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

Authors

Claire Patiou[1], Christelle Blassiau[1], Isabelle Hatin[2], Lars A. Eicholt[3], Enora Corler[2], Chloé Ponitzki[1], Laurence Debacker[1], Erich Bornberg-Bauer[3,4], Olivier Namy[2], Sylvain Legrand[1], Vincent Castric[1], Eléonore Durand[1]

[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

[4]Department of Protein Evolution, Max-Planck Institute for Biology, Max-Planck-Ring 5, 72076 Tuebingen, Germany

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

Abstract

Ancestrally non-genic sequences are now widely recognized as potential reservoirs for the *de novo* emergence of new genes. Across clades, some *de novo* genes were proven to have substantial phenotypic effects, and to contribute to the emergence of novel biological functions. Yet, still very little is known about the starting material from which *de novo* genes emerge, especially in plants. To fill this gap, we generated Ribosome Profiling data from the closely related species *Arabidopsis halleri*, *A. lyrata* and *A. thaliana* and characterized genome-wide patterns of translation across them. Synteny analysis revealed 211 Open Reading Frames (ORFs) that have emerged *de novo* within the Arabidopsis genus and already exhibit signs of active translation. Most of these *de novo* translated ORFs were species- and even accession-specific, indicating their transient nature, with patterns of polymorphism consistent with neutral evolution in natural populations. They were also significantly shorter and less expressed than conserved Coding DNA Sequences (CDS), and their GC content increased with phylogenetic conservation. While most of them were located in intergenic regions and are thus newly discovered, 34 were previously annotated as CDS in at least one genome, and are promising putative genes. Our results demonstrate the abundance of translation events outside of conserved CDS, and their role as starting material for the emergence of novel genes in plants.

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). Yet, this historical paradigm was challenged by the gradual 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 proposed in *Drosophila* (Begun et al. 2006; Levine et al. 2006), yeast (Cai et al. 2008) and primates (Toll-Riera et al. 2008; Knowles and McLysaght 2009). Since these first discoveries, important *de novo* genes have been identified and functionally characterized in varied taxa over the past decade, ranging from the anti-freeze *afgp* gene in cod (Baalsrud et al. 2018), to the DNA repair *BSC4* gene in yeast (Cai et al. 2008), the spermatogenesis related *goddard* gene in flies (Lange et al. 2021) or the *NCYM* gene in humans (Suenaga et al. 2014). In comparison, little is currently known about how *de novo* genes arise in plants. In the model plant *A. thaliana,* the first identified *de novo* gene was the qua-quine-starch (*QQS*) gene, which encodes 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 is the Sun Wu-Kong (*SWK*) *de novo* gene, which encodes a 65 amino-acid protein involved in seed germination under osmotic stress (Jin et al. 2025). In rice, the *de novo* gene *GSE9* was proven to contribute to the shape of rice grains and thus the differences between *indica* and *japonica* varieties (Chen et al. 2023). Beyond such specific examples, the global genome-wide repertoires of *de novo* genes have been characterized in a limited number of model species only, like *A. thaliana, Oryza sativa, Prunus persica, Arachis hypogaea*, *Rosa banksiae* or *Triticeae* species (Silveira et al. 2013; Cui et al. 2015; Li et al. 2016; Zhang et al. 2019; Song et al. 2022; Poretti et al. 2023; Cao et al. 2024; Li et al. 2025). For instance, in *A. thaliana*, a total of 782 putative *de novo* genes

have been annotated, and were found to be shorter, more highly methylated and more likely to be expressed in the male germline (pollen) than other gene categories (Wu et al. 2014; Li et al. 2016). Within closely related species from the *Oryza* genus, 175 putative *de novo* genes were found to have emerged in *O. sativa* (Zhang et al. 2019). Overall, it is now clear that *de novo* genes are important drivers of phenotypic innovation (Chen et al. 2013; Blevins et al. 2021; Montañés et al. 2023; Peng and Zhao 2024; Xia et al. 2025). Yet, in plants, relatively few comparable studies exist and quantitative studies on the dynamics of translation gain over the course of evolution are missing.

Most studies on *de novo* gene birth 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 ancestral non-coding sequences and how they eventually acquire “gene-like” characteristics. So far, studies in *Mus musculus* (Ruiz-Orera et al. 2018) and *Saccharomyces paradoxus* (Durand et al. 2019) seem to indicate that small, translated ORFs located in intergenic regions and without protein-coding gene orthologs in neighbouring species could provide the basis for *de novo* gene emergence. These ORFs were found to abound in several species, including yeast with 447 translated ORFs (Durand et al. 2019), or mice with 689 high-coding score ORFs (Ruiz-Orera et al. 2018), where they underwent a high rate of turnover and evolved mostly neutrally. Hence, these ORFs are viewed as a “reservoir” from which a few novel functional protein-coding genes might eventually emerge (Ruiz-Orera et al. 2018; Durand et al. 2019; Ruiz-Orera and Albà 2019). Yet, the extent to which these observations can be generalized beyond well-studied model species remains to be determined.

Studying the early steps of *de novo* gene emergence 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 varied sequences may undergo translation in a

given genome, which have been referred to as “translons” (Świrski et al. 2025). Selecting the most promising gene candidates among them requires evidence that translational activity is actually taking place. Such evidence is hard to obtain precisely because *de novo* genes are not expected to harbor the hallmarks of canonical protein-coding sequences (such as greater length and sustained expression), and because their expression tends to be low (Lei et al. 2024). Obtaining translational evidence is nonetheless possible by using Ribosome Profiling, a well-established sequencing technology that provides a detailed snapshot of translational activity at the genome-wide scale (Ingolia et al. 2019; Kiniry et al. 2020). Yet, few ribosome profiling datasets have been produced outside of a few well-studied model organisms (Gilmore et al. 2015; Kim et al. 2020; Zhao et al. 2024). In plants, only few high-quality ribosome profiling datasets are available, all from model species such as *O. sativa* (Zhu et al. 2023), *Zea mays* (Chotewutmontri et al. 2018) and *A. thaliana* (Hsu et al. 2016; Wu and Hsu 2022). Pervasive translation activity has been detected in intergenic regions of *A. thaliana* (Hanada et al. 2007; Hanada et al. 2013) but its potential role in *de novo* gene emergence remains to be explored. Second, confidently asserting that a given ORF has emerged *de novo* requires the identification of a clear orthologous non-coding region in a sister species. However, homology detection lacks power beyond a certain degree of divergence, resulting in a potential “phylostratigraphic bias” (Moyers and Zhang 2015; Moyers and Zhang 2016; Moyers and Zhang 2017; Moyers and Zhang 2018). Yet, later studies suggested that this issue affects predictions to a much lesser extent than previously predicted (Domazet-Lošo et al. 2017; Weisman et al. 2020). Furthermore, a solid way to circumvent this problem, and increase the reliability of *de novo* gene detection, is to use synteny information to identify orthologous intergenic regions that have remained syntenic (Vakirlis et al. 2020). Because genomes can sometimes be rapidly and drastically rearranged over the course of evolution, synteny analysis benefits from comparing species that are sufficiently closely related so that synteny relationships remain preserved (Tang et al. 2008). Previous studies have tackled either of those two problems, either by producing ribosome-protected mRNA fragments (RPF) in a few distantly related

species (Ruiz-Orera et al. 2018; Xiao et al. 2018), or by establishing detailed synteny relationships between closely related species, but without corresponding ribosome profiling data (Vakirlis et al. 2020; Roginski et al. 2024). One study by Durand et al. (2019) addressed both problems simultaneously, by characterizing the translation profile of closely related *Saccharomyces paradoxus* populations, demonstrating a high turnover of translational activity in ORFs emerging *de novo* within species (Durand et al. 2019). This approach, combining genomic and translation data in a comparative framework, has shown great potential in deciphering the first steps of *de novo* gene birth at a short evolutionary timescale. Yet, it has so far been under-exploited, in particular 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). We focused on three closely related species of the *Arabidopsis* genus: *A. halleri*, *A. lyrata* and *A. thaliana*. For each of the three species we produced ribosome profiling data from root and leaf tissue to obtain high-level evidence for the discovery of novel protein-coding ORFs (Prensner et al. 2023). The recent divergence between these species provides high power to the synteny analysis (Tang et al. 2008), resulting in a unique opportunity to reliably identify the non-coding sequences that are homologous to the putative *de novo* ORFs. Specifically, we find that at least 211 ORFs may have emerged *de novo* within the *Arabidopsis* genus and show signs of ongoing translation. These ORFs share many characteristics with untranslated intergenic ORFs: they are short, have a low GC content, prevalently use non-canonical START codons, and evolve in the absence of selective pressure. Nonetheless, a minority possess characteristics closer to those of conserved genes, and make up the most promising putative gene candidates.

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 and protein-coding gene annotations are available (*A. thaliana*; (Lamesch et al. 2012), *A. lyrata*; (Hu et al. 2011), *A. halleri;* (Pavan et al. 2025) and *Capsella rubella*; (Slotte et al. 2013)). *A. halleri* and *A. lyrata* diverged less than one million years ago, and they diverged about five million years ago from *A. thaliana* (Roux et al. 2011). *Capsella* is an outgroup genus with an approximate eight million years deep divergence (Acarkan et al. 2000; Koch and Kiefer 2005). 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 nine nucleotides, regardless of overlap with gene annotations. This yielded a total of 19, 23, 3 and 10 million intergenic ORF annotations in *A. halleri*, *A. lyrata, A. thaliana* and *C. rubella* respectively (**Table S1**). These intergenic ORF annotations were combined with the pre-existing, high-quality gene annotations, which represent a total of 34, 30, 27 and 26 thousand CDS annotations in the same species respectively (**Table S1**). Second, we sought to obtain direct experimental evidence for translation activity. To that end, we simultaneously grew plants of the four species under identical 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). In a pilot experiment, we prepared ribosome profiling libraries and obtained a total of 90 to 140 million raw sequencing reads per replicate (**Table S2**). Despite experimental ribosomal depletion, rRNA contamination remained in our libraries, and rRNA sequences were subsequently computationally removed. In the end, we could use between 0.8 and 16 million mapped reads per replicate for the detection of translation signals.

To evaluate the quality of ribosome profiling libraries, we examined metagene plots produced by RibORF (Ji 2018) within the CDS of canonical protein-coding genes. A high quality ribosome profiling 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 excluded the *A. thaliana* and *C. rubella* libraries from further downstream analyses due to their suboptimal overall quality. For *A. lyrata* and *A. halleri,* we retained the 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 instead relied on high-quality ribosome profiling datasets from Hsu et al. 2016 and Wu and Hsu 2022, comprising a total of 40 to 110 million mapped reads (**Table S3**). 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 ribosome profiling datasets can be used to obtain reliable evidence for translational activity. The signal of periodicity was apparent for 27 to 32-nucleotide long 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-mers (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*.

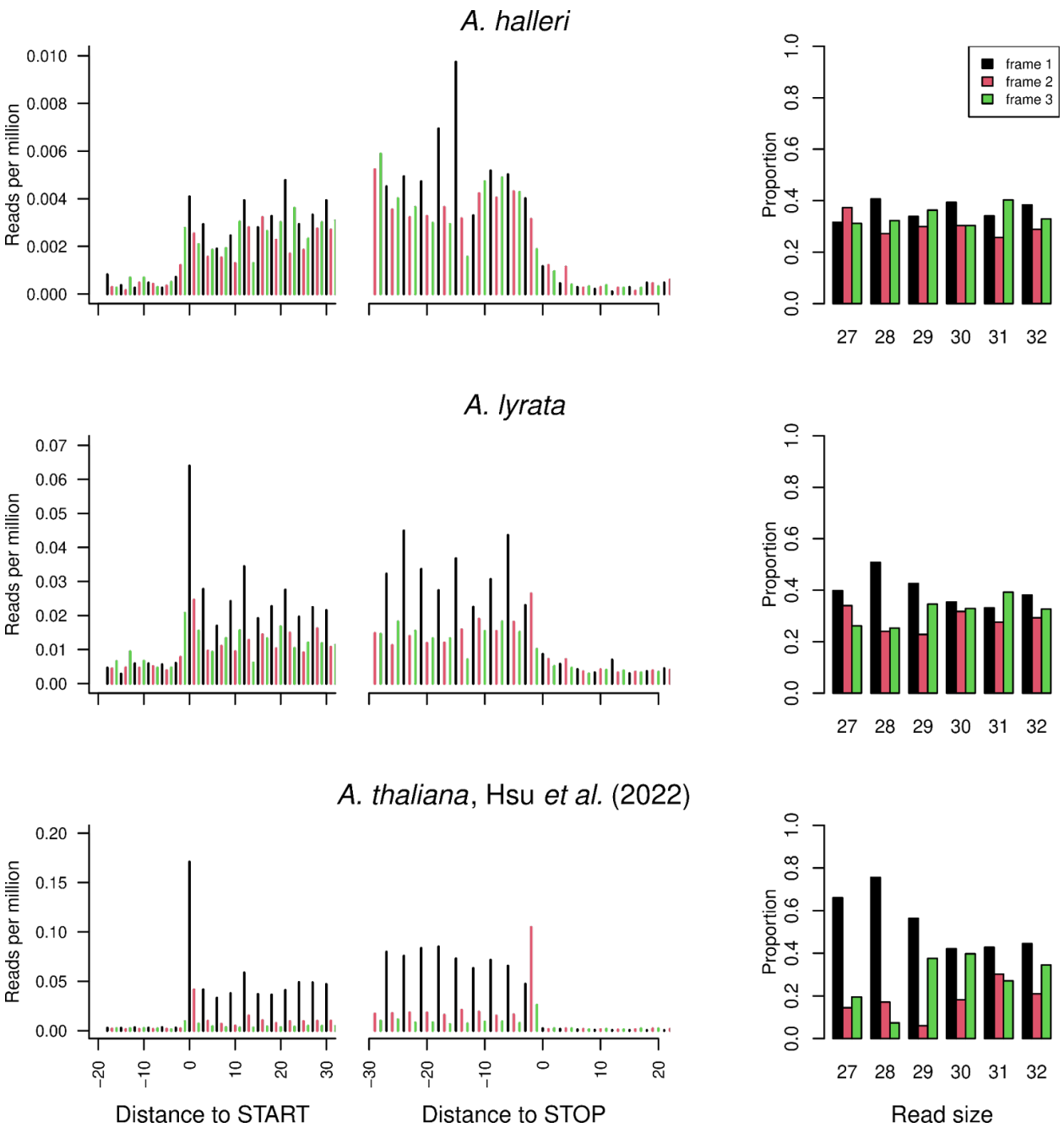


Figure 1: Metagene plots of 28-kmers from ribosome profiling, after P-site shifting (reads were shifted according to their k-mer specific P-site offset, and trimmed to be 1 base-long), used for the annotation of translation events; and prevalence of k-mers depending on which frame they are found in, produced with modified RibORF scripts (Ji 2018). Metagene plots are created by concatenating the accumulation of ribosome profiling footprints 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 (**Table S2**), 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 ribosome profiling data, especially in the context of the discovery of recently emerged *de novo* genes (Ruiz-Orera et al. 2014; Ruiz-Orera et al. 2018; Ruiz-Orera and Albà 2019; Papadopoulos et al. 2021). For this reason, we annotated ORFs with translation signatures using three independent approaches specifically developed for the analysis of ribosome profiling 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 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 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, representing 60 to 80% of all CDS) by at least one of the three annotation tools (ORFscore, RibORF or RiboCode) (**Table S1**). Among them, 20,592 and 13,266 CDS were found to be translated by the three tools, respectively in *A. thaliana* (75%) and *A. lyrata* (44%). In *A. halleri*, the difference was more pronounced between the tools, and we detected 3,467 translated CDS by the three tools (10%) and 11,996 translated CDS by two tools (34%) (**Table S1**). This observation is probably due to the less apparent periodicity of the ribosome profiling 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 not all canonical CDS are expected to be identified as being translated by this single experiment, due to inherent technical (low power of ribosome profiling data to assess translation on a per-locus basis)

and biological limitations (*e.g.* limited number of sampled tissues, 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.

Beside the expected signals of translation overlapping canonical CDS, we identified translation signals for 3 to 23 million Intergenic ORFs (IGORFs, that is, ORFs with no overlap to any previous genome annotation) depending on the species considered (**Table S1**). Compared to CDS, we expect a smaller proportion of the millions of annotated IGORFs to display translation signals, since they belong to supposedly non-functional, intergenic regions. Indeed, only a small proportion show 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 S1**). 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 *A. halleri*, 1,548 in *A. lyrata* and 142 in *A. thaliana*; **Figure 2** and **Table S1**). 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 annotation 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.

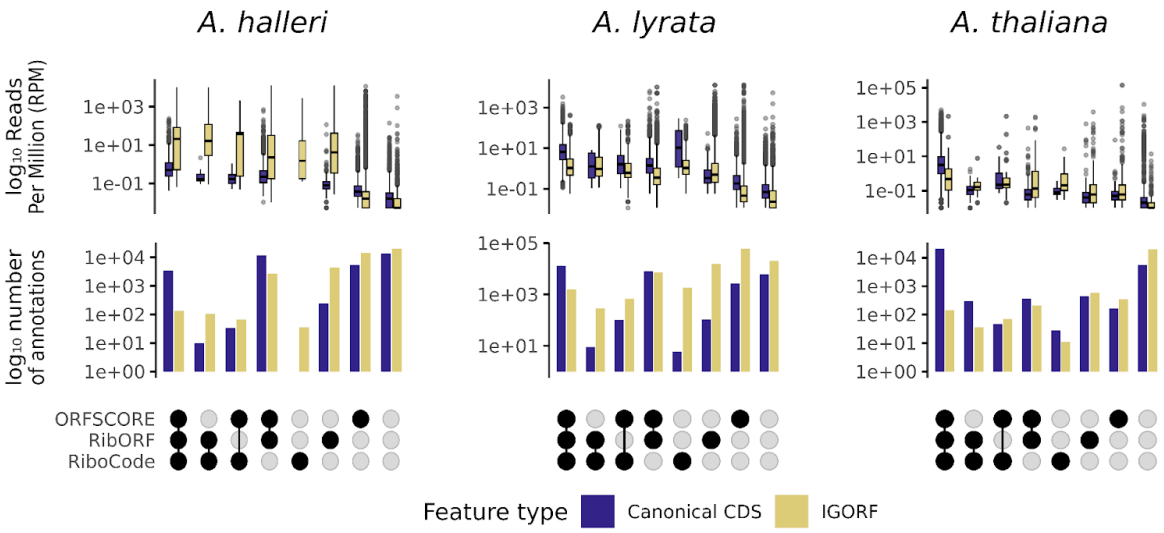


Figure 2: Number of annotations detected as translated and their translation levels in $\log_{10}$ Reads Per Million (RPM, where each read is a ribosome footprint); according to the tools that detected them as translated, and to the category to which they belong: canonical CDS or IGORFs. Black circles 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 (**Table S1**), and data from *A. thaliana* come from two studies: (Hsu et al. 2016; Wu and Hsu 2022).

Synteny analysis identifies many *de novo* ORFs being translated

We then sought to identify which of those translated ORFs (either CDS or IGORFs), have appeared *de novo* within the *Arabidopsis* genus, *i.e.* after the divergence from *C. rubella* which we used as an outgroup. To do so, we identified orthologous micro-syntenic regions, which we aligned and filtered to obtain syntenic, orthologous regions of the ORF in neighbouring species (**Table S4, Figure S3**). Additionally, homology searches were performed to exclude any ORF that had homology to proteins outside of *Arabidopsis*, and/or outside of syntenic regions. The *de novo* origin of the remaining ORFs was then assessed on the basis of their sequence conservation within the syntenic alignments. 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 because 1) they were fully absent, 2) they contained a frameshift mutation or 3) they lacked a START and a STOP codon at the same positions.

Based on these stringent thresholds, a total of 42, 163 and 14 *de novo* ORFs with detectable translational activity were identified in *A. halleri, A. lyrata* and *A. thaliana* respectively (**Table S4**). Of these, 13 to 23% were already annotated as CDS in at least one species: 9 in *A. halleri*, 21 in *A. lyrata* and 4 in *A. thaliana*, which all contained less than five exons, with 70% displaying only one (**Table S5**). Having removed redundant ORFs which had been annotated across several Arabidopsis species, we ended up with 211 *de novo* emerged ORFs (**Figure 3**). More than half of these *de novo* ORFs are species-specific (30 in *A. halleri*, 118 in *A. lyrata* and 7 in *A. thaliana*, **Table S5**), while 27% are found in at least two and 6% in the three species. Despite a divergence time of approximately eight 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* (Table **S4**), indicative of the power granted by synteny analysis in this context.

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, 34 species-specific ORFs were identified as translated by several tools, and hence likely display translational potential.

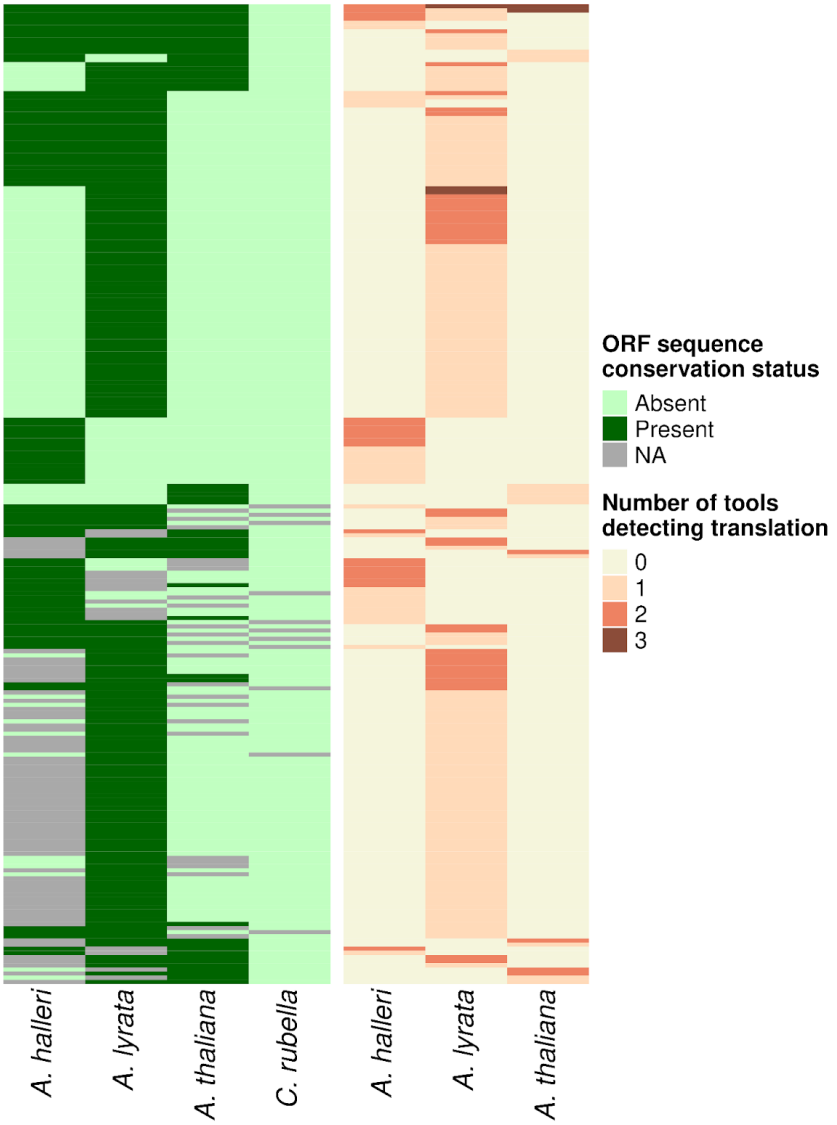


Figure 3: Heatmap of the conservation of *de novo* ORF sequence (left) and translation detection (right). Each line represents an individual ORF, and each column a species. Redundancy within each species was removed by keeping only the longest of overlapping *de novo* ORFs, and between species by removing 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.

Evolutionary trajectory of *de novo* translated ORFs

We then evaluated the evolutionary trajectory of *de novo* translated ORFs by comparing their features to those of conserved CDS (canonical CDS with an identifiable ortholog in every other species of interest) on one hand, and to untranslated intergenic ORFs (unt_IGORFs; with no coverage by ribosome footprints) on the other hand. We excluded ORFs with ribosome profiling reads but no translation signals, so as not to include ORFs that were not classified as translated due to possible detection power limitations. The high divergence of non-coding genomic regions among our four study species rendered ancestral sequence reconstruction unreliable. Hence, we were unable to determine at which specific point of the phylogeny the ORFs emerged, and we refer to conservation classes rather than to actual evolutionary ages. We classified *de novo* translated ORFs into several 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 (**Table 1**, **Figure 4 C**).

Translated in:	Conservation category	Already annotated as CDS*	Newly identified as IGORF
A. halleri	Category A	7	2
	Category B	0	4
	Category D	3	27
	TOTAL:	**10**	**33**
A. lyrata	Category A	9	10
	Category B	6	20
	Category C	6	112
	TOTAL:	**21**	**142**

Table 1: Number of translated *de novo* ORFs identified by species, conservation category, and annotation type (*= already annotated as a CDS in at least one *Arabidopsis* species). Conservation categories are as follows: A) conserved in all *Arabidopsis* species, B) conserved in *A. halleri* and *A. lyrata*, C) *A. lyrata*-specific and D) *A. halleri*-specific

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 have subsequently been lost in *A. halleri* and *A. lyrata* (or their common ancestor), or have emerged after the divergence of *A. thaliana.*The following analyzes will be focused on *A. lyrata*, as the limited number of *de novo* ORFs per conservation class do not allow for sufficient statistical power in *A. halleri* (although the trends are similar; **Figure 4 A,C**).

First, we examined the usage of the canonical and alternative START codons of the genetic code (ATG, ACG, CTG, GTG, TTG). As gene prediction methods generally consider canonical start codons only, conserved CDS almost exclusively start with ATGs. In *de novo* ORFs, the TTG START codon was found to be used as often as the canonical ATG, and other alternative START codons were also common, representing up to 25% of all *de novo* translated ORFs. These patterns are also observed when distinguishing between conservation classes of untIGORFs (**Figure S4**) and are hence reflective of the nucleotide composition of these sequences (**Figure 4 A,C**). Secondly, we observed that, as expected, conserved CDS were on average longer (mean length = 1,111 nt) than unt_IGORFs (mean length= 52 nt), while *de novo* translated ORFs were intermediate (mean length = 100 nt) (**Figure 4 A,C**). A statistically significant difference in length was apparent between the most conserved A and least conserved C/D *de novo* conservation classes in *A. lyrata,* with a shorter length of the more conserved ORFs (mean length = 66 nt for A and 104 nt for C/D, *p-value* = 2.6×10^{-3}). This is most likely a consequence of the stringent way we defined conservation, as longer ORFs are more likely to be interrupted by mutations compared to shorter ones, thus constituting larger mutational targets (Iyengar and Bornberg-Bauer 2023). Comparing the length of unt_IGORFs across conservation classes yields the same observation, and hence confirms this interpretation (**Figure S4**). Third, we observed a continuous decrease of GC content, from a median of 44% in conserved CDS, to a significantly lower 35% in the *de novo* ORFs that are specific to either *A. halleri* or *A. lyrata* (C/D conservation class). Though lower in GC content than conserved CDS, *de novo* ORFs from the B and C/D conservation classes still had a significantly higher GC content than unt_IGORFs, with a median of 30% (**Figure 4 A,C**). Oppositely, no such pattern could be observed when comparing unt_IGORFs of different conservation classes (**Figure S4**). When comparing *de novo* ORFs and untranslated IGORFs belonging to the same conservation classes, we observe that C/D *de novo* ORFs have a statistically significantly higher GC content than their untranslated IGORFs counterparts (**Figure S5**), suggesting that the GC content patterns in translated *de novo* ORFs might be an effect of translation. Finally, in line

with previous observations (Carvunis et al. 2012), we observed that translation activity (quantified by ribosome profiling reads depth in reads per million, RPM) of *de novo* ORFs was less intense than that of conserved CDS (on average between 1.29 RPM versus 2.43 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 many conserved CDS. We verified that the overall difference in translation activity was not caused by the difference in the length of the annotations. We did that by using a linear least squares regression, which showed that length alone could explain only about 2.6% of the observed variance in read depth (adjusted R-squared=0.02619, *p-value*<2.2×10^{-16}). Furthermore, the differences in translation activity were also apparent when length differences were controlled for by calculating read depth only within the first 60 nucleotides of each feature, where the translation signal is most pronounced. Using a linear least squares regression, there was no evidence that length could explain the observed variance in read depth within the first 60 nucleotides (adjusted R-squared=-0.0001479, *p-value*=0.3675).

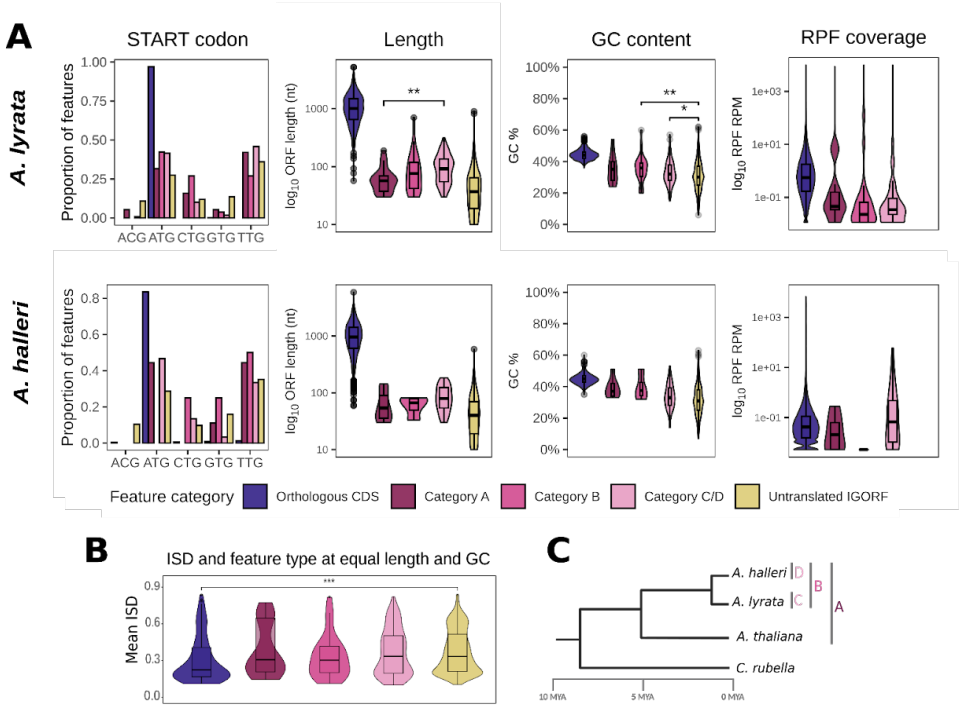


Figure 4: Characteristics of conserved canonical CDS, *de novo* ORFs and untranslated IGORFs. *De novo* ORFs were classified according to their conservation category: A, B, C or D; the meaning of which is showcased in the phylogeny. For length and GC content, all statistically significant differences 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).

Following the analysis of the DNA features of polypeptide-encoding sequences, we next examined their predicted polypeptides, specifically their levels of intrinsic structural disorder (ISD). We predicted ISD using AIUPred (Erdős et al. 2021) on the *de novo* ORFs, as well on a random sub-sample of 1,000 orthologous CDS and 1,000 unt_IGORFs for comparison (**Figure S6**). In our results, unt_IGORFs and *de novo* ORFs exhibited significantly higher levels of ISD than orthologous CDS (Mann-Whitney U test, *p-value*<2.2×10^{-16}, **Figure S6**).

However, in the literature, ISD has been shown to be negatively connected to length, a technical factor affecting disorder prediction algorithms (Peng et al. 2006), and positively to GC content, due to the higher chance of triplets encoding disorder-promoting amino acids (Basile et al. 2017). Indeed, across unt_IGORFs and conserved CDS, disorder decreases with length (unt_IGORFs: Spearman ρ*(disorder,length)*= -0.24, *p-value*= 1.22×10^{-14}; conserved CDS: ρ*(disorder,length)*= -0.14, *p-value*=1.29×10^{-5}), but increases with GC content (unt_IGORFs: Spearman ρ*(disorder,GC)*= 0.13, *p-value*= 2.53×10^{-5}; conserved CDS: ρ*(disorder,GC)*= 0.15, *p-value*=1.67×10^{-6}) (**Figure S6**). Yet, both within and across translated *de novo* ORFs conservation categories, no such significant relationship could be established between disorder and length or GC, with the exception of species-specific ORFs (category C/D), where GC content was unexpectedly negatively correlated to ISD (ρ*(disorder,GC)*= -0.20, *p-value*=0.01474). This observation might be the result of their shorter length, which likely affects prediction accuracy or prevents formation of structural elements (Dill 1990; Peng et al. 2006). When considering all feature categories at once, the length and GC content explain a significant—though modest—proportion of the variance in ISD (adjusted $R^2 = 0.0353$, $p < 2.2 \times 10^{-16}$ for length; adjusted $R^2 = 0.0844$, $p < 2.2 \times 10^{-16}$ for GC content)(**Figure S6**). Together, these results show a confounding effect of GC content and protein length on ISD. To control for these two factors jointly, we predicted ISD for all conserved CDS and kept only those matching the length and GC content of unt_IGORFs

and *de novo* ORFs (<139 amino acid residues long, 21-39% GC content). We retained 160 CDS, and yet still observe a statistically significant difference in ISD between categories (**Figure 4 B**). Hence, *de novo* ORFs and unt_IGORFs appear to be more disordered than conserved CDS even when length and GC are matched.

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

Out of the 211 translated *de novo* ORFs we identified, a few stand out as having notable gene-like characteristics, making them strong candidates for *de novo* gene emergence. A total of 34 out of the 211 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 AL6G18580 (**Figure 5, A**) has no homology to any previously described protein in any species, displays translation signals by two tools (ORFSCORE and RibORF), starts with the canonical start codon ATG, and is 225 nucleotides long (74 amino acids). The homologous sequence was present in *C. rubella,* where it was clearly non-coding (no conservation of the START and STOP codon, and of the frame), indicating that this ORF emerged *de novo* within the *Arabidopsis* genus. While the START and STOP codons are conserved in *A. halleri* and *A. thaliana*, a few indels interrupt the reading frame in these species. Hence, this particular *de novo* ORF 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. Supplementary BLASTs (**Table S6**) revealed that this locus had previously been predicted as a non-coding RNA in *A. lyrata* (Hupalo and Kern 2013), and has homology with the *A. thaliana* primary transcript of miR162a. Further alignments (**Figure S7**) revealed that AL6G18580 is homologous to the upstream region of the *A. thaliana* miR162a miRNA, encompassing the start of the precursor and the future mature miRNA, leaving out the miRNA*. Given the high conservation degree of miR162a (conserved beyond *Brassicaceae*), the upstream region of the miR162a locus may have recently acquired a translated *de novo* CDS in *A. lyrata*. This

is reminiscent of the discovery of polypeptide-encoding ORFs within the 5' region of miRNA primary transcripts in several plant species, including *A. thaliana* (Lauressergues et al. 2015; Lauressergues et al. 2022). AlphaFold2 confidently predicts that the potential encoded polypeptide forms an alpha-helix (predicted local distance difference test (pLDDT) between 70 and 90), facing five beta-sheets with varying prediction confidence (from pLDDT < 50 to >90, **Figure 5, A**).

Another example of *de novo* CDS is the *A. halleri* - specific 829626 gene (**Figure 5, B**), which has no homology to any previous annotation (**Table S6**), was identified as translated by two tools (ORFSCORE and RibORF), and is 132 nucleotides in length (43 amino acids). While no high-quality syntenic alignment could be obtained in *A. thaliana*, alignments in *C. rubella* and *A. lyrata* showcase the absence of STOP codon and frame conservation in the former, and the absence of START codon and frame conservation in the latter. Again, this ORF may have emerged either specifically in *A. halleri*, or in a more ancestral *Arabidopsis* species and have been lost secondarily in *A. lyrata* by point mutations and indels. The potential encoded polypeptide of the 829626 gene is predicted to form a single alpha helix with a very high confidence (pLDDT > 90) (**Figure 5, B**).

Beside such "*de novo* CDS", most of the *de novo* ORFs we identified were actually intergenic, and some of them have features similar to those of the *de novo* CDS. For example, the *A. halleri*-specific intergenic ORF named IGORF_37 (**Figure 5, C**) was detected as translated by ORFSCORE and RibORF, similarly to many CDS from *A. halleri*, and is 147 nucleotides long (46 amino acids). 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 detectable homology to any known protein-coding gene, non-coding locus or sequenced genomes. The structure of the potential encoded polypeptide in IGORF_37 is predicted to be a single alpha-helix with very high confidence (pLDDT > 90) (**Figure 5, C**).

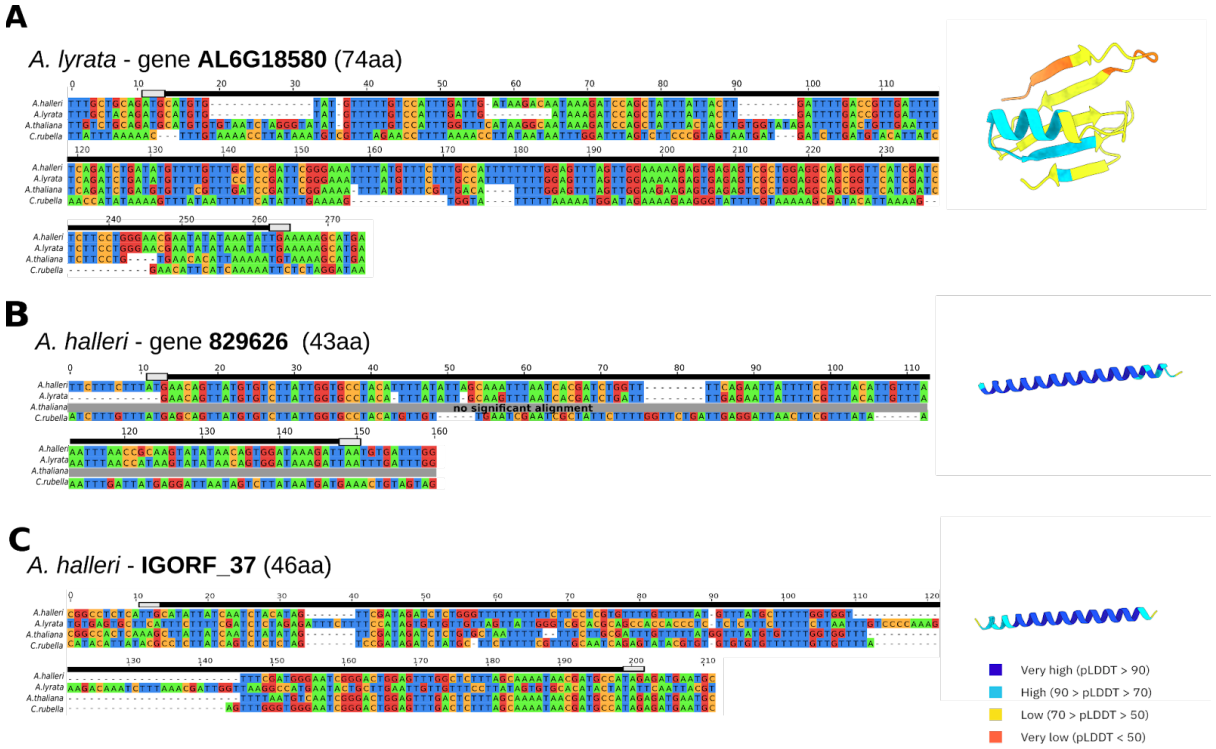
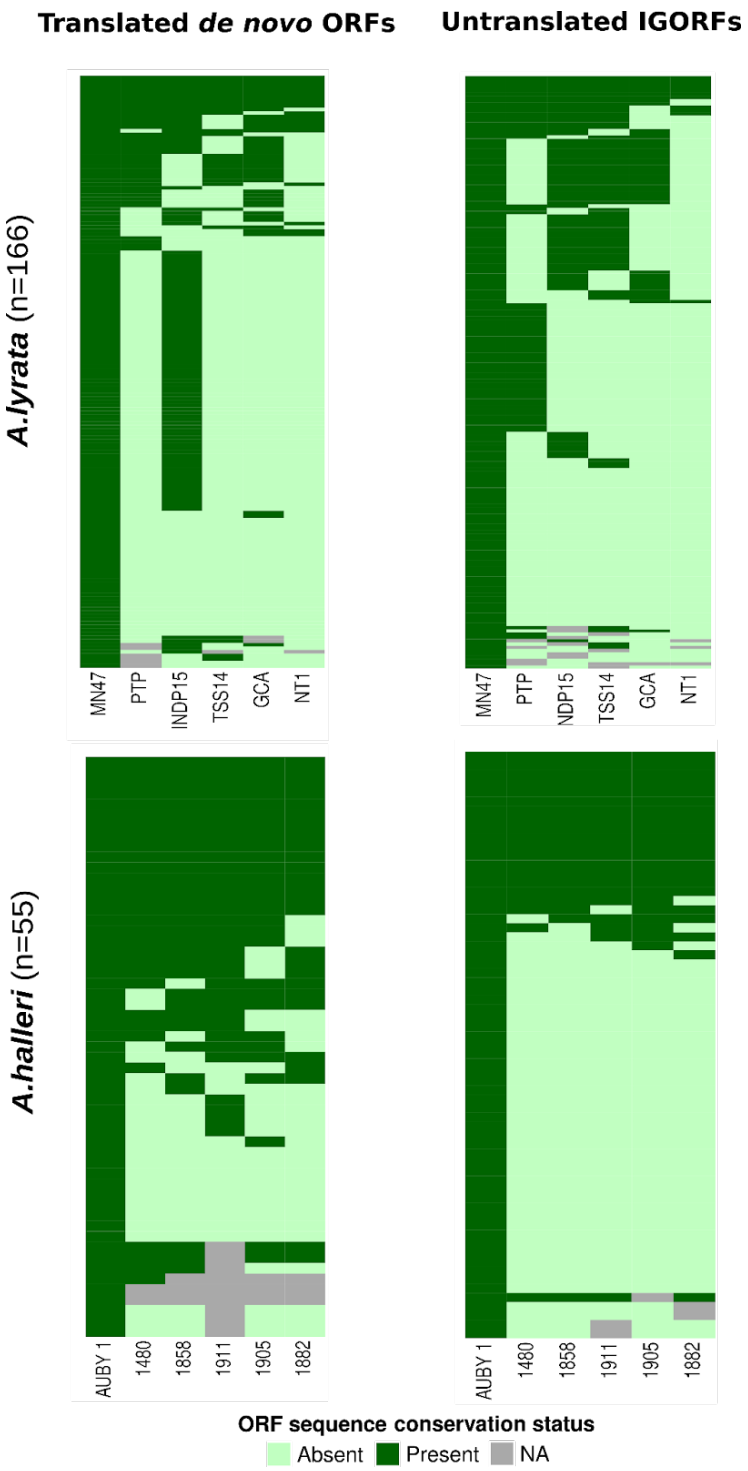


Figure 5: Alignments of potential *de novo* gene candidates, either *A. lyrata* (AL6G18580, **A**) or *A. halleri* (829626, **B** and IGORF_37, **C**) specific; as well as the corresponding AlphaFold2 structure prediction of the encoded polypeptides . Alignments are colored by nucleotides. The START 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. Residues are colored by pLDDT score (see legend).

The repertoires of *de novo* ORFs are highly labile within species

To gain insight into the micro-evolutionary conservation of the *de novo* translated ORFs, we then analyzed five high-quality genome assemblies from natural accessions of both *A. halleri* and *A. lyrata* that we either collected from the literature (Padilla-García et al. 2025; Scott et al. 2025) or assembled *de novo*. We identified syntenic regions based on the transfer of gene annotations from the AUBY1 (Pavan et al. 2025) and MN47 (Kolesnikova et al. 2023) reference accessions, respectively. The conservation of each ORF within syntenic regions that were conserved between accessions was determined using the previous filters and alignments (**Figure S3**). Even within species, a high degree of lability was apparent (**Figure 6**). Specifically, very few *de novo* ORFs were conserved across all accessions, with

up to 21% being accession-specific, and only 29% found in more than two accessions in *A. lyrata*. In *A. halleri,* a similar 25% of the *de novo* ORFs were accession-specific and a slightly higher proportion (62%) were found in more than two. We then compared these patterns of conservation to those of untranslated IGORF, which we used as a neutral control. We found that *de novo* translated ORFs tended to be more conserved across accessions than untranslated IGORFs in *A. halleri* (permutation test; *p-value*=0.014), although this trend was not apparent in *A. lyrata* (*p*-value=0.442). On average, while *de novo* ORFs repertoires are highly labile within species, at least in *A. halleri,* translated *de novo* ORFs are on average significantly more likely to be conserved across accessions than untranslated IGORFs.



482

Figure 6: Heatmap of the conservation of translated *de novo* ORF and untranslated IGORF sequences across accessions from the same species. Each line represents an individual ORF, and each column an accession. If the ORF is conserved in an accession, 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".

483

No evidence for purifying selection on *de novo* ORFs

Finally, we sought to evaluate the strength of purifying selection acting on translated *de novo* ORFs. To do this, we collected genomic resequencing data from a range of natural populations of *A. halleri* and *A. lyrata* (with a total of 37 *A. halleri* individuals, and 39 *A. lyrata* individuals from across their species range (**Table S7**). We first evaluated total nucleotide diversity per site, and found that in both species, conserved CDS displayed the lowest levels of diversity (on average 5.60×10^{-5} in *A. lyrata*, and 7.52×10^{-5} in *A. halleri*), and untranslated IGORFs the highest (on average 1.67×10^{-4} in *A. lyrata*, and 2.4×10^{-4} in *A. halleri*) (**Figure S8**). Translated *de novo* ORFs displayed levels of nucleotide diversity similar to those of untranslated IGORFs, with an average of 1.67×10^{-4} in *A. halleri* and of 9.12×10^{-5} in *A. lyrata* (p=0.25 in *A. halleri* and p=0.39 in *A. lyrata*), and much higher than those at conserved CDS (p=3.6×10^{-6} in *A. lyrata;* **Figure S8**). The ratio of non-synonymous (π_N) to synonymous nucleotide diversity (π_S) revealed strong purifying selection on conserved CDS (π_N/π_S=0.32). The average ratio for untranslated IGORFs was close to one (0.86 in *A. halleri* and 0.90 in *A. lyrata*), corresponding to the expectation under neutral evolution (**Figure 7**). The π_N/π_S ratio for translated *de novo* ORFs was 0.95 in *A. lyrata*, and 0.86 in *A. halleri*, which landed them right in the distribution of untranslated IGORFs (p=0.68 and p=0.57, respectively). Hence, in both species the strength of purifying selection on *de novo* translated ORFs taken as a group does not detectably differ from that on our neutral control (untranslated IGORFs), and is thus either null or at most very weak.

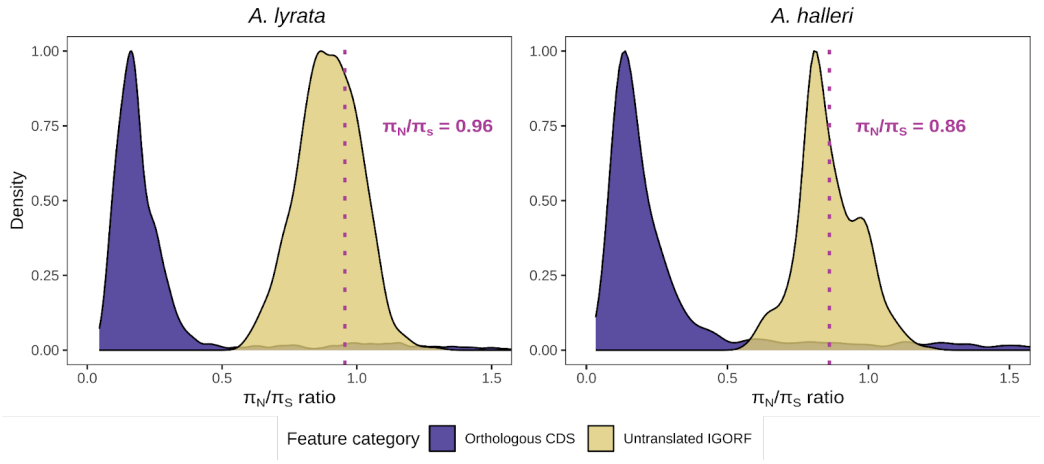


Figure 7: The π_N/π_S ratio of *de novo* ORFs (dotted vertical line; a single value computed after concatenation) matches that of untranslated intergenic ORFs (random subset of 1,000 ORFs) and departs from that of conserved canonical CDS (random subset of 1,000 CDS), indicating that the strength of purifying selection on them is either null or at most very weak .

Discussion

Pervasive translation of intergenic ORFs in Arabidopsis

Our study revealed that translation of IGORFs is common in *Arabidopsis* species, even when stringently restricting the analysis to those with the highest evidence of translation. The extent of this pervasive translation varied across our three study species, possibly for both technical and biological reasons. For instance, the higher number of IGORFs with translation signals detected in *A. lyrata* (87,433) compared to *A. halleri* (21,668) might be attributed to a better resolution within the ribosome profiling dataset in *A. lyrata,* which is characterized by better triplet periodicity. In contrast, the lower number of translated IGORFs found in *A. thaliana* (1,396), in spite of the excellent quality of the ribosome profiling dataset for that species, might be attributed to its more compact genome compared to *A. lyrata* and *A. halleri*, which is largely due to shorter intergenic regions (Hu et al. 2011). The genome-wide number of annotated ORFs in *A. halleri* and *A. lyrata* is about seven times higher than in *A. thaliana* (19-23 millions *versus* about 3 millions), so genome size alone can only partly explain the observed differences in numbers of IGORFs with

translation signals. Additional factors might contribute to the lower number of IGORFs in *A. thaliana*, including the greater exhaustiveness of the *A. thaliana* annotation, which results in ORFs considered as intergenic in *A. halleri* or *A. lyrata* to be considered as annotated CDS in *A. thaliana.* Additionally, the higher quality ribosome sequencing data used to define translation in *A. thaliana* might result in a more stringent training of tools towards strict canonical properties, less sensitive to non-canonical, pervasive translation signals typically found in intergenic regions (Papadopoulos et al. 2024). Finally, a last possibility is that *A. thaliana* may indeed exhibit a lower intrinsic translational activity outside of well-conserved genes, but this would need to be formally tested to examine the possible reasons behind this pattern. Hence, at this stage, we caution against a too direct interpretation of the differences of the number of IGORFs with translation signals among species, since differences in experimental methodology, plant tissues and/or developmental stages do not allow for a straightforward comparison. Yet, we note that the numbers of IGORFs with translation signals we observed are consistent with previous reports in *A. thaliana*, with 2,996 ORFs found to be transcribed in intergenic regions (Hanada et al. 2007) and 2,099 highly expressed under certain experimental conditions (Hanada et al. 2013). They are also consistent with reports across distant clades such as the yeast *S. cerevisiae*, where 1,266 IGORFs were either occasionally or highly translated (Papadopoulos et al. 2021), the yeast *S. paradoxus*, where 447 intergenic ORFs were translated (Durand et al. 2019), the mouse *M. musculus*, where 1,980 non-conserved ORFs were translated (Ruiz-Orera et al. 2018), or the fly *D. melanogaster*, where 343 small ORFs were found to be translated (Patraquim et al. 2020).

The exhaustive characterization of non-canonical translation events remains a major challenge. Indeed, ribosome profiling may introduce biases, both during laboratory experiments and computational analyses (Bartholomäus et al. 2016). For example, translation elongation inhibitors do not homogeneously stop ribosomes along transcripts and may be affected by codon composition (Orelle et al. 2013), making it more difficult to

pin-point the exact positions at which translation is happening (Bartholomäus et al. 2016). This makes the precise identification of which ORFs are actively being translated difficult, especially in non-coding regions, and as we observed, different analysis tools can result in widely different predictions. Dosage of RNAse I is delicate and of crucial importance, and we therefore relied on the recommended dosage for *A. thaliana* (Wu and Hsu 2022). Still, we observed high levels of rRNA contamination (up to 40%), which might have partly been the product of too high of a RNAse I dosage (Becker et al. 2013; Bartholomäus et al. 2016), or to inefficiency of the experimental ribosomal depletion, with varying results across samples. These biases affect all coding sequences, but render non-canonical translation events particularly difficult to detect, where different prediction methods differ the most, and ORFs are not saturated in coverage at usual sequencing depths (Lei et al. 2024). All of these factors contribute to diminishing the statistical power to uncover translation events outside of genic regions. To palliate these biases, a recommended strategy is to combine datasets to maximize the chances of detecting translation events (Wacholder et al. 2023). Since no other public datasets were available for *A. lyrata* and *A. halleri,* we instead maximized sequencing depths, as recommended by (Lei et al. 2024), reaching up to 639 million reads (as opposed to the typical range of 150 to 200 million reads in available *A. thaliana* ribosome profiling datasets (Hsu et al. 2016; Wu and Hsu 2022)). Finally, it has previously been reported that non-canonical translation, including that of *de novo* genes, tends to happen in a tissue-specific manner (Wu et al. 2014; McLysaght and Hurst 2016). For instance, in *A. thaliana*, the *de novo* gene *SWK* was found to be highly expressed specifically in seeds and siliques (Jin et al. 2025). Therefore, an exciting next step will be to determine the extent to which non-canonical translation signals vary across developmental stages, across plants collected from diverse environments and across a broad panel of genetic origins. Overall, beyond these variations and inherent technical limitations, it is clear from our results that in all three species translation activity is not restricted to canonical genic sequences, as is now recognized in a range of model and non-model organisms

(Ingolia et al. 2014; Ruiz-Orera et al. 2018; Durand et al. 2019; Chen et al. 2020; Sruthi et al. 2022; Papadopoulos et al. 2024).

Frequent emergence of *de novo* translated ORFs

Such 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 previously been described as the "transient translatome" (Wacholder et al. 2023). The presence of numerous species-specific translated IGORFs (27 in *A. halleri*, 112 in *A. lyrata* and 7 in *A. thaliana*) supports the idea of frequent gains of transcription and translation in regions considered to be "non-coding", which would become temporary, initial substrates for the *de novo* emergence of genes (Schmitz et al. 2018; Durand et al. 2019; Grandchamp et al. 2024). Phylogenetic conservation of these translated *de novo* IGORFs is rather low, both across species (most *de novo* ORFs are species-specific), or across accessions within species (up to 25% being accession-specific), and their levels of nucleotide diversity and π_N/π_S ratio indicate that they globally evolve neutrally, which is in line with results in mice (Ruiz-Orera et al. 2018). However translated *de novo* ORFs were significantly more retained across accessions of *A. halleri* than untranslated IGORFs. On one hand, translated *de novo* ORFs might be located in regions of the genome that experience lower intrinsic mutation rates and are thus more likely to be found across accessions, as opposed to untranslated IGORFs that would be found in genomic contexts that favor the rapid gain and loss of coding sequences. Mutation rates vary across genomes, including in plants where some regions experience elevated levels of DNA repair resulting in lower mutation rates (Baer et al. 2007; Quiroz et al. 2023); and they have been linked to the emergence of novel coding sequences in yeast through heightened GC content, proximity to promoters or to recombination hot spots (Vakirlis et al. 2018). On the other hand, translated *de novo* ORFs might be located within regions of the genome with higher gene density, which tend to be more stable over evolutionary times because

deletions would likely affect nearby genes and thus be selected against. In the “grow slow and moult” model of emergence, *de novo* genes that are in close proximity to pre-existing protein coding genes benefit from increased transcription and translation without being functional yet, initially leaving them free to evolve randomly while progressively rendering them visible to natural selection (Andreatta et al. 2015; Bornberg-Bauer et al. 2015; Kleppe and Bornberg-Bauer 2018). Whether in regions of lower mutation rate, or stronger purifying selection, recently emerged translated *de novo* ORFs would therefore be more likely to be shared across accessions, while their sequence in themselves would not yet be of functional relevance and evolve without the constraints of negative selection. They thus represent a reservoir of potential candidates that can be tested by natural selection and on rare occasions may acquire physiological relevance.

The first steps of *de novo* gene emergence

The processes by which some translated *de novo* ORFs move away from noise and eventually acquire gene-like characteristics remain somewhat unclear. Here we found that translated *de novo* ORFs showcased many similarities with untranslated intergenic ORFs: they had a much shorter length than conserved CDS, a high prevalence of non-canonical START codons and no sign of purifying selection, which corresponds to what has previously been observed in yeast, mice and fruit flies (Carvunis et al. 2012; Ruiz-Orera et al. 2018; Vakirlis et al. 2018; Papadopoulos et al. 2021; Grandchamp et al. 2023; Wacholder et al. 2023). Also in line with previous observations in distant clades, translated *de novo* ORFs were overall translated at lower levels than conserved CDS. Strikingly, the GC content of the translated *de novo* ORFs was intermediate between that of conserved CDS and that of untranslated IGORFs, and increased continuously from the least to the most conserved ORFs. A similar continuum of GC content was observed in *A. thaliana* and *Drosophila* (Li et al. 2016; Peng and Zhao 2024), but a different trend was observed in yeast, where *de novo* genes tend to emerge in regions that are richer in GC content than in conserved CDS

(Vakirlis et al. 2018) and in humans, where *de novo* genes can arise from long non-coding RNAs (lncRNAs) exhibiting significantly higher GC content than protein-coding genes globally (An et al. 2023). As the number of biological systems where the emergence of *de novo* genes is studied increases in the future, it will be exciting to finally have the possibility to test whether the trends observed across species can be generalized as global mechanisms for *de novo* gene emergence.

GC content is also strongly correlated to other molecular features, including ISD (Basile et al. 2017; Nielly-Thibault and Landry 2019). Similarly to what we observed in conserved CDS and unt_IGORFs, a positive relationship between GC content and disorder, and a negative one between length and disorder have been documented in the literature (Peng et al. 2006; Basile et al. 2017). In *Drosophila*, the higher GC content and shorter length of *de novo* genes are both contributing factors to the observed increase in disorder compared to other gene categories (Zheng and Zhao 2022; Middendorf et al. 2024). While the shorter length of translated *de novo* ORFs in our data may contribute to their higher disorder, their lower GC content should act in the opposite direction, contributing to lower disorder. Furthermore, in species-specific *de novo* ORFs, we observe a significant negative relationship between GC and disorder. This highlights the difficulty of performing and interpreting such analyzes on very short features like *de novo* ORFs, where factors such as length may overpower the influence of GC on disorder. Here, in line with previous observations in *Drosophila* (Middendorf et al. 2024), we conclude that translated *de novo* ORFs are more disordered than conserved CDS, even when correcting for their different length and GC content .

A few translated *de novo* ORFs are promising gene candidates

While most translated *de novo* ORFs possess features similar to those of unt_IGORFs, a few ones are closer to the genic end of the spectrum and are promising *de novo* gene candidates. In fact, a small subset of the *de novo* translated ORFs we identified had already been annotated as proper CDS (9 in *A. halleri*, 21 in *A. lyrata* and 4 in *A. thaliana*). All contained less than four exons, with 80% displaying only one, and so far none has been characterized at the functional level. Several of them display unambiguous evidence of active translation, and are predicted to have the potential to adopt simple 3D structures such as alpha helices or beta sheets, in line with structure predictions of *de novo* genes in other species (Chen et al. 2024; Middendorf et al. 2024; Peng and Zhao 2024). Having established that these CDS actually correspond to *de novo* genes opens the exciting possibility to explore their functional properties in more detail, especially given the wealth of molecular resources available in the model *A. thaliana*. In addition to these ORFs that had already been spotted as CDS, our analyses revealed a substantial number of previously unannotated intergenic ORFs displaying similarly strong evidence of translation, with similar length and structure predictions as the *de novo* CDS. While previous plant studies have mostly focused on the functional potential of previously annotated *de novo* emerged genes with already well-established gene-like features (Zhang et al. 2019; Jin et al. 2025), our results highlight the value of investigating the properties of intergenic translated ORFs as raw material for the emergence of new genes *de novo*.

Overall, our findings demonstrate that pervasive translation takes place within intergenic regions in *Arabidopsis* species, providing a vast reservoir of transiently, lowly expressed, mostly neutrally evolving short ORFs. Among those, a select few may ultimately give rise to functional genes, by progressively becoming longer, more highly expressed, and acquiring a higher GC content. Future work combining ribosome profiling, mass spectrometry, and functional assays across a variety of tissues and developmental stages will be essential to

validate this repertoire of *de novo* ORFs and determine their physiological relevance, particularly how these *de novo* peptides interact with other proteins and potentially integrate into molecular interaction networks.

Material and methods

Plant Materials and Growth Conditions

Three seeds of either *Arabidopsis thaliana* accession Columbia-0, *A.halleri*, *A. lyrata* or *C. rubella* were deposited on each of ten 0.7% agar-filled micro-centrifuge tub lid and imbibed in germination medium as described in (Conn et al. 2013) then left in obscurity at 4 °C for two days. They 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), and was replaced once every week. Floral buds, leaves and roots were collected and flash frozen in liquid nitrogen progressively along the growth of the plants over the course of three months (July to September) for 6 to 10 individuals per species, and stored at -80°C.

Lysates preparation

For each species, we used tissues collected from three individuals to prepare three ribosome profiling libraries per species: one per individual, except for *A. thaliana* where nine individuals were merged into three groups of three because of a too low amount of material per individual. We also pooled each tissue sample collected on different dates to reduce the effect of the day of sampling and of the environment, as well as to ensure that we have sufficient material for the experiment. Tissue types were separately ground into a powder using liquid nitrogen and a mortar and pestle. We then suspended 0.25 g of each tissue powder in 4 mL of lysis buffer, as described in (Wu and Hsu 2022). The lysates were vortexed twice, transferred into 2 mL tubes and then agitated at 4°C for 10 min. All lysates

were then spun at 5 000 × g at 4°C for 3 min; and the supernatant was aliquoted into refrigerated 1.5 mL tubes. The aliquots were spun at 20 000 × g at 4°C for 10 min and the supernatants were transferred to refrigerated 1.5 mL tubes. Aliquots of 200 µL were stored at -80°C and a single 10 µL 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 contained on average lower RNA concentrations than those from leaves, root and leaf aliquots were pooled per individual so as to obtain 75% of material coming from roots, and 25% of material coming from leaves. In the end, we have produced three lysate replicates per species

Ribosome footprints

The lysates were treated with 50 units RNase I Ambion (Invitrogen, AM2295) per 40 µg 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 90 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. 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 was 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 eluded in 500 uL of extraction buffer (0,3 M CH_3COONa 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 0,45 µm Corning column (Dutscher; 2515806); and centrifuged at maximal speed for 5 mn. 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 21,000 x g, 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 was adjusted with an equal, 100% ethanol volume, and two volumes (180 µL) of this adjusted buffer were added to the rRNA-depleted ribosome profiling samples (90 µL). Then, 1 volume of 100% ethanol (270 µL) was added to this mix to favor the retrieval of small RNAs between 17 and 200 nucleotides (ribosome footprints are typically approximately 30 nucleotides long). The total 540 µL was then deposited on a column, and centrifuged for 30s at 16,000 x g at room temperature. Then, 400 µL of RNA Prep Buffer were added to the column, which was 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 were added and the column was centrifuged for 1 min at 16,000g at room temperature. The column was put on a 2mL tube and the ribosome footprints were eluded on the column with 15 µL of distilled, nuclease-free water and then centrifuged for 1 min at 16,000g at room temperature again. The final microRNA concentration was measured by fluorescence with the micro-RNA Quant-iT kit (Thermofisher Scientific; Q32882).

Library preparation and sequencing

The libraries were prepared according to the Biolabs NEBNext Multiplex Small RNA library Prep kit's standard protocol with 12 PCR cycles. The resulting libraries were purified with the Monarch PCR & cleanup kit (Biolabs) and concentrations were assessed with the Quant-iT DNA HS kit. Sequencing was performed with a NextSeq system by the I2BC High-throughput sequencing facility, Gif-sur-Yvette.

Ribo-seq Data Analysis

Firstly, adapters were trimmed from the reads and then filtered on quality and length using Cutadapt (v3.2) (Martin 2011). We removed known non-coding RNAs such as rRNA, tRNA and snoRNA through bowtie2 (v2.4.1, Langmead and Salzberg 2012) alignments on TAIR10 non-coding RNA annotations. Depleted sequences were then aligned on the corresponding reference genome per species: TAIR10 for *A. thaliana* (Lamesch et al. 2012), AUBY1 for *A. halleri* (Pavan et al. 2025) and MN47 for *A. lyrata* (Hu et al. 2011). The alignment was performed using the splice-aware aligner STAR (v2.7.5a) (Dobin et al. 2013), tolerating up to one 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. We inferred the P-site offsets by visual assessment of the metagene plotsThe mapped reads were subsequently shifted by the assigned offsets using

RibORF. We then annotated all ORFs across entire genomes using RibORF (Ji 2018). 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 three tools. First we used ORFSCORE (Bazzini et al. 2014), where ribosome footprints accumulation is computed in each possible frame, with the assumption that coding nucleotide sequences would exhibit greater accumulation in the first frame. Annotations were produced using R (v4.2.2), 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 signals. Second, we used RibORF, which relies on Machine Learning ("Support Vector Machine"), 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, we used RiboCode (Xiao et al. 2018), 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 one mismatch and up to 20 multi-mapping positions. Translation signal was predicted after P-site shifting the aligned reads, 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

Translated CDS annotations for each species were used to identify gene orthologs between all pairs of species using Inparanoid (Remm et al. 2001), 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 conserved orthologs, and which includes intergenic sequences. 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 extended 1kb upstream and downstream (or until the closest conserved CDS annotation), 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. In parallel, the ORFs falling within syntenic regions were used for a BLASTX search (Camacho et al. 2009) against the proteomes of five *Brassicaceae* species downloaded from Genbank (Benson et al. 2012): *Camelina sativa*, *Brassica napus*, *Brassica oleraceae*, *Eutrema salsugineum* and *Raphanus sativus*. ORFs that showcased any significant hit were discarded (threshold 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 ≥ 50% of the ORF’s length, 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 ≥ 0.6. 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 (Katoh 2002). Once the multiple alignments were obtained, they were used for conservation analyzes with the help of home-made python scripts. Multiple alignments were further filtered, with a PID of at least 0.5 in at least three species, and a maximal gap content of 40%, all within the ORF region. Finally, a minimal length of 30 nucleotides was required and smaller ORFs were discarded. The *de novo* status was inferred when a region lacking ORF conservation (START, STOP and reading

frame) could be identified, or if an ORF had not been bioinformatically predicted at those positions, in either *C. rubella* or in *A. thaliana*. Furthermore, if in any pair species an ORF was absent but its position overlapped with a gene annotation, the conservation of the frame was checked, and the ORF was considered as conserved if the gene's frame was the same as that of the reference ORF.

Intrinsic disorder and structure prediction

Alphafold2 (v.2.3.0) (Jumper et al. 2021) predictions were performed on HPC cluster Palma II. Predictions with highest confidence (ranked_0) were used for analysis. Disorder was predicted locally using AIUpred (no smoothing) (Erdős and Dosztányi 2024). Mean and median disorder per sequence were calculated using Python v3.10.13 with pandas v2.1.1 (McKinney 2011). Structures were visualized using UCSF ChimeraX-1.7.1 (Pettersen et al. 2021).

Cross-accession conservation

For *A. halleri*, we gathered five high-quality genome assemblies (BN_1480 from Italy and CX_1858, DD_1911, DB_1905 and DO_1882 from Romania; raw long reads available at ENA project # PRJEB70880 (Batista et al. 2024)). For *A. lyrata*, we gathered five high-quality genome assemblies (PTP, JBTAFC000000000), INDP15 and TSS14.18 (ENA project: PRJEB80457 (Padilla-García et al. 2025)), GCA_026151145 (GenBank assembly accession: GCA_026151145.1) from Austria, and NT1 (GenBank assembly accession: GCA_963992905.1 from Siberia (Scott et al. 2025)). Annotations from the reference genomes of *A. halleri* (AUBY1) and *A. lyrata* (MN47) were lifted off to the other assemblies using the liftoff tool (v1.6.3) (Shumate and Salzberg 2021). Using the same parameters as

870 previously, all potential ORF candidates were bioinformatically annotated with RibORF (Ji
871 2018), on the basis of their sequence only, in each accession. The coordinates of the
872 syntenic region from the reference genome were transposed to each accession, and
873 conserved syntenic regions in different accessions were aligned using MAFFT (Katoh 2002).
874 The conservation of the ORF sequence was then assessed from these multiple alignments,
875 using the same procedure as previously. To test whether translated *de novo* ORFs or
876 untranslated intergenic ORFs were more conserved across accessions, we created a
877 "conservation score" that is, per ORF, the number of accessions within which each ORF is
878 found. A Mann-Whitney U test was performed across the two ORF categories for the
879 observed conservation patterns, and "*de novo"* and "untranslated IGORF" labels were then
880 randomly permuted 1,000 times, with new conservation scores and a new Mann-Whitney U
881 test being computed for each iteration. The observed test score was then compared to the
882 distribution of test scores under the null hypothesis (*de novo* and untranslated IGORFs have
883 similar conservation pattern), and a corresponding p-value was assessed.
884

885 Polymorphism analysis

886

887 Genomic data from a total of 37 *A. halleri* individuals from central Europe (**Table S7**)
888 (Suda et al. 2024; Pavan et al. 2025) and from 39 *A. lyrata* individuals from North America
889 and central Europe (**Table S7**) (Mattila et al. 2017; Hämälä et al. 2018; Marburger et al.
890 2019; Takou et al. 2021; Pavan et al. 2025) were collected. Paired-end reads were mapped
891 to the AUBY1 (*A. halleri*) or MN47 (*A. lyrata*) reference genome using bowtie 2 (v2.4.1)
892 (Langmead and Salzberg 2012) and duplicates were marked and removed using Picard
893 Tools (v2.21.4) (Anon 2019). Variant calling and subsequent genotype joining were
894 performed using GATK (v4.1.9.0) (Van der Auwera and O'Connor 2020). Variant calling files
895 were split into invariant and filtered variant sites (filtered with VCFtools (Danecek et al.
896 2011), with parameters *--mac 1, --remove-indels, --minDP 10, --min-alleles 1, --max-alleles*

2, --max-missing 0.75 and *--minQ 30*), and concatenated back into one single variant calling file with BCFtools (Danecek et al. 2021). Average nucleotide diversity per site was computed from the concatenated variant calling file, within each feature type, using pixy (Korunes and Samuk 2021). Bedtools intersect (Quinlan and Hall 2010) was used to retrieve regions of the variant calling file corresponding to each individual annotation, which were then converted into fasta sequences for each allele of each individual using BCFtools consensus (Danecek et al. 2021) with default parameters. The resulting sequences were aligned using MAFFT (Katoh 2002). Because of the limited number of polymorphic sites in *de novo* ORFs and unt_IGORFs resulting from their short lengths, all alignments from each category were concatenated. To get an equivalent number of polymorphic sites, five randomly selected conserved CDS alignments were concatenated and the procedure was repeated 1,000 times. The $\pi N/\pi S$ ratios were computed from these concatenated alignments using dNdSpiNpiS_1.0 (Gayral et al. 2013) with parameters *-gapN_site=10*, *-gapN_seq=0.6*, *-remove_frameshift=yes* and *-allow_internal_bc=yes*; and the $\pi N/\pi S$ ratios of conserved CDS from the 1,000 samples were plotted as a distribution.

Plots were produced using R (v 4.2.2) (R Core Team 2025), 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). *De novo* ORF annotations are available on Zenodo at https://zenodo.org/records/18230755, as well as the DeNoFo standardized annotation (Dohmen et al. 2025).

Acknowledgements

We acknowledge the sequencing and bioinformatics expertise of the I2BC High-throughput sequencing facility, supported by France Génomique (funded by the French

National Program "Investissement d'Avenir" ANR-10-INBS-09), as well as Anne Lopes for sharing with us her expertise in Ribosome Profiling data analysis. We thank the high performance computing service and Bilille at the University of Lille for providing computing resources, as well as Mathieu Genete and Clément Mazoyer for providing bioinformatic support. This work was performed using the infrastructure and technical support of the "Plateforme Serre, cultures et terrains expérimentaux – Université de Lille" for the greenhouse/field facilities, and we thank Pierre Saumitou-Laprade and Flavia Pavan for their help in setting up the hydroponic culture protocol. 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). E.B.-B. acknowledges funding from HFSP (Human Frontiers of Science Programme, RGP004/2023); University of Münster Topical Programme. Some parts of this research were conducted while visiting the Okinawa Institute of Science and Technology (OIST) through the Theoretical Sciences Visiting Program 314 (TSVP).

Funding

This work was supported by the ERC (NOVEL project, grant #648321) and the ANR (project TE-MoMa, grant ANR-18-CE020020-01), as well as the "Ecole doctorale SMRE" and the Graduate Program "Science for a Changing Planet" of the University of Lille. L.A.E. has been supported by EMBO Scientific Exchange Grant 10944.

Bibliography

Acarkan A, Roßberg M, Koch M, Schmidt R. 2000. Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. *Plant J.* 23:55–62.

An NA, Zhang J, Mo F, Luan X, Tian L, Shen QS, Li X, Li C, Zhou F, Zhang B, et al. 2023. De novo genes with an lncRNA origin encode unique human brain developmental functionality. *Nat. Ecol. Evol.* 7:264–278.

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.

Anon. 2019. Picard toolkit. *Broad Inst. GitHub Repos.* [Internet]. Available from: https://broadinstitute.github.io/picard/

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.

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* 8:619–631.

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.

Batista RA, Durand E, Mörchen M, Azevedo-Favory J, Simon S, Dubin M, Kumar V, Lacoste E, Cruaud C, Blassiau C, et al. 2024. Modifiers of genetic dominance at the Arabidopsis self-incompatibility locus retain proto-miRNA features and act through non-canonical gene silencing pathways. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.05.01.591913

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.

Becker AH, Oh E, Weissman JS, Kramer G, Bukau B. 2013. Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. *Nat. Protoc.* 8:2212–2239.

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.

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.

Cao Y, Hong J, Zhao Y, Li X, Feng X, Wang H, Zhang L, Lin M, Cai Y, Han Y. 2024. *De novo* gene integration into regulatory networks via interaction with conserved genes in peach. *Hortic. Res.* 11:uhae252.

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487:370–374.

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, et al. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science* 367:1140–1146.

Chen J, Li Q, Xia S, Arsala D, Sosa D, Wang D, Long M. 2024. The Rapid Evolution of De Novo Proteins in Structure and Complex. *Genome Biol. Evol.* 16:evae107.

Chen R, Xiao N, Lu Y, Tao T, Huang Q, Wang S, Wang Z, Chuan M, Bu Q, Lu Z, et al. 2023. A de novo evolved gene contributes to rice grain shape difference between indica and japonica. *Nat. Commun.* 14:5906.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* 14:645–660.

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.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10:giab008.

Dill KA. 1990. Dominant forces in protein folding. *Biochemistry* 29:7133–7155.

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.

Dohmen E, Aubel M, Eicholt LA, Roginski P, Luria V, Karger A, Grandchamp A. 2025. DeNoFo: a file format and toolkit for standardized, comparable *de novo* gene annotation.Nikolski M, editor. *Bioinformatics* 41:btaf539.

Domazet-Lošo T, Carvunis AR, Albà MM, Šestak MS, Bakarić R, Neme R, Tautz D, Kim Y. 2017. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol. Biol. Evol.* 34:843–856.

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.

Erdős G, Pajkos M, Dosztányi Z. 2021. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* 49:W297–W303.

Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, et al. 2013. Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap.Welch JJ, editor. *PLoS Genet.* 9:e1003457.

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.

Grandchamp A, Kühl L, Lebherz M, Brüggemann K, Parsch J, Bornberg-Bauer E. 2023. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in *Drosophila melanogaster*. *Genome Res.* 33:872–890.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32:2847–2849.

Hämälä T, Mattila TM, Savolainen O. 2018. Local adaptation and ecological differentiation under selection, migration, and drift in *Arabidopsis lyrata* *: LOCAL ADAPTATION UNDER GENE FLOW AND DRIFT. *Evolution* 72:1373–1386.

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.

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.

Hupalo D, Kern AD. 2013. Conservation and Functional Element Discovery in 20 Angiosperm Plant Genomes. *Mol. Biol. Evol.* 30:1729–1744.

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.Harris K, editor. *Mol. Biol. Evol.* 40:msad079.

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 translatome 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.

Koch MA, Kiefer M. 2005. Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species— *Capsella rubella* , *Arabidopsis lyrata* subsp. *petraea* , and *A. thaliana*. *Am. J. Bot.* 92:761–767.

Kolesnikova UK, Scott AD, Van De Velde JD, Burns R, Tikhomirov NP, Pfordt U, Clarke AC, Yant L, Seregin AP, Vekemans X, et al. 2023. Transition to Self-compatibility Associated With Dominant *S* -allele in a Diploid Siberian Progenitor of Allotetraploid *Arabidopsis kamchatica* Revealed by *Arabidopsis lyrata* Genomes.Purugganan M, editor. *Mol. Biol. Evol.* 40:msad122.

Korunes KL, Samuk K. 2021. PIXY : Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol. Ecol. Resour.* 21:1359–1368.

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.

Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, Findlay GD, Bornberg-Bauer E. 2021. Structural and functional characterization of a putative de novo gene in Drosophila. *Nat. Commun.* 12:1667.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.

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.

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. 2024. A systematic evaluation of computational methods for predicting translated non-canonical ORFs from ribosome profiling data. *J. Genet. Genomics* 51:105–108.

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 Y, Li R, Shang J, Zhao K, Sui Y, Liu Z, Yan H, Bao M, Liang M, Zhao Q, et al. 2025. A de novo-originated gene drives rose scent diversification. *Cell*:S0092867425009250.

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.

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.

Marburger S, Monnahan P, Seear PJ, Martin SH, Koch J, Paajanen P, Bohutínská M, Higgins JD, Schmickl R, Yant L. 2019. Interspecific introgression mediates adaptation to whole genome duplication. *Nat. Commun.* 10:5218.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10.

Mattila TM, Tyrmi J, Pyhäjärvi T, Savolainen O. 2017. Genome-Wide Analysis of Colonization History and Concomitant Selection in Arabidopsis lyrata. *Mol. Biol. Evol.* 34:2665–2677.

McKinney W. 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. Available from: https://api.semanticscholar.org/CorpusID:61539023

McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: What, how and why. *Nat. Rev. Genet.* 17:567–578.

Michał Krassowski, Arts M, Lagger C, Max. 2022. krassowski/complex-upset: v1.3.5. Available from: https://zenodo.org/record/3700590

Middendorf L, Ravi Iyengar B, Eicholt LA. 2024. Sequence, Structure, and Functional Space of *Drosophila* De Novo Proteins.Alvarez Carreño C, editor. *Genome Biol. Evol.* 16:evae176.

Montañés JC, Huertas M, Messeguer 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. 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.

Orelle C, Carlson S, Kaushal B, Almutairi MM, Liu H, Ochabowicz A, Quan S, Pham VC, Squires CL, Murphy BT, et al. 2013. Tools for Characterizing Bacterial Protein Synthesis Inhibitors. *Antimicrob. Agents Chemother.* 57:5994–6004.

Padilla-García N, Le Veve A, Čermák V, İltaş Ö, Contreras-Garrido A, Legrand S, Aury J-M, Horvath R, Lafon Placette C. 2025. The Demographic History of Populations and Genomic Imprinting have Shaped the Transposon Patterns in *Arabidopsis lyrata*.Mascagni F, editor. *Mol. Biol. Evol.* 42:msaf093.

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.

Patraquim P, Mumtaz MAS, Pueyo JI, Aspden JL, Couso J-P. 2020. Developmental regulation of canonical and small ORF translation from mRNAs. *Genome Biol.* 21:128.

Pavan F, Favory JA, Lacoste E, Beaumont C, Louis F, Blassiau C, Cruaud C, Labadie K, Gallina S, Genete M, et al. 2025. The evolutionary history and functional specialization of microRNA genes in *Arabidopsis halleri* and *A. lyrata*. *Plant Cell* 37:koaf168.

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.

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 7:208.

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF ChimeraX : Structure visualization for researchers, educators, and developers. *Protein Sci.* 30:70–82.

Poretti M, Praz CR, Sotiropoulos AG, Wicker T. 2023. A survey of lineage-specific genes in *Triticeae* reveals de novo gene evolution from genomic raw material. *Plant Direct* 7:e484.

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 ? *Mol. Cell. Proteomics* 22:734–763.

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

Quiroz D, Lensink M, Kliebenstein DJ, Monroe JG. 2023. Causes of Mutation Rate Variability in Plant Genomes. *Annu. Rev. Plant Biol.* 74:751–775.

R Core Team. 2025. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing Available from: https://www.R-project.org/

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 N* ovo Emerged Gene *Se* arch in Eukaryotes with DENSE.Alvarez Carreño C, editor. *Genome Biol. Evol.* 16:evae159.

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, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. *eLife* 3:1–24.

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.

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.

Scott AD, Kolesnikova UK, Glushkevich A, Steinmann L, Tikhomirov NP, Pfordt U, Bohutínská M, Burns R, Seregin AP, Kolar F, et al. 2025. Multiple Autopolyploid *Arabidopsis lyrata* Populations Stabilized by Long-Range Adaptive Introgression Across Eurasia.Hodgins K, editor. *Mol. Biol. Evol.* 42:msaf153.

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations.Valencia A, editor. *Bioinformatics* 37:1639–1643.

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.

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.

Song H, Guo Z, Zhang X, Sui J. 2022. *De novo* genes in *Arachis hypogaea* cv. Tifrunner: systematic identification, molecular evolution, and potential contributions to cultivated peanut. *Plant J.* 111:1081–1095.

Sruthi KB, Menon A, P A, Vasudevan Soniya E. 2022. Pervasive translation of small open reading frames in plant long non-coding RNAs. *Front. Plant Sci.* 13:975938.

Suda RA, Kubota S, Kumar V, Castric V, Krämer U, Morinaga S-I, Tsuchimatsu T. 2024. Population Genomics Reveals Demographic History and Climate Adaptation in Japanese *Arabidopsis halleri*. *Plant Cell Physiol.*:pcae113.

Suenaga Y, Islam SMR, Alagu J, Kaneko Y, Kato M, Tanaka Y, Kawana H, Hossain S, Matsumoto D, Yamamoto M, et al. 2014. NCYM, a Cis-Antisense Gene of MYCN, Encodes a De Novo Evolved Protein That Inhibits GSK3β Resulting in the Stabilization of MYCN in Human Neuroblastomas.Eilers M, editor. *PLoS Genet.* 10:e1003996.

Świrski MI, Tierney JAS, Albà MM, Andreev DE, Aspden JL, Atkins JF, Bassani-Sternberg M, Berry MJ, Biffo S, Boris-Lawrie K, et al. 2025. Translon: a single term for translated regions. *Nat. Methods* 22:2002–2006.

Takou M, Hämälä T, Koch EM, Steige KA, Dittberner H, Yant L, Genete M, Sunyaev S, Castric V, Vekemans X, et al. 2021. Maintenance of Adaptive Dynamics and No Detectable Load in a Range-Edge Outcrossing Plant Population.Harris K, editor. *Mol. Biol. Evol.* 38:1820–1836.

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.

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.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Mar Alba M. 2008. Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Mol. Biol. Evol.* 26:603–612.

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 der Auwera G, O'Connor B. 2020. Genomics in the Cloud. 1st edition. O'Reilly: O'Reilly Media, Incorporated

Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, Carvunis A-R. 2023. A vast evolutionarily transient translatome contributes to phenotype and fitness. *Cell Syst.* 14:363-381.e8.

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.

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.

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 D-D, Wang X, Li Y, Zeng L, Irwin DM, Zhang Y-P. 2014. "Out of Pollen" Hypothesis for Origin of New Genes in Flowering Plants: Study from Arabidopsis thaliana. *Genome Biol. Evol.* 6:2822–2829.

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 translatome 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.

Zheng EB, Zhao L. 2022. Protein evidence of unannotated ORFs in Drosophila reveals diversity in the evolution and properties of young proteins. *eLife* 11:e78772.

Zhu X-T, Zhou R, Che J, Zheng Y-Y, Tahir Ul 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.