2009)), and accordingly we used 27 to 32 nucleotide-long reads for the detection of translation signals (**Figure S2**).

The final metagene plots for the samples selected in *A. halleri*, *A. lyrata* and *A. thaliana* confirmed the expected periodicity signal, together with a sharp increase in coverage at the start position and a corresponding sharp decrease at the stop position for each species (**Figure 1**), indicating that these RPF datasets can be used to obtain reliable evidence for translational activity. The signal of periodicity was apparent for 27 to 32 k-mers in the *A. thaliana* dataset, but somewhat less apparent in *A. lyrata* and *A. halleri*, where the pattern was mostly restricted to 28-kmers (the expected size of ribosomal footprints). Overall, based on the aforementioned characteristics, we conclude that the highest quality dataset is that of *A. thaliana*, followed by *A. lyrata* and *A. halleri*.

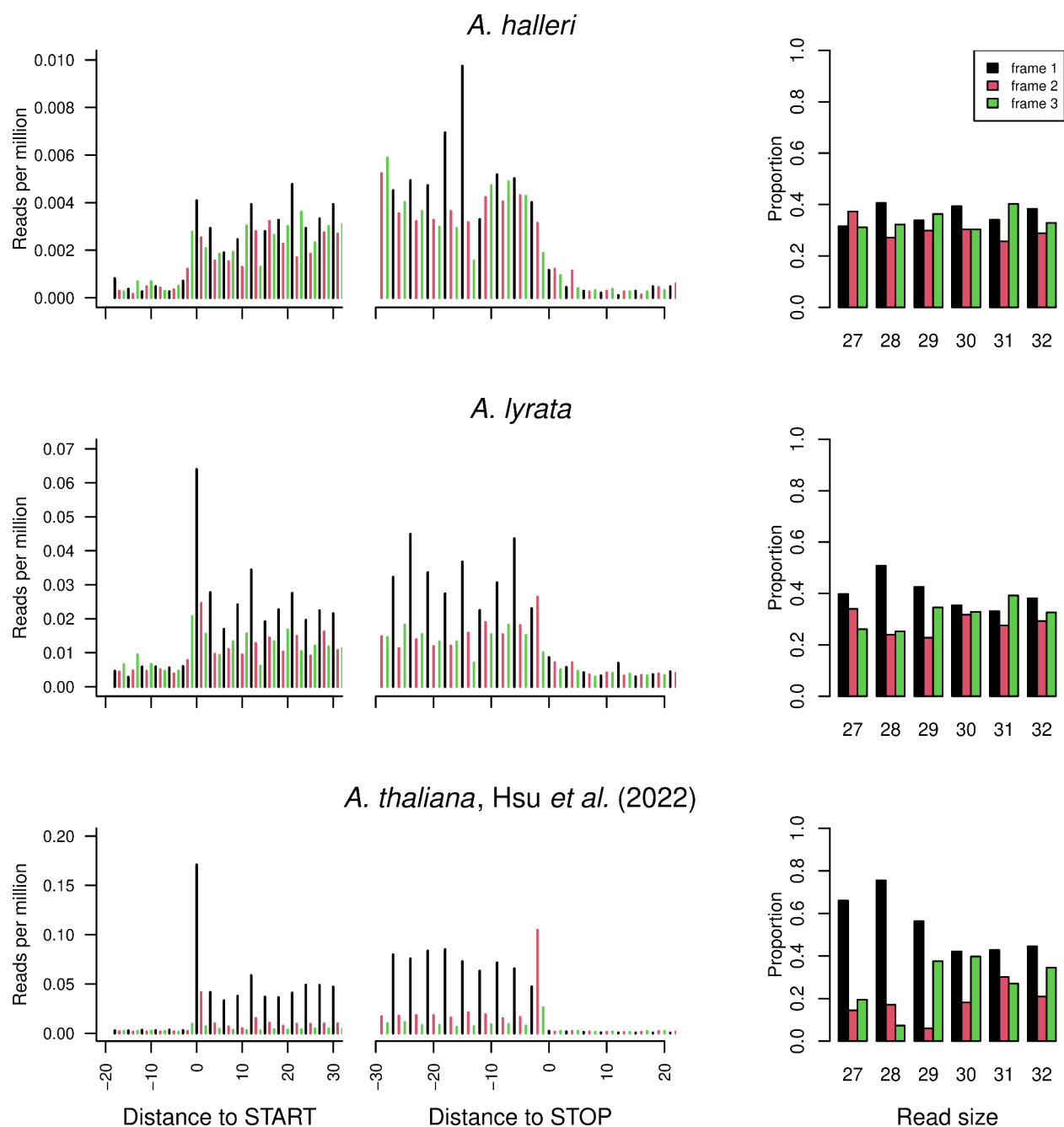


Figure 1: Metagenic plots of 28-kmers from RPF, after P-site shifting (reads were shifted according to their k-mer specific P-site offset, and trimmed to be 1 base-long), used for the annotation of translation events; and prevalence of k-mers depending on which frame they are found in, produced with modified RibORF scripts (Ji 2018). Metagenic plots are created by concatenating the RPF read accumulation within all known canonical protein-coding genes, measured in reads per million (RPM) relative to the start position of CDS. Data from *A. halleri* and *A. lyrata* come from the AHA_rep-2 and ALY_rep-1 samples of this study, and data from *A. thaliana* come from (Wu and Hsu 2022). Read accumulation is colored based on its frame relative to the CDS: black for the first frame, red for the second frame and green for the third frame. The relative abundance of each k-mer, between 27 and 32 nucleotides in size, is also colored based on the frame in which the reads are located.

Identifying individual ORFs with translation signals

The identification of translation signals at individual ORFs remains a challenging task even with high-quality RPF data, especially in the context of the discovery of recently emerged *de novo* genes because of their typically low level of expression (Carvunis et al. 2012; Zhao et al. 2014). For this reason, we annotated ORFs with translation signatures using three independent approaches specifically developed for the analysis of RPF data. First, we used the least stringent “ORFscore” (Bazzini et al. 2014), which is based on the preferential accumulation of reads in the first reading frame, taking the bias in canonical genes as a reference to distinguish the true signal from noise, where reads are expected to accumulate across all possible reading frames equally. A second, more stringent method, RibORF (Ji 2018), uses “Support Vector Machine”, a type of Machine Learning trained on annotated canonical genes to identify and recognize triplet periodicity. Finally, the third and most stringent method, RiboCode (Xiao et al. 2018), uses a modified Wilcoxon signed-rank test to annotate triplet periodicity *de novo*. To determine the accuracy of these annotations, we first examined CDS that were annotated as protein-coding genes. In all species, the majority of annotated CDS were found to be translated (around 21,000 to 24,000 across species) by at least one of the three tools (ORFscore, RibORF or RiboCode). Among them, 20,592 and 13,266 CDS were found to be translated by the three tools, respectively in *A. thaliana* and *A. lyrata*. In *A. halleri*, the difference is more pronounced between the tools and we detected 3,467 translated CDS by the three tools and 11,996 translated CDS by two tools (**Table S3**). This observation is probably due to the less apparent periodicity of the RPF dataset in *A. halleri* (**Figure 1**). Hence, the samples from the different species differ in the proportion of CDS detected as translated and the number of tools detecting CDS as being translated. It should also be noted that, as expected, not all canonical CDS are identified as being translated by this single experiment, due to limitations of our data both in technical (e.g. low power inherent to RPF data, where translation is hard to assess on a per-locus basis) and biological terms (e.g. only two tissues sampled, developmental stage-specific expression). Overall, we conclude that the signals we detect are powerful but correspond to an under-estimation of the actual translation potential that may exist.

Next, we took advantage of the fact that these analyses were performed across the entire genomes (*i.e.* irrespective of the annotation of coding or non-coding sequences) to distinguish two categories of translation signals. On one hand, those overlapping canonical CDS, taken directly from available gene annotations. On the other hand, intergenic ORFs (IGORFs), *i.e.*

ORFs with no overlap with gene annotations. In total, we identified translation signals for between 21,000 and 24,000 canonical CDS and between 3 and 23 million IGORFs depending on the species considered (**Table S3**).

Compared to CDS, we expect a smaller proportion of IGORFs to display translation signals, since they belong to supposedly non functional, intergenic regions. Indeed, among the millions of IGORFs annotated on the basis of their nucleotide sequence across genomes, only a small proportion shows signs of active translation. Yet, this small fraction represents up to 1,396 and 87,433 IGORFs (depending on the species), with detectable translation signatures by at least one tool (**Table S3**). The CDS and IGORFs that have the most robust translation signatures, i.e. those identified by all three tools, are translated at a relatively higher level compared to other features, illustrating the importance of sequencing depth in annotating translation events (**Figure 2**). Across species, 1,825 IGORFs were detected to have such robust translation signatures (135 in *halleri*, 1,548 in *lyrata* and 142 in *thaliana*; **Figure 2** and **Table S3**). Furthermore, at the same level of expression, CDS are more often detected as translated than IGORFs, reflecting the training of tools on canonical gene annotations, requiring lower depth to identify periodicity within canonically translated regions. Among all tools, RiboCode comes out as the most stringent, with many CDS and IGORFs found to be translated by both ORFscore and RibORF. For downstream analyses, in an effort to be as exhaustive as possible, we focused on CDS and IGORFs that were detected as translated by at least one tool.

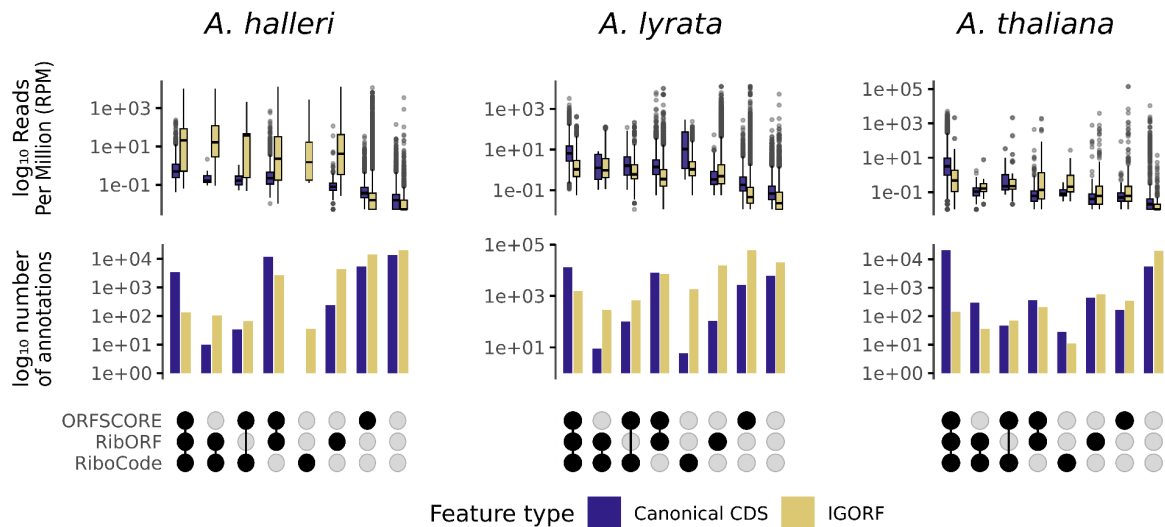


Figure 2: Complex upset plot representing the $\log_{10}$ number of annotations detected as translated and their translation levels in $\log_{10}$ Reads Per Million (RPM, where each read is a ribosome footprint); according to the tools that detected them as translated, and to the category to which they belong: canonical CDS or IGORFs. Black circles represent the detection of translation, grey circles represent its absence. Data from *A. halleri* and *A. lyrata* come from the AHA_rep-2 and ALY_rep-1 samples of this study, and data from *A. thaliana* come from 2 studies: (Hsu et al. 2016; Wu and Hsu 2022).

Synteny analysis identifies many *de novo* ORFs being translated

We used the repertoires of translated CDS and IGORFs obtained across the genomes of the three *Arabidopsis* species to identify which of those, referred to as translated ORFs, might have appeared *de novo* within the *Arabidopsis* genus. To do so, we used conserved CDS as orthology anchors to finely define syntenic regions between pairs of species. For non-conserved translated CDS and IGORFs found within those chromosomal intervals in one species, we used global sequence alignment to search for micro-syntenic regions in the orthologous non-genic sequence of the other species. ORFs located in chromosomal regions that are too divergent to allow sequence alignment were removed (**Table S4, Figure S3**). This yielded between ~2,000 to 8,000 aligned potential *de novo* ORFs. Blast searches were also performed against the protein sequences from other Brassicaceae species, as well as the whole genomes of each focal species to remove ORFs with homologs in other Brassicaceae species, or homology outside of their annotated syntenic region. ORFs with translation signal were then assessed to have emerged *de novo* or not on the basis of their sequence conservation within the alignments, regardless of their translation status in the other species. Specifically, we annotated as having emerged *de novo* every ORF that was absent from the syntenic region in the outgroup (either *C. rubella* or *A. thaliana*), either 1) because they are fully absent, 2) because they contained a frameshift mutation or 3) lacked a START and a STOP codon at the same positions. With these analyses, our goal was to identify ORFs that have emerged *de novo* within the *Arabidopsis* genus, i.e. after the divergence from *C. rubella* or *A. thaliana*, which we used as outgroups.

A total of 228, 779 and 28 *de novo* ORFs with translational activity were identified in *A. halleri*, *A. lyrata* and *A. thaliana* respectively (**Table S4**). Of these, 2 to 17% were already annotated as CDS: 5 in *A. halleri*, 18 in *A. lyrata* and 8 in *A. thaliana*, all were species-specific and contained less than four exons, with 80% displaying only one. Having removed redundancy between species, we ended up with 738 *de novo* emerged ORFs (**Figure 3**). More than half of these *de novo* ORFs are species-specific (181 in *A. halleri*, 582 in *A. lyrata* and 24 in *A. thaliana*, **Table S5**), while 27% are found in at least two and 6% in three species. Despite a divergence time of

nearly ten million years (Roux et al. 2011), most ORFs had a recognizable homologous non-coding sequence in *C. rubella*, and those that did not almost always had one in *A. thaliana*. Most of these *de novo* ORFs were found in *A. lyrata*, closely followed by *A. halleri*, which can be attributed to the greater genome size of those species (i.e. much longer intergenic regions) compared to *A. thaliana*, as well as the better periodicity of the *A. lyrata* RPF dataset compared to the *A. halleri* one. The lesser amount of *de novo* ORFs in *A. thaliana* might be attributed to the thorough annotation of its genome, however in this case, if the rate of *de novo* gene birth was the same in all three species, we would have found many *de novo* ORFs already annotated as CDS. This is not the case here: CDS and IGORFs together, we found less *de novo* ORFs in *A. thaliana* than in the two other *Arabidopsis* species. The ORFs detected in multiple species, as opposed to just one, tend to be identified as translated by more tools (**Figure 3**), which could be interpreted as evidence for a more robust translation signal, i.e a greater functional potential of conserved ORFs. Still, 110 species-specific ORFs were identified as translated by several tools, and hence likely display translational potential.

De novo ORFs of varying conservation levels display a continuum of features

We then compared features of these *de novo* translated ORFs to those of “orthologous CDS” (canonical CDS with an identifiable ortholog in every other species of interest, used as synteny anchors) on one hand, and to “untranslated intergenic ORFs (unt_IGORFs) (with no coverage by ribosome footprints) on the other hand. The nature of our data, characterized by highly divergent, non-coding genomic regions between only a handful of species, rendered ancestral sequence reconstruction (ASR) unreliable. Hence we were unable to determine at which specific point of the phylogeny the ORFs emerged. As a result, we refer to conservation classes rather than to actual evolutionary age. Specifically, *de novo* translated ORFs were classified into different conservation categories: A) conserved in all *Arabidopsis* species, B) conserved in *A. halleri* and *A. lyrata*, C) *A. lyrata*-specific and D) *A. halleri*-specific (**Figure 4**). It should be noted that some ORFs were specific to *A. thaliana*, in which case our inability to conclude as to their true origin led us to exclude them from further analyses: they could have

either emerged before the divergence of *A. thaliana*, and has subsequently been lost in *A. halleri* and *A. lyrata*, or have emerged after the divergence of *A. thaliana*.



Figure 3: Complex heatmap of the conservation of *de novo* ORF sequence and translation. Each line represents an individual ORF, and each column a species. Redundancy within each species was removed by keeping only the longest of overlapping *de novo* ORFs, and between species by removing duplicates of a same ORF found to be translated in several species. An ORF sequence was considered as conserved when its START and STOP positions, as well as its reading frame were conserved. If the ORF is conserved in a species, it is annotated as “present”, if not it is annotated as “absent”. If no homologous syntenic region could be aligned, the ORF was annotated as “NA”. As translation detection tools rarely annotate the exact same ORF as being translated, a given tool was considered as having detected translation for an ORF either if the exact ORF was detected as translated, or if an overlapping ORF was detected as translated.

Firstly, we examined the usage of the three alternative START codons of the genetic code. We found a prevalence of non-canonical START codons, specifically TTG, in the *de novo* ORFs that closely resembled that of unt_IGORFs, with TTG being even more commonly employed than ATG in *A. halleri* (**Figure 4**). In contrast, as gene prediction algorithms only consider canonical start codons, orthologous CDS exclusively harbor ATGs. Secondly, we observed that as expected orthologous CDS were on average much longer (mean length = 1,111nt) than unt_IGORFs (mean length= 52 nt) (**Figure 4**). Statistically significant differences in length were apparent between *de novo* conservation classes, most strikingly in *A. lyrata*, with a shorter length of the more conserved ORFs. This is most likely a consequence of the way conservation is stringently defined, as longer ORFs are more likely to be lost by mutation compared to shorter ones. Third, we observed that the more conserved *de novo* ORFs tend to have a higher GC content (median ~36%), similar to those of CDS (median ~44%), but above that of the unt_IGORFs, which had the lowest GC content (median ~30%) (**Figure 4**). This pattern followed a continuum along conservation classes. Finally, in line with previous observations (Carvunis et al. 2012; Zhao et al. 2014), we observed that translation activity (quantified by coverage by aligned RPF reads) of *de novo* ORFs was less intense than that of orthologous CDS (on average between 0.035 RPM versus 0.6 RPM in *A. lyrata*). Still the two distributions overlap largely, and a substantial fraction of *de novo* ORFs exhibits a level of translation activity that is comparable to that of orthologous CDS, indicative of ongoing translation. We verified that this difference in coverage was not caused by the difference in the length of the annotations; by testing the influence of length on expression level. Using a linear least squares regression, we found that length alone was not sufficient to explain differences in expression levels ($estimate=-0.002224$, $p.value= 0.9272$). Furthermore, the differences in coverage were also apparent when length differences were controlled for by calculating coverage on the first 60 nucleotides of each feature only, where the translation signal was most pronounced. Again, length alone was not sufficient to explain differences in expression levels ($estimate=-0.01254$, $p.value=0.893623$). All the aforementioned patterns were tested in all three species, for ORFs that were detected as translated by either one, two or three tools, and were found to be robust.

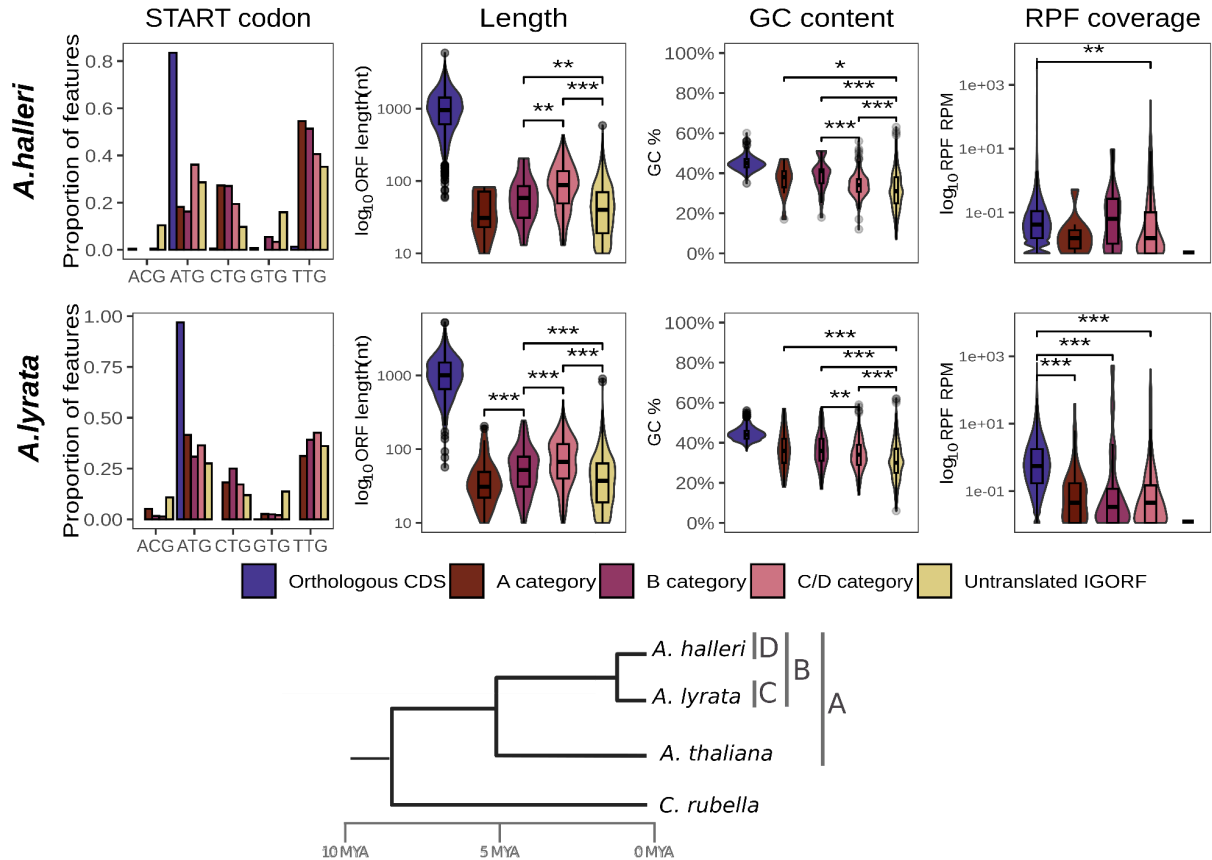


Figure 4: Characteristics of conserved canonical CDS, *de novo* ORFs and untranslated IGORFs. *De novo* ORFs were classified according to their conservation category: A, B, C or D; the meaning of which is showcased in the phylogeny. For length and GC content, all statistically significant differences were represented, except for the ones between orthologous CDS and all other categories (Mann-Whitney U-test, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$). For RPF coverage, all statistically significant differences were represented.

De novo ORF - encoded peptides are highly disordered

We then examined the level of Intrinsic Structural Disorder (ISD) of polypeptides encoded by *de novo* ORFs. ISD was estimated using IUPred (Dosztányi 2018), on all *de novo* ORFs, as well on a sub-sample of 1,000 orthologous CDS and 1,000 unt_IGORFs. Statistically significant differences were observed between unt_IGORFs and orthologous CDS, as well as between *de novo* ORFs and orthologous CDS, with both unt_IGORFs and *de novo* ORFs exhibiting the highest levels of ISD. *De novo* ORFs were not different from unt_IGORFs. Further analyses however demonstrate a statistically significant negative correlation between the

polypeptide length and ISD, (**Figure S4**), which yet only explains a small part of the observed variance (*adjusted* $R^2 = 0.03532$, *p-value* < $2.2e-16$). On the contrary, GC content seems to be positively associated with ID, at least in some feature categories (in conserved CDS: *adjusted* $R^2=0.03359$, *p-value* = $5.374e-09$). When sub-sampling CDS of similar size and GC content as *de novo* ORFs, we no longer observe a statistically significant difference in ISD between categories. Hence, while we do observe that on average *de novo* ORFs tend to be more highly disordered than CDS; it is difficult to further interpret the meaning of this difference, given that both length and GC content strongly influence the predicted ID, but in opposite directions.

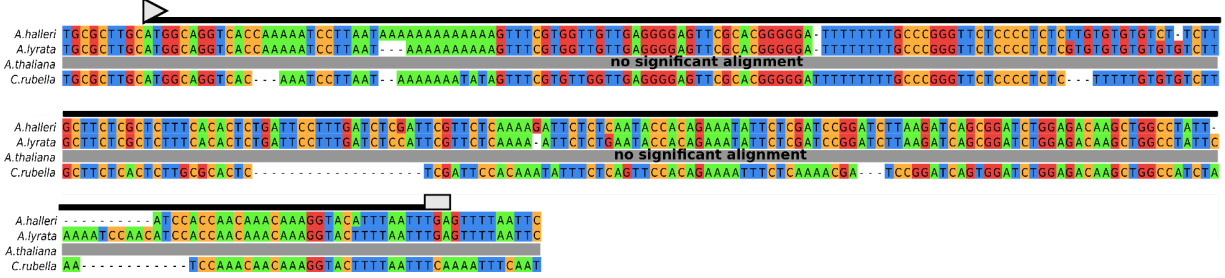
A select few *de novo* ORFs have the potential to be proto-genes

Out of hundreds of translated *de novo* ORFs, a few notable ones stand out as having a number of gene-like characteristics, making them ideal candidates for *de novo* gene emergence. A first set of these translated *de novo* ORFs had already been annotated as CDS, and we therefore refer to them as “*de novo* CDS”. For example, the *A. lyrata*-specific CDS annotated as AL7G14370 (**Figure 5**) has no homology to previously described proteins, displays translation signals by all three tools, and is 285 nucleotides in length. The homologous sequence was present in *C. rubella*, where it was clearly non-coding (no conservation of the STOP codon and of the frame), indicating that this *de novo* ORF emerged within the *Arabidopsis* genus. It was entirely missing from *A. thaliana*, but the START and STOP codons were conserved in *A. halleri*, though with a few indels leading to loss of the reading frame. Hence, this particular *de novo* ORFs could have emerged either specifically in *A. lyrata* or in a more ancestral *Arabidopsis* species and have been lost secondarily in *A. thaliana* and *A. halleri* by indels.

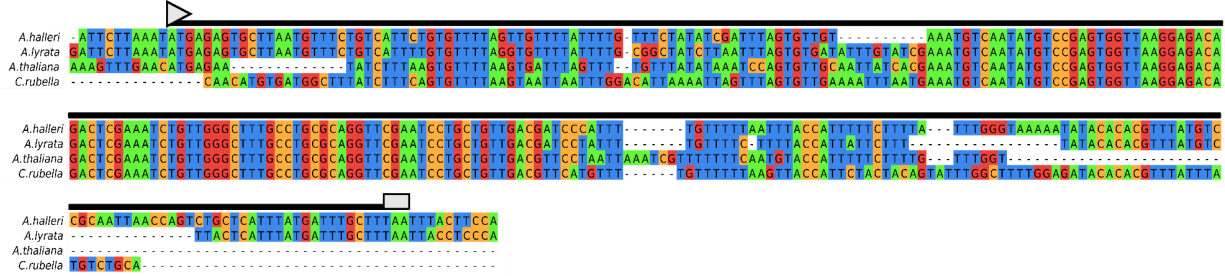
Not all of the *de novo* ORFs we identified, however, had previously been annotated as CDS. We classified them as “intergenic ORFs”, although some of them have features similar to those of the *de novo* CDS. For example, the *A. halleri*-specific intergenic ORF named IGORF_162 (**Figure 5**) features both a canonical start codon (ATG) and a detection of translation by two separate tools; both features similar to those of CDS from *A. halleri*. This *de novo* ORF is 264 nucleotides long (hence shorter than the commonly used 300 nucleotides threshold for gene annotation; Dinger et al. 2008). The orthologous sequence exhibits indels in every other species we studied, leading to a frameshift and the loss of either the START or the STOP codon. The

sequence exhibits no homology to any known protein-coding gene. The IGORF was found to be located directly downstream of a canonical protein-coding gene, a feature suggestive of potential transcriptional and translational readthrough, possibly enabling its sustained expression. Another newly annotated intergenic ORF named IGORF_262 (**Figure 5**) was found to be *A. lyrata*-specific. Despite its shorter length of 111 nucleotides, it displays a canonical ATG start codon and was detected as translated by three tools, making it a promising gene candidate. In *A. thaliana*, indels led to the loss of the frame and a point mutation led to the loss of the STOP codon. In *A. halleri* and *C. rubella*, no homologous sequence could be aligned. In conclusion, while most translated *de novo* ORFs possess features similar to those of unt_IGORFs, a few ones stand out as the best candidates for *de novo* gene emergence. They can be described as intermediates between unt_IGORFs and conserved CDS: they are longer than unt_IGORFs but shorter than conserved CDS, have not been annotated as CDS (or if they have, have not previously been functionally characterized), yet harbor canonical start codons and are independently detected as translated by several tools.

A. lyrata - gene **AL7G14370** (285nt)



A. halleri - **IGORF_162** (264nt)



A. lyrata - **IGORF_262** (111nt)

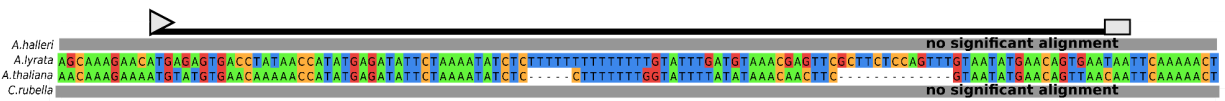


Figure 5: Alignments of potential *de novo* gene candidates, both *A. halleri* or *A. lyrata* specific. Alignments are colored by nucleotides. The START codon is represented by a grey arrow, and the STOP codon by a grey box. The ORF itself is represented by a black line above the alignments. When no homologous syntenic region could be found for a given species, the sequence was replaced by a "no significant alignment" box.

Discussion

Pervasive translation with a rapid evolutionary turnover

Our study revealed that translation of IGORFs is prevalent in *Arabidopsis* species, even when restricting the analysis to features with the highest evidence of translation (*i.e.* detection by 3 tools). Although these numbers vary greatly among our three study species, possibly due to differences in the quality of the RPF datasets, they are broadly consistent with observations in the yeast *Saccharomyces cerevisiae*, where 1,266 IGORFs were found to be either occasionally or highly translated (Papadopoulos et al. 2021). While we already detected a large number of *de novo* ORFs, this repertoire likely represents an underestimation of the true rate of *de novo* gene emergence due to several methodological limitations. Ribosome profiling may introduce biases, both during laboratory experiments and computational analyses (Bartholomäus et al. 2016), rendering non-canonical translation events particularly difficult to detect (Lei et al. 2022). To palliate this bias, a possible strategy is to combine datasets so as to maximize their chances of detecting translation events (Wacholder et al. 2023), as sequencing depths are often insufficient for an exhaustive discovery of translated ORFs (Lei et al. 2022). Since no other public datasets were available for *A. lyrata* and *A. halleri*, we instead decided to maximize sequencing depths, as recommended by (Lei et al. 2022), reaching up to 639 million reads (as opposed to the typical range of 150 to 200 million reads in available *A. thaliana* RPF datasets (Hsu et al. 2016; Wu and Hsu 2022)). Furthermore, multiple analysis tools, each based on different algorithmic principles, were used in parallel to annotate translated features. While using different tools and maximizing tissue diversity greatly improved detection sensitivity, many more *de novo* ORFs may exhibit translation signals at other developmental stages, in different environments, or in different individuals. Indeed, a recent study in *A. thaliana* found that, collectively, *de novo* genes are expressed in a wide variety of tissues, including flowers, siliques and seeds (Jin et al. 2025).

Such transitory pervasive transcription (Neme and Tautz 2016; Schmitz et al. 2018) and translation (Ingolia et al. 2014; Ruiz-Orera et al. 2018; Blevins et al. 2021; Wacholder et al. 2023; Papadopoulos et al. 2024) has been described as the “transient translome” (Wacholder et al. 2023). The overall prevalence of species-specific translated ORFs (181 in *A. halleri*, 582 in *A. lyrata* and 24 in *A. thaliana*) further supports the idea of a high turnover of transcripts and of their translation, which would become temporary substrates for *de novo* gene emergence

(Durand et al. 2019; Grandchamp et al. 2024). Sequence conservation itself is also rather low, as most *de novo* ORFs are species-specific. Hence, while many ORF candidates may appear by mutation within intergenic regions, few will be transcribed and translated, even fewer will be conserved across species, and only a select few might develop the functional potential to ultimately give rise to genes (Neme and Tautz 2014; Neme and Tautz 2016). While we lack the necessary power to assess the *de novo* gene emergence rate in *Arabidopsis*, in the literature, a study in plants (Zhang et al. 2019) has reported a strikingly high rate of *de novo* gene emergence in rice, with 51.5 *de novo* genes per Myr. Not only was the *de novo* emergence rate high, but so was the retention rate; both features which were linked to high protein diversity and phenotypic plasticity.

The first steps of *de novo* gene emergence

Pioneering studies investigating *de novo* gene emergence have mostly focused on already-characterized, functionally potent genes (Cai et al. 2008; Knowles and McLysaght 2009; Xiao et al. 2009). Following (Carvunis et al. 2012; Durand et al. 2019; Ruiz-Orera and Albà 2019; Papadopoulos et al. 2021), we instead chose to focus on recently emerged translated ORFs, in order to characterize more specifically the starting material from which very young “proto-genes” might be found (Carvunis et al. 2012). Indeed, only a small subset of the *de novo* translated ORFs we identified had already been annotated as CDS (5 in *A. halleri*, 18 in *A. lyrata* and 8 in *A. thaliana*), and all were species-specific. All contained less than four exons, with 80% displaying only one, and so far none has been specifically characterized at the functional level. The large majority of *de novo* translated ORFs were intergenic, and are therefore expected to resemble intergenic ORFs. Accordingly, we note several lines of similarity between characteristics of *de novo* ORFs and those of untranslated IGORFs. In particular, the prevalence of non-canonical START codons, the short length and low GC content of *de novo* ORFs are either similar to those of untranslated IGORFs, or intermediate between them and those of conserved CDS. The lower expression level of *de novo* ORFs compared to orthologous CDS entails that whatever phenotype may be associated with the *de novo* translation products, would be less visible to natural selection than those encoded by more highly-expressed features. Hence, whatever selective pressure may be acting on *de novo* ORFs would be minimal at best, leaving them to evolve largely neutrally. An interesting perspective to

our work would be to compare the strength of purifying selection acting on intergenic ORFs with or without evidence of translation, which could be done by analyzing nucleotide polymorphism in natural populations. However, the fact that most *de novo* ORFs are short (measuring for the most part between 10 and 100 nucleotides) limits our ability to perform such analyses.

The processes by which random ORFs move away from the random noise and eventually acquire gene-like characteristics such as higher GC content, greater length and higher transcription remain obscure. In rice, the late ORF–early transcription model has been described as the most prevalent mechanism for *de novo* gene emergence (Zhang et al. 2019), according to which non-coding regions would pervasively become transcribed and only later acquire an ORF by mutation, which would then already be visible to selection. Similarly to what has previously been suggested in *S. cerevisiae* (Carvunis et al. 2012; Papadopoulos et al. 2021), *de novo* ORFs could undergo, in plants, a progressive “refinement” towards gene-like features, including changes in START codon usage and expression levels (Li et al. 2016; Zhang et al. 2019). In a positive feedback-loop, increase in expression would render ORFs progressively more visible to selection, which in turn would lead to either their elimination or their functionalization, and a greater level of expression. Under this scenario, new functions would emerge progressively, in an iterative process. Another explanatory scenario has been colloquially named “Born this way”, whereby pre-existing non-canonical ORFs that present the most “gene-like” features are the ones retained by natural selection, while the others quickly fade back into expressional noise (Aubel et al. 2024). In our study, while we do observe statistically significant differences in ISD between conserved CDS and *de novo* ORFs at first glance, those differences disappear when length and GC content are matched. Indeed, length has been found as being a major factor in ISD prediction (Aubel et al. 2023) and the observed apparent differences in ISD would be an indirect consequence of the greater length of conserved CDS compared to young *de novo* ORFs. Our results alone do not allow us to rule in favor of one scenario or the other, but characterizing the properties of the same ORFs on a finer evolutionary timescale, within populations, would allow us to gain further insight into the scenario under which gene-like characteristics are acquired.

From random intergenic ORF to canonical gene

In many species, including *A. thaliana*, the GC content within protein coding genes is lower in introns than in exons, and more globally in sequences that are not translated compared

to those that are (Carels and Bernardi 2000; Singh et al. 2016). When comparing orthologous CDS (exons) to untranslated IGORFs, we observe a corresponding pattern in which coding sequences tend to have a higher GC content than non-coding ones. This difference will influence the frequencies of START and STOP codons along the genomes. In GC-rich regions of the genome, as AT content is lower, the prevalence of START (ATG,TTG,...) and STOP (TAA,TAG,TGA) codons is correspondingly lower, meaning putative reading frames would be longer (McLysaght and Hurst 2016). Inversely, within GC-poor regions, like intergenic regions in *Arabidopsis* species, putative reading frames would be shorter, and more numerous. This corresponds to what we observe both in the number and length of IGORFs compared to canonical CDS. This makes the intergenic regions of *Arabidopsis* species the ideal reservoir for novel ORFs. However, it remains unclear whether this trend can be generalized as a global mechanism for *de novo* gene emergence. Indeed, while similar patterns have been observed in terms of GC content in *A. thaliana* and *Drosophila*, (Li et al. 2016; Peng and Zhao 2024), the opposite trend has been observed in *de novo* genes in yeast, which tend to emerge in GC-rich regions, close to promoters or within recombination hot spots (Vakirlis et al. 2018). Strong correlations have been identified between GC content and other key genomic features, including length and intrinsic disorder (Basile et al. 2017; Nielly-Thibault and Landry 2019). Yet, in our data, the relationship between GC content and ISD remains hard to assess, as there appears to be a significant positive relationship between the two characteristics within some ORF categories, and no association in others.

In line with previous observations (Basile et al. 2017; Willis and Masel 2018; Aubel et al. 2023), we found that *de novo* ORFs tend to harbor high levels of intrinsic disorder. This can at least in part be attributed to their shorter length, directly preventing the formation of secondary structures. This, however, does not mean that they are devoid of functional potential. An extensive body of literature now highlights the functional potential of intrinsically disordered regions (IDRs), which through their highly labile and context-dependent conformation, can be crucial in molecular communication and environment-induced phenotypic responses (Holehouse and Kragelund 2023). The conservation of the function of IDRs is even possible in spite of a lack of sequence conservation, through the conservation of their structure (Holehouse and Kragelund 2023). Within eukaryotic proteins at large, IDRs confer conformational flexibility to canonical proteins, especially those involved in transcriptional and post-translational regulation (Ahrens et al. 2017). Furthermore, small ORFs at large have been hypothesized to contribute greatly to phenotypic variability through mechanisms such as regulation of protein levels,

function, recruitment (Orr et al. 2021), signaling or formation of protein complexes (Andrews and Rothnagel 2014; Brar and Weissman 2015; Schlesinger and Elsässer 2022).

Proximity to genes is one of the factors invoked in the literature as a mechanism by which *de novo* ORFs could benefit from increased transcription and translation, rendering them visible to natural selection (Andreatta et al. 2015; Kleppe and Bornberg-Bauer 2018). This mechanism has been referred to as “transcriptional/translational readthrough”, and involves the emergence of newly coding, highly disordered sequences at the C-termini of highly expressed proteins, which provide conformational flexibility without disrupting protein function (Kleppe and Bornberg-Bauer 2018). This possibility is exemplified by the *A. halleri*-specific IGORF_162 presented above, as this IGORF is located directly downstream of a canonical, protein-coding gene. *De novo* genes have been consistently described as being shorter than other genes, and the “grow slow and moult” hypothesis suggests that one of the ways *de novo* genes could progressively become longer as they age is through the acquisition of new domains at their C-termini (Bornberg-Bauer et al. 2015; Kleppe and Bornberg-Bauer 2018).

Overall, our findings suggest that pervasive translation takes place within intergenic regions in *Arabidopsis* species, providing a vast reservoir of transiently, lowly expressed short ORFs. Among those, a select few may ultimately give rise to functional genes, by progressively becoming longer, more highly expressed, and more GC-rich. Further production of ribosome profiling and/or mass spectrometry data, in other tissue types and developmental stages, could further extend and validate this repertoire, while providing insight into the tissue-specificity of *de novo* genes. Further biochemical and functional validation would now be necessary to evaluate their physiological relevance, in particular to determine how they integrate into and eventually disturb cellular interaction networks.

Material and methods

Plant Materials and Growth Conditions

Three seeds of either *Arabidopsis* Columbia-0; *Arabidopsis halleri*; *Arabidopsis lyrata* or *Capsella rubella* were deposited on 0.7% agar and imbibed in germination medium; as described in (Conn et al. 2013); then left in obscurity at 4 °C for 2 days. The microcentrifuge lids were then transferred to hydroponic growth tanks in a greenhouse, where the seeds germinated and seedlings were grown hydroponically in a growth medium composed of 1 mM Ca(NO₃)₂, 0.5 mM MgSO₄, 3 mM KNO₃, 0.5 mM NH₄H₂PO₄, 0.1 µM CuSO₄, 0.1 µM NaCl, 1 µM KCl, 2µM MnSO₄, 25 µM H₃BO₃, 0.1 µM (NH₄)₆Mo₇O₂₄, 20 µM FeEDDHA, and 1 µM (A. lyrata, A. thaliana and C. rubella) or 10 µM (A. halleri) ZnSO₄. The pH of the solution was maintained at 5.0 using MES acid buffer (2 mM). Floral buds, leaves and roots were collected and flash frozen in liquid nitrogen progressively along the growth of the plants; over the course of 3 months (July to September). In total; 10 A. thaliana, 8 A. halleri, 8 A. lyrata and 6 C. rubella individuals were sampled from. Flash frozen samples were stored by tissue type; individual and sampling date, at -80°C.

Lysates preparation

Three individuals from each species were chosen for subsequent data production; based on the amount of material that had been gathered. Roots and leaves from the same single individual (at the exception of A. thaliana, where 9 individuals were merged into 3 groups of 3) were gathered together regardless of their sampling date. Tissue types were separately ground into a powder, using liquid nitrogen, a mortar and pestle. The obtained powder amounted to an approximate total of 1g of tissue, per individual. Approximately 0.25g of each tissue type was kept for ribosome footprint production, per individual. The powders were then suspended in 4mL of lysis buffer, as described in (Wu and Hsu 2022). The obtained lysates were vortexed twice; distributed in 2mL tubes; and then agitated at 4°C for 10mn. All lysates were then spun at 5 000 × g at 4°C for 3 mn; and the supernatant was aliquoted into refrigerated 1.5mL tubes. The aliquots were spun at 20 000 × g at 4°C for 10 mn; and the supernatants were then transferred to refrigerated 1.5mL tubes. Aliquots of 200µL were produced and stored at -80°C; and a single

10µL aliquot was kept for the assessment of RNA concentrations by a Qubit RNA HS assay (Thermo Fisher Scientific). As it was apparent that root aliquots seemed to contain on average much less material than leaf ones; root and leaf aliquots were pooled per individual; so as to obtain 75% of material coming from leaves, and 25% of material coming from roots.

Ribosome footprints

The lysates were treated with 50 units RNase I Ambion (Invitrogen, AM2295) per 40µg of RNA; and digested for 1 hour on a nutator at 25°C. The samples were then moved to ice; and 30 µL of SUPERase-IN (Fisher scientific, 10773267) were added to interrupt digestion. A sucrose cushion solution was prepared as described in (Baudin-Baillieu et al. 2016); and digested lysates were subsequently deposited on the cushions and centrifuged at 4 °C, 657,000×g in a TLa110 rotor for 2 h 15 min. Pellets were then washed twice with 500 µl lysis buffer and resuspended in 750 µl lysis buffer; after which they were solubilized by pipetting up and down and transferred in a 2 ml microtube. An RNA extraction test was first performed using the Zymo RNA clean and concentrator kit (Zymo Research; R1017) on aliquots of all three replicates of *A. lyrata*, and on one aliquot of the second replicate of *A. halleri*. Standard protocol procedure was followed in order to extract RNAs between 17 and 200 nucleotides long. However, after subsequent quantification of those RNA extracts, it became apparent that they were too lowly concentrated in order to undergo gel migration (from less than 1 to 7 ng/µL; when we expected 10-100 ng/µL); and this extraction protocol was subsequently not used anymore. Instead, for the remaining aliquots, total RNA was extracted using a standard hot phenol procedure as described in (Baudin-Baillieu et al. 2016). Then, the totality of extracted RNAs from each aliquot were charged on 17% polyacrylamide; 7M TAE-urea gels, which were pre-migrated for 30 min at 100V. The TAE 1x migration of samples was performed at 100 V for approximately 5 h; with size markers at 28 and 30 nucleotides. After migration, gels were stained with SYBR-GOLD and gel slices between the 28 and 30 nt markers were excised. The excised gel band was pulverized and eluted in 500 µL of extraction buffer (0,3 M CH₃COONa pH 5,5 ; 1 mM EDTA); and then spun on a wheel at 10 turns/minute at +4°C for a whole night. The gel residues were retained on a Spin-X cellulose acetate filter 0,45µm Corning column (Dutscher; 2515806); and centrifuged at maximal speed for 5mn. Per tube, 2µL of glycogen (ThermoFisher; R0551) were added with 2 to 2,5 volumes of 100% EtOH; and were then

precipitated overnight at -20°C. Tubes were then centrifuged at +4°C at 21,000xg, and the pellets were dissolved in 15 µL of distilled water. The final microRNA concentration was measured by fluorescence with the micro RNA Quant-iT kit (Thermofisher Scientific).

Ribosomal RNAs depletion

Samples were depleted in ribosomal RNAs according to the depletion protocol of the TruSeq® Stranded Total RNA Library Prep Plant kit (Illumina, 20020610). The obtained footprints were subsequently purified according to the following Zymo RNA clean and concentrator kit (Zymo Research; R1017) modified protocol. An RNA Binding Buffer is adjusted with an equal, 100% ethanol volume; and 2 volumes (180 µL) of this adjusted buffer are added to the rRNA-depleted RPF samples (90 µL). Then, is added to this mix 1 volume of 100% ethanol (270 µL) to favor the retrieval of small RNAs between 17 and 200 nucleotides (ribosome footprints are typically approximately 30 nucleotides long). The total 540 µL are then deposited on a kit's column, and centrifuged for 30s at 16,000xg at room temperature. Then, 400 µL of RNA Prep Buffer are added to the column, which is again centrifuged in identical conditions; followed by the addition of 700 µL of RNA Wash Buffer and a third identical centrifugation. Finally, 400 µL of RNA Wash Buffer are added; and the column is centrifuged for 1mn at 16,000xg at room temperature. The column is put on a 2mL tube and the ribosome footprints are eluted on the column with 15 µL of distilled, nuclease-free water; and then centrifuged for 1mn at 16,000xg at room temperature again. Again, the final microRNA concentration is measured by fluorescence with the micro-RNA Quant-iT kit (Thermofisher Scientific; Q32882).

Library preparation and sequencing

The banks were prepared according to the Biolabs NEBNext Multiplex Small RNA library Prep kit's standard protocol; with 12 PCR cycles. The resulting banks were purified with the Monarch PCR & cleanup kit (Biolabs); and concentrations were assessed with the Quant-iT DNA HS kit.

Ribo-seq Data Analysis

Firstly, adapters were trimmed from the reads and then filtered on quality and length (with parameters `--minimum-length 25`, `--maximum-length 35`, `--error-rate 0.1` and `-q 20`) using Cutadapt (v3.2) (Martin 2011). Taken from TAIR10 annotations, rRNA, tRNA, snoRNA, etc. sequences were removed through bowtie2 alignments (v2.4.1) (Langmead and Salzberg 2012). Depleted sequences were then aligned to either the TAIR10 (*A. thaliana*), AUBY1 (*A. halleri*) or MN47 (*A. lyrata*) reference genome, depending on the species of origin of the sample. The alignment was performed using the splice-aware aligner STAR (v2.7.5a) (Dobin et al. 2013), tolerating up to 1 mismatch in the alignment, and up to 20 multi-mapping positions. Raw sequencing data from (Hsu et al. 2016; Wu and Hsu 2022) (GEO accession GSE81332 and BioProject ID PRJNA854638) were downloaded from NCBI; trimmed from their adapters and filtered on quality and length using Cutadapt (same parameters as previously); and further processed identically to the data produced in this study.

The quality of the Ribosome Profiling datasets was assessed by producing protein coding gene meta-plots using RibORF (Ji 2018); allowing for the selection of the samples presenting the best periodicity signals (Replicate 2 of *A. halleri* and Replicate 1 of *A. lyrata*, Figure S1). These metaplots also allowed for the assignment of k-mer specific P-site offsets, showcased by an asterisk for 28-mers in Figure S1. The distribution of read size was also assessed, as showcased in Figure S2. The mapped reads were subsequently shifted by the assigned offsets using RibORF. All potential ORF candidates were bioinformatically annotated with RibORF (Ji 2018), on the basis of their sequence only, from the available genomes. A minimal length of 9 nucleotides was specified, and non-canonical START codons were allowed: TTG, CTG, GTG and ACG. Signatures of translation within those candidates were then detected by 3 tools, starting with ORFSCORE (Bazzini et al. 2014), where ribosome footprints accumulation in each possible frame is computed for different features, with the assumption that coding ones would showcase greater accumulation in the first frame. Annotations were produced using R (v 4.2.2), with packages ORFik (Tjeldnes et al. 2021), rtracklayer (Lawrence et al. 2009), GenomicRanges and GenomicFeatures (Lawrence et al. 2013). Predicted ORFs that fell within the 95% confidence interval of the CDS ORFscore distribution were kept as showcasing translation signal. Secondly, RibORF, which relies on Machine Learning (“Support Vector Machine”), was trained to recognize triplet periodicity on canonical gene annotations specific to each species. Predictions were then performed on candidates with a cut-off that was adjusted to each dataset

to keep a FDR < 0.05. Finally, RiboCode (Xiao et al. 2018) was used, which allows for the de novo annotation of periodicity signals using a modified Wilcoxon signed-rank test. New alignments were performed to the merge of the reference transcriptomes, and artificial intergenic transcripts (as to include intergenic ORFs) using STAR again, tolerating up to 1 mismatch and up to 20 multi-mapping positions. Translation signal was predicted after p-site shifting the aligned reads with the previously assessed p-sites, among newly predicted candidates, with a minimal size of 9 nucleotides and allowing for the same non-canonical START codons as RibORF. The default p-value (0.001) was used to distinguish translated ORFs.

Syntenic region annotation and alignment

Reference proteomes for each species were used as input for Inparanoid (Remm et al. 2001) and pairs of orthologs were identified for each pair of species, while removing paralogs and keeping ortholog pairs with a score ≥ 100 and a bootstrap value ≥ 99 . Syntenic regions were consequently annotated with homemade R scripts, where a syntenic region was defined as a genomic region anchored by two genes which each had an ortholog in the pair species. ORFs that fell within syntenic regions were identified with the bedtools suite “intersect” (Quinlan and Hall 2010). An initial local pairwise alignment was performed between each ORF and its corresponding syntenic region using EMBOSS water (Madeira et al. 2024), and used to identify the coordinates of the ORF-homologous region in the pair species. Coordinates of the ORF and its homologous region were then both extended by 1kB upstream and downstream, and were globally aligned using the ggsearch36 tool from the fasta36 software (v3.8i) (Pearson and Lipman 1988) with an e-value threshold of 0.01. Parallely, the ORFs falling within syntenic regions were used for a BLASTX search (Camacho et al. 2009) against the proteomes of 5 *Brassicaceae* species downloaded from Genbank (Benson et al. 2012): *Camelina sativa*, *Brassica napus*, *Brassica oleraceae*, *Eutrema salsugineum* and *Raphanus sativus*. The ORFs that showcased any significant hit were discarded (default e-value of 0.001). The ORFs that were kept were subjected to a BLASTn search (Camacho et al. 2009) against genomes of every species of interest. The ones that exhibited a significant hit (default e-value) outside of the expected syntenic region, and whose length was $\geq 50\%$, were discarded. Among the remaining sequences, the Percent Identity (ID) between each ORF and all its syntenic aligned regions were computed. Two elements were required to select ORFs with sufficient homology: that the global pairwise alignment was significant with *C. rubella*, or if not with *A. thaliana*, and that the

PID in the ORF region within the pairwise alignment was $\geq 60\%$. ORFs that passed those criteria were used for Multiple Sequence Alignment (MSA); between the extended ORF sequence and every correctly aligned, extended syntenic region using MAFFT (Kato 2002). Once the multiple alignments were obtained, they were used for conservation analyses with the help of home-made python scripts. The de novo status was inferred when a region lacking ORF conservation (START, STOP and reading frame) could be identified in either *C. rubella*, or if impossible in *A. thaliana*.

Intrinsic disorder prediction

AlphaFold2 (v.2.3.0: Database cut-off data: 2022-12-19) (Jumper et al. 2021) predictions were performed on HPC cluster Palma II. Predictions with highest confidence (ranked_0) were used for analysis. Secondary elements were annotated using Stride (Frishman and Argos 1995). Disorder was predicted locally using AIUpred (no smoothing) (Erdős and Dosztányi 2024) and IUpred2A (long & short) (Mészáros et al. 2018). Mean and median disorder per sequence were calculated using custom python scripts. Plots were produced using R (v 4.2.2), with packages ggplot2 (Wickham 2009), ggsignif (Constantin and Patil 2021), ggpubr (Kassambara 2023), patchwork (Pedersen 2024), ComplexUpset (Michał Krassowski et al. 2022) and ComplexHeatmap (Gu et al. 2016), and alignment visualizations with Jalview (Waterhouse et al. 2009).

Acknowledgements

Structure predictions for this publication were performed on the HPC cluster PALMA II of the University of Münster, subsidized by the Deutsche Forschungsgemeinschaft (INST 211/667-1)

Bibliography

- Ahrens JB, Nunez-Castilla J, Siltberg-Liberles J. 2017. Evolution of intrinsic disorder in eukaryotic proteins. *Cell. Mol. Life Sci.* 74:3163–3174.
- Andreatta ME, Levine JA, Foy SG, Guzman LD, Kosinski LJ, Cordes MHJ, Masel J. 2015. The recent de Novo origin of protein C-termini. *Genome Biol. Evol.* 7:1686–1701.
- Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15:193–204.
- Aubel M, Buchel F, Heames B, Jones A, Honc O, Bornberg-Bauer E, Hlouchova K. 2024. High-throughput Selection of Human *de novo* -emerged sORFs with High Folding Potential. Carreño CA, editor. *Genome Biol. Evol.* 16:evae069.
- Aubel M, Eicholt L, Bornberg-Bauer E. 2023. Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. *F1000Research* 12:347.
- Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, Jentoft S. 2018. De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. *Mol. Biol. Evol.* 35:593–606.
- Bartholomäus A, Del Campo C, Ignatova Z. 2016. Mapping the non-standardized biases of ribosome profiling. *Biol. Chem.* 397:23–35.
- Basile W, Sachenkova O, Light S, Elofsson A. 2017. High GC content causes orphan proteins to be intrinsically disordered. Dunbrack RL, editor. *PLOS Comput. Biol.* 13:e1005375.
- Baudin-Baillieu A, Hatin I, Legendre R, Namy O. 2016. Translation Analysis at the Genome Scale by Ribosome Profiling. In: Devaux F, editor. *Yeast Functional Genomics*. Vol. 1361. *Methods in Molecular Biology*. New York, NY: Springer New York. p. 105–124. Available from: https://link.springer.com/10.1007/978-1-4939-3079-1_7
- Bazzini AA, Johnstone TG, Christiano R, MacKowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 33:981–993.
- Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* 172:1675–1681.
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. *Nucleic Acids Res.* 41:D36–D42.
- Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nat. Commun.* 12:1–13.
- Bornberg-Bauer E, Hlouchova K, Lange A. 2021. Structure and function of naturally evolved de

- novo proteins. *Curr. Opin. Struct. Biol.* 68:175–183.
- Bornberg-Bauer E, Schmitz J, Heberlein M. 2015. Emergence of de novo proteins from “dark genomic matter” by “grow slow and moult.” *Biochem. Soc. Trans.* 43:867–873.
- Brar GA, Weissman JS. 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* 16:651–664.
- Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* 179:487–496.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
- Carels N, Bernardi G. 2000. Two Classes of Genes in Plants. *Genetics* 154:1819–1825.
- Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charleat B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487:370–374.
- Chotewutmontri P, Stiffler N, Watkins KP, Barkan A. 2018. Ribosome Profiling in Maize. In: Lagrimini LM, editor. *Maize*. Vol. 1676. *Methods in Molecular Biology*. New York, NY: Springer New York. p. 165–183. Available from: http://link.springer.com/10.1007/978-1-4939-7315-6_10
- Conn SJ, Hocking B, Dayod M, Xu B, Athman A, Henderson S, Aukett L, Conn V, Shearer MK, Fuentes S, et al. 2013. Protocol: Optimising hydroponic growth systems for nutritional and physiological analysis of *Arabidopsis thaliana* and other plants. *Plant Methods* 9:1–11.
- Constantin A-E, Patil I. 2021. ggsignif: R Package for Displaying Significance Brackets for “ggplot2.” *PsyArxiv* [Internet]. Available from: <https://psyarxiv.com/7awm6>
- Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. 2015. Young genes out of the male: An insight from evolutionary age analysis of the pollen transcriptome. *Mol. Plant* 8:935–945.
- Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. McEntyre J, editor. *PLoS Comput. Biol.* 4:e1000176.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.
- Dosztányi Z. 2018. Prediction of protein disorder based on IUPred. *Protein Sci.* 27:331–340.
- Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. 2019. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* 29:932–943.
- Erdős G, Dosztányi Z. 2024. AIUPred: combining energy estimation with deep learning for the

- enhanced prediction of protein disorder. *Nucleic Acids Res.* 52:W176–W181.
- Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Bioinforma.* 23:566–579.
- Gilmore SA, Voorhies M, Gebhart D, Sil A. 2015. Genome-Wide Reprogramming of Transcript Architecture by Temperature Specifies the Developmental States of the Human Pathogen *Histoplasma*. Heitman J, editor. *PLOS Genet.* 11:e1005395.
- Grandchamp A, Czuppon P, Bornberg-Bauer E. 2024. Quantification and modeling of turnover dynamics of *de novo* transcripts in *Drosophila melanogaster*. *Nucleic Acids Res.* 52:274–287.
- Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32:2847–2849.
- Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, Nishi R, Ohashi C, Iida K, Tanaka M, et al. 2013. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc. Natl. Acad. Sci. U. S. A.* 110:2395–2400.
- 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.* 17:632–640.
- Holehouse AS, Kragelund BB. 2023. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* [Internet]. Available from: <https://www.nature.com/articles/s41580-023-00673-0>
- Hsu PY, Calviello L, Wu HYL, Li FW, Rothfels CJ, Ohler U, Benfey PN. 2016. Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 113:E7126–E7135.
- Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43:476–481.
- Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* 8:1365–1379.
- Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Resolution Using Ribosome Profiling. *Science* 324:218–223.
- Ingolia NT, Hussmann JA, Weissman JS. 2019. Ribosome profiling: Global views of translation. *Cold Spring Harb. Perspect. Biol.* 11:1–20.
- Iyengar BR, Bornberg-Bauer E. 2023. Neutral models of *de novo* gene emergence suggest that gene evolution has a preferred trajectory. *bioRxiv* 40:2023.02.05.527172.
- Jacob F. 1977. Evolution and Tinkering. *Science* 196:1161–1166.

- Ji Z. 2018. RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling. *Curr. Protoc. Mol. Biol.* 124:e67.
- Jin G-T, Xu Y-C, Hou X-H, Jiang J, Li X-X, Xiao J-H, Bian Y-T, Gong Y-B, Wang M-Y, Zhang Z-Q, et al. 2025. A de novo Gene Promotes Seed Germination Under Drought Stress in Arabidopsis. Josephs E, editor. *Mol. Biol. Evol.* 42:msae262.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.
- Kassambara A. 2023. ggpubr: “ggplot2” Based Publication Ready Plots. Available from: <https://CRAN.R-project.org/package=ggpubr>
- Katoh K. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.
- Kim W, Hwang S, Lee N, Lee Y, Cho S, Palsson B, Cho B-K. 2020. Transcriptome and translome profiles of Streptomyces species in different growth phases. *Sci. Data* [Internet] 7. Available from: <https://www.nature.com/articles/s41597-020-0476-9>
- Kiniry SJ, Michel AM, Baranov PV. 2020. Computational methods for ribosome profiling data analysis. *Wiley Interdiscip. Rev. RNA* 11:1–22.
- Kleppe AS, Bornberg-Bauer E. 2018. Robustness by intrinsically disordered C-termini and translational readthrough. *Nucleic Acids Res.* 46:10184–10194.
- Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res.* 19:1752–1759.
- Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* 40:1202–1210.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.
- Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 25:1841–1842.
- Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. Prlic A, editor. *PLoS Comput. Biol.* 9:e1003118.
- Lei T, Chang Y, Yao C, Zhang H. 2022. A systematic evaluation revealed that detecting translated non-canonical ORFs from ribosome profiling data remains challenging.
- Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* 103:9935–9939.

- Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant J.* 58:485–498.
- Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, Ge S, Guo YL. 2016. On the origin of de novo genes in Arabidopsis thaliana populations. *Genome Biol. Evol.* 8:2190–2202.
- Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* 4:865–875.
- Madeira F, Madhusoodanan N, Lee J, Eusebi A, Niewielska A, Tivey ARN, Lopez R, Butcher S. 2024. The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res.* 52:W521–W525.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10.
- McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: What, how and why. *Nat. Rev. Genet.* 17:567–578.
- Mészáros B, Erdős G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46:W329–W337.
- Michał Krassowski, Arts M, Lagger C, Max. 2022. krassowski/complex-upset: v1.3.5. Available from: <https://zenodo.org/record/3700590>
- Montañés JC, Huertas M, Messegue X, Albà MM. 2023. Evolutionary Trajectories of New Duplicated and Putative De Novo Genes. *Mol. Biol. Evol.* 40:1–16.
- Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.* 32:258–267.
- Moyers BA, Zhang J. 2016. Evaluating Phylostratigraphic Evidence for Widespread de Novo Gene Birth in Genome Evolution. *Mol. Biol. Evol.* 33:1245–1256.
- Moyers BA, Zhang J. 2017. Further simulations and analyses demonstrate open problems of phylostratigraphy. *Genome Biol. Evol.* 9:1519–1527.
- Moyers BA, Zhang J. 2018. Toward reducing phylostratigraphic errors and biases. *Genome Biol. Evol.* 10:2037–2048.
- Neme R, Tautz D. 2014. Evolution: Dynamics of de novo gene emergence. *Curr. Biol.* 24:R238–R240.
- Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife* 5:1–20.
- Nielly-Thibault L, Landry CR. 2019. Differences between the raw material and the products of de novo gene birth can result from mutational biases. *Genetics* 212:1353–1366.
- Ohno S. 1970. Evolution by Gene Duplication. Berlin, Heidelberg: Springer Berlin Heidelberg

Available from: <http://link.springer.com/10.1007/978-3-642-86659-3>

Orr MW, Mao Y, Storz G, Qian SB. 2021. Alternative ORFs and small ORFs: Shedding light on the dark proteome. *Nucleic Acids Res.* 48:1029–1042.

Papadopoulos C, Arbes H, Cornu D, Chevrollier N, Blanchet S, Roginski P, Rabier C, Atia S, Lespinet O, Namy O, et al. 2024. The ribosome profiling landscape of yeast reveals a high diversity in pervasive translation. *Genome Biol.* 25:268.

Papadopoulos C, Callebaut I, Gelly J-C, Hatin I, Namy O, Renard M, Lespinet O, Lopes A. 2021. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Res.* 31:2303–2315.

Pavan F, Azevedo Favory J, Lacoste E, Beaumont C, Louis F, Blassiau C, Cruaud C, Labadie K, Gallina S, Genete M, et al. 2024. The evolutionary history and functional specialization of microRNA genes in *Arabidopsis halleri* and *A. lyrata*. Available from: <http://biorxiv.org/lookup/doi/10.1101/2024.05.03.592357>

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* 85:2444–2448.

Pedersen TL. 2024. patchwork: The Composer of Plots. Available from: <https://patchwork.data-imaginist.com>

Peng J, Zhao L. 2024. The origin and structural evolution of de novo genes in *Drosophila*. *Nat. Commun.* 15:810.

Prensner JR, Abelin JG, Kok LW, Clauser KR, Mudge JM, Ruiz- J, Bassani-sternberg M, Deutsch EW, Heesch SV. 2023. What can Ribo-seq and proteomics tell us about the non-canonical proteome ? :734–763.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Remm M, Storm CEV, Sonnhammer ELL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314:1041–1052.

Roginski P, Grandchamp A, Quignot C, Lopes A. 2024. DE Novo emerged gene SEarch in Eukaryotes with DENSE. Genomics Available from: <http://biorxiv.org/lookup/doi/10.1101/2024.01.30.578014>

Roux C, Castric V, Pauwels M, Wright SI, Saumitou-Laprade P, Vekemans X. 2011. Does Speciation between *Arabidopsis halleri* and *Arabidopsis lyrata* Coincide with Major Changes in a Molecular Target of Adaptation? Ingvarsson PK, editor. *PLoS ONE* 6:e26872.

Ruiz-Orera J, Albà MM. 2019. Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. *Trends Genet.* 35:186–198.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. 2015. Origins of De Novo Genes in Human and

- Chimpanzee. *PLoS Genet.* 11:1–24.
- 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.* 2:890–896.
- Schlesinger D, Elsässer SJ. 2022. Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. *FEBS J.* 289:53–74.
- Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* 2:1626–1632.
- Silveira AB, Trontin C, Cortijo S, Barau J, Del Bem LEV, Loudet O, Colot V, Vincentz M. 2013. Extensive Natural Epigenetic Variation at a De Novo Originated Gene. *PLoS Genet.* 9:3–10.
- Singh R, Ming R, Yu Q. 2016. Comparative Analysis of GC Content Variations in Plant Genomes. *Trop. Plant Biol.* 9:136–149.
- Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* 45:831–835.
- Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12:692–702.
- Tjeldnes H, Labun K, Torres Cleuren Y, Chyżyńska K, Świrski M, Valen E. 2021. ORFik: a comprehensive R toolkit for the analysis of translation. *BMC Bioinformatics* 22:336.
- Vakirlis N, Carvunis AR, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* 9:1–23.
- Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2018. A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* 35:631–645.
- Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet.* 15:1–23.
- Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, Carvunis A-R. 2023. A vast evolutionarily transient translome contributes to phenotype and fitness. *bioRxiv*:2021.07.17.452746.
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189–1191.
- Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer New York Available from: <https://link.springer.com/10.1007/978-0-387-98141-3>
- Willis S, Masel J. 2018. Gene birth contributes to structural disorder encoded by overlapping genes. *Genetics* 210:303–313.
- Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted

- by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* 1.
- Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* 3:1245–1252.
- Wu HYL, Hsu PY. 2022. A custom library construction method for super-resolution ribosome profiling in Arabidopsis. *Plant Methods* 18:1–8.
- Xiao W, Liu H, Li Y, Li X, Xu C, Long M, Wang S. 2009. A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS ONE* 4:1–12.
- Xiao Z, Huang R, Xing X, Chen Y, Deng H, Yang X. 2018. De novo annotation and characterization of the translome with ribosome profiling data. *Nucleic Acids Res.* 46:E61.
- Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat. Ecol. Evol.* 3:679–690.
- Zhao G, Liu Z, Quan J, Lu J, Li L, Pan Y. 2024. Ribosome Profiling and RNA Sequencing Reveal Translation and Transcription Regulation under Acute Heat Stress in Rainbow Trout (*Oncorhynchus mykiss*, Walbaum, 1792) Liver. *Int. J. Mol. Sci.* 25:8848.
- Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* 343:769–772.
- Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Res.* 18:1446–1455.
- Zhu X-T, Zhou R, Che J, Zheng Y-Y, Tahir UI Qamar M, Feng J-W, Zhang J, Gao J, Chen L-L. 2023. Ribosome profiling reveals the translational landscape and allele-specific translational efficiency in rice. *Plant Commun.* 4:100457.

Supplementary material

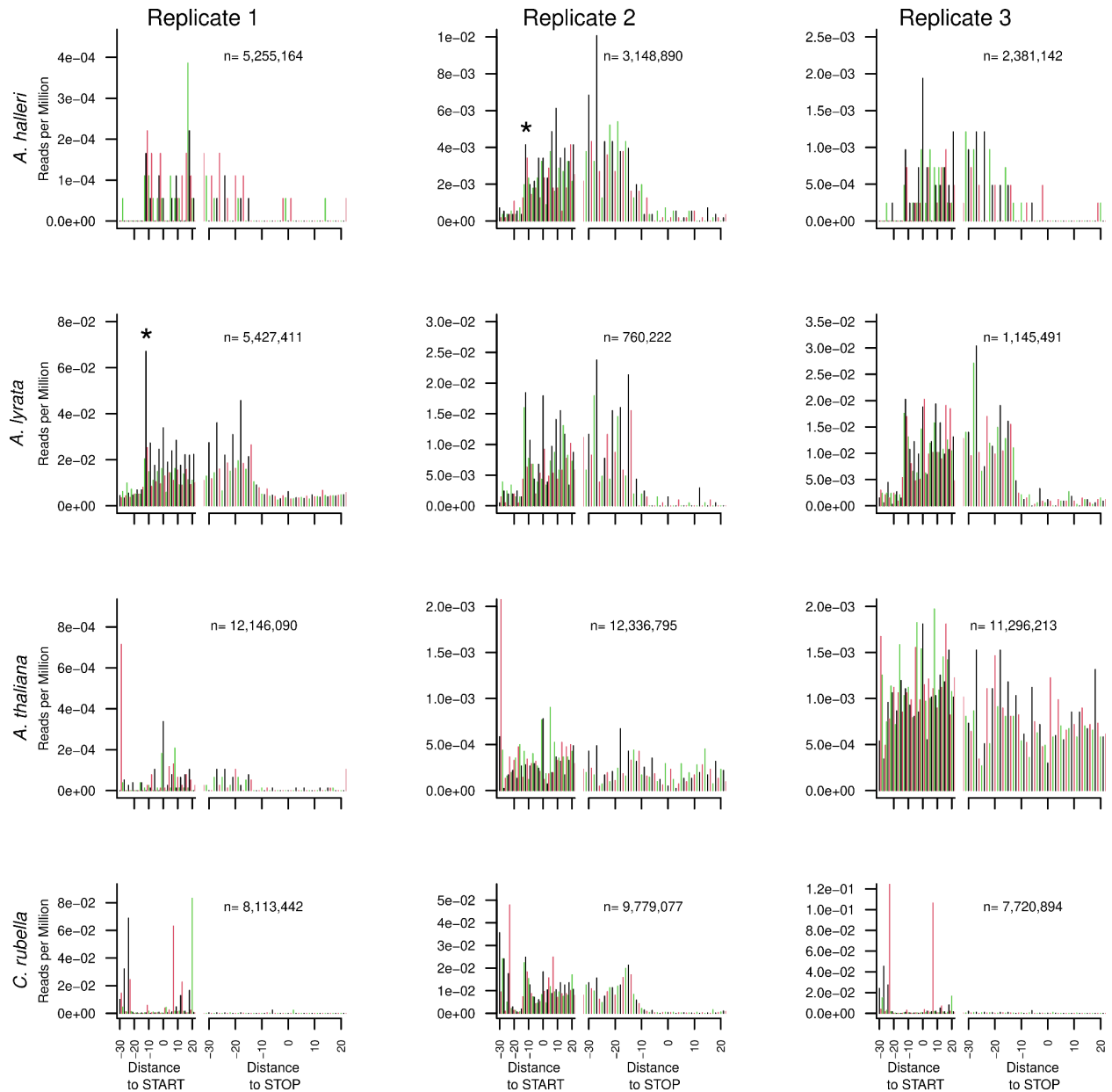


Figure S1: Metagene plots of 28-kmers from all RPF libraries produced in this study, before P-site shifting, where each line corresponds to a species, and each column to a replicate. Produced using data from all samples described in Table S1. Read accumulation is colored based on its frame relative to the CDS: black for the first frame, red for the second frame and green for the third frame. The total number of mapped reads per library is given by "n=". The estimated p-site offsets of sample AHA_rep-2 and ALY_rep-1, for 28-kmers, are represented by an asterisk "*".

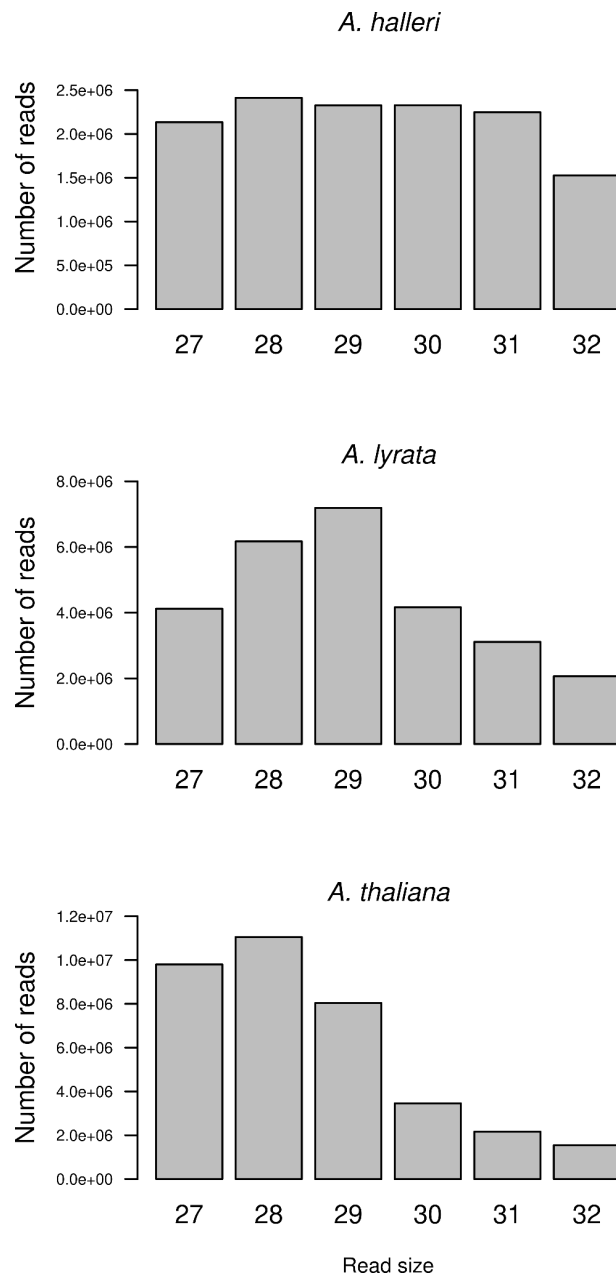


Figure S2: RPF k-mer distributions in the same libraries as those from Table S2.

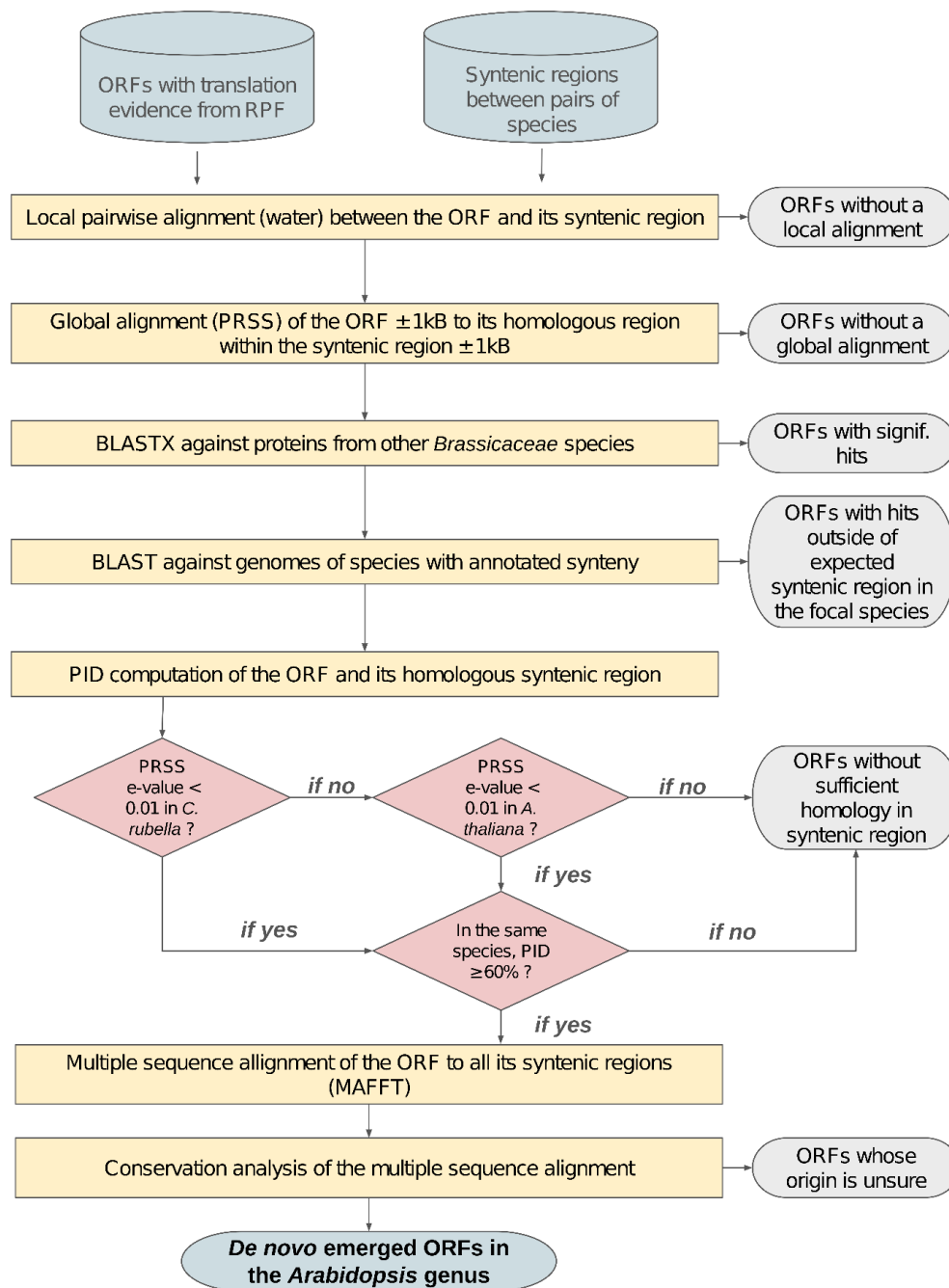


Figure S3: Flowchart of the bioinformatic pipeline used for the annotation of translated de novo ORFs. “ORFs with translation evidence from RPF” include both previously annotated CDS, and IGORFs.

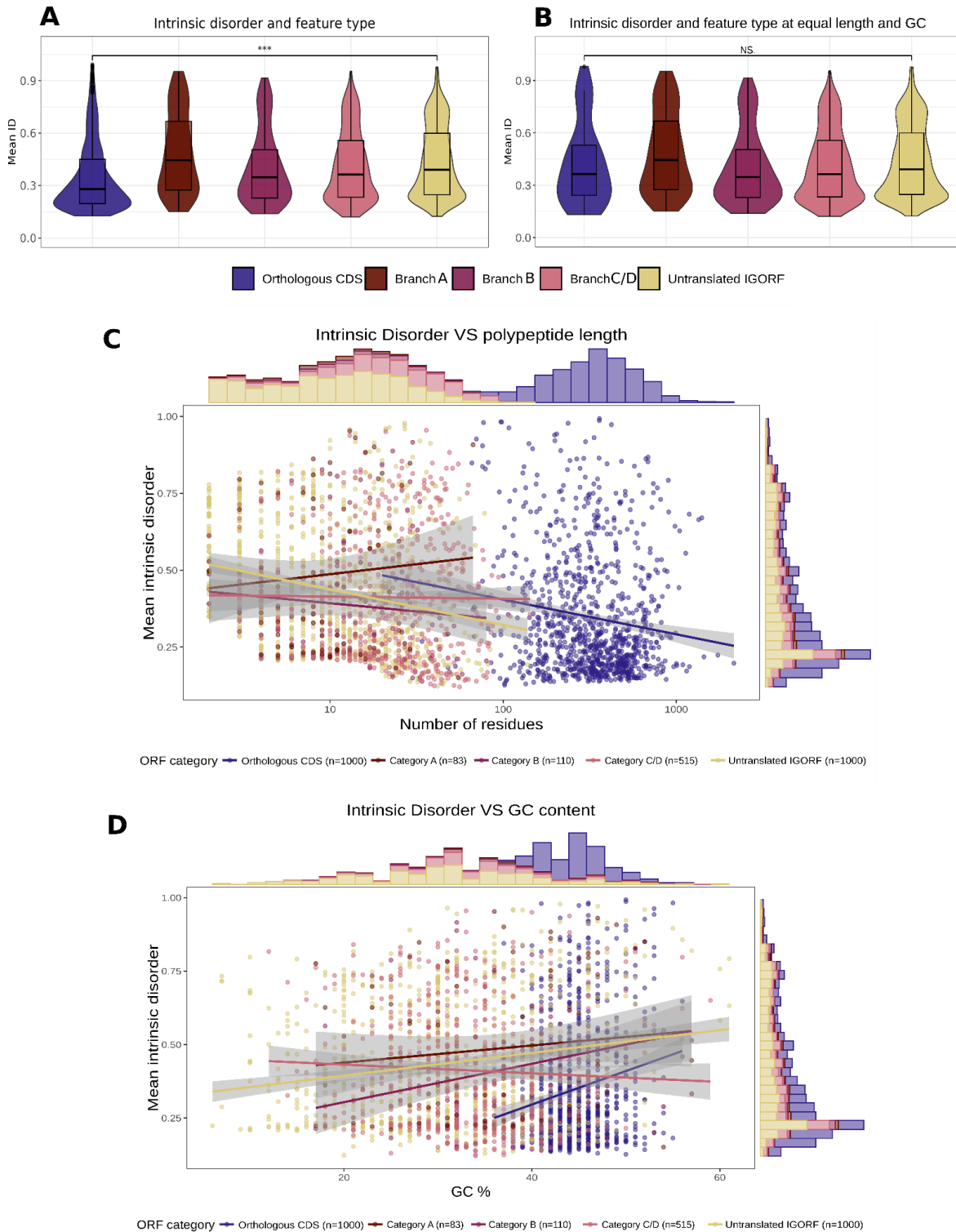


Figure S4: A) Measures of ISD in the translation products of conserved canonical CDS, different conservation classes of *de novo* ORFs and untranslated IGORFs. The only statistically significant difference (between orthologous CDS and untranslated IGORFs) is represented (Mann-Whitney U-test, *** = $p < 0.001$). B) Measures of ISD in the translation products of conserved canonical CDS, different conservation classes of *de novo* ORFs and untranslated IGORFs; but where CDS were selected to match other categories in length and GC content. There is no longer any statistical

difference. C) Linear regression of Mean Intrinsic Disorder in the translation products of the same previous categories, against the number of amino acid residues. Across categories, a significant negative association was found between the 2 variables (*adjusted $R^2 = 0.03532$, $p\text{-value} < 2.2e-16$*). D) Linear regression of Mean Intrinsic Disorder in the translation products of the same previous categories against the level of GC content in the DNA sequences encoding them. A significant positive association was found in orthologous CDS and IGORFs (in conserved CDS: *adjusted $R^2=0.03359$, $p\text{-value} = 5.374e-09$*).

	Sample number	Raw reads	Reads after adapter trimming	Reads after quality control and size selection	Reads after small nuclear RNA depletion	Reads mapped once to genome	Reads mapped 2-20 times to genome	Mapped too many times (>20)	Unmapped	% of depleted reads mapped 1-20 times to genome	Reference
<i>A. halleri</i>	AHA_rep-1	109,726,477	100,737,210	34,920,869	18,633,216	281,388	4,973,776	11,653,581	1,707,550	28.20%	This study
	AHA_rep-2	109,984,173	105,770,804	29,170,453	19,722,159	587,077	2,561,813	6,379,039	10,172,228	15.97%	This study
	AHA_rep-3	140,313,302	133,578,840	24,139,697	13,927,608	133,513	2,247,629	8,370,767	3,142,856	17.10%	This study
<i>A. lyrata</i>	ALY_rep-1	105,132,663	103,787,753	71,073,073	42,007,507	3,978,159	1,449,252	986,710	33,858,990	12.92%	This study
	ALY_rep-2	105,002,821	100,887,241	22,519,299	12,470,433	529,556	230,666	81,703	6,762,931	6.10%	This study
	ALY_rep-3	101,421,943	97,158,242	22,427,522	12,715,120	872,255	273,236	208,522	6,985,862	9.01%	This study
<i>A. thaliana</i>	ATH_rep-1	99,112,478	96,653,379	24,608,498	13,171,494	270,408	11,875,682	9,871	998,251	92.21%	This study
	ATH_rep-2	106,347,575	104,676,787	29,382,988	17,873,730	709,442	11,627,353	20,657	5,491,661	69.02%	This study
	ATH_rep-3	92,828,713	91,385,713	24,592,438	13,201,302	1,372,527	9,923,686	20,087	1,863,135	85.57%	This study
<i>C. rubella</i>	CRU_rep-1	115,874,453	108,068,275	18,040,300	10,811,907	450,022	7,663,420	22,853	2,675,581	75.04%	This study
	CRU_rep-2	91,888,901	89,545,731	26,377,761	17,750,598	2,488,358	7,290,719	44,646	7,926,844	55.09%	This study
	CRU_rep-3	101,780,575	97,333,280	17,652,587	10,198,453	361,403	7,359,491	29,396	2,448,136	75.71%	This study

Table S1: Summary of the RPF sequencing data produced in this study.

	Sample number	Raw reads	Reads after adapter trimming	Reads after quality control and size selection	Reads after small nuclear RNA depletion	Reads mapped once to genome	Reads mapped 2-20 times to genome	Mapped too many times (>20)	Unmapped	% of depleted reads mapped 1-20 times to genome	Reference
<i>A. lyrata</i>	ALY_2_S2_all	609,771,764	601,833,807	409,856,100	242,989,105	22,947,667	8,584,665	5,704,766	195,785,778	12.98%	This study
<i>A. halleri</i>	ALY_rep-1	639,109,284	614,402,572	168,360,788	114,000,850	3,409,600	14,812,250	36,770,015	58,882,173	15.98%	This study
<i>A. thaliana</i>	Root_1	163,396,264	159,778,193	142,176,011	107,573,216	73,679,997	19,213,813	18,778	14,625,936	86.35%	Hsu <i>et al.</i> , 2016
	Root_2	166,148,265	162,112,008	138,493,159	115,260,206	71,582,384	13,005,304	18,222	30,625,465	73.39%	Hsu <i>et al.</i> , 2016
	Root_3	207,346,547	202,257,844	137,022,278	97,879,433	69,194,868	19,289,223	39,751	9,299,834	90.40%	Hsu <i>et al.</i> , 2016
	Shoot_1	145,235,719	142,483,567	121,767,293	102,210,257	33,727,185	11,065,144	20,256	57,388,016	43.82%	Hsu <i>et al.</i> , 2016
	Shoot_2	179,566,257	175,969,706	163,389,872	129,057,937	87,963,741	22,356,181	23,358	18,682,755	85.48%	Hsu <i>et al.</i> , 2016
	Shoot_3	134,241,510	131,386,200	117,584,678	84,372,098	59,378,977	11,417,693	20,030	13,535,155	83.91%	Hsu <i>et al.</i> , 2016
	Seedlings	185,867,719	183,322,013	87,331,306	66,719,312	32,452,595	32,075,787	44,822	2,141,266	96.72%	Hsu <i>et al.</i> , 2022

Table S2: Summary of the RPF sequencing data, both from this study and from publicly available datasets, used to annotate translation events.

Annotation number			Detection of translation by:			Translated annotation number	
Species	Canonical genes	IGORF	RiboCode	RibORF	ORFSCORE	Canonical genes	IGORF
A. halleri	34,805	19,117,874	YES	YES	YES	3,467	135
			NO	YES	YES	11,952	2,727
			YES	NO	YES	34	66
			YES	YES	NO	10	106
			NO	NO	YES	5,466	14,195
			NO	YES	NO	238	4,424
			YES	NO	NO	0	35
TOTAL:					21,167	21,688	
A. lyrata	30,443	23,503,084	YES	YES	YES	13,266	1,548
			NO	YES	YES	8,210	7,114
			YES	NO	YES	101	676
			YES	YES	NO	9	292
			NO	NO	YES	2,701	60,249
			NO	YES	NO	108	15,684
			YES	NO	NO	6	1,870
TOTAL:					24,401	87,433	
A. thaliana	27,559	3,327,672	YES	YES	YES	20,592	142
			NO	YES	YES	364	208
			YES	NO	YES	47	70
			YES	YES	NO	295	36
			NO	NO	YES	169	346
			NO	YES	NO	451	583
			YES	NO	NO	28	11
TOTAL:					21,946	1,396	

Table S3: Summary of the data used to produce Figure 2: the number of available annotations, the number of tools detecting them as translated, and the resulting number of annotations found as translated.

Species pair		Translated ORFs found within syntenic regions	Without significant hits in protein coding genes of other <i>Brassicaceae</i> species	Without significant hits in genomic regions outside of synteny in focal species	With an extended alignment in <i>C. rubella</i> or <i>A. thaliana</i> of E-value < 0.001 and a PID > 60%	With a <i>de novo</i> emergence
<i>A. halleri</i>	<i>A. lyrata</i>	2,880	1,416	864	432	228
	<i>A. thaliana</i>	3,334				
	<i>C. rubella</i>	3,100				
<i>A. lyrata</i>	<i>A. halleri</i>	7,152	5,454	3,520	1,553	779
	<i>A. thaliana</i>	8,614				
	<i>C. rubella</i>	8,504				
<i>A. thaliana</i>	<i>A. halleri</i>	2,192	376	247	60	28
	<i>A. lyrata</i>	2,362				
	<i>C. rubella</i>	2,888				

Table S4: Summary of the data produced by synteny analyzes as shown in Figure S3.

RPF data from	Conservation category	Already annotated as CDS*	Newly identified as IGORF
<i>A. halleri</i>	Category A	2	9
	Category B	2	34
	Category D	1	180
	TOTAL:	5	223
<i>A. lyrata</i>	Category A	3	74
	Category B	2	118
	Category C	13	569
	TOTAL:	18	761
<i>A. thaliana</i>	Category A	3	1
	A. thaliana - specific	5	19
	TOTAL:	8	20
ALL SPECIES TOTAL:		31	1,004
SUM:		1,035	
(without redundance between species)		738	

* in at least one species for conserved ones

Table S5: Summary of the number of ORFs found to be translated and have emerged *de novo* in each species, according to their conservation and annotation (CDS or IGORF) category.

CHAPTER II

Evolution of plant microRNA-encoded peptides (miPEPs) in the *Arabidopsis* genus

Authors

Claire Patiou¹, Christelle Blassiau¹, Isabelle Hatin², Flavia Pavan¹, Clément Mazoyer¹, Chloé Ponitzki¹, Laurence Debacker¹, Olivier Namy², Vincent Castric¹, Sylvain Legrand¹, Eléonore Durand¹

¹ Univ. Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France

² Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France

Author for correspondence : eleonore.durand@univ-lille.fr

Abstract

Long thought of as non-coding, the ability of microRNA (miRNA) primary transcripts (primiRNAs) to encode peptides (miPEPs) has been recently uncovered as a key regulation mechanism of plant miRNA expression. By specifically enhancing the expression of the miRNA with which they are transcribed, miPEPs were shown to have a wide variety of phenotypic effects on many traits, including plant development. Despite their biological importance, the lack of exhaustive primiRNA annotations means little is known about their broader prevalence and conservation, especially outside of model species. Leveraging the power of long-RNA sequencing technology, we annotated primiRNAs in *A. thaliana*, *A. halleri* and *A. lyrata*. Using Ribosome Profiling (RPF) datasets from the same species, we obtained a repertoire of potentially translated microRNA-ORFs (miORFs). The 126 primiRNAs we identified in *A. halleri* and *A. lyrata* were on average seven times longer than the miRNA precursors they contained. Up to 39 of them (about 31%) contained miORFs with translation signals, with a higher prevalence in conserved primiRNAs. While ribosomes scanned the entirety of primiRNAs, translated miORFs preferably accumulated in their 5' region. MiORFs displayed low levels of conservation, but their existence itself was on the contrary often conserved. Our analyzes confirm the extent of coding potential in *Arabidopsis* primiRNAs, especially within their 5' region, and the importance of translation events themselves rather than the exact peptide sequence.

Introduction

Determining the molecular basis of phenotypic variation is one of evolutionary biology's most important endeavors. Among a plethora of biological and environmental mechanisms, one way to explain the differences between species is the differences in their genomes. Yet, the gene-coding sequences of recently diverged species tend to be so similar that they are highly unlikely to be able to, alone, account for all of the observed phenotypic differences. A seminal 1975 paper (King and Wilson 1975) suggested that differences in protein coding genes alone between humans and chimpanzees were not enough to explain phenotypic differences between the two species. To fully explain those differences, it was necessary to take into account regulatory elements. Historically, transcription factors were the first discovered regulatory elements (Jacob and Monod 1961), that bind to protein-coding genes' promoters and either inhibit or enhance gene expression (Yu and Gerstein 2006). "Non-coding" RNAs are a second class of regulatory elements and include a number of molecules, including but not limited to long non-coding RNAs (lncRNAs), small interfering RNAs (siRNAs) and microRNAs (miRNAs). The latter are found in a wide range of eukaryotes, including the animal and plant kingdoms, across which their biogenesis and modes of action vary. They play key roles in determining the specific expression of protein-coding genes in time and space, and are involved in essential biological processes like development or environmentally-induced responses (Xie et al. 2005; Cuperus et al. 2011; Millar 2020).

In plants, mature miRNAs are 21-22 base-pair long (in most cases) one-stranded RNA molecules capable of degrading the messenger RNA of a target gene, or inhibiting its translation by almost perfect sequence complementarity. Most plant miRNA genes (67%) are independent transcription units with exons, introns, and their own promoters, and are transcribed by RNAPol II, capped and poly-adenylated (Zhang et al. 2015; Stepien et al. 2017). They are initially transcribed as primary transcripts (pri-miRNAs) and are then recruited into Dicer bodies (D-bodies) within the nucleus, by what is known as the "microprocessor": an assembly of Dicer-Like 1 (DCL1), Hyponastic Leaves 1 (HYL1) and Serrate (SE), as well as other auxiliary proteins (Zhang et al. 2015; Ding and Zhang 2023). This longer pri-miRNA (from a few hundred to over a thousand bases long) is processed by the microprocessor into a precursor (pre-miRNA) exhibiting a stem-loop structure (of varying length), itself processed into a miRNA duplex. This duplex is exported to the cytoplasm, where an Argonaute protein (AGO) selects the mature miRNA strand by recognizing its 5' terminal nucleotide (Mi et al. 2008), and then integrates it

into the RNA-induced Silencing Complex (RISC) (Wang et al. 2019). This RISC complex is then able to target specific messenger RNAs (mRNAs), thanks to their sequence complementarity to the miRNA, and inhibits expression of the gene by mRNA degradation or mRNA translation inhibition. MiRNAs are involved in a wide range of physiological processes, including leaf morphogenesis, stress response and root development (Millar 2020); and are critical to global plant development, with *dcl1* mutants being systematically lethal (Laubinger et al. 2008).

miRNAs have long been considered strictly non-coding, and hence unable of encoding and being translated into functional polypeptides. Yet, Laressergues and his team revealed, in 2015 in *Arabidopsis thaliana* and *Medicago truncatula*, the presence and function of peptides encoded by short ORFs, located within the 5' region of miRNA primary transcripts (Laressergues et al. 2015) which we thereafter refer to as "miORFs". The miRNA-encoded peptides (miPEPs) they discovered were able to positively regulate the transcription of the same miRNA locus from which they originated. Inducing the overexpression of miPEPs, or applying synthetic miPEPs to plants, could lead to visible phenotypic effects such as a reduction in lateral root production, due to the quantitatively greater presence of mature miRNA (Laressergues et al. 2015). The detailed processing leading to a dual coding and non-coding role of pri-miRNAs was characterized by (Laressergues et al. 2022), according to which miRNA genes could produce two forms of transcripts. The first one is short and contains only the miORF. It would be exported outside the nucleus to be translated into a miPEP, which would then be reintegrated into the nucleus and positively regulate the expression of the miRNA from which it originated. The second, longer transcript, in which only the stem-loop is excised and conserved, leads to the mature miRNA. Several studies were conducted to further characterize the functional potential of these miPEPs, in other species, and within other tissues. These have identified peptides such as AtmiPEP164b, which inhibits root growth in *Barbarea vulgaris* (Ormancey et al. 2021); miPEP165a, which promotes root growth and influences flowering time in *A. thaliana* (Ormancey et al. 2020); or miPEP172c, which stimulates nodulation in *Glycine max*. However, the majority of candidate miPEPs have not yet been the target of any functional validation, and a more exhaustive repertoire of candidates has yet to be characterized. Globally, additional genome-wide studies are needed to quantify the overall abundance of miPEPs expressed in a wide range of miRNA families, as previous analyses have so far been limited to just a few.

Quantifying the abundance of miPEPs is difficult, because it requires primary transcript annotations, where miORFs seem to be located. Instead, miRNAs are typically annotated on the

basis of their precursor only, rather than on their full primary transcript. Databases such as miRBase (Griffiths-Jones et al. 2006) list potential miRNA precursors, with e.g. a total number of 326 miRNA genes currently annotated in the *A. thaliana* genome. On the other hand, primary transcripts have no distinguishing features, other than containing one or more future precursors, making their *in silico* annotation challenging. Combining long-read sequencing and ribosome profiling technology was shown to be a powerful method to uncover miPEP repertoires, allowing both the full-length annotation of primary transcripts and providing the required evidence for translation of the predicted ORFs they contain. (Lauressergues et al. 2022) used this approach and revealed that many short, alternatively spliced miORF-containing pri-miRNAs were loaded into ribosomes, but this analysis was limited by the lack of a clear periodicity signal in the ribosome profiling data they used, and the limited range of miRNAs investigated.

An important limitation to our understanding of the overall significance of miPEPs is that, to the best of our knowledge, no study has investigated the evolutionary dynamics of the miPEP repertoires. First, plant miRNA repertoires are highly dynamic, with a number of highly conserved miRNA genes alongside numerous lineage-specific miRNA genes (Fahlgren et al. 2010; Nozawa et al. 2012; Pavan et al. 2024). Second, intergenic ORFs also have their own evolutionary dynamics (Chapter 1). In mice, for example, most peptides derived from the expression of species-specific short ORFs (sORFs) appear to evolve neutrally (Ruiz-Orera et al. 2018). Fixation of mutations at the sORF level, enabling a progressive gain in translation, could thus be achieved by simple genetic drift, as long as the polypeptide product is not toxic (Ruiz-Orera and Albà 2019). Furthermore, while the presence of a type of sORFs, upstream ORFs (uORFs), is often conserved between species, their sequence as such does not appear to be evolutionarily constrained, and high rates of divergence have been observed in *Drosophila* (Zhang et al. 2018). Repertoires of sORFs are thus considered highly labile, and evolve rapidly ((Durand et al. 2019), Chapter 1). Hence, while the evolution of miRNA genes on the one hand, and intergenic sORFs on the other, has already been characterized, that of miORFs in themselves has not been investigated yet.

Previous work in *A. thaliana* has demonstrated the existence in a few miRNA families of functional “miRNA-miPEP” systems made of miPEPs translated from pri-miRNAs. To further characterize the prevalence of translated miORFs, as well as the conservation and evolutionary dynamics of miRNA-miPEP systems, we chose three closely related species from the *Arabidopsis* genus as study models: *A. thaliana*, *A. lyrata* and *A. halleri*, as well as an outgroup

species, *Capsella rubella*. We annotated the repertoires of both primiRNAs and miORFs by producing both long read RNA-seq and Ribosome Profiling data, leveraging the high-quality genome assemblies available for each species (*A. thaliana*; (Lamesch et al. 2012), *A. lyrata*; (Hu et al. 2011), *A. halleri*; (Pavan et al. 2024) and *Capsella rubella*; (Slotte et al. 2013)). Doing so, we uncovered primiRNAs that are much longer than their precursors. We confirmed that translated miORFs are indeed common within primiRNAs, but are highly divergent between species.

Results

Long-read Nanopore sequencing yields miRNA primary transcripts

The first step to annotate miORF candidates is to obtain primiRNA annotations. However, while the experimental and bioinformatic protocols for miRNA precursor annotation are well established due to their typical hairpin structure and pattern of sRNA reads accumulation (Guigon et al. 2019), annotation of the primary transcripts of miRNAs remains challenging. On one hand, miRNA precursors can be annotated using the pile-like accumulation of small RNAseq reads at the positions of the mature and star miRNAs contained in the hairpin precursor, with relatively little variation in both of their lengths. Recent work in our team has yielded a total of 418, 276 and 463 premiRNA annotations in *A. thaliana*, *A. lyrata* and *A. halleri* respectively (**Table S2**) (Pavan et al. 2024). On the other hand, there is no consistent size expected for primiRNAs, and the abundance of possible promoters, transcription start and termination sites (TSS and TTS respectively) renders the identification of their boundaries on the basis of sequence alone impossible. Hence, annotating primary transcripts requires RNA sequencing data, where the full transcripts would be sequenced before downstream processing and splicing. For these reasons, we used a long-read cDNA-based sequencing approach (Oxford Nanopores Technologies) from leaves, roots and flower buds of our four species of interest, yielding between one and 19 million mapped reads per species (**Table S1**). Using sequencing reads that mapped to previously identified miRNA precursors (373, 211 and 353 respectively and excluding those that overlapped known coding DNA sequences (CDS), **Table S2, Figure 1**), we were able to annotate 64 primiRNAs in *A. halleri*, 62 in *A. lyrata* and 94 in *A. thaliana*, defined as the longest RNA transcripts overlapping the miRNA precursors annotation. The primary transcripts vary greatly in length, ranging from 131 to 4385 nucleotides long with an

average of 1340, and are on average 7 times longer than their corresponding hairpin precursors.

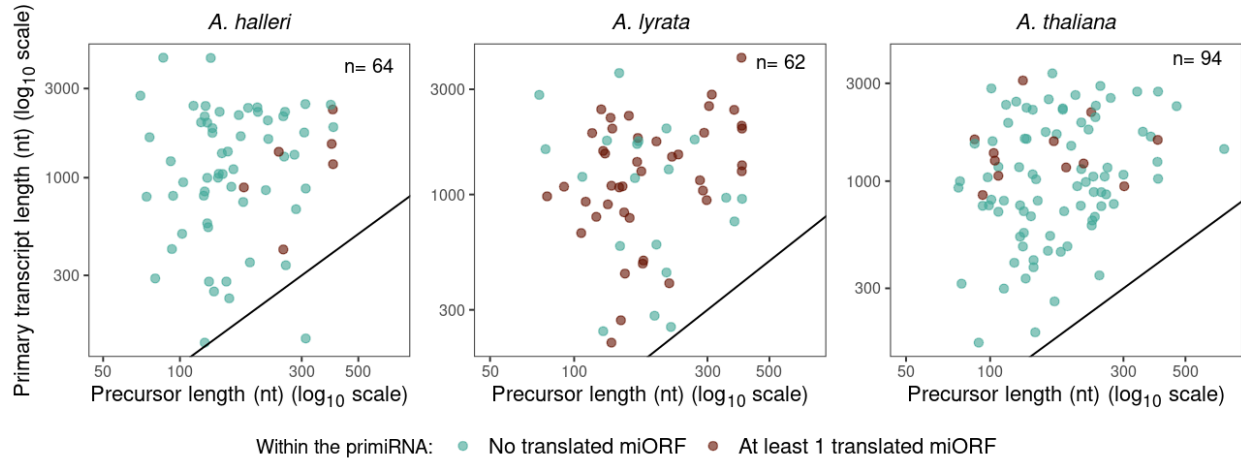


Figure 1: Length of primiRNAs plotted against the length of their corresponding premiRNAs (in nucleotides, log₁₀ scale). Total number of annotated primiRNAs per species indicated by “n”. The diagonal black line represents $x=y$. PrimiRNAs with at least one miORF detected as translated are differentiated by color.

An overwhelming majority of primary transcripts contained only one miRNA precursor, yet several examples of polycistronic transcripts could be identified. In *A. thaliana*, the precursors miR850 and miR863, miR169k and miR169l, and miR776 and a novel miRNA, were located within the same primary transcript, confirming previously obtained evidence (Merchan et al. 2009). Furthermore, both miR775 and miR779 belonged to the same primary transcript as nearby small nucleolar RNAs (snoRNAs), as was already described by (Qu et al. 2015). Additionally, a novel polycistronic transcript was annotated for miR4228, residing on the same primary transcripts as recently annotated C/D box snoRNAs (Azevedo-Favory et al. 2021). Furthermore, in line with (Stepien et al. 2017), we observed events of splicing of primiRNA transcripts when visualizing the alignments of long-read sequencing data. For example, 20 *A. thaliana* primary transcripts exhibited alternative splicing, with miRNA precursors found either within or outside spliced sequences (introns).

ORFs are abundant in pri-miRNAs, which are scanned by ribosomes

We next identified ORFs that may encode polypeptides within miRNA primary transcripts. To do so, we started by broadly annotating all possible ORFs within pri-miRNA nucleotide sequences using a set of relaxed criteria, requiring a minimum length of nine nucleotides, and allowing for non-canonical start codons. Within those miORF candidates, we then investigated the overlap with translation signals from RPF data (data from Chapter 1). A total of 5,166, 4,509 and 5,133 miORFs were annotated across the entire set of pri-miRNAs in *A. lyrata*, *A. halleri* and *A. thaliana* respectively (**Figure 2**). We chose to distinguish between ORFs that did not display any RPF read, and those that display at least one, for two main reasons. Firstly, we expect ORFs outside of canonical protein coding genes to be expressed at relatively low levels (see chapter 1), and in an effort to keep as many potential candidates, we used relaxed coverage criteria. Secondly, as ribosomes might simply scan the primary transcript without necessarily translating it actively, ribosomal footprints would not accumulate in great numbers but instead sparsely cover the transcript. Hence, miORFs with only one read can still belong to a miRNA primary transcript undergoing ribosomal scanning. Within the miORFs that harbour at least one RPF read, the translation signal was evaluated by three separate tools: ORFSCORE (Bazzini et al. 2014), RibORF and RiboCode (Xiao et al. 2018). If translation was detected by at least one of those three tools, the miORF was classified as having “Translation Signal” (TS). In some cases, RPF reads were detected, but no TS was found: in such cases, whether translation is truly happening or not is uncertain, and might be explained by either a lack of power, or a true absence of translation. We classify those miORFs as having at least one RPF, but no TS. Finally, some miORFs displayed neither RPF nor TS, and were distinguished from the two previous categories. In total, we identified about four thousand putative miORFs with neither RPF nor TS, between 160 and 1,700 with RPF but no TS, and between six and 149 miORFs with both RPF and TS (**Figure 2**). While high quality RPF data is available for *A. thaliana* and *A. lyrata* (with stark periodicity signals, Chapter 1), the lesser quality of the *A. halleri* RPF dataset implies that our repertoire is probably more incomplete in this species. Overall, our results thus indicate that several miORFs with TS could be identified as candidates for potential peptide-encoding abilities, across *Arabidopsis* species. Among all pri-miRNAs transcripts, between 10 and 70% display miORFs with evidence of TS (**Table S2, Figure 1**).

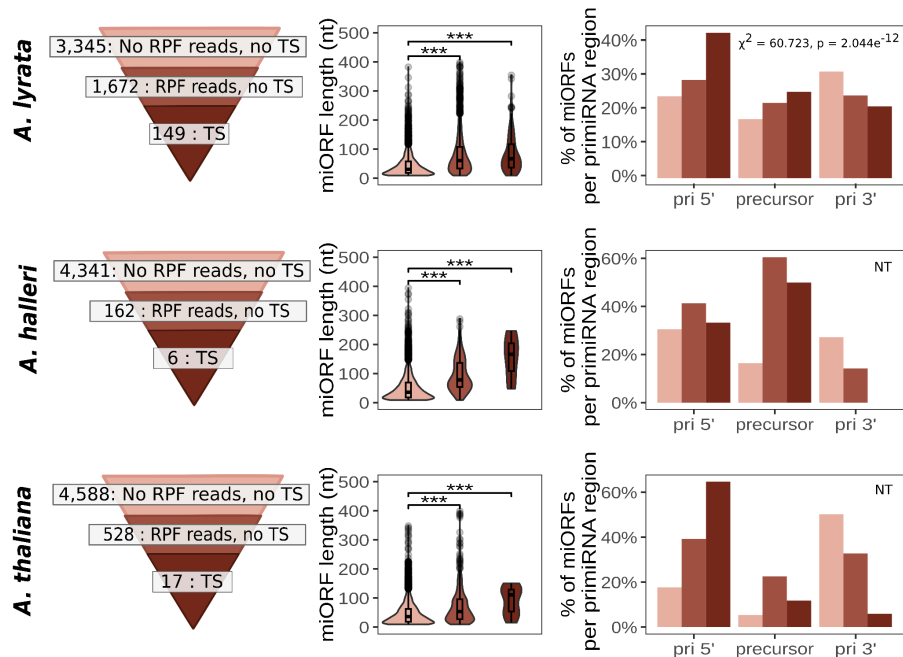


Figure 2: Characteristics of the miORF repertoire, per species. miORFs are classified into one of three categories: no RPF reads and no translation signal (TS), RPF reads but no TS, and RPF reads and TS. Only statistically significant differences in length are represented (Mann-Whitney U-test, ** = $p < 0.01$, *** = $p < 0.001$). The differential accumulation of miORFs within primiRNA regions was tested with a χ^2 test in *A. lyrata*, but not in *A. halleri* and *A. thaliana* due to low miORFs with TS sample size (NT = Not Tested).

As expected, miORFs were much shorter than canonical CDS (the latter being around 1,000 nucleotides long on average), and their length varies depending on the level of evidence for their translation. Across the three species, miORFs without any RPF reads have a mean length of around 10 nucleotides, those with at least one RPF read of around 50 nucleotides, and finally those with TS of around 50 to 150 nucleotides (**Figure 2**). Although longer miORFs may truly be more likely to be translated, this trend may also be related to a bias of RPF data, whereby the detection of TS is harder than when ORFs are short than when they are long (Lei et al. 2022). Still, miORFs as short as 12 nucleotides were detected as being translated, and their abundance in *A. lyrata* indicates that a given primiRNA may contain several actively translated miORFs (with an average of 2.4 miORFs with TS per primiRNA).

We then examined the localization of miORFs along the primary transcripts. We observed RPF reads aligned throughout the entire length of primiRNAs, suggesting that ribosomes eventually

scan the entire primary transcripts (**Figure 2**). However, we observed a clear trend for miORFs with TS to be preferably located in the 5' region of the primiRNAs. This pattern was especially clear in the species with the greatest number of miORF annotations, *A. lyrata* (**Figure 2**). This pattern was also apparent in *A. thaliana*, and to a lesser extent in *A. halleri*, but we could not test for enrichment in the 5' region because of the smaller sample size. When normalizing the miORFs accumulation by the length of the primiRNA regions, the precursor is the region that proportionally harbors the highest density of miORFs, all categories included. Still, the density of miORFs was higher in the 5' than the 3' region. Hence, while miORFs have previously been described as lying upstream of miRNA precursors, we find evidence that they may also lie within, or downstream of precursors; although they would then be less likely to be translated.

Conserved miRNAs have conserved coding potential, through both conserved and divergent miORFs

We then compared the coding potential of primiRNAs according to the level of evolutionary conservation of the miRNA precursor they contain. Among the 126 annotated primiRNAs, 38 were classified as *A. lyrata*-specific, 51 as *A. halleri*-specific, and finally 37 as conserved between the two species (**Figure 3B**). Strikingly, all primiRNAs containing conserved precursors between the two species exhibit at least one RPF read, with about half harboring TS detection by one or more tools (**Figure 3B**). Hence, all of the annotated conserved primiRNAs are at least scanned, and about half translated, by ribosomes.

Next, we focused specifically on primiRNAs that contain precursors that are conserved between *A. lyrata* and *A. halleri*. We investigated the conservation of the miORFs they contained, by aligning the nucleotide sequence of primiRNAs between the two species. We considered a miORF to be conserved if its START/STOP positions and reading frame were conserved. We considered a miORF as present if at least one RPF read mapped to it, or if it was detected as being translated by at least one tool. In conserved primiRNAs, the vast majority of miORFs that harbor at least one RPF read are actually detected as translated by at least one tool (**Figure 3A**). Interestingly, only a fifth of miORFs ($n=159$) are conserved between *A. halleri* and *A. lyrata*. Among the 591 that are not conserved, 150 overlap a miORF with a different sequence in the pair species, which itself is overlapped by at least one RPF read or found to be translated. Again, the majority of those overlapping miORFs in the pair species are detected as translated

by at least one tool. Therefore, while conserved primiRNAs may encode different miPEPs between species, the event of translation itself, within a same broad region, seems to be conserved. In conclusion, all of the 37 conserved primiRNAs between *A. halleri* and *A. lyrata* showcased the potential to be at least scanned by ribosomes, with half being translated. In spite of sequence divergence of the primary transcripts the event of translation in itself is conserved in about half of the cases.

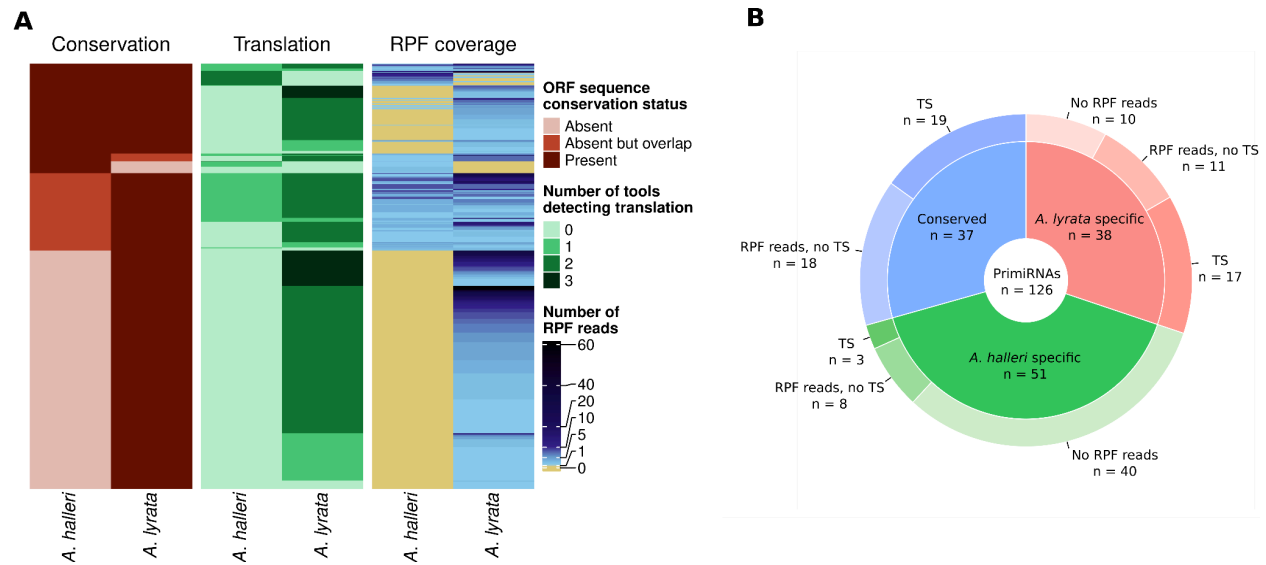


Figure 3: Complex heatmap of the conservation of *de novo* ORF sequence, translation and RPF coverage. Each line represents an individual ORF, and each column a species. Redundancy within each species was removed by keeping only the longest of overlapping *de novo* ORFs, and between species by removing duplicates of a same ORF found to be translated in several species. An ORF sequence was considered as conserved when its START and STOP positions, as well as its reading frame were conserved. If the ORF is conserved in a species, it is annotated as “present”, if not it is annotated as “absent”. If the ORF was not present at the exact same position in the pair species; but its coordinates were overlapped with another miORF annotation (that exhibited at least one RPF read), then the ORF was annotated as “Absent but overlap”. As translation detection tools rarely annotate the exact same ORF as being translated, a given tool was considered as having detected translation for an ORF either if the exact ORF was detected as translated, or if an overlapping ORF was detected as translated. PrimiRNAs of *A. lyrata* and *A. halleri* were classified according to the conservation of their precursors, and the presence within those primiRNAs of miORFs with RPF reads and potential TS.

The primary transcripts of young and old miRNAs showcase coding potential across species

To gain further insight into the conservation and expression of miORFs within the *Arabidopsis* genus, we decided to investigate in detail examples of evolutionarily old and young miRNA genes, and their corresponding expressed miORFs (**Figure 4A**). First we examined miR3440, a “young” miRNA gene, which is shared between *A. lyrata* and *A. halleri* only, and is missing from *A. thaliana*. The precursor, as in all following examples, was found to be at a syntenic position between *A. lyrata* and *A. halleri*. A 2,839 kb primary transcript could be annotated in *A. lyrata*, but we could not detect it in *A. halleri*, possibly due to its low expression and transient presence. 225 miORFs were annotated across the entire primary transcript, 15 of which were found to display TS and to lie within the 5' region of the pri-miRNA in *A. lyrata* (73 and 2 in *A. halleri*, respectively). These observations indicate that the region has a high potential of being translated, but since several of these ORFs form clusters and overlap, it remains difficult to pinpoint which ORF exactly is translated. The ORF candidate with the greatest number of RPF reads within this pri-miRNA sequence is ALY_miORF_1582, which is among the longest ORFs, was identified as translated by all three tools in *A. lyrata* (no evidence of translation could be found in *A. halleri*), and has a perfectly conserved amino-acid sequence between *A. halleri* and *A. lyrata* (**Figure 4A**). Hence, the miR3440-ALY_miORF_1582 system is an example of a pri-miRNA harboring a conserved miORF in its 5' region, with evidence of translation in at least one species, and a conserved precursor downstream which is shared between *A. halleri* and *A. lyrata*.

Another young miRNA gene, miR4222, is also only shared between *A. lyrata* and *A. halleri*, and has an annotated primary transcript in *A. lyrata* (**Figure 4B**). The precursor lies in a syntenic region but in opposite orientation in *A. lyrata* and *A. halleri*, suggesting that an inversion occurred within the region of miR4222. Again, hundreds of miORF candidates were annotated along the entire pri-miRNA transcript in both species. Whereas only a few RPF reads mapped to the transcript in *A. halleri*, five miORFs were shown to have TS in *A. lyrata*, within the 3' region of the primary transcript. The longest one, ALY_miORF_3780, was found to be translated by two tools, and harbors 30 RPF reads. While the precursor is well conserved and aligned between the two species, the downstream region containing ALY_miORF_3780 is difficult to align and the miORF is therefore likely absent in *A. halleri*.

We then examined two older miRNA genes. miR167d is conserved across flowering plants, and its primary transcript could be annotated in both *A. lyrata* and *A. thaliana* (**Figure 4C**). Open reading frames could be annotated all along its length, both within and outside the precursor, in all three species except within the *A. thaliana* precursor. Its mature and star sequences are separated by a large, 230 nucleotides long stem-loop, within which many miORFs are overlapped by RPF reads in *A. lyrata* and *A. thaliana*. One of those is the *A. lyrata* miORF_151, which is overlapped by five RPF reads, but does not showcase TS. Starting with a non-canonical “GTG”, the miORF sequence itself is perfectly conserved between *A. lyrata* and *A. thaliana*. In *A. halleri*, the reading frame is lost due to a two-nucleotides deletion; yet a downstream alternative “TTA” stop codon leads to a divergent miORF with a conserved start and a divergent stop. Interestingly, although no TS is detected in the *A. thaliana* primary transcript either, RPF reads accumulate at the ALY_miORF_151 sequence. Towards the 3’ end of the primary transcript of *A. lyrata*, another miORF, ALY_AL1G49310, previously annotated as a gene, exhibits both RPF reads and translation signal.

Similarly, miR160c is also conserved across flowering plants, and its primary transcript could also be annotated in every *Arabidopsis* species (**Figure 4D**). The primiRNA annotations vary in coordinates and length between species, with the longest annotation in *A. lyrata* and the shortest in *A. halleri*. Still, miORF candidates were annotated throughout the entire primiRNAs in each species. While the miRNA is located on the forward strand in both *A. lyrata* and *A. halleri*, it is on the reverse strand in *A. thaliana*. Many miORF candidates lie in the precursor sequences, and RPF reads are found in both *A. lyrata* and *A. halleri*. An example miORF candidate is ALY_miORF_4906, which overlaps the entire miR160c precursor. While it does not have TS itself, it overlaps with other miORFs that have TS: one in *A. lyrata*, AL8G11460, which was found to be translated by two tools and corresponds to a pre-existing gene annotation; the other in *A. halleri*, 671623, which was found to be translated by one tool, and also corresponds to a pre-existing gene annotation.

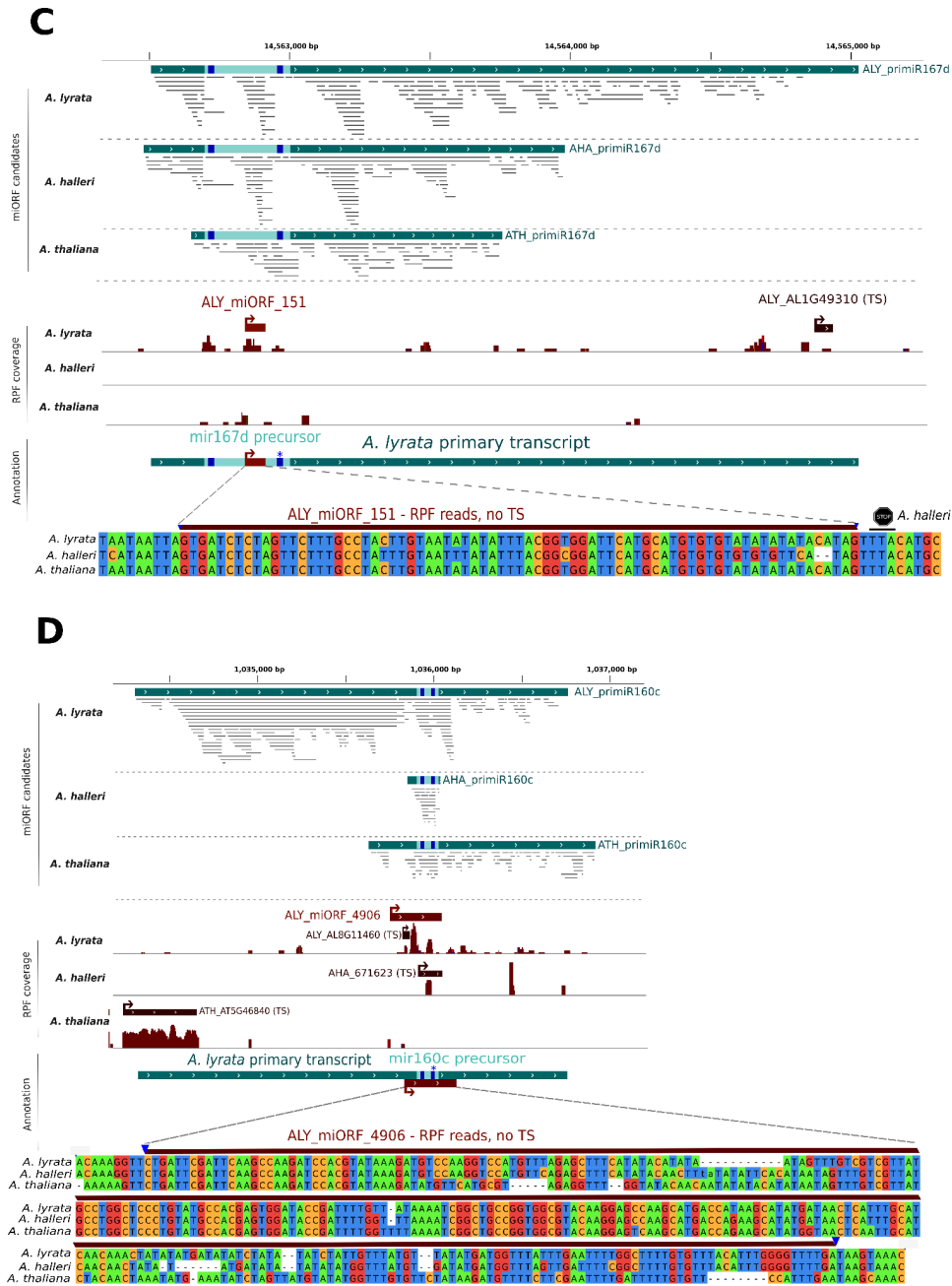


Figure 4: Visual representations of the genomic locations of primiRNAs across species, with miORF candidates (and the corresponding annotated/predicted primiRNAs based on precursor extensions), the RPF coverage and corresponding miORF annotations (either with reads and no TS, or with reads and TS), a schematic representation of the primiRNA, premiRNA, mature miRNA and star positions (indicated in dark blue, with an “*” indicating the star) and miORF of interest; and finally the DNA sequence alignment at the given ORF of interest. A) primi3440 and its candidate ALY_miORF_1582, B) primi4220 and its candidate ALY_miORF_3780, C) primiR167d and its candidate ALY_miORF_151, D) primiR160c and its candidate ALY_miORF_4906.

The miRNA-miPEP system can emerge from inverted gene duplications

To our knowledge, no study to date has investigated the potential origins of the miRNA-miPEP system. Among many potential mechanisms, we hypothesized that the inverted duplication of a CDS might give rise to a miRNA and a miORF simultaneously, each of them corresponding to a part of the former CDS. We were able to investigate two such instances in *A. thaliana*, with miR163 and miR822 (**Figure 5**), where both miRNAs might have originated from the inverted duplication of a CDS. In both cases, evidence of active translation was found for a miORF located within the 5' primiRNA region. In the case of miR163, the miRNA is a result of the inverted duplication of the second exon of the AT1G66700 CDS; and its upstream miORF of the direct duplication of the first exon. In the case of miR822, the miRNA is the result of the partial inverted duplication of the second exon of the AT2G40050 CDS; and its upstream miORF of the partial, direct duplication of a downstream region of the second exon. While miPEP163 has been functionally validated as a positive regulator of miR163's expression; no miPEP has yet been characterized for miR822 yet. This makes miR822 an excellent candidate for further investigation and functional validation of its potential miPEP.

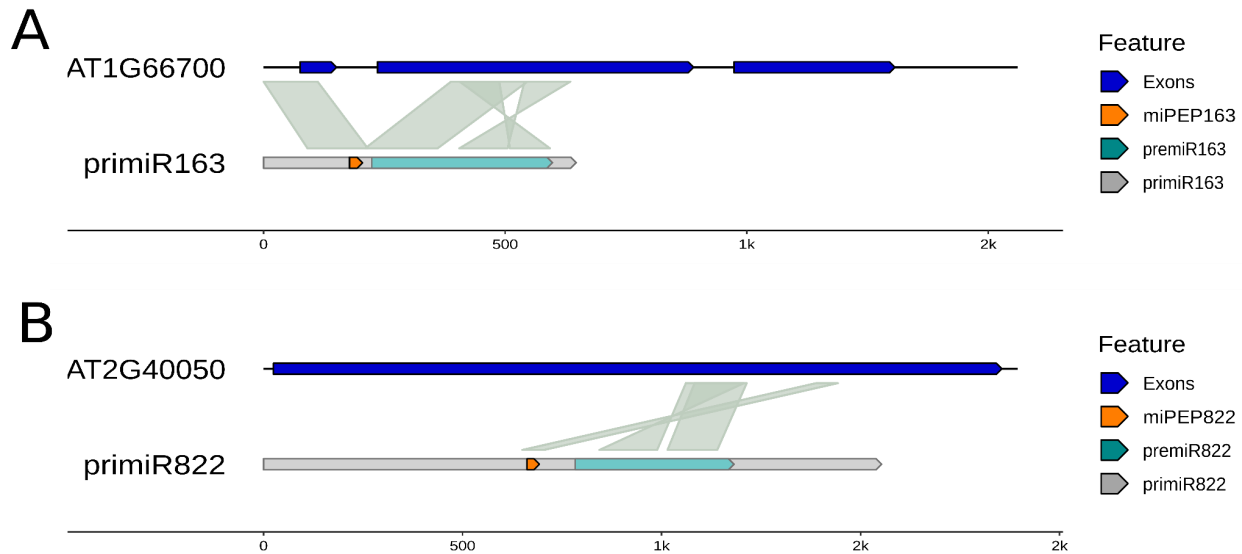


Figure 5: Representation of the sequence homology between protein coding genes and the miRNAs that were derived from them. A) The introns and exons of AT1G66700 are represented on the top panel, the black line representing the whole gene annotation. Regions of sequence homology represented by the gray intervals linking the top and bottom panels. miPEP163 is homologous to the 5' UTR and part of the first exon of AT1G66700, and the miR163 precursor is the result of both a direct and 2 inverted duplications of part of the second exon of AT1G66700. B) Similarly to AT1G66700, the gene structure

of one isoform of AT2G40050 is represented on the top panel (AT2G40050.1). In this isoform, there is only one exon, and miPEP822 corresponds to the direct duplication of a region of this exon, while the miR822 precursor is the result of both a direct and an inverted duplication of another region of this exon.

Discussion

miRNA loci produce a wide variety of transcripts - some of which can be coding

After the initial discovery of miPEPs (Lauressergues et al. 2015), a crucial question remained open as to the molecular mechanisms by which the miRNA-miPEP system operates: how could a pri-miRNA-encoded ORF be translated; when only the miRNA-miRNA* duplex gets exported to the cytoplasm, and not the pri-miRNA itself? More recently, evidence was found that miRNA loci that harbored upstream miORFs could lead to the production of two types of transcripts (Lauressergues et al. 2022). A first, long stem-loop including one that follows the canonical miRNA processing pathway, and leads to the production of a mature miRNA; and a second, short one that includes the miORF, but not the stem-loop. This second transcript type would therefore not undergo processing, and instead be directed to the cytoplasm where it would be translated into a miPEP, which then can be reintegrated into the cytoplasm to act as a regulator of its own miRNA's expression. Having annotated primary transcripts using long-read RNA sequencing, in three *Arabidopsis* species, and on a genome-wide scale, we have shown that primary transcripts are much longer than their precursor counterparts, with an average length of 1,350 nucleotides across species. Such length implies that potentially many distinct genomic features could be present on the same transcript as the miRNA precursor, including ORFs, and accordingly we revealed an abundance of miORF annotations all along the miRNA primary transcripts. Still, one would expect primary transcripts to remain within the nucleus, removing the possibility of translation; or if they did get exported to the cytoplasm, to have their stem-loop render the progression of ribosomes along the transcript impossible. Hence, while we chose to annotate pri-miRNA transcripts as the longest molecule we observed, it remains possible that a single miRNA locus can in fact be transcribed into several potential transcripts, some of which would have the potential to be coding.

One of the sources of transcript diversity is alternative splicing, which has been described in intergenic miRNAs that possess their own promoters, and similarly to messenger RNAs, are

transcribed by RNAPol II, capped and poly-adenylated (Zhang et al. 2015). In fact, splicing has been shown to be tightly linked to miRNA processing, with actors of both pathways like DCL1 and SE (Serrate) interacting in close proximity (Laubinger et al. 2008; Szarynska et al. 2009). Splicing has even been described as prevalent in plant miRNAs, concerning about half of known miRNAs in *Arabidopsis* (Stepien et al. 2017), with some authors even suggesting that “few [miRNA loci] have no introns” (Zhang et al. 2015). In *A. thaliana* alone, we found evidence of splicing (including alternative splicing events) in 20 primary transcripts annotated from our own data, and in 12 other primary transcripts from the available literature (Zielezinski et al. 2015); for a total of 32 intergenic primary transcripts with evidence of splicing (around a third of our 94 annotations). To better characterize the extent of alternative splicing and the diversity of pri-miRNA isoforms, it would now be necessary to increase the sequencing depth and include a larger diversity of plant tissues and developmental stages.

Regulating regulators: the high prevalence and functional importance of miORFs

To this date, 35 miPEPs have been functionally validated as enhancers of miRNA expression in *A. thaliana* (Lauressergues et al. 2015; Ormancey et al. 2020; Ormancey et al. 2021; Badola et al. 2022; Lauressergues et al. 2022). Among them, most were encoded by primary transcripts of miRNAs that are widely conserved across *Brassicaceae* species. Additional miPEPs had been predicted in younger miRNA genes, but their encoding miORFs had not been validated by translational evidence (Lauressergues et al. 2015). Here, we showcased that across miRNA conservation categories, miORFs are widely abundant throughout the entirety of miRNA primary transcripts, and that the pri-miRNAs are pervasively scanned by ribosomes. Ribosomes will preferably translate the miORFs that are localized within the 5' region of pri-miRNAs, upstream of the precursor; in line with the original model suggested by (Lauressergues et al. 2015). Hence, the regulation of miRNA transcription by pri-miRNA-encoded peptides is a common and general mechanism in the *Arabidopsis* genus.

This high prevalence begs the question of the functional importance of the regulation of miRNA expression by miPEPs. Typically, older miRNA genes are considered to be of greater functional importance than younger ones; and one of their characteristics is their relatively higher degree of expression (Fahlgren et al. 2010; Pavan et al. 2024). Strikingly, in our data, all of the pri-miRNAs that were conserved between *A. lyrata* and *A. halleri* harbored at least one miORF with at least one RPF read, and half with detectable TS; while in contrast that is the case of only

22% and 6% of the *A. halleri*-specific pri-miRNAs respectively. Interestingly, even within conserved miRNAs, it seems to be the event of translation that is conserved, and not the miORF sequence itself. Across species, translation happens within the same conserved pri-miRNAs, but not at the exact same miORF. This suggests a model where what matters is not the sequence of the miPEP itself, but rather its mere presence and ability to enhance the miRNA expression (Lauressergues et al. 2022).

An intriguing observation is that translated, potentially functional miORFs are more common in older miRNA genes. A first hypothesis could be that miRNAs gradually acquire functional importance throughout evolution, in a fashion similar to that of *de novo* genes in the “continuum” model of gene emergence (Carvunis et al. 2012; Pavan et al. 2024). Within the pri-miRNAs of functionally important miRNAs, some of the many pervasively scanned and translated miORFs would be under increasing selective pressure and would therefore be progressively refined to maintain consistent translation. In being translated at higher frequency and with greater accuracy, they would in turn enhance the expression of their functional miRNA and would, in a positive feedback-loop, be under greater selective constraints. A second hypothesis could be that functional miORFs harbor such characteristics because of the origin of their miRNA locus. While lowly functional miRNAs emerge at high rates, predominantly from transposable elements-related sequences (Pavan et al. 2025), some authors have proposed that miRNAs can also arise from the inverted duplication of CDS, which in turn become their instant targets (Nozawa et al. 2012). In the latter model, novel miRNAs targeting functionally critical CDS would first be deleterious and their expression maintained to low levels, before their expression level eventually increases. We were able to investigate two such instances, with miR163 and miR822. Strikingly, conservation analyzes revealed that the miORF that was translated, and for miR163 functionally validated as an enhancer (Lauressergues et al. 2022) corresponded to truncated and diverged exons of the CDS from which their miRNA had putatively emerged. Hence, in cases where the miRNA-miORF pair is the result of the rearrangement of already functional regions, the miRNA and its miORF would have already been shaped by selective constraints so as to not be too deleterious.

Overall, we have found that miORFs can be annotated pervasively in pri-miRNAs of several *Arabidopsis* species and that a significant portion displays evidence of active translation, especially in the 5' region of pri-miRNAs. These miORFs with translational potential are especially common in conserved pri-miRNAs, where their sequence itself is not conserved; but rather their existence. These miORFs may be encoded by alternative types of pri-miRNA

transcripts; through mechanisms like alternative splicing or alternative transcription start and termination sites. Their functional potential can be explained by a progressive functionalization concomitant to that of their miRNA; or by the mutational mechanisms that gave rise to their miRNA locus. Many other scenarios can be suggested for the emergence of the miRNA-miPEP system; some of which could include the emergence of the miRNA from the rearrangement of a CDS with an upstream ORF. MiRNAs themselves could emerge from snoRNAs (Scott and Ono 2011); or from MITEs (Pavan et al. 2025) and acquire down the line coding potential in ORF-containing flanking regions. Further validation experiments will now be necessary to validate the potential function of the predicted miPEPs, and supplementary investigation might shed light onto the population dynamics of the miRNA-miPEP system.

Material and Methods

Plant Materials and Growth Conditions

Three seeds of either *Arabidopsis thaliana* (accession Col-0); *Arabidopsis halleri*; *Arabidopsis lyrata* or *Capsella rubella* were deposited on 0.7% agar and imbibed in germination medium as described in (Conn et al. 2013); then left in obscurity at 4 °C for 2 days. The microcentrifuge lids were then transferred to hydroponic growth tanks in a greenhouse, where the seeds germinated and seedlings were grown hydroponically in a growth medium composed of 1 mM Ca(NO₃)₂, 0.5 mM MgSO₄, 3 mM KNO₃, 0.5 mM NH₄H₂PO₄, 0.1 µM CuSO₄, 0.1 µM NaCl, 1 µM KCl, 2µM MnSO₄, 25 µM H₃BO₃, 0.1 µM (NH₄)₆Mo₇O₂₄, 20 µM FeEDDHA, and 1 µM (*A. lyrata*, *A. thaliana* and *C. rubella*) or 10 µM (*A. halleri*) ZnSO₄. The pH of the solution was maintained at 5.0 using MES acid buffer (2 mM). Floral buds, leaves and roots were collected and flash frozen in liquid nitrogen progressively along the growth of the plants; over the course of 3 months (July to September). In total; 10 *A. thaliana*, 8 *A. halleri*, 8 *A. lyrata* and 6 *C. rubella* individuals were sampled from. Flash frozen samples were stored by tissue type; individual and sampling date, at -80°C.

Ribosome Profiling data preparation and analysis

Three individuals from each species were chosen for subsequent data production, based on the amount of material that had been gathered. Roots and leaves from the same single individual were gathered and were separately ground into a powder and were used to produce lysates. Root and leaf lysates were pooled per individual, ensuring a composition of 75% leaf-derived material, and 25% of material coming from roots. These lysates were then treated with 50 units RNase I Ambion (Invitrogen, AM2295) per 40µg of RNA to create ribosomal footprints.

These footprints were subsequently centrifuged and total RNA was extracted using a standard hot phenol procedure as described in (Baudin-Baillieu et al. 2016). The extracted RNAs were charged and migrated on 17% polyacrylamide; 7M TAE-urea gels. After migration, gel slices between the 28 and 30 nt markers were excised. The excised gel band was pulverized and eluded; and then spun on a wheel for a whole night. The gel residues were centrifuged several

times and the pellets were dissolved in 15 μ L of distilled water. The solubilized pellets were depleted in ribosomal RNAs according to the depletion protocol of the TruSeq® Stranded Total RNA Library Prep Plant kit (Illumina, 20020610). The obtained footprints were subsequently purified according to the following Zymo RNA clean and concentrator kit (Zymo Research; R1017) modified protocol; and used to prepare libraries according to the Biolabs NEBNext Multiplex Small RNA library Prep kit's standard protocol; with 12 PCR cycles. The resulting banks were purified with the Monarch PCR & cleanup kit (Biolabs).

From the obtained Ribosome Profiling libraries, adapters were trimmed from the reads and then filtered on quality and length. Taken from TAIR10 annotations, rRNA, tRNA, snoRNA, etc. sequences were removed through bowtie2 alignments (v2.4.1) (Langmead and Salzberg 2012). Depleted sequences were then aligned to their reference genome using the splice-aware aligner STAR (v2.7.5a) (Dobin et al. 2013), tolerating up to 1 mismatch in the alignment, and up to 20 multi-mapping positions. Raw sequencing data from (Hsu et al. 2016; Wu and Hsu 2022) (GEO accession GSE81332 and BioProject ID PRJNA854638) were downloaded from NCBI; and processed identically to the data produced in this study. All potential ORF candidates were bioinformatically annotated with RibORF (Ji 2018), on the basis of their sequence only, from the available genomes. A minimal length of 9 nucleotides was specified, and non-canonical START codons were allowed: TTG, CTG, GTG and ACG. Signatures of translation within those candidates were then detected by 3 tools: either ORFSCORE (Bazzini et al. 2014), RibORF or RiboCode (Xiao et al. 2018). Translation signal was predicted after p-site shifting the aligned reads with the previously assessed p-sites, among newly predicted candidates, with a minimal size of 9 nucleotides and allowing for non-canonical START codons. For a more comprehensive account of the RPF methodology, please refer to the Material and Methods section of the first chapter of this thesis.

Long-read Nanopore sequencing

Long RNAs were extracted from a set of biological replicates, i.e. individuals, from each species. In *A. lyrata*, leaves and roots were used from 4 separate biological replicates, and young leaves from a single individual. Leaves and roots were used from 3 biological replicates in *A. halleri*; and from 2 biological replicates in *C. rubella*. Independently, floral buds were used from a single individual of *A. thaliana* (col0). For all species except *A. thaliana*, libraries were

prepared from RNA extracted according to the "Direct cDNA Sequencing V14" protocol from the Nanopore ligation sequencing kit V14. They were then sequenced with R10 flowcells on MiniON. For *A. thaliana*, libraries were prepared from RNA extracted according to the manufacturer's instructions with the direct cDNA Sequencing kit (SQK-CDS109, Oxford Nanopores Technologies). They were then sequenced with R9 flow cells on a MiniON sequencing device.

miRNA precursor annotation

MiRNA precursor annotations for *A. lyrata*, *A. halleri* and *A. thaliana* were taken from previously published data (Pavan et al. 2024b). Shortly, small RNA sequencing data was either produced or taken from the public literature depending on the species, and was used to predict miRNA precursors using a combination of miRkwood (Guigon et al. 2019) and Shortstack 4.0.2 (Johnson et al. 2016). The annotated precursors were subsequently experimentally validated using Argonaute immunoprecipitation experiments, with anti-AGO1 and anti-AGO4 antibodies.

miRNA primary transcript annotation

Long RNAseq reads from *A. lyrata*, *A. halleri* and *C. rubella* were basecalled using high accuracy models (HAC) with both Dorado (Oxford Nanopores Technologies, version 0.7.0) and Guppy (Oxford Nanopores Technologies, version 6.5); and reads from *A. thaliana* were basecalled using HAC models with Guppy (Oxford Nanopores Technologies, version 4). An important rRNA contamination within certain samples required the removal of sequences with similarity to rRNAs, using SortMeRNA (v.4.3.6) (Kopylova et al. 2012). Raw sequencing data from (Parker et al. 2020) (European Nucleotide Archive, PRJEB32782) were downloaded from ENA and further processed identically to the data produced in this study. Reads were mapped to their respective genomes (*A. thaliana*; (Lamesch et al. 2012), *A. lyrata*; (Hu et al. 2011), *A. halleri*; (Pavan et al. 2024) and *Capsella rubella*; (Slotte et al. 2013)) using minimap2 (v.2.24) (Li 2018) with parameters `--MD -t 9 -G 1000 -ax splice`. After mapping, reads that had an identity percentage to the reference genome of less than 90%, or that had a mapping quality of 0, were discarded. Aligned reads were pooled per species, and were assembled into consensus transcriptomes with StringTie (v.2.2.1) (Pertea et al. 2015) with parameters `-L -m 100 -t -u -s 1`. The resulting transcript annotations were intersected with pre-existing miRNA precursor

annotations using the bedtools suite “intersect” (Quinlan and Hall 2010), and the transcripts overlapping premiRNAs were annotated as primiRNAs.

ORF conservation analysis

When a precursor within an annotated primiRNA was found to be conserved between *A.lyrata* and *A. halleri* (according to (Pavan et al. 2024b)), it was selected for miORF conservation analyses. If a primary transcript had been annotated in both species, their sequences were globally aligned using the ggsearch36 tool from the fasta36 software (v3.8i) (Pearson and Lipman 1988) with an e-value threshold of 0.05. If a primary transcript had been annotated in only one of two species, the precursor of the species lacking the annotation was then extended on both extremities by the corresponding lengths of the 3' and 5' annotated primiRNA regions; and the extended precursor was then aligned to the annotated primiRNA in the same manner. Once the pairwise alignments were obtained, they were used for conservation analyzes with the help of home-made python scripts. An ORF was considered to be conserved if its START and STOP positions, as well as its reading frame, were conserved in both species. The origins of the miRNA-miPEP system were investigated using miRNAs whose origin had been described by (Fahlgren et al. 2010), and the regions of homology were identified by performing BLASTn searches (Camacho et al. 2009) using the primiRNA annotations as queries. Plots were produced using R (v 4.2.2), with packages ggplot2 (Wickham 2009), ggsignif (Constantin and Patil 2021), ggpubr (Kassambara 2023), patchwork (Pedersen 2024), ComplexHeatmap (Gu et al. 2016), gggenomes (Hackl et al. 2024) and webr (Moon 2020), and alignment visualizations with Jalview (Waterhouse et al. 2009).

Bibliography

- Azevedo-Favory J, Gaspin C, Ayadi L, Montacci C, Marchand V, Jobet E, Rompais M, Carapito C, Motorin Y, Sjez-Vsquez J. 2021. Mapping rRNA 2'-O-methylations and identification of C/D snoRNAs in *Arabidopsis thaliana* plants. *RNA Biol.* 18:1760–1777.
- Badola PK, Sharma A, Gautam H, Trivedi PK. 2022. MicroRNA858a, its encoded peptide, and phytosulfokine regulate *Arabidopsis* growth and development. *Plant Physiol.* 189:1397–1415.
- Baudin-Baillieu A, Hatin I, Legendre R, Namy O. 2016. Translation Analysis at the Genome Scale by Ribosome Profiling. In: Devaux F, editor. *Yeast Functional Genomics*. Vol. 1361. *Methods in Molecular Biology*. New York, NY: Springer New York. p. 105–124. Available from: https://link.springer.com/10.1007/978-1-4939-3079-1_7
- Bazzini AA, Johnstone TG, Christiano R, MacKowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 33:981–993.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
- Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charleatoux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487:370–374.
- Conn SJ, Hocking B, Dayod M, Xu B, Athman A, Henderson S, Aukett L, Conn V, Shearer MK, Fuentes S, et al. 2013. Protocol: Optimising hydroponic growth systems for nutritional and physiological analysis of *Arabidopsis thaliana* and other plants. *Plant Methods* 9:1–11.
- Constantin A-E, Patil I. 2021. ggsignif: R Package for Displaying Significance Brackets for “ggplot2.” *PsyArxiv* [Internet]. Available from: <https://psyarxiv.com/7awm6>
- Cuperus JT, Fahlgren N, Carrington JC. 2011. Evolution and functional diversification of MIRNA genes. *Plant Cell* 23:431–442.
- Ding N, Zhang B. 2023. microRNA production in *Arabidopsis*. *Front. Plant Sci.* 14:1–11.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.
- Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. 2019. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* 29:932–943.

- Dziasek K, Santos-González J, Wang K, Qiu Y, Zhu J, Rigola D, Nijbroek K, Köhler C. 2024. Dosage-sensitive maternal siRNAs determine hybridization success in *Capsella*. *Nat. Plants* 10:1969–1983.
- Fahlgren N, Jogdeo S, Kasschau KD, Sullivan CM, Chapman EJ, Laubinger S, Smith LM, Dasenko M, Givan SA, Weigel D, et al. 2010. MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell* 22:1074–1089.
- Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 34:140–144.
- Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32:2847–2849.
- Guigon I, Legrand S, Berthelot J-F, Bini S, Lanselle D, Benmounah M, Touzet H. 2019. miRkwood: a tool for the reliable identification of microRNAs in plant genomes. *BMC Genomics* 20:532.
- Hackl T, Ankenbrand MJ, Adrichem B van. 2024. gggenomes: A Grammar of Graphics for Comparative Genomics. Available from: <https://CRAN.R-project.org/package=gggenomes>
- Hsu PY, Calviello L, Wu HYL, Li FW, Rothfels CJ, Ohler U, Benfey PN. 2016. Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 113:E7126–E7135.
- Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43:476–481.
- Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3:318–356.
- Ji Z. 2018. RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling. *Curr. Protoc. Mol. Biol.* 124:e67.
- Johnson NR, Yeoh JM, Coruh C, Axtell MJ. 2016. Improved Placement of Multi-mapping Small RNAs. *G3 GenesGenomesGenetics* 6:2103–2111.
- Kassambara A. 2023. ggpubr: “ggplot2” Based Publication Ready Plots. Available from: <https://CRAN.R-project.org/package=ggpubr>
- King M-C, Wilson AC. 1975. Evolution at Two Levels in Humans and Chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. *Science* 188:107–116.
- Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28:3211–3217.
- Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The *Arabidopsis* Information Resource

- (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* 40:1202–1210.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.
- Laubinger S, Sachsenberg T, Zeller G, Busch W, Lohmann JU, Ratsch G, Weigel D. 2008. Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 105:8795–8800.
- Lauressergues D, Couzigou JM, San Clemente H, Martinez Y, Dunand C, Bécard G, Combier JP. 2015. Primary transcripts of microRNAs encode regulatory peptides. *Nature* 520:90–93.
- Lauressergues D, Ormancey M, Guillotin B, San Clemente H, Camborde L, Duboé C, Tourneur S, Charpentier P, Barozet A, Jauneau A, et al. 2022. Characterization of plant microRNA-encoded peptides (miPEPs) reveals molecular mechanisms from the translation to activity and specificity. *Cell Rep.* 38.
- Lei T, Chang Y, Yao C, Zhang H. 2022. A systematic evaluation revealed that detecting translated non-canonical ORFs from ribosome profiling data remains challenging.
- Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100.
- Merchan F, Boualem A, Crespi M, Frugier F. 2009. Plant polycistronic precursors containing non-homologous microRNAs target transcripts encoding functionally related proteins. *Genome Biol.* 10:R136.
- Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C, et al. 2008. Sorting of Small RNAs into *Arabidopsis* Argonaute Complexes Is Directed by the 5' Terminal Nucleotide. *Cell* 133:116–127.
- Millar AA. 2020. The function of miRNAs in plants. *Plants* 9.
- Moon K-W. 2020. webr: Data and Functions for Web-Based Analysis. Available from: <https://CRAN.R-project.org/package=webr>
- Nozawa M, Miura S, Nei M. 2012. Origins and evolution of microRNA genes in plant species. *Genome Biol. Evol.* 4:230–239.
- Ormancey M, Guillotin B, San Clemente H, Thuleau P, Plaza S, Combier JP. 2021. Use of microRNA-encoded peptides to improve agronomic traits. *Plant Biotechnol. J.* 19:1687–1689.
- Ormancey M, Le Ru A, Duboé C, Jin H, Thuleau P, Plaza S, Combier JP. 2020. Internalization of miPEP165a into *arabidopsis* roots depends on both passive diffusion and endocytosis-associated processes. *Int. J. Mol. Sci.* 21:1–16.
- Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, Hall AJW, Barton GJ, Simpson GG. 2020. Nanopore direct RNA sequencing maps the complexity of

- arabidopsis mRNA processing and m6A modification. *eLife* 9:1–35.
- Pavan F, Azevedo Favory J, Lacoste E, Beaumont C, Louis F, Blassiau C, Cruaud C, Labadie K, Gallina S, Genete M, et al. 2024a. The evolutionary history and functional specialization of microRNA genes in *Arabidopsis halleri* and *A. lyrata*. Available from: <http://biorxiv.org/lookup/doi/10.1101/2024.05.03.592357>
- Pavan F, Lacoste E, Castric V, Legrand S. 2025. Scenarios for the emergence of new miRNA genes in the plant *Arabidopsis halleri*. *Plant Cell Physiol.*:pcaf004.
- Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* 85:2444–2448.
- Pedersen TL. 2024. patchwork: The Composer of Plots. Available from: <https://patchwork.data-imaginist.com>
- Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33:290–295.
- Qu G, Kruszka K, Plewka P, Yang SY, Chiou TJ, Jarmolowski A, Szweykowska-Kulinska Z, Echeverria M, Karlowski WM. 2015. Promoter-based identification of novel non-coding RNAs reveals the presence of dicistronic snoRNA-miRNA genes in *Arabidopsis thaliana*. *BMC Genomics* 16:1–15.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
- Ruiz-Orera J, Albà MM. 2019. Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. *Trends Genet.* 35:186–198.
- 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.* 2:890–896.
- Scott MS, Ono M. 2011. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. *Biochimie* 93:1987–1992.
- Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* 45:831–835.
- Stepien A, Knop K, Dolata J, Taube M, Bajczyk M, Barciszewska-Pacak M, Pacak A, Jarmolowski A, Szweykowska-Kulinska Z. 2017. Posttranscriptional coordination of splicing and miRNA biogenesis in plants. *Wiley Interdiscip. Rev. RNA* 8.
- Szarzynska B, Sobkowiak L, Pant BD, Balazadeh S, Scheible WR, Mueller-Roeber B, Jarmolowski A, Szweykowska-Kulinska Z. 2009. Gene structures and processing of *Arabidopsis thaliana* HYL1-dependent pri-miRNAs. *Nucleic Acids Res.* 37:3083–3093.
- Wang J, Mei J, Ren G. 2019. Plant microRNAs: Biogenesis, homeostasis, and degradation.

Front. Plant Sci. 10:1–12.

- Wang Z, Butel N, Santos-González J, Borges F, Yi J, Martienssen RA, Martinez G, Köhler C. 2020. Polymerase IV Plays a Crucial Role in Pollen Development in *Capsella*. *Plant Cell* 32:950–966.
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189–1191.
- Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer New York Available from: <https://link.springer.com/10.1007/978-0-387-98141-3>
- Wu HYL, Hsu PY. 2022. A custom library construction method for super-resolution ribosome profiling in Arabidopsis. *Plant Methods* 18:1–8.
- Xiao Z, Huang R, Xing X, Chen Y, Deng H, Yang X. 2018. De novo annotation and characterization of the translome with ribosome profiling data. *Nucleic Acids Res.* 46:E61.
- Xie Z, Allen E, Fahlgren N, Calamar A, Givan SA, Carrington JC. 2005. Expression of Arabidopsis MIRNA genes. *Plant Physiol.* 138:2145–2154.
- Yu H, Gerstein M. 2006. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci.* 103:14724–14731.
- Zhang H, Dou S, He F, Luo J, Wei L, Lu J. 2018. Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during Drosophila development. Chen X, editor. *PLOS Biol.* 16:e2003903.
- Zhang S, Liu Y, Yu B. 2015. New insights into pri-miRNA processing and accumulation in plants. *Wiley Interdiscip. Rev. RNA* 6:533–545.
- Zielezinski A, Dolata J, Alaba S, Kruszka K, Pacak A, Swida-Barteczka A, Knop K, Stepień A, Bielewicz D, Pietrykowska H, et al. 2015. mirEX 2.0 - an integrated environment for expression profiling of plant microRNAs. *BMC Plant Biol.* 15:144.

Supplementary material

	Tissue type	Sample number	Library type	Nanopore chemistry	Raw reads	Reads after rRNA depletion	Reads primarily mapped to genome	Mapped reads after quality control (ID>90% and mapping quality > 0)	% of depleted reads mapped to genome after QC	Reference
<i>A. halieri</i>	Roots +Leaves	AHA_1	cDNA direct - from total RNA	R10	4,345,815	1,502,912	1,134,177	994,583	66.18%	This study
	Roots +Leaves	AHA_2	cDNA direct - from total RNA	R10	4,461,851	890,106	705,876	604,971	67.97%	This study
	Roots +Leaves	AHA_3	cDNA direct - from total RNA	R10	1,509,664	429,451	374,790	338,940	78.92%	This study
				Total	10,317,330	2,822,469	2,214,843	1,938,494	68.68%	
<i>A. lyrata</i>	Roots +Leaves	ALY_1	cDNA direct - from total RNA	R10	1,742,421	451,827	372,868	295,364	65.37%	This study
	Roots +Leaves	ALY_2	cDNA direct - from total RNA	R10	3,837,653	1,318,687	1,010,113	790,938	59.98%	This study
	Roots +Leaves	ALY_3	cDNA direct - from total RNA	R10	202,770	116,900	101,344	67,196	57.48%	This study
	Roots +Leaves	ALY_4	cDNA direct - from total RNA	R10	2,961,037	806,336	716,008	555,281	68.86%	This study
	Young leaves	ALY_5	cDNA direct - from total RNA	R10	2,378,347	661,679	561,884	377,450	57.04%	This study
				Total	11,122,228	3,355,429	2,762,217	2,086,229	62.17%	
<i>A. thaliana</i>	Flower buds	ATH_1	PCR - from total RNA	R9	6,974,240	6,338,478	6,181,797	5,586,133	80.10%	This study
	Flower buds	ATH_2	PCR - from total RNA	R9	2,101,632	1,909,037	1,846,204	1,622,210	77.19%	This study
	Flower buds	ATH_3	PCR - from polyA purified RNA	R9	3,108,951	3,044,649	2,881,616	2,531,200	81.42%	This study
	Flower buds	ATH_4	PCR - from polyA purified RNA	R9	1,546,328	1,515,358	1,375,509	1,172,744	75.84%	This study
	Flower buds	ATH_5	cDNA direct - from polyA purified RNA	R9	8,389,309	7,709,291	7,658,641	6,937,803	89.99%	This study
	Flower buds	ATH_6	RNA direct - from total RNA	R9	1,718,132	1,709,113	1,676,297	813,836	47.37%	Parker <i>et al.</i> , 2020
				Total	23,838,592	22,225,926	21,620,064	18,663,926	83.97%	
<i>C. rubella</i>	Roots +Leaves	CRU_1	cDNA direct - from total RNA	R10	6,635,748	3,592,759	3,190,399	2,822,418	78.56%	This study
	Roots +Leaves	CRU_2	cDNA direct - from total RNA	R10	6,550,075	2,840,576	2,599,640	2,348,760	82.69%	This study
				Total	13,185,823	6,433,335	5,790,039	5,171,178	80.38%	

Table S1: Summary of the long-read RNA sequencing data produced in this study and publicly available; used for primiRNA annotation.

	Annotated premiRNAs	Annotated premiRNAs outside CDS	Annotated premiRNAs outside CDS	Annotated premiRNAs outside CDS with miORFs translated in the reference species	miORFs	% of premiRNAs with miORFs
<i>A. halleri</i>	463	373	64	6	6	9.38%
<i>A. lyrata</i>	276	211	62	42	180	67.74%
<i>A. thaliana</i>	418	353	94	12	18	12.77%

Table S2: Summary of the premiRNA, primiRNA and miORF annotations across *Arabidopsis* species.

Perspectives

This second chapter is a preliminary draft of a future paper, and several supplementary analyzes are planned and will be integrated into the final version of the paper, which are detailed below:

Exploitation of the full long-read RNA sequencing data

Nanopore long-read data has been produced, but not yet exploited in *C. rubella*. While miRNA precursor repertoires were previously annotated in *Arabidopsis* species, this was not the case for *C. rubella*. Making use of public small RNA sequencing data from *C. rubella* (Fahlgren et al. 2010; Wang et al. 2020; Dziasek et al. 2024), a novel miRNA precursor repertoire will be annotated and combined with our long-read RNA sequencing data to annotate full pri-miRNAs.

More precise and exhaustive characterization of alternative transcription

Signs of prevalent splicing have been found within a third of *A. thaliana* pri-miRNA annotations. However, there has been no quantitative, automated analysis of alternative splicing put into place yet. In further analyses, scripts and pipelines will be developed to do so, not just in *A. thaliana* but also in all remaining species. In regards to miORFs, splicing is crucial in the creation or destruction of reading frames. It has been previously reported that the skipping of an exon within miR156c results in the formation of an ORF; while the inclusion of an exon in miR166b also results in the formation of an ORF (Szczepanska et al. 2009). In our own data, only looking at the 20 pri-miRNAs we annotated to harbor splicing, 585 miORFs were annotated when skipping all introns, while 733 were when retaining all introns. In miR839, TS were found for miORFs that could exist only in the event of intron retention, as well as in miR822; while in others such as miR168a and miR319a, miORFs entirely contained within introns were found to have RPF reads. In addition to splicing, alternative TSS, TTS and poly-adenylation sites have also been described within plant pri-miRNAs (Stepien et al. 2017). We were indeed able to identify the presence of many transcripts of varying lengths coming from the same miRNA locus, as is the case in miR172b, miR164b or miR839. Interestingly, in these cases, some transcript variants did not contain the stem-loop, but contained many miORF annotations, some of them upstream, and others downstream of the precursor position, as had been described in

(Lauressergues et al. 2022). Again, scripts will be developed to quantitatively assess alternative transcription events outside of alternative splicing.

Extension of conservation analyzes: more analyzes and more species

Current conservation analyzes have not yet characterized how miRNA conservation specifically relates to miORF conservation and expression. Statistical analyzes will be performed to assess whether a link truly exists between higher miRNA conservation and higher rates of miORF expression. Preliminary analyzes were performed in *A. thaliana* and *A. halleri*, comparing the percentage of identity (PID) along the primary transcripts, for precursors that were conserved between the two species (**Figure 5**). We found that while the mature miRNA and miRNA* positions had a very high PID (95%), the PID gradually decreases along the transcripts the further from the miRNA/miRNA* duplex we get. Strikingly, the 5' primiRNA region, while less conserved than the precursor and the miRNA/miRNA* duplex, was significantly more conserved than the 3' primiRNA region, and than both flanking and random intergenic regions, even when matched for length. Hence, the lack of miORF sequence conservation is not particularly surprising, even between conserved miRNAs, as we have assessed that most miORFs with TS are localized within the 5' regions of primiRNAs, which were found to be much less conserved than the miRNA/miRNA* duplex through those analyzes. All of the aforementioned analyzes will be performed on all species with available long-read RNA sequencing data, i.e. *A. thaliana*, *A. halleri*, *A. lyrata* and *C. rubella*.

Genome-wide investigation of the miRNA-miPEP system's origins

We have been able to investigate in detail the mutational mechanisms that led to the emergence of two examples of a miRNA-miPEP system. Down the line, those analyzes will be extended to all miRNAs for which we have access to the mechanisms that led to their emergence, as characterized by (Pavan et al. 2025). In doing so, we expect to possibly uncover examples of other emergence mechanisms, and perhaps even suggest new models for the emergence of the miRNA-miPEP system.

Conclusions and perspectives

In this thesis, I explored the fundamental role that short, non-canonical ORFs can play in the emergence of novel phenotypes, both through the emergence of novel genes, as well as novel regulatory elements. Using Ribosome Profiling sequencing, I have annotated both canonical and non-canonical translation events, at a genome wide scale, across *Arabidopsis* species. This study uncovered hundreds of translated short ORFs, which could then provide insight into the very first steps of *de novo* gene emergence, as well as into the coding potential of miRNA primary transcripts. Using synteny analysis, I uncovered various conservation levels of translated *de novo* ORFs and characterized the properties of this starting material for *de novo* gene birth within the *Arabidopsis* genus. In regards to the emergence of potential new regulatory elements, I annotated 220 miRNA primary transcripts with both a coding and non-coding role, by combining long-read RNA-based sequencing, miRNA genes annotations and the repertoire of short ORFs detected as translated. With ribosome profiling data, I have identified a repertoire of newly annotated potential miPEPs; and investigated its conservation. To further assess the evolutionary dynamics and functional potential of those ORFs, both experimental data acquisition and further analyzes are planned for the near future; and will be detailed in the perspectives section below.

Conclusion

1. *de novo* gene emergence from translated short ORFs

In the first chapter of my thesis, leveraging the accuracy of ribosome profiling (RPF) data and synteny analyses, I annotated a numerous and diverse repertoire of *de novo*-emerged short ORFs within the *Arabidopsis* genus. Acting as the raw material for the *de novo* gene birth process, those ORFs proved to exhibit many characteristics similar to those of untranslated, neutrally evolving intergenic ORFs (IGORFs): a high prevalence of non-canonical START codons, a shorter length, a lower degree of GC content, an overall lower expression level and a higher degree of disorder relatively to canonical CDS. Interestingly, there seemed to be a continuum of GC-content levels between conserved CDS, progressively less conserved *de novo* ORFs and untranslated IGORFs, which is in line with the “continuum” model of *de novo* gene emergence. Translated *de novo* ORFs seemed to emerge frequently, with most sequences

being species-specific. The presence of a few indels disrupting the reading frame in neighbouring species, suggests a high rate of birth and death. Among them, a few ORFs stood out as being the most promising candidates for proto-gene emergence: they exhibited greater lengths, were detected as translated by several tools, and some had even already been annotated as genes. The exact mechanisms by which these promising candidates came to acquire those characteristics remains to be elucidated: either they progressively increased in length, expression levels and conservation (the *continuum* model); or they originated directly with those characteristics and were the ones whose expression was maintained while candidates without those characteristics quickly lost expression (the *born this way* model). However, these two models are difficult to distinguish and are not mutually exclusive. Promising examples included the *A. halleri*-specific intergenic ORF “IGORF_162”, which was located directly downstream of a protein-coding gene. This ORF might have benefited from the expression of this upstream gene thanks to “translational read-through”, and raises the interesting question of whether translated *de novo* ORFs are preferably localized within specific regions of the genome.

2. The importance of alternative miRNA primary transcripts and their translation in the regulation of mature miRNA levels

In the second chapter of my thesis, I showed that the coding potential of primiRNA transcripts was ubiquitous across *Arabidopsis* species; and seemed to particularly be common in transcripts that contained conserved miRNA precursors. Specifically, I found that miRNA expression could be regulated in at least two ways : through the presence of translated miORFs, which can act as enhancers of miRNA transcription; and through the structure of primiRNA transcripts, whose splicing and processing may enhance or inhibit mature miRNA levels.

While the sequence of the miORFs themselves was conserved in only about a third of cases, the fact that translation events themselves were happening, either within the same approximate region of the primiRNA or in a different part, was highly conserved. Preliminary analyzes between *A. thaliana* and *A. halleri* primiRNAs showed that the miORFs’ sequence was not significantly more conserved than other positions of the primary transcripts, and their conservation levels were best explained by the conservation level of their associated premiRNA.

Indeed, when performing a linear regression between the identity percentage (PID) of miRNA precursors (premiRNAs) and of miORFs, the premiRNA PID alone could account for half of the observed variance in miORF PID (**Figure 1**, $R^2 = 0.48$, $p < 0.001$). Hence, what seems to matter for the upregulation of the miRNA is not the sequence and properties of the miPEP; but rather the fact that this miPEP specifically interacts physically with the miORF it is encoded by (Lauressergues et al. 2022). Having observed that conserved miRNAs are particularly likely to harbor translated miORFs, it might be possible that they are under selective pressure to be expressed at higher levels, and that one of the ways they achieve this is by encoding specific up-regulating miPEPs.

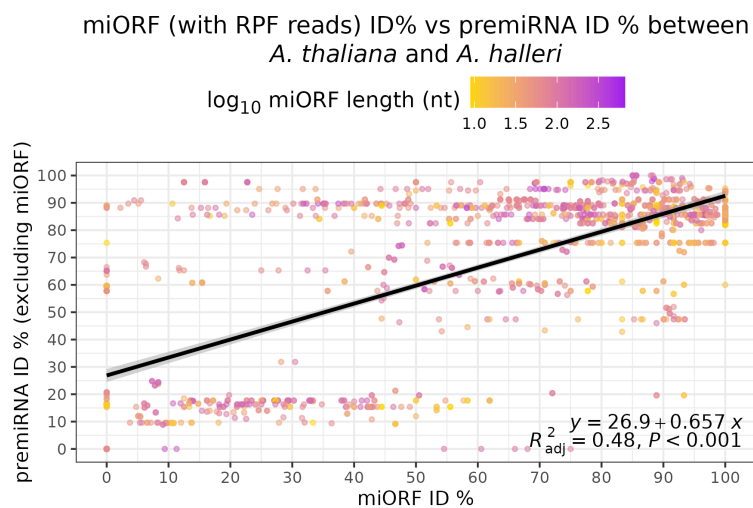


Figure 1: Linear regression of the identity (ID) percentage between *A. thaliana* and *A. halleri*, within the miORFs (with at least 1 RPF read), VS within the premiRNAs (excluding miORFs), on the same premiRNAs. The PID is a proxy for the level of conservation between the two species. miORFs do not tend to be significantly more conserved than premiRNAs, and half of the variance in their conservation level can be explained by the conservation level of the premiRNA with which they are associated.

To be able to produce expression-enhancing miPEPs, miRNA loci have been reported as being translated into two types of transcripts: a long, stem-loop including one, and a short, alternatively spliced other (Lauressergues et al. 2022). This second, alternatively spliced transcript type is the one exported to the cytoplasm, where it can be translated into a miPEP. Hence, premiRNA splicing may also play a role in the accumulation of mature miRNA molecules, and the associated phenotypes (Lauressergues et al. 2022). Most plant miRNAs are independent transcription units, that possess exons, introns, and their own promoters, and that are transcribed by RNAPol II, capped and poly-adenylated (Zhang et al. 2015; Stepien et al. 2017). In being such, their primary transcripts are highly similar to messenger RNAs, and thus the question of the importance of splicing naturally comes up. Throughout the past decade, increasing evidence has accumulated that splicing plays a key role in the processing of premiRNAs; through a range of factors. Splicing actors, like proteins from the Cap Binding Complex (CBC, e.g. CBP20 and CBP80/ABH1) play a role in mature miRNA accumulation

through processing. Processing actors like Serrate (SE) (Laubinger et al. 2008) or Hyponastic Leaves 1 (HYL1) (Szarzynska et al. 2009) are known to also be involved in splicing. As such, pri-miRNA splicing and processing are effectively coupled, and dependent on each other. While both spliced and unspliced pri-miRNAs can be processed, some spliced variants are not functional, meaning splicing can be a necessity for proper mature miRNA accumulation (Zhang et al. 2015). Additional actors were later identified as key components of both splicing and mature miRNA accumulation, like SMA1 (Li et al. 2018) or Serrate-associated Protein 1 (SEAP1) (Li et al. 2021), further underlying the simultaneity of the transcription, splicing and processing of miRNA primary transcripts (Ding and Zhang 2023).

Splicing has even been described as a key regulation mechanism of mature miRNA levels (Schwab et al. 2013). Examples like the intronic miR400 exhibit stress-induced alternative splicing, which generates transcripts with partially retained precursors leading to lower levels of mature miRNA (Yan et al. 2012); or the miR162a which is located within a protein coding gene intron, whose distinct alternative splicing in stress conditions favours functional isoforms that retain the miRNA-coding intron and a distal polyA site, leading to higher levels of mature miRNAs (Barciszewska-Pacak et al. 2016). In other examples, like miR163 and miR161; a splicing event within the pri-miRNA, triggered by stress, is absolutely necessary for the mature miRNA to be able to accumulate (Bielewicz et al. 2013).

Interestingly, the position of the precursor stem-loop within the wider primary transcript has been described as key to the progression of the miRNA processing (Zhang et al. 2015). Notably, as most miRNA loci are described as possessing at least one intron, the relative position of the stem-loop relative to the start of the intron is of great importance. Whether the miRNA is localized upstream, or within/downstream this intron determines whether the interaction of SE with the CBC is going to have a positive or a negative effect on mature miRNA accumulation (Stepien et al. 2017). In cases where the miRNA is upstream of the intron, splicing is mostly going to have a positive effect on mature miRNA accumulation. In cases where the miRNA is either within or downstream of this intron, splicing is mostly going to have a negative effect on mature miRNA accumulation (**Figure 2**). In the literature, most stem-loops have been described as being upstream of the start of the intron; including in miR163 and miR172a (Schwab et al. 2013), both of which were proven to possess a peptide-encoding miORF which was validated as a positive regulator of their transcription (Lauressergues et al. 2022).

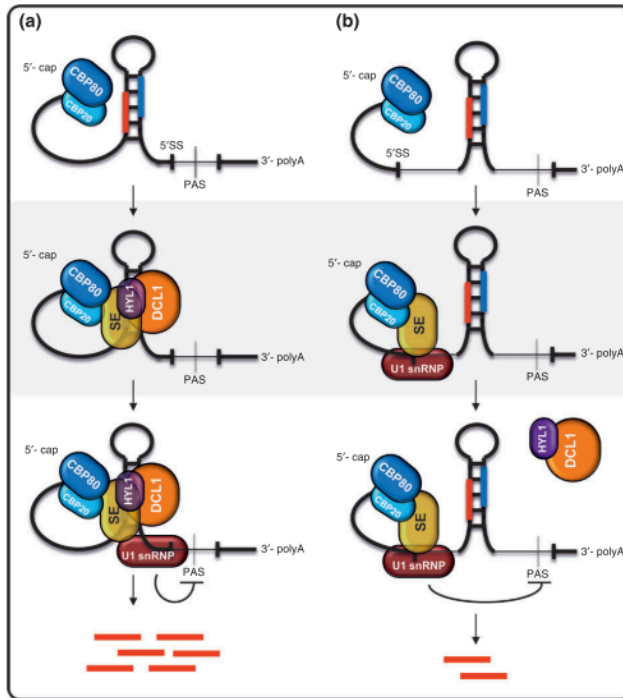


Figure 2: Proposed model of the crosstalk between the miRNA processing actors (SE, HYL1, DCL1) and the splicing actors (CBP20 and 80, U1 snRNP). 5'SS designates the 5' splice site, i.e. the start of the first intron, and PAS designates the polyadenylation sites. a) represents a miRNA stem-loop upstream of the first intron, b) represents a miRNA within the first intron. The red bars represent the mature miRNA. Figure from (Stepien et al. 2017).

Having produced pri-miRNA annotations, and knowing about the influence of precursor position on mature miRNA accumulation, I investigated the position of precursors within pri-miRNAs. I found out that older, more conserved miRNAs were significantly more likely to be located within the first exon, when compared to younger ones (**Figure 3**).

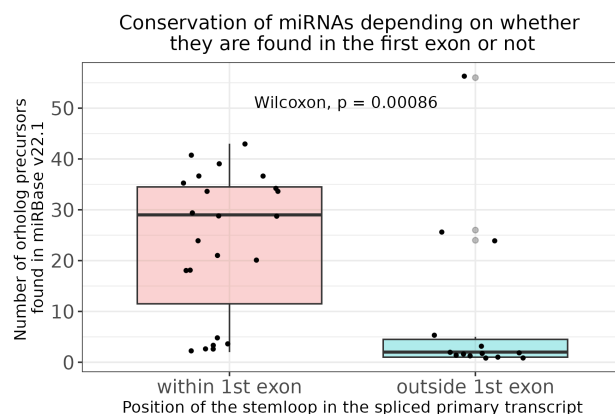


Figure 3: Degree of conservation of miRNAs (approximated by the number of ortholog precursors found in miRBase v.22.1) depending on the position of the stem-loop relative to the first exon of the pri-miRNA. The difference is statistically significant (Wilcoxon-ranked test, $p < 0.001$).

My hypothesis thus became the following: we know that older miRNA genes tend to be expressed at higher levels than younger ones (Ma et al. 2010; Cuperus et al. 2011; Pavan et al. 2024), and we also observe that older miRNA genes seem to be more likely to be located within the first exon of their pri-miRNA. Perhaps, one of the ways miRNA genes would gradually become more functional, and consequently more conserved, is through the selection and conservation of specific splicing sites? To correct for the expression levels, I compared the position of older and younger miRNAs that were expressed at similar levels. Again, I was able to verify the previous trend in which older miRNAs are more likely to be located within the first exon of the pri-miRNA (**Figure 4**). Further experimental validation would be needed to prove that the localization of the stem-loop is itself enough to explain significant differences in expression levels, and this is so far only speculative.

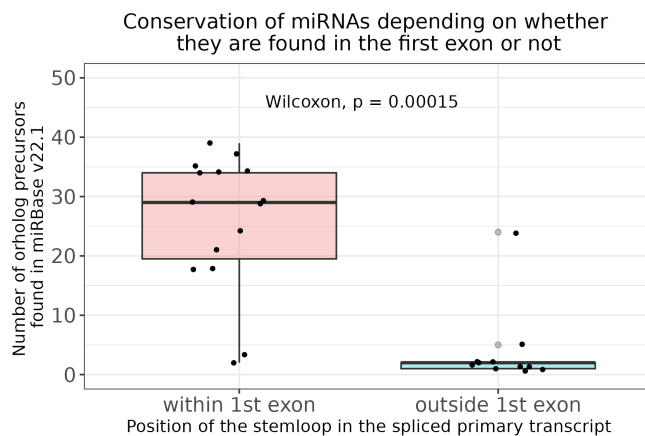


Figure 4: Degree of conservation of miRNAs (approximated by the number of ortholog precursors found in miRBase v.22.1) depending on the position of the stem-loop relative to the first exon of the pri-miRNA; but this time comparing only pri-miRNAs within a similar expression range. The difference is statistically significant (Wilcoxon-ranked test, $p < 0.001$).

In conclusion, I propose two ways by which conserved miRNAs could become more expressed than younger ones : through the greater prevalence of translated miORFs, which may act as positive regulators of miRNA transcription; and through the upstream localization of the stem-loop relative to the first intron, which favours splicing conducive to mature miRNA accumulation.

Perspectives

This thesis has given us insight into the roles short, translated ORFs may play in the emergence of novel protein-coding genes and regulatory elements. On one hand, they can act as the raw material for *de novo* gene birth, with hundreds of unconserved *de novo* ORFs exhibiting several characteristics similar to that of untranslated intergenic ORFs, yet a few dozen being promising proto-gene candidates. On the other hand, they are pervasively found in miRNA primary transcripts, where they tend to be preferably localized in the 5' region, and are rarely conserved between species. This work raises new questions, such as the prevalence of translated short ORFs in more transcriptionally active tissues, especially reproductive ones, the functional potential of the polypeptides encoded by proto-gene candidates, or the nature of evolutionary forces acting on them. Furthermore, several scenarios remain to be investigated in regards to the origins and the evolutionary dynamics of the miRNA-miPEP system, and many newly annotated miPEPs have yet to be experimentally validated.

1. *de novo* ORF-encoded peptides: population dynamics and functional potential

In the near future, we will analyze the genome-wide translational landscape of more tissues that are known to be more highly translationally active than mature leaves, i.e. seedlings and flower buds of *A. lyrata* and *A. halleri*. The material has been collected, and we will produce new ribosome profiling data from these tissues which, according to what was described in (Lecampion et al. 2016), may be more promising in terms of quality, that is both clear periodicity and high expression levels (collaboration with Olivier Namy's team, Paris-Saclay). We are also currently generating Mass Spectrometry data (MS) from *A. lyrata* and *A. halleri* to confirm the existence of the predicted *de novo* peptides we have identified (collaboration with Fabrice Bray, University of Lille). Alphafold2 predictions were also performed by Lars Eicholt (Erich Bornberg-Buaer lab, Institute for Evolution and Biodiversity, Muenster), and will allow us to further investigate the structural properties of *de novo* peptides, specifically those of the most promising candidates.

We will also perform new analyzes to investigate whether the genomic context may influence the emergence of *de novo* ORFs. The genomic localization of *de novo* ORFs is of particular interest, as *de novo* genes have been characterized in the past as emerging preferentially within

recombination hotspots in yeast (Vakirlis et al. 2018). The availability of population genomic data for *A. thaliana*, *A. halleri* and *A. lyrata* (Le Veve et al. 2023), will allow us to assess the nature of evolutionary forces acting on *de novo* ORFs (e.g. by performing McDonald–Kreitman tests, (Sawyer 1994)); something that has seldom been investigated in the *de novo* gene emergence field (Begun et al. 2006; Zhao et al. 2014; Li et al. 2016). The literature suggests that the majority of sORFs evolves neutrally (Ruiz-Orera et al. 2018; Durand et al. 2019), and this would allow us to test if we observe the same pattern in plants .

2. The origins, conservation and functional potential of the miRNA-miPEP system

The new RPF and MS data, as well as populational data, will be used for further analyzes enabling us to better understand the importance of miPEPs as potential regulatory elements. We will annotate their repertoire, validate their existence and presence in cells with MS, and characterize their evolutionary dynamics in relation to that of their precursor. We will also further investigate the link between miRNA and miORF conservation, and the evolutionary dynamics of the miRNA-miPEP system both within the *Arabidopsis* genus, and at a wider scale within *Brassicaceae*. PremiRNA and primiRNA annotations will also be produced in *C. rubella*, providing us with an outgroup species for conservation analyzes; and alternative transcription will be investigated across species to better characterize primiRNA isoforms.

Additionally, when investigating the potential origins of the miPEP-miRNA system, I identified examples of miORF-miRNA pairs that derived from the partial, inverted duplication of pre-existing CDS. Yet, other models of origin might be possible, and miRNAs have been reported to predominantly emerge from pre-existing transposable elements (Pavan et al. 2025). A deeper, broader investigation would most likely yield examples of those different mechanisms. We might identify miORFs that emerged *de novo*, or that correspond to former upstream ORFs. When investigating miRNA primary transcripts, I found evidence of polycistronic transcripts that contained both miRNA precursors, and either H/ACA or C/D box snoRNAs, such as primiR775 and primiR779 which had already been described in the literature (Qu et al. 2015). Most plant snoRNAs come from long, polycistronic transcripts (Qu et al. 2015) and are involved in a range of functions including but not limited to rRNA modification, pre-rRNA processing, mRNA methylation and mRNA binding (Zhu et al. 2016). This raises the question of whether these two

miRNAs (miR775 and miR779) could be somehow related to snoRNA functions? Indeed, a number of observations seem to link both snoRNAs and miRNAs, such as the presence of snoRNA-associated ribonucleoproteins (snoRNPs) in AGO complexes, the association of snoRNA-derived small RNAs with AGO7 and RISC complexes (Taft et al. 2009), or the ability of some miRNA precursors to bind to core snoRNPs (Scott and Ono 2011). Therefore, an evolutionary spectrum between snoRNAs and miRNAs was suggested by Scott and Ono (2011), according to which some snoRNAs might exhibit miRNA features, some miRNAs might exhibit snoRNA features, and some regulatory elements might act as dual-function sno-miRNAs (Scott and Ono 2011). During my PhD, I identified a particularly interesting example of a potential dual sno-miRNA: miR4228. This miRNA precursor is conserved in many Brassicaceae species (highly so within the *Arabidopsis* genus), has several predicted targets found by TargetFinder, is predicted by miRkwood in 17 different *A. thaliana* sRNAseq samples, and is enriched in AGO2 Immunoprecipitation experiments (Annacondia and Martinez 2021). It is on the same primary transcript as two annotated C/D box snoRNAs, and itself was annotated as a C/D box snoRNA, on the basis of the presence of C and D boxes (Azevedo-Favory et al. 2021). Yet, no methylation targets could be validated for all three predicted snoRNAs within the miR4228 pri-miRNA. Hence, I suspect that miR4228 and its neighbouring snoRNAs (which were enriched in AGO1, AGO2, AGO5 and AGO7 immunoprecipitation experiments) might act as dual sno-miRNA regulatory elements.

Finally, it would be particularly interesting to perform experimental validations of the functional potential of both previously described and newly annotated miPEPs, especially the species-specific ones. Functional validation could be performed by exposing plants with synthetic miPEPs and analyzing the effect of this treatment, in line with (Ormancey et al. 2020; Ormancey et al. 2021; Lauressergues et al. 2022). Examples of experimental set-ups include performing qPCRs to check for an elevated expression of the corresponding miRNA, or phenotypic observations for strong effects on traits. Specificity could be tested in many ways, for example by engineering mutated miPEPs, or applying unconserved miPEPs from conserved pri-miRNAs.

References

- Annacondia ML, Martinez G. 2021. Reprogramming of RNA silencing triggered by cucumber mosaic virus infection in Arabidopsis. *Genome Biol.* 22:340.
- Azevedo-Favory J, Gaspin C, Ayadi L, Montacié C, Marchand V, Jobet E, Rompais M, Carapito C, Motorin Y, Sáez-Vásquez J. 2021. Mapping rRNA 2'-O-methylations and identification of C/D snoRNAs in Arabidopsis thaliana plants. *RNA Biol.* 18:1760–1777.
- Barciszewska-Pacak M, Knop K, Jarmołowski A, Szweykowska-Kulińska Z. 2016. Arabidopsis thaliana microRNA162 level is posttranscriptionally regulated via splicing and polyadenylation site selection. *Acta Biochim. Pol.* 63:811–816.
- Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. *Genetics* 172:1675–1681.
- Bielewicz D, Kalak M, Kalyna M, Windels D, Barta A, Vazquez F, Szweykowska-Kulinska Z, Jarmolowski A. 2013. Introns of plant pri-miRNAs enhance miRNA biogenesis. *EMBO Rep.* 14:622–628.
- Cuperus JT, Fahlgren N, Carrington JC. 2011. Evolution and functional diversification of MIRNA genes. *Plant Cell* 23:431–442.
- Ding N, Zhang B. 2023. microRNA production in Arabidopsis. *Front. Plant Sci.* 14:1–11.
- Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. 2019. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* 29:932–943.
- Laubinger S, Sachsenberg T, Zeller G, Busch W, Lohmann JU, Ratsch G, Weigel D. 2008. Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 105:8795–8800.
- Lauressergues D, Ormancey M, Guillotin B, San Clemente H, Camborde L, Duboé C, Tourneur S, Charpentier P, Barozet A, Jauneau A, et al. 2022. Characterization of plant microRNA-encoded peptides (miPEPs) reveals molecular mechanisms from the translation to activity and specificity. *Cell Rep.* 38.
- Le Veve A, Burghgraeve N, Genete M, Lepers-Blassiau C, Takou M, De Meaux J, Mable BK, Durand E, Vekemans X, Castric V. 2023. Long-Term Balancing Selection and the Genetic Load Linked to the Self-Incompatibility Locus in *Arabidopsis halleri* and *A. lyrata*. Wilson M, editor. *Mol. Biol. Evol.* 40:msad120.
- Lecampion C, Floris M, Fantino JR, Robaglia C, Laloi C. 2016. An Easy Method for Plant Polysome Profiling. *J. Vis. Exp.*:54231.
- Li M, Yu H, Liu K, Yang W, Zhou B, Gan L, Li S, Zhang C, Yu B. 2021. Serrate-Associated

- Protein 1, a splicing-related protein, promotes miRNA biogenesis in Arabidopsis. *New Phytol.* 232:1959–1973.
- Li S, Xu R, Li A, Liu K, Gu L, Li M, Zhang H, Zhang Y, Zhuang S, Wang Q, et al. 2018. SMA1, a homolog of the splicing factor Prp28, has a multifaceted role in miRNA biogenesis in Arabidopsis. *Nucleic Acids Res.* 46:9148–9159.
- Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, Ge S, Guo YL. 2016. On the origin of de novo genes in arabidopsis thaliana populations. *Genome Biol. Evol.* 8:2190–2202.
- Ma Z, Coruh C, Axtell MJ. 2010. *Arabidopsis lyrata* Small RNAs: Transient *MIRNA* and Small Interfering RNA Loci within the *Arabidopsis* Genus. *Plant Cell* 22:1090–1103.
- Ormancey M, Guillotin B, San Clemente H, Thuleau P, Plaza S, Combier JP. 2021. Use of microRNA-encoded peptides to improve agronomic traits. *Plant Biotechnol. J.* 19:1687–1689.
- Ormancey M, Le Ru A, Duboé C, Jin H, Thuleau P, Plaza S, Combier JP. 2020. Internalization of miPEP165a into arabidopsis roots depends on both passive diffusion and endocytosis-associated processes. *Int. J. Mol. Sci.* 21:1–16.
- Pavan F, Azevedo Favory J, Lacoste E, Beaumont C, Louis F, Blassiau C, Cruaud C, Labadie K, Gallina S, Genete M, et al. 2024. The evolutionary history and functional specialization of microRNA genes in *Arabidopsis halleri* and *A. lyrata*. Available from: <http://biorxiv.org/lookup/doi/10.1101/2024.05.03.592357>
- Pavan F, Lacoste E, Castric V, Legrand S. 2025. Scenarios for the emergence of new miRNA genes in the plant *Arabidopsis halleri*. *Plant Cell Physiol.*:pcaf004.
- Qu G, Kruszka K, Plewka P, Yang S-Y, Chiou T-J, Jarmolowski A, Szweykowska-Kulinska Z, Echeverria M, Karlowski WM. 2015. Promoter-based identification of novel non-coding RNAs reveals the presence of dicistronic snoRNA-miRNA genes in Arabidopsis thaliana. *BMC Genomics* 16:1009.
- 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.* 2:890–896.
- Sawyer SA. 1994. Inferring Selection and Mutation from DNA Sequences: The McDonald-Kreitman Test Revisited. In: Golding B, editor. Non-Neutral Evolution. Boston, MA: Springer US. p. 77–87. Available from: http://link.springer.com/10.1007/978-1-4615-2383-3_7
- Schwab R, Speth C, Laubinger S, Voinnet O. 2013. Enhanced microRNA accumulation through stemloop-adjacent introns. *EMBO Rep.* 14:615–621.
- Scott MS, Ono M. 2011. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. *Biochimie* 93:1987–1992.
- Stepien A, Knop K, Dolata J, Taube M, Bajczyk M, Barciszewska-Pacak M, Pacak A, Jarmolowski A, Szweykowska-Kulinska Z. 2017. Posttranscriptional coordination of

splicing and miRNA biogenesis in plants. *Wiley Interdiscip. Rev. RNA* 8.

Szarzynska B, Sobkowiak L, Pant BD, Balazadeh S, Scheible WR, Mueller-Roeber B, Jarmolowski A, Szweykowska-Kulinska Z. 2009. Gene structures and processing of *Arabidopsis thaliana* HYL1-dependent pri-miRNAs. *Nucleic Acids Res.* 37:3083–3093.

Taft RJ, Glazov EA, Lassmann T, Hayashizaki Y, Carninci P, Mattick JS. 2009. Small RNAs derived from snoRNAs. *Rna* 15:1233–1240.

Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2018. A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* 35:631–645.

Yan K, Liu P, Wu CA, Yang GD, Xu R, Guo QH, Huang JG, Zheng CC. 2012. Stress-Induced Alternative Splicing Provides a Mechanism for the Regulation of MicroRNA Processing in *Arabidopsis thaliana*. *Mol. Cell* 48:521–531.

Zhang S, Liu Y, Yu B. 2015. New insights into pri-miRNA processing and accumulation in plants. *Wiley Interdiscip. Rev. RNA* 6:533–545.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* 343:769–772.

Zhu P, Wang Y, Qin N, Wang F, Wang J, Deng XW, Zhu D. 2016. *Arabidopsis* small nucleolar RNA monitors the efficient pre-rRNA processing during ribosome biogenesis. *Proc. Natl. Acad. Sci. U. S. A.* 113:11967–11972.

Acknowledgements

To my thesis jury and CSI committee members

I express my most sincere gratitude to the people that have accepted to be part of my defense jury. I could not be more grateful for Li Zhao and Nikolaos Vakirlis for having welcomed me to the field of *de novo* genes, giving me the incredible opportunity to attend the SMBE meeting on *de novo* gene birth and attending my defense. I sincerely thank Karine Alix and Juliette de Meaux for making this defense possible despite administrative contingencies, and providing precious external feedback to my work. I also sincerely thank Ingrid Lafontaine and Jacinthe Azevedo-Favory for their support throughout my PhD during all CSIs, which have been incredibly helpful; as well as for attending my defense.

To my financers

I am grateful for the financial support provided throughout my PhD project, by the the University of Lille, the SMRE doctoral school, the Graduate Program “Science for a changing planet”, the Agence Nationale de la Recherche (projet TE-MoMa, subvention ANR-18-CE02-0020-01), the IRePSE group and the I-SITE ULNE Foundation. I am also grateful to the Texas A&M University and again, the Graduate Program “Science for a changing planet”, respectively for providing me with a travel award allowing me to present at an international conference, and for providing me with an international mobility grant allowing me to take part in a scientific exchange in an international laboratory.

À mes encadrants

Je remercie du fond du cœur Eléonore, sans laquelle ce projet n'aurait jamais vu le jour, et qui a été un pilier essentiel de soutien et de compréhension, aussi bien scientifique qu'humain, tout au long de ma thèse. Je remercie également Sylvain et Vincent, qui ont été très présents et d'un précieux soutien tout au long de ma thèse, m'aidant à sortir de plus d'une impasse. Au-delà des aspects scientifiques sur lesquels j'ai excellemment été guidée, qui ont été une profonde source de joie par les découvertes et apprentissages qui m'ont été permis, je

sors de cette thèse avec un profond attachement humain vis-à-vis de chacun d'entre vous. J'estime avoir eu d'excellent encadrants, ne pouvant pas avoir souhaité meilleur encadrement. Vous êtes incroyables, et j'espère que de nombreux autres doctorants connaîtront la même joie que j'ai connu pendant ma thèse.

To my collaborators

I express my most sincere gratitude to the entire "Genomics, Structure and Translation" lab for welcoming me with open arms during my stay there, and in particular Olivier Namy and Isabelle Hatin, who taught me so much about Ribosome Profiling. I am also extremely grateful to the whole Erich Bornberg-Bauer lab, who made my stay in Münster incredibly gratifying, and gave me the opportunity to have amazing scientific discussions. I thank Erich for your guidance, understanding and help, as well as Lars for your irreplaceable contributions to both my scientific work and personal life. I thank you Anna for being so incredibly helpful in so many ways, I promise I will get you a sweater back. I also thank Marie, Margaux, Alina, Bharat, Alun, Elias and Sabine for all our time spent together, both inside and outside the laboratory, you are a huge part of the reason why I am so very much looking forward to coming back to Münster.

To my fellow lab members

I am incredibly grateful to every single lab member who welcomed me warmly to the lab, you have quite literally turned my life around, and became my second family. Je tiens à remercier Chloé, Christelle et Laurence pour m'avoir permis de réaliser ce projet, je n'aurais pas pu le faire sans vous. Merci à Clément et Mathieu, qui m'ont permis d'apprendre tant sur la bioinformatique, et sans lesquels ce projet n'aurait pas non plus pu être possible. Je remercie Pierre et Xavier, rouages essentiels du laboratoire sans lesquels j'aurais été totalement perdue. Je remercie Camille pour avoir été une source de divertissement infinie. I thank Christelle and Rita for organizing our weekly seminars and Journal Clubs which I adored so much. I thank my most beloved royal subjects, also known as both former and current PhD students, who I will dearly miss. Merci Audrey, Zoé, Emilie, François, Agathe, Flavia, Guillaume, votre présence et accueil m'ont été incroyablement précieux, les mots ne suffisent pas à décrire à quel point je suis heureuse de vous avoir connus. Merci en particulier à Flavia, mon acolyte de l'ombre, mon modèle, ma muse. Merci à Fabien, Achille, Camille, Matteo, Diane, Claire, et Arthur, pour tous ces moments passés ensemble qui sont tout autant de souvenirs si précieux à mes yeux. Merci

à Justine pour toutes nos discussions, soirées jeux de société et visionnage de films Ghibli. Thank you Sarah, for your support when I needed it most. I sincerely wish everyone the absolute best for the future, may all your projects come to fruition.

À mes proches

Merci Matthieu pour tout ce que tu m'as donné, pour ton soutien et ton écoute inflexibles, pour avoir apporté tant de joie dans ma vie et m'avoir donné la force nécessaire de surmonter les épreuves que j'ai connues, pour m'avoir me faire sentir aimée. Merci Brice pour mes escapades parisiennes, pour ton amitié continue au fil des années, pour nos discussions lunaires. Merci aux organisateurs et participants de l'Atelier Critique, notamment Sylvain et Alexandre, pour m'avoir aidée à partiellement étancher ma soif infinie de discussions philosophiques. Merci aux professionnels de santé qui m'ont accompagné au long de ma thèse, notamment Alexandre Péral et Mickaël Rassel, qui m'ont permis d'apaiser mes maux et de soigner mes troubles. Merci enfin à ma mère et à mon frère, dont la solidarité et la présence ont été d'une précieuse aide au cours de moments difficiles, et dont je suis très heureuse de pouvoir être proche.

Abstract (english)

Understanding the genetic basis for the origin and evolution of novel phenotypes and, ultimately, biological diversity is one of the most fundamental questions in evolutionary biology. An important source of genetic novelty includes the emergence of new protein-coding genes and of new regulatory elements. Within genomes, Open Reading Frames (ORFs) have the potential to encode phenotype-inducing polypeptides. While “canonical” ORFs, located within well-conserved genes, have extensively been characterized; short, non-canonical ORFs outside genes have long been considered “non-coding” by default, and therefore not functional. Recently, the role of non-canonical ORFs both as regulatory elements of microRNAs (miRNAs) and as starting material for the emergence of new genes *de novo* (i.e. from ancestrally non-coding sequences) has been demonstrated in a variety of species and genomic contexts. However, few studies have investigated the evolution of non-canonical ORFs over short evolutionary time-scales. In this thesis, I focused on three closely related species of the *Arabidopsis* genus: *A. halleri*, *A. lyrata* and *A. thaliana*, and investigated (i) the role of short, non-canonical ORFs in the emergence of both new genes, by taking the example of genes arising *de novo*, and (ii) of new regulatory elements, by taking the example of miRNA-encoded peptides. By producing Ribosome Profiling data, I was able to evaluate the genome-wide emergence of non-canonical translation events across *Arabidopsis* species. By producing long-read RNA sequencing data in the same species, I was further able to evaluate their prevalence in miRNA primary transcripts (pri-miRNAs). In the first chapter of my thesis, I was able to annotate a numerous and diverse repertoire of translated *de novo*-emerged short ORFs using stringent synteny analysis and Ribosome Profiling. Acting as the starting material for *de novo* gene birth, those ORFs proved to exhibit characteristics similar to those of untranslated, neutrally evolving intergenic ORFs. Translated *de novo* ORFs seemed to be under a high birth rate, and pervasively expressed. Among them, I characterized the process by which a few ORFs that stood out as the most promising gene candidates have evolved from previously non-coding nucleotide sequences. In the second chapter of my thesis, I showed that the coding potential of pri-miRNAs was ubiquitous across *Arabidopsis* species and seemed to be particularly common in transcripts that contained conserved miRNA precursors. While the sequence of the miRNA ORFs themselves was conserved in only about a third of the cases, the fact that translation events themselves were happening; either within the same approximate region of the pri-miRNA or not, was highly conserved. Globally, my thesis highlighted the under-estimated functional potential of non-canonical coding regions. Short, translated ORFs can encode polypeptides that may play crucial roles in the emergence and regulation of protein-coding genes.

Résumé (français)

Comprendre les bases génétiques de l'origine et de l'évolution de nouveaux phénotypes et, ultimement, de la diversité biologique est l'une des questions les plus fondamentales de la biologie évolutive. L'émergence de nouveaux gènes, codant pour des protéines, et de nouveaux éléments régulateurs constitue une source importante de nouveauté génétique. Au sein des génomes, les cadres de lecture ouverts (ORF) ont le potentiel de coder des polypeptides induisant un phénotype. Alors que les ORF « canoniques », situés à l'intérieur de gènes bien conservés, ont largement été caractérisés, les ORF courts et non canoniques situés en dehors de ces gènes ont longtemps été considérés comme « non codants » par défaut, et donc non fonctionnels. Récemment, le rôle des ORFs non canoniques en tant qu'éléments régulateurs des microARN (miARN) et en tant que matériel de départ pour l'émergence de nouveaux gènes *de novo* (c'est-à-dire à partir de séquences ancestralement non codantes) a été démontré chez plusieurs espèces, et dans différents contextes génomiques. Cependant, peu d'études l'ont démontré de manière exhaustive au sein d'un genre, chez les espèces végétales. Dans cette thèse, je me suis concentrée sur trois espèces étroitement liées du genre *Arabidopsis* : *A. halleri*, *A. lyrata* et *A. thaliana*, et j'ai étudié le rôle (i) des ORF courts et non canoniques dans l'émergence de nouveaux gènes, en prenant l'exemple des gènes apparaissant *de novo*, et (ii) de nouveaux éléments régulateurs, en prenant l'exemple des peptides codés par des transcripts de miARNs. En produisant des données de ribosome profiling, j'ai pu évaluer à l'échelle de génomes entiers l'évolution des événements de traduction non canoniques chez les espèces d'*Arabidopsis*. En produisant des données de séquençage de longs fragments d'ARNs chez ces mêmes espèces, j'ai pu évaluer leur prévalence dans les transcrits primaires des miARN (primiARN). Dans le premier chapitre de ma thèse, j'ai annoté un large répertoire d'ORFs non canoniques traduits, et ayant émergé *de novo* à l'aide d'analyses de synténie stringentes et de Ribosome Profiling. Agissant comme matériel de départ pour le processus de naissance de gènes *de novo*, ces ORFs présentent de nombreuses caractéristiques similaires à celles des ORFs intergéniques non traduits qui évoluent de manière neutre. Les ORF *de novo* traduits semblent émerger fréquemment au sein des lignées et être exprimés de manière omniprésente. Parmi eux, j'ai identifié des gènes *de novo* candidats qui présentent des caractéristiques qui se distinguent des autres ORFs et qui se rapprochent des gènes canoniques. Dans le deuxième chapitre de ma thèse, j'ai montré que le potentiel codant des primiARN était omniprésent chez les espèces d'*Arabidopsis* et semblait particulièrement fréquent dans les transcrits contenant des précurseurs de miRNA conservés. Alors que la séquence des ORF des miARN eux-mêmes n'était conservée que dans un tiers des cas, le fait que les événements de traduction eux-mêmes se produisaient, approximativement aux mêmes positions ou non, était hautement conservé. Globalement, ma thèse a mis en évidence le potentiel fonctionnel sous-estimé des régions codantes non canoniques. Des ORFs courts et traduits peuvent produire des polypeptides qui peuvent jouer un rôle majeur dans l'émergence et la régulation de gènes codant pour des protéines.