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 Bonrberg-Bauer[3], 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

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 of these *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 produced Ribosome Profiling data in the closely related species *Arabidopsis halleri* and *A. lyrata,* that we combined with the model *A. thaliana,* and compared genome-wide patterns of translation across the three species. Synteny analysis revealed 211 Open Reading Frames (ORFs) that show signs of active translation and that have emerged *de novo* within the *Arabidopsis* genus. Most of these *de novo* translated ORFs were species-specific and neutrally evolving, indicating their transient nature. 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, 34 were previously annotated as CDS in at least one genome, and are amongst the most promising gene candidates. Our results illuminate the abundance of translation events outside of conserved CDS, and their role as starting material for the emergence of novel genes.

Introduction

According to François Jacob, "The probability of a functional protein appearing *de novo* by random association of amino acids is almost zero" (Jacob 1977). This quote embodies the paradigm that historically dominated evolutionary genetics until the early 21st century. According to this paradigm, evolution is a "tinkerer", simply reusing pre-existing coding material to generate new genes. These tinkering processes include gene and genome duplication, horizontal gene transfer, exon shuffling, gene fusion/fission, retroposition or the activity of mobile genetic elements, among others (Ohno 1970; Long et al. 2003; Zhou et al. 2008). In contrast, *de novo* gene emergence implies a mutational origin from ancestrally non-coding nucleotide sequences. Like Jacob's quote suggests, the molecular characteristics of genes were considered far too complex to be the result of chance alone, and *de novo* gene emergence was seen at best as a very minor contributor to genetic novelty (McLysaght and Hurst 2016; Van Oss and Carvunis 2019). Yet, this 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 discovered in *Drosophila* (Begun et al. 2006; Levine et al. 2006), yeast (Cai et al. 2008) and primates (Knowles and McLysaght 2009). Since these first discoveries, *de novo* genes have been identified, over the past decade, in an increasingly wide range of organisms from varied taxa. Some were extensively characterized at the functional level and have key fitness consequences, including the antifreeze glycoprotein in cod, which emerged as a vital component of survival in polar environments (Baalsrud et al. 2018). Overall, it is now clear that *de novo* genes are important drivers of the evolutionary process (Chen et al. 2013; Blevins et al. 2021; Montañés et al. 2023; Peng and Zhao 2024; Xia et al. 2025).

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

al. 2023; Jin et al. 2025), the continuum theory has been supported by evidence from an array of organisms (Ruiz-Orera et al. 2018; Schmitz et al. 2018; Van Oss and Carvunis 2019; Heames et al. 2020; Papadopoulos et al. 2021; Peng and Zhao 2024), gradually deciphering the still unclear mechanisms by which *de novo* genes acquire their molecular properties as they become older. Starting from untranscribed and untranslated non-genic regions, whether the Open Reading Frame (ORF) sequence or the transcription of the region emerged first has long remained unresolved. Recent modeling, grounded in experimental data from yeast and drosophila, suggests that gene emergence more often follows a 'transcription-first' trajectory, with ORF formation occurring subsequently (Iyengar and Bornberg-Bauer 2023). Following the acquisition of transcription, either progressive lengthening, or selection of the longest transcripts would give rise to "proto-genes", the starting material from which *de novo* genes emerge (Carvunis et al. 2012). Subsequent steps include the acquisition of translation and the avoidance of both aggregation and proteasomal degradation, enabling the polypeptide to adopt an structural conformation, whether ordered or disordered (Wilson and Masel 2011). Most studies have focused on novel protein-coding sequences that had already been annotated as genes, and hence harbor clearly established "gene-like" characteristics (Ruiz-Orera et al. 2015; Vakirlis et al. 2018; Bornberg-Bauer et al. 2021; Peng and Zhao 2024). However, little is known about the very first steps of *de novo* gene emergence, especially the properties of the 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 cerevisiae* (Durand et al. 2019) seem to indicate that small, translated ORFs located outside of annotated protein-coding sequences and without coding homologues in neighbouring species could provide the basis for *de novo* gene emergence through neutral evolution. The function of these ORFs is expected to be limited, though previous studies have found them to be associated with morphogenesis in plants (Hanada et al. 2013) or with immunity in humans (Chen et al. 2020), and they are rather viewed as a "reservoir" from which a few novel functional protein coding genes might ultimately emerge (Ruiz-Orera et al. 2018; Durand et al. 2019; Ruiz-Orera and Albà 2019). Yet, these results remain to be generalized to supplementary taxa.

Studying the early steps of this process is challenging for at least two reasons. First, depending on the set of thresholds chosen to distinguish non-genic from genic elements (Dinger et al. 2008), millions of potential peptide-encoding sequences can be annotated in the intergenic regions of any genome. Selecting the most promising gene candidates among them requires translational evidence, which is hard to obtain precisely because they are not expected to

harbor the hallmarks of canonical protein-coding sequences (including greater length and sustained expression), and because their expression tends to be low. 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). Second, confidently asserting that a given ORF has emerged *de novo* requires the identification of a clear homologous non-coding region. However, detecting homology lacks power beyond a certain degree of divergence, resulting in a potentially serious "phylostratigraphic bias" (Moyers and Zhang 2015; Moyers and Zhang 2016; Moyers and Zhang 2017; Moyers and Zhang 2018). The best way to circumvent this problem is to use synteny information and only compare intergenic regions flanked by orthologous genes that have remained syntenic (Vakirlis et al. 2020). Because genomes can be rapidly and drastically rearranged over the course of evolution, this involves 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 one or a handful of distantly related species (Ruiz-Orera et al. 2018; Xiao et al. 2018), or by establishing detailed synteny relationships between closely related species, but without RPF data from each of them (Vakirlis et al. 2020; Roginski et al. 2024). To our knowledge, a single paper has combined synteny analyzes with the production of RPF datasets from closely related species, demonstrating a high turnover of translational activity of emerging *de novo* ORFs in yeast (Durand et al. 2019). Hence, combining genomic and translation data in a comparative framework has shown great potential in deciphering the first steps of *de novo* gene birth at a short evolutionary timescale, but has so far been under-exploited.

The genome-wide study of *de novo* genes, as well as the functional characterization of a few candidates, has focused mostly on a handful of model organisms, including Drosophila (Peng and Zhao 2024), yeast (Vakirlis et al. 2018) and primates (Ruiz-Orera et al. 2015). In comparison, little is currently known about how *de novo* genes arise in plants, and very few have been functionally verified. One early example of a plant *de novo* gene is the qua-quine-starch (QQS) gene in *A. thaliana*, which was found to encode a 59 amino-acid polypeptide involved in the regulation of starch synthesis in response to external and internal perturbations (Li et al. 2009). A more recent example, also in *A. thaliana*, is the Sun Wu-Kong

(SWK) *de novo* gene, which was found to encode a 8kD protein involved in seed germination under osmotic stress (Jin et al. 2025). In rice, the *de novo* gene GSE9 was proven to contribute to rice grain shape, partly responsible for the differences between *indica* and *japonica* varieties (Chen et al. 2023). However, beyond such specific examples, only few studies have investigated the global emergence of *de novo* genes in plants, and most of them also focused on single model species like *A. thaliana, Oryza sativa, Prunus persica, Arachis hypogaea* and *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). In *A. thaliana*, 782 *de novo* genes have been annotated, and found to be shorter, more highly methylated and more likely to be expressed in pollen than other gene categories (Li et al. 2016). Yet, a strong limitation is that these studies focused on genes that had already been characterized as such, and none has investigated the earlier steps of *de novo* gene birth, such as the role short, translated ORFs might play. Translation itself is under-characterized, with only a handful of high-quality RPF datasets available in plants, all from distantly related species such as *O. sativa* (Zhu et al. 2023), *Zea mays* (Chotewutmontri et al. 2018) or *A. thaliana* (Hsu et al. 2016; Wu and Hsu 2022). Again however, these studies focused mostly on genes that had already been annotated. The fact that pervasive translation activity can be detected in intergenic regions has been demonstrated in *A. thaliana* (Hanada et al. 2007; Hanada et al. 2013), but the role this may play in the context of *de novo* gene emergence has not been explored. Hence, it has so far not been possible to assess the importance of short, translated intergenic ORFs in *de novo* gene birth in plants.

Here, we aimed to characterize the first steps of *de novo* gene emergence in the *Arabidopsis* genus by specifically investigating the extent to which short ORFs could serve as a reservoir for the appearance of new genes (Ruiz-Orera et al. 2018; Durand et al. 2019). We focused on three species of the *Arabidopsis* genus: *A. halleri*, *A. lyrata* and *A. thaliana*, and produced RPF 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, resulting in a unique opportunity to reliably identify the non-coding sequences that are homologous to the putative *de novo* ORFs. Overall, our analysis reveals dozens of translated *de novo* ORFs within intergenic syntenic regions, shedding new light on the very first steps of the emergence of *de novo* genes in plants.

Results

Ribosome Profiling sequencing yields evidence of ongoing translation

Our overall strategy was based on four closely related Brassicaceae species for which high-quality reference genome assemblies are available (*A. thaliana*; (Lamesch et al. 2012), *A. lyrata*; (Hu et al. 2011), *A. halleri;* (Pavan et al. 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 S3**). Second, we sought to obtain direct experimental evidence for translation activity. To that end, we grew plants of the four species under hydroponic conditions and extracted total RNAs from a controlled mixture of leaf and root tissue (with three biological replicates per species; *i.e.* a total of 12 samples). We prepared RPF libraries and obtained a total of 90 to 140 million raw sequencing reads per replicate (**Table S1**). Reads were filtered based on their length (between 25 and 35 nucleotides) and quality (minimum quality score of 20), followed by a bioinformatic step to remove sequences with similarity to rRNAs. Substantial rRNA contamination remained in our libraries (50 to 70% of reads after quality control, **Table S1**), despite the ribo-depletion included in the libraries preparation protocol. The mapping rates of sequencing reads to their respective reference genomes varied across species (with lower mapping success in *A. halleri,* possibly due to contamination of the hydroponic culture medium by *Acidovorax* bacteria, **Table S1**). In the end, we could use between 0.8 and 16 million mapped reads per replicate for the detection of translation signals.

We first evaluated the quality of RPF libraries by examining metagene plots produced by RibORF (Ji 2018). A high quality RPF dataset would exhibit, within the CDS of protein-coding genes, 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 RPF datasets from (Hsu et al. 2016; Wu and Hsu 2022), comprising a total of 40 to 110 million mapped reads (**Table S2**). Ribosomal footprints are expected to be around 28 nucleotides long (*i.e.* the length of the mRNA segment protected by the ribosome (Ingolia et al. 2009)), and accordingly we used 27 to 32 nucleotide-long reads for the detection of translation signals (**Figure S2**).

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

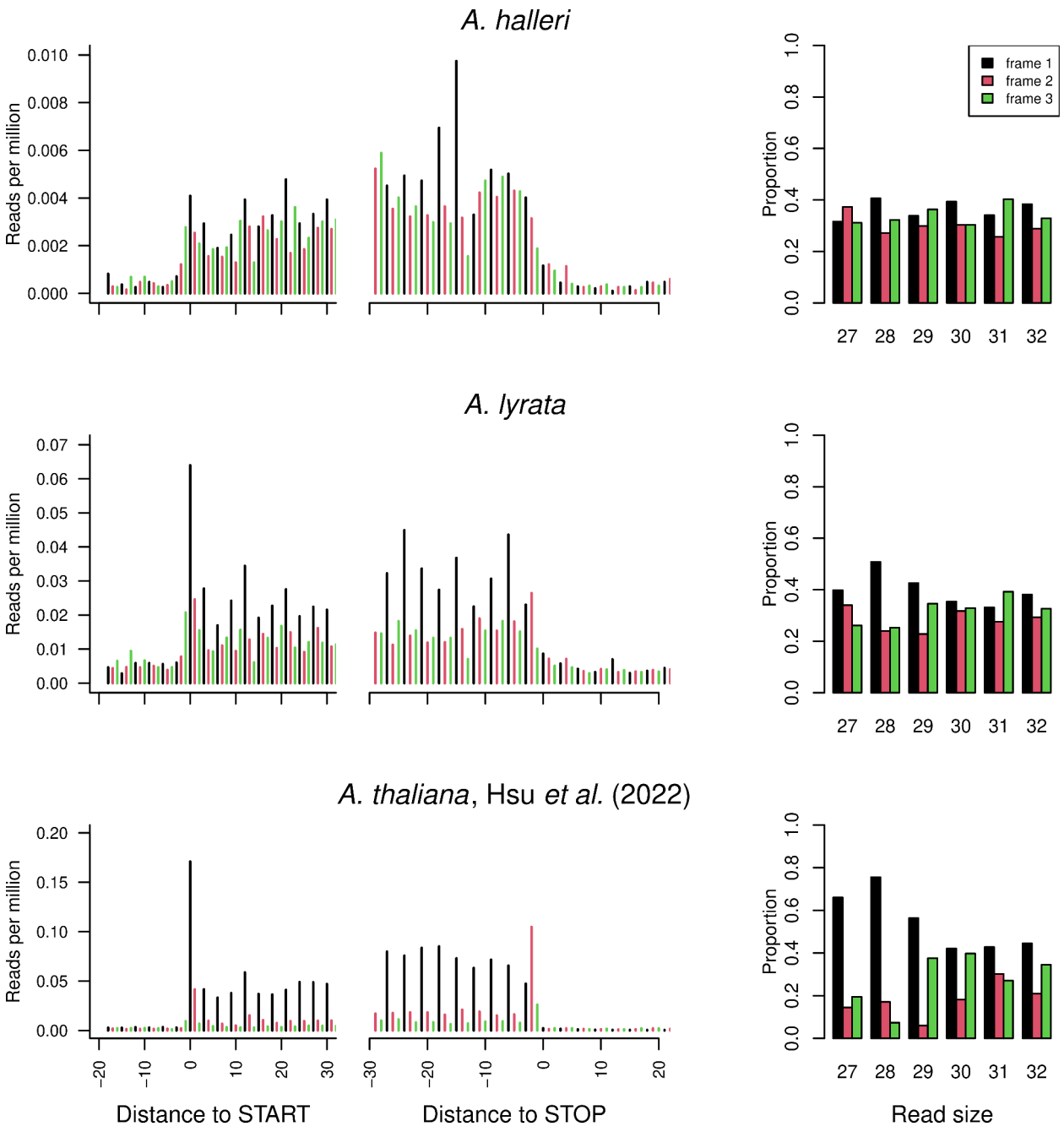


Figure 1: Metagene plots of 28-kmers from ribosome profiling, after P-site shifting (reads were shifted according to their k-mer specific P-site offset, and trimmed to be 1 base-long), used for the annotation of translation events; and prevalence of k-mers depending on which frame they are found in, produced with modified RibORF scripts (Ji 2018). Metagene plots are created by concatenating the RPF accumulation within all known canonical protein-coding genes, measured in reads per million (RPM) relative to the start position of CDS. Data from *A. halleri* and *A. lyrata* come from the AHA_rep-2 and ALY_rep-1 samples of this study (**Table S1**), and data from *A. thaliana* come from (Wu and Hsu 2022). Read accumulation is colored based on its frame relative to the CDS: black for the first frame, red for the second frame and green for the third frame. The relative abundance of each k-mer, between 27 and 32 nucleotides in size, is also colored based on the frame in which the reads are located.

Identifying individual ORFs with translation signals

The identification of translation signals at individual ORFs remains a challenging task even with high-quality RPF data, especially in the context of the discovery of recently emerged *de novo* genes (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 RPF data. First, we used the least stringent "ORFscore" (Bazzini et al. 2014), which is based on the preferential accumulation of reads in the first reading frame, taking the bias in canonical genes as a reference to distinguish 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 tools (ORFscore, RibORF or RiboCode). Among them, 20,592 and 13,266 CDS were found to be translated by the three tools, respectively in *A. thaliana* (75%) and *A. lyrata* (44%). In *A. halleri*, the difference is 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 S3**). This observation is probably due to the less apparent periodicity of the RPF dataset in *A. halleri* (**Figure 1**). Hence, the samples from the different species differ in the proportion of CDS detected as translated and the number of tools detecting CDS as being translated. It should also be noted that not all canonical CDS are expected to be identified as being translated by this single experiment, due to inherent technical (low power of RPF data to assess translation on a per-locus basis) and biological limitations (*e.g.* only two tissues sampled, developmental stage-specific expression). Overall, we conclude that the signals we detect are powerful, but correspond to an under-estimation of the actual translation potential that may exist.

Beside the expected signals of translation overlapping canonical CDS, we identified translation signals for 3 to 23 million Intergenic ORFs (IGORFs, that is, ORFs without any overlap to any previous genome annotation) depending on the species considered (**Table S3**). 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 S3**). The CDS and IGORFs that have the most robust translation signatures, *i.e.* those identified by all three tools, are translated at a relatively higher level compared to other features, illustrating the importance of sequencing depth in annotating translation events (**Figure 2**). Across species, 1,825 IGORFs were detected to have such robust translation signatures (135 in *A. halleri*, 1,548 in *A. lyrata* and 142 in *A. thaliana*; **Figure 2** and **Table S3**). Furthermore, at the same level of expression, CDS are more often detected as translated than IGORFs, reflecting the training of tools on canonical gene annotations, requiring lower depth to identify periodicity within canonically translated regions. Among all 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.

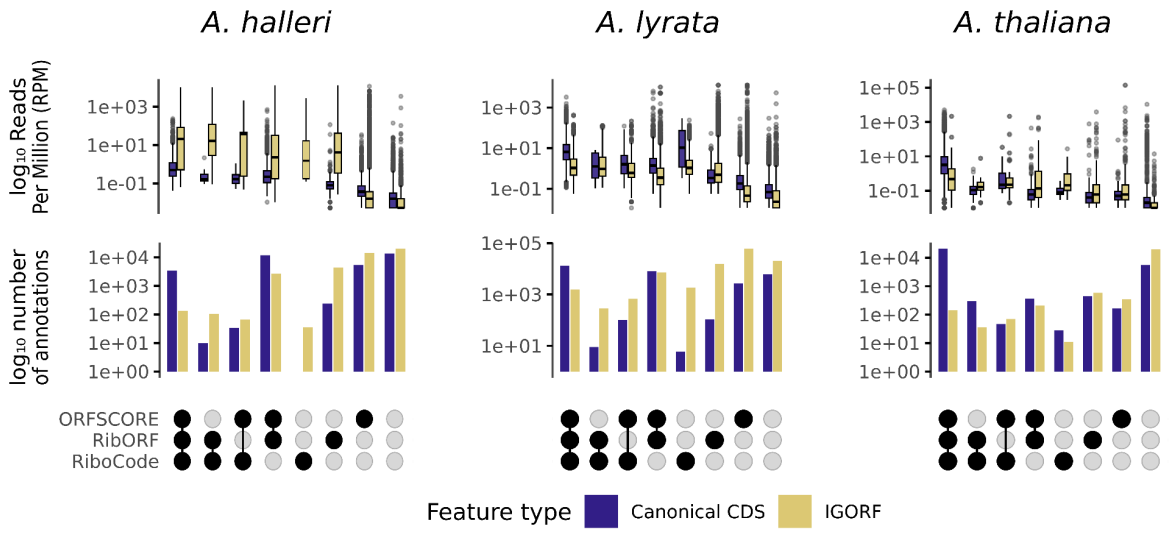


Figure 2: Upset plot representing the number of annotations detected as translated and their translation levels in $\log_{10}$ Reads Per Million (RPM, where each read is a ribosome footprint); according to the tools that detected them as translated, and to the category to which they belong: canonical 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).

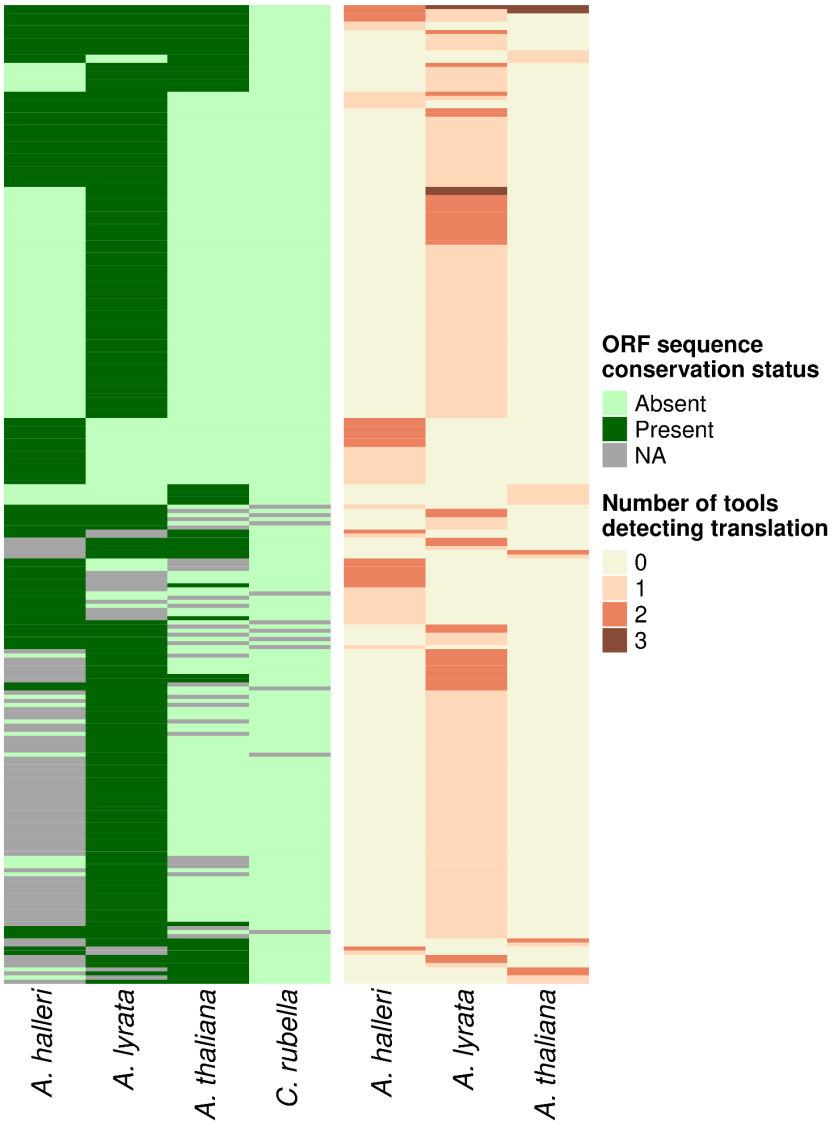
Synteny analysis identifies many de novo ORFs being translated

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* or *A. thaliana,* which we used as outgroups. To do so, we used conserved CDS as orthology anchors to finely define syntenic regions between pairs of species. For non-conserved, translated CDS and IGORFs found within these syntenic regions, we then used global sequence alignment to identify the corresponding orthologous, micro-syntenic region in each other species. ORFs located in chromosomal regions that are too divergent to allow sequence alignment were discarded (**Table S4, Figure S3**). This yielded between 2,192 and 8,614 aligned potential *de novo* ORFs (**Table S4**). Blast searches were also performed against the protein sequences from a database of 5 Brassicaceae species, as well as the whole genomes of each focal species to remove ORFs with homologs in other Brassicaceae species, or homology outside of their annotated syntenic region. 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 1) because they are fully absent, 2) because they contained a frameshift mutation or 3) lacked a START and a STOP codon at the same positions.

Based on these stringent thresholds, a total of 42, 163 and 14 *de novo* ORFs with 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*, all of which contained less than five exons, with 70% displaying only one (**Table S5**). Having removed redundancy between 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 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.* Most of these *de novo* ORFs were found in *A. lyrata*, closely followed by *A. halleri*, which can be attributed to the greater genome size of those species (i.e. much longer intergenic regions) compared to *A. thaliana*, as well as the better periodicity of the *A. lyrata* RPF dataset compared to the *A. halleri* one. The lesser amount of *de novo* ORFs in *A. thaliana* might be attributed to the thorough annotation of its

genome, however in this case, if the rate of *de novo* gene birth was the same in all three species, we would have found many *de novo* ORFs already annotated as CDS. This is not the case here: CDS and IGORFs together, we found less *de novo* ORFs in *A. thaliana* than in the two other *Arabidopsis* species. The ORFs detected in multiple species, as opposed to just one, tend to be identified as translated by more tools (**Figure 3**), which could be interpreted as evidence for a more robust translation signal, *i.e.* a greater functional potential of conserved ORFs. Still, 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.



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

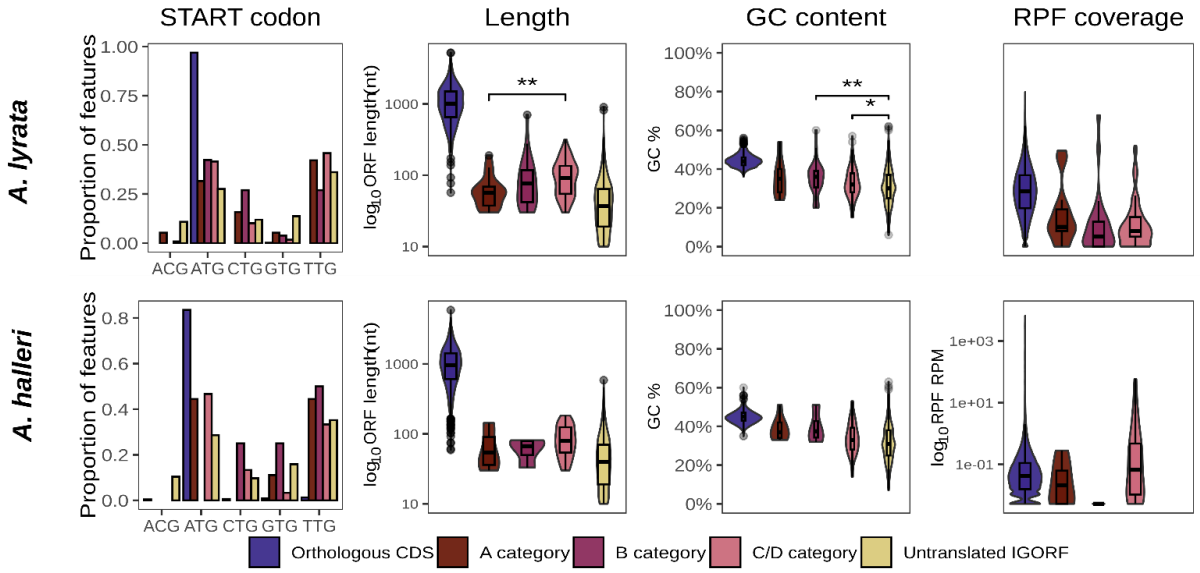
We then compared features of these *de novo* translated ORFs to those of conserved CDS (canonical CDS with an identifiable ortholog in every other species of interest, used as synteny anchors) on one hand, and to "untranslated intergenic ORFs (unt_IGORFs) (with no coverage by ribosome footprints) on the other hand. The nature of our data, characterized by highly divergent, non-coding genomic regions between only a handful of species, rendered ancestral sequence reconstruction (ASR) unreliable. Hence we were unable to determine at which specific point of the phylogeny the ORFs emerged. As a result, we refer to conservation classes rather than to actual evolutionary age. Specifically, *de novo* translated ORFs were classified into different conservation categories: A) conserved in all *Arabidopsis* species, B)

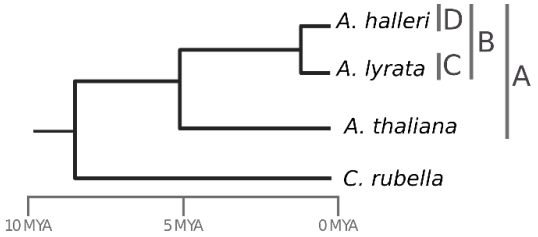
conserved in *A. halleri* and *A. lyrata*, C) *A. lyrata*-specific and D) *A. halleri*-specific (**Figure 4**). It should be noted that some ORFs were specific to *A. thaliana*, in which case our inability to conclude as to their true origin led us to exclude them from further analyses: they could have either emerged before the divergence of *A. thaliana*, and have subsequently been lost in *A. halleri* and *A. lyrata*, or have emerged after the divergence of *A. thaliana.*

The following analyzes will concern *A. lyrata* only, as the limited number of *de novo* ORFs per conservation class in *A. halleri* do not allow for sufficient statistical power. However, despite this limitation, the trends observed in *A. lyrata* are still visible in *A. halleri*, and the results are nonetheless showcased (**Figure 4**). Firstly, we examined the usage of the three alternative START codons of the genetic code. We found a prevalence of non-canonical START codons in the *de novo* ORFs, specifically TTG, that closely resembled that of unt_IGORFs (**Figure 4**). In contrast, as gene prediction algorithms only consider canonical start codons, orthologous CDS exclusively harbor ATGs. Secondly, we observed that, as expected, orthologous CDS were on average much longer (mean length = 1,111 nt) than unt_IGORFs (mean length= 52 nt), and than *de novo* ORFs (mean length = 100 nt) (**Figure 4**). A statistically significant difference in length was apparent between the A and C/D *de novo* conservation classes in *A. lyrata,* with a shorter length of the more conserved ORFs. This is most likely a consequence of the way conservation is stringently defined, as longer ORFs are more likely to be lost by mutation compared to shorter ones, thus being larger mutational targets. Third, we observed that the more conserved *de novo* ORFs tend to have a higher GC content (median ~35%) than less conserved ones, lower than that of CDS (median ~44%), but significantly above that of the unt_IGORFs, which had the lowest GC content (median ~30%) (**Figure 4**). Finally, in line with previous observations (Carvunis et al. 2012), we observed that translation activity (quantified by coverage by aligned RPF) of *de novo* ORFs was less intense than that of orthologous 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 orthologous CDS, indicative of ongoing translation. We verified that this difference in coverage was not caused by the difference in the length of the annotations; by testing the influence of length on expression level. Using a linear least squares regression, we found that length alone was not sufficient to explain differences in expression levels (*estimate=-0.002224, p.value= 0.9272*). Furthermore, the differences in coverage were also apparent when length differences were controlled for by calculating coverage on the first 60 nucleotides of each feature only, where the translation signal was most pronounced. Again, length alone was not

sufficient to explain differences in expression levels (*estimate=-0.01254, p.value=0.893623*). To further explore the possible factors (conservation or translation) behind these patterns, we examined the same properties between conservation classes of unt_IGORFs (**Figure S4**). Across conservation classes, untranslated IGORFs predominantly harboured ATG and TTG START codons, like their translated *de novo* counterparts. Similarly, the same length pattern was observed, with ORFs from classes A and B being statistically significantly longer than ORFs belonging to the C/D class, reflecting how smaller ORFs are smaller mutational targets, less likely to be lost by mutation in other species. When it comes to GC content however, no significant difference was observed between conservation classes. 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**). While, by lack of sample size, no statistically significant difference could be found in A and B classes, the pattern is nevertheless present: translated *de novo* ORFs seem to systematically harbor a higher GC content. In conclusion, while the conservation characteristics of translated *de novo* ORFs, similar to those of unt_IGORFs, seem to explain both the prevalence of non-canonical START codons, and the length patterns observed across conservations classes; the observed trends in GC content are specific to translated *de novo* ORFs, and might be an effect of translation.

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).





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**

De novo ORF - encoded peptides are highly disordered

We then examined the level of Intrinsic Structural Disorder (ISD) of polypeptides encoded by de novo ORFs. ISD was predicted using AIUPred (Erdős and Dosztányi 2024), on all 211 *de novo* ORFs, as well on a random sub-sample of 1,000 orthologous CDS and 1,000 unt_IGORFs. Statistically significant differences were observed between unt_IGORFs and orthologous CDS, as well as between *de novo* ORFs and orthologous CDS, with both unt_IGORFs and *de novo* ORFs exhibiting the highest levels of ISD. Further analyses however demonstrate a statistically significant negative correlation between the polypeptide length and

ISD, (**Figure S6**), which yet only explains a small part of the observed variance (*adjusted* R^2 = *0.03532, p-value < 2.2e-16)*. On the contrary, GC content seems to be positively associated with ID, at least in some feature categories (in conserved CDS: *adjusted* R^2*=0.08435, p-value < 2.2e-16)*. When sub-sampling 86 CDS of similar length and GC content as unt_IGORFs (less than the maximal 139 amino acid residues, and between 6 and 61% GC content) , we no longer observe a statistically significant difference in ISD between categories. Hence, while we do observe that on average, *de novo* ORFs tend to be more highly disordered than CDS; it is difficult to further interpret the meaning of this difference. Indeed, length has been reported to influence ID (Peng et al. 2006), and is negatively linked to it in our data; and as previously reported in *eukaryotic* orphan genes (Basile et al. 2017), we also observe a positive link between ID and GC content.

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

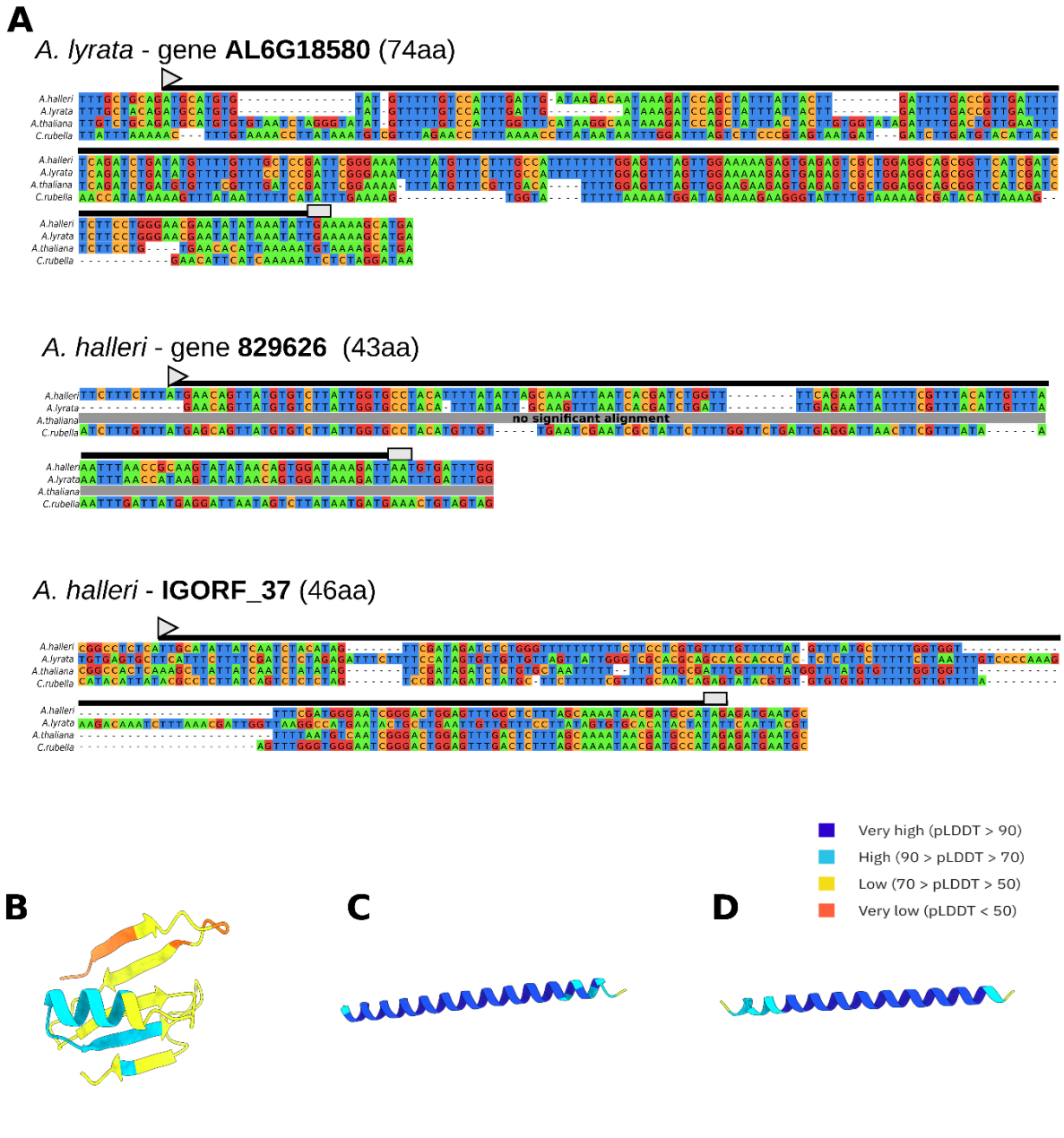
Out of 211 translated *de novo* ORFs, a few notable ones stand out as having a number of gene-like characteristics, making them ideal candidates for *de novo* gene emergence. A first set of these translated *de novo* ORFs had already been annotated as CDS, and we therefore refer to them as "*de novo* CDS". For example, the *A lyrata*-specific CDS annotated as AL6G18580 (**Figure 5**) has no homology to previously described proteins, displays translation signals by two tools, and is 225 nucleotides in length (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 de novo ORF emerged within the *Arabidopsis* genus. While the START and STOP codons were conserved in *A. halleri* and *A. thaliana*, a few indels lead to the loss of the reading frame. 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 has previously been predicted as a non-coding RNA in *A. lyrata*, 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 in its 3' extremity the start of the precursor and the future mature miRNA, leaving out the miRNA* and hence eliminating the possibility for a stem-loop structure to form within the transcript. For the potential encoded polypeptide AL6G18580 (**Figure 5, B**), AlphaFold2 confidently predicts 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).

Another example of *de novo* CDS is the *A. halleri* - specific 829626 gene (**Figure 5**), which has no homology to any previous annotation (**Table S6**), was identified as translated by two tools, 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, C**).

Most of the *de novo* ORFs we identified were classified as “intergenic ORFs”, although some of them have features similar to those of the *de novo* CDS. For example, the *A. halleri*-specific intergenic ORF named IGORF_37 (**Figure 5**) was detected as translated by two separate tools; similarly to CDS from *A. halleri*. This *de novo* ORF is 147 nucleotides long (46 amino acids, hence shorter than the commonly used 300 nucleotides threshold for gene annotation; Dinger et al. 2008). The orthologous sequence exhibits indels in every other species we studied, leading to a frameshift and the loss of either the START or the STOP codon. The sequence exhibits no homology to any known protein-coding gene, 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, D**). In conclusion, while most translated *de novo* ORFs possess features similar to those of unt_IGORFs, a few ones stand out as the best candidates for *de novo* gene emergence. They can be described as intermediates between unt_IGORFs and conserved CDS: they are longer than unt_IGORFs but shorter than conserved CDS, have not been annotated as CDS (or if they have, have not previously been functionally characterized), and yet are independently detected as translated by several tools. Structural predictions reveal the potential for a few and simple structures, aligning with structure prediction results of *de novo ORFs* in other species (Chen et al. 2024; Middendorf and Eicholt 2024; Peng and Zhao 2024).

Figure 5: (A) Alignments of potential *de novo* gene candidates, either *A. lyrata* (AL6G18580) or *A. halleri* (829626, IGORF_37) specific. Alignments are colored by nucleotides. The START codon is represented by a grey arrow, and the STOP codon by a grey box. The ORF itself is represented by a black line above the alignments. When no homologous syntenic region could be found for a given species, the sequence was replaced by a "no significant alignment" box. **(B-D)** AlphaFold2 structure predictions of the predicted corresponding polypeptides are represented for ORFs AL6G18580 **(B)**, 829626 **(C)**, and IGORF_37 **(D)**. Residues are colored by pLDDT score (see legend).



The repertoires of de novo ORFs are highly labile within species

To gain further insight into the conservation of *de novo* ORFs, we then analyzed their within-species distribution, and collected 5 high-quality genome assemblies of both *A. halleri* and *A. lyrata*. The reference annotations from the *A. halleri* (AUBY1) and *A. lyrata* (MN47) reference accessions were respectively transposed to the corresponding accessions, within which syntenic regions were annotated. We then assessed the conservation status of each ORF that fell within syntenic regions that were conserved between accessions, using the previous filters and alignments (cf. **Figure S3**). The resulting conservation pattern is presented in **Figure 6**, which showcases for both *A. halleri* and *A. lyrata* a high degree of lability. While, in *A. lyrata*, approximately 21% of translated *de novo* ORFs were accession-specific, and 29% were found in at least 3 accessions; in *A. halleri,* respectively 25% and 62% of ORFs fell within this category. As a point of comparison, the conservation of untranslated IGORFs was also assessed between accessions of the same species. A simple “conservation score” was devised, computing the number of accessions within which each ORF was conserved, and allowing for the comparison of conservation score distributions between heatmaps. Randomly permuting “translated *de novo* ORF” and “untranslated IGORF” labels 1000 times, a Wilcoxon rank sum test was performed each time, resulting in a W statistic distribution from which a p-value could be deducted. Whereas no significant difference between *de novo* ORFs and untranslated IGORFs was observed in *A. lyrata*, a p-value of 0.014 was assigned to the observed difference in conservation score distribution between the two ORF categories in *A. halleri*. This indicates that, on average, while *de novo* ORFs repertoires are highly labile within species, at least in *A. halleri,* translated *de novo* ORFs are significantly more likely to be conserved across accessions than random untranslated IGORFs.

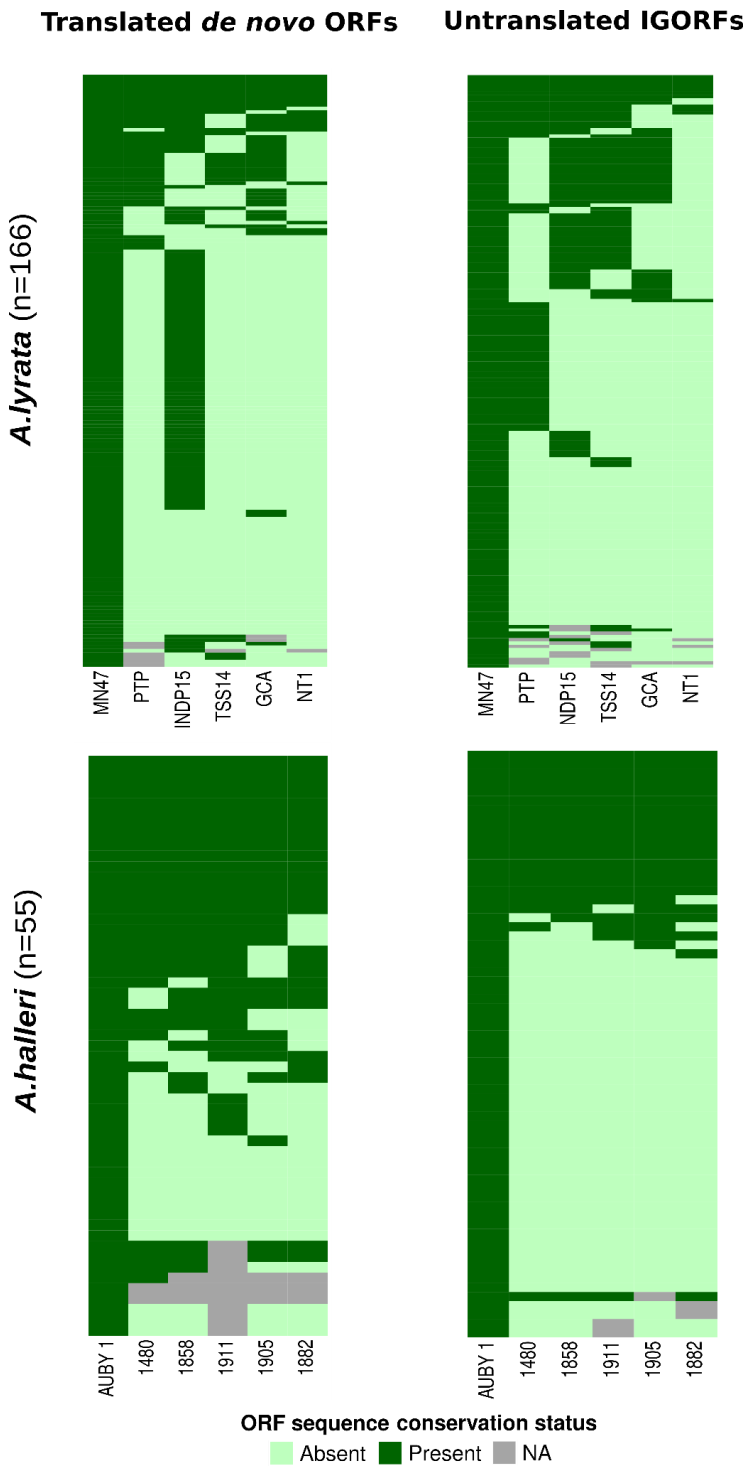


Figure 6: Heatmap of the conservation of translated *de novo* ORF and untranslated IGORF sequences across accessions from the same species. Each line represents an individual ORF, and each column an accession. If the ORF is conserved in an accession, it is annotated as "present", if not it is annotated as "absent". If no homologous syntenic region could be aligned, the ORF was annotated as "NA".

De novo ORFs evolve neutrally - similarly to untranslated intergenic ORFs

Given some of the observed characteristics of translated *de novo* ORFs (prevalence of non-canonical START codons, lower ribosome profiling coverage), reminiscent of those of untranslated intergenic ORFs, we expected *de novo* ORFs to evolve neutrally, or under relatively minimal selection. To test this hypothesis, we collected genomic data from a range of European and North-american *A. halleri* and *A. lyrata* populations (with a total of 37 *A. halleri* individuals, and 39 *A. lyrata* individuals (**Table S7**)). We then detected the polymorphism within each species, and calculated pN/pS ratios for different feature categories. Because *de novo ORFs* tend to be much shorter than orthologous CDS, computing ratios ORF per ORF would have been unfeasible due to the lack of polymorphic sites. Therefore, we concatenated *de novo* ORFs until a length equivalent to that of the shortest orthologous CDS was reached, and then computed ratios. Because untranslated IGORFs suffer the same length problem, the same method was applied to them. In both *A. lyrata* and *A. halleri,* orthologous CDS present a mean ratio of ~0.25, indicative of selective negative pressure; while untranslated IGORFs have a mean ratio of ~1.8, implying mostly neutral, possibly slightly positive selection (**Figure 7**). Translated *de novo* ORFs, on the other hand, exhibit a mean ratio of 1.6 in *A. lyrata*, and 1.3 in *A. halleri*, indicative of neutral evolution. In *A. lyrata,* while a significant difference is observed between orthologous CDS and untranslated IGORFs (*W = 14783*, *p-value < 2.2e-16)*, as well as between orthologous CDS and *de novo* ORFs (*W = 5552*, *p-value < 2.2e-16*), no significant difference is found between untranslated intergenic ORFs and *de novo* ORFs (*W = 3727*, *p-value = 0.8288*). The exact same pattern is observed in *A. halleri*, implying that in both species, orthologous CDS evolve under negative selective pressures, while *de novo* ORFs and untranslated IGORFs both evolve neutrally.

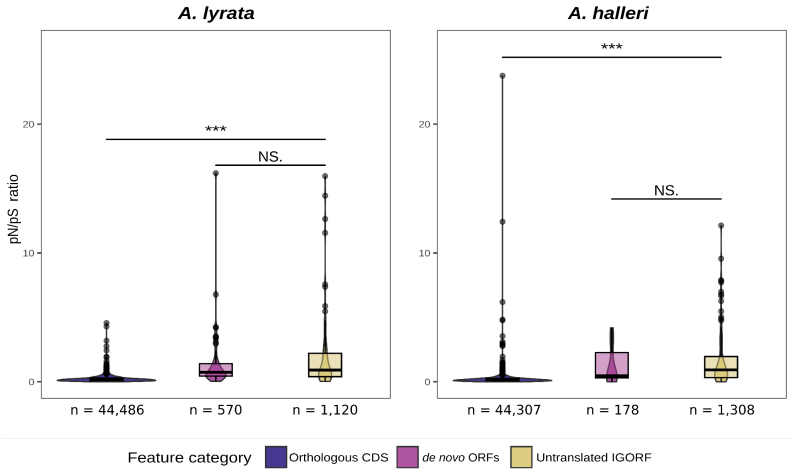


Figure 7: pN/pS ratio of conserved canonical CDS, *de novo* ORFs and untranslated IGORFs. *De novo* ORFs were concatenated across conservation classes, as were untranslated IGORFs. The total number of polymorphic sites used for the calculation of pN/pS ratios is indicated by n. Statistically significant differences are represented, except for those between orthologous CDS and *de novo* ORFs, where $p < 2.2e-16$ (Mann-Whitney U-test, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$).

Discussion

Pervasive translation with a rapid evolutionary turnover

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

Such transitory pervasive transcription (Neme and Tautz 2016; Schmitz et al. 2018) and translation (Ingolia et al. 2014; Ruiz-Orera et al. 2018; Blevins et al. 2021; Wacholder et al.

2023; Papadopoulos et al. 2024) has been described as the “transient translatome” (Wacholder et al. 2023). The overall prevalence of species-specific translated ORFs (30 in *A. halleri*, 118 in *A. lyrata* and 7 in *A. thaliana*) further supports the idea of a high turnover of transcripts and of their translation, which would become temporary substrates for *de novo* gene emergence (Schmitz et al. 2018; Durand et al. 2019; Grandchamp et al. 2024). Sequence conservation itself is also rather low, as most *de novo* ORFs are species-specific, with even up to 25% being accession-specific. Hence, while many ORF candidates may appear by mutation within intergenic regions, few will be transcribed and translated, even fewer will be conserved across species, and only a select few might develop the functional potential to ultimately give rise to genes (Neme and Tautz 2014; Neme and Tautz 2016). While we lack the necessary power to assess the *de novo* gene emergence rate in *Arabidopsis*, a study in rice (Zhang et al. 2019) has reported a strikingly high rate of *de novo* gene emergence, with 51.5 *de novo* genes per Myr. Not only was the *de novo* emergence rate high, but so was the retention rate; both features which were linked to high protein diversity and phenotypic plasticity.

The first steps of *de novo* gene emergence

Pioneering studies investigating *de novo* gene emergence have mostly focused on already-characterized, functionally potent genes (Cai et al. 2008; Knowles and McLysaght 2009; Xiao et al. 2009). Following Carvunis et al. 2012; Durand et al. 2019; Ruiz-Orera and Albà 2019; Papadopoulos et al. 2021, we instead chose to focus on recently emerged translated ORFs, in order to characterize more specifically the starting material from which very young “proto-genes” might be found (Carvunis et al. 2012). Indeed, only a small subset of the *de novo* translated ORFs we identified had already been annotated as CDS (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 specifically characterized at the functional level. The large majority of *de novo* translated ORFs were intergenic, and are therefore expected to resemble intergenic ORFs. Accordingly, we note several lines of similarity between characteristics of *de novo* ORFS and those of untranslated IGORFs. In particular, the prevalence of non-canonical START codons, the short length and low GC content of *de novo* ORFs are either similar to those of untranslated IGORFs, or intermediate between them and those of conserved CDS. The lower expression level of *de novo* ORFs compared to orthologous CDS entails that whatever phenotype may be

associated with the *de novo* translation products, would be less visible to natural selection than those encoded by more highly-expressed features. Indeed, we observed that translated *de novo* ORFs, like untranslated intergenic ORFs, exhibit non-synonymous to synonymous mutation ratios close to 1; much higher than that of CDS, and indicative of neutral evolution. Under neither strong positive nor negative selective pressures, translated *de novo* ORFs hence evolve neutrally, which is in line with Ruiz-Orera et al. 2018 results in mice.

The processes by which random ORFs move away from random noise and eventually acquire gene-like characteristics such as higher GC content, greater length and higher transcription remain somewhat unclear. In rice, the late ORF–early transcription model has been described as the most prevalent mechanism for *de novo* gene emergence (Zhang et al. 2019), according to which non-coding regions would pervasively become transcribed and only later acquire an ORF by mutation, which would then already be visible to selection. Similarly to what has previously been suggested in *S. cerevisiae* (Carvunis et al. 2012; Papadopoulos et al. 2021), *de novo* ORFs could undergo, in plants, a progressive “refinement” towards gene-like features, including changes in START codon usage and expression levels (Li et al. 2016; Zhang et al. 2019). In a positive feedback-loop, increase in expression would render ORFs progressively more visible to selection, which in turn would lead to either their elimination or their functionalization. Under this scenario, new functions would emerge progressively, in an iterative process. Another explanatory scenario has been colloquially named “Born this way”, whereby pre-existing non-canonical ORFs that present the most “gene-like” features may be more likely to be retained by natural selection, while the others quickly fade back into expressional noise (Aubel et al. 2024). In our study, while we do observe statistically significant differences in ISD between conserved CDS and *de novo* ORFs at first glance, those differences disappear when length and GC content are matched. Indeed, length has been found as being a major factor in ISD prediction (Middendorf and Eicholt 2024) and the observed apparent differences in ISD would be an indirect consequence of the greater length of conserved CDS compared to young *de novo* ORFs. Within-species analyses allow us to gain further insight into the scenario under which gene-like characteristics are acquired. While translated *de novo* ORFs evolve neutrally, like unt_IGORFs do, they were found to be significantly more conserved than unt_IGORFs between high-quality accessions of *A. halleri*. Even though both translated *de novo* and untranslated intergenic ORFs repertoires are highly labile, with up to ⅔ of them being accession-specific, *de novo* ORFs can spread more efficiently across populations than unt_IGORFs.

From random intergenic ORF to canonical gene

In many species, including *A. thaliana*, the GC content within protein coding genes is lower in introns than in exons, and more globally in sequences that are not translated compared to those that are (Carels and Bernardi 2000; Singh et al. 2016). When comparing orthologous CDS (exons) to untranslated IGORFs, we observe a corresponding pattern in which coding sequences tend to have a higher GC content than non-coding ones. This difference will influence the frequencies of START and STOP codons along the genomes. In GC-rich regions of the genome, as AT content is lower, the prevalence of START (ATG,TTG,...) and STOP (TAA,TAG,TGA) codons is correspondingly lower, meaning putative reading frames would be longer (McLysaght and Hurst 2016). Inversely, within GC-poor regions, like intergenic regions in *Arabidopsis* species, putative reading frames would be shorter, and more numerous. This corresponds to what we observe both in the number and length of IGORFs compared to canonical CDS. This makes the intergenic regions of *Arabidopsis* species the ideal reservoir for novel ORFs. However, it remains unclear whether this trend can be generalized as a global mechanism for *de novo* gene emergence. Indeed, while similar patterns have been observed in terms of GC content in *A. thaliana* and *Drosophila*, (Li et al. 2016; Peng and Zhao 2024), the opposite trend has been observed in *de novo* genes in yeast, which tend to emerge in GC-rich regions, close to promoters or within recombination hot spots (Vakirlis et al. 2018). Strong correlations have been identified between GC content and other key genomic features, including length and intrinsic disorder (Basile et al. 2017; Nielly-Thibault and Landry 2019). Yet, in our data, the relationship between GC content and ISD remains hard to assess, as there appears to be a significant positive relationship between the two characteristics, at least in orthologous CDS and unt_IGORFs.

In line with previous observations (Basile et al. 2017; Willis and Masel 2018; Middendorf and Eicholt 2024), we found that *de novo* ORFs tend to harbor high levels of intrinsic disorder. This can at least in part be attributed to their shorter length, directly preventing the formation of secondary structures. This, however, does not mean that they are devoid of functional potential. Extensive experimental evidence highlights the functional potential of intrinsically disordered regions (IDRs), whose highly labile and context-dependent conformations can play key roles in molecular communication and environment-induced phenotypic responses (Holehouse and Kragelund 2023). The conservation of the function of IDRs is even possible in spite of a lack of sequence conservation, through the conservation of their physico-chemical properties

(Holehouse and Kragelund 2023). Within eukaryotic proteins at large, IDRs confer conformational flexibility to canonical proteins, especially those involved in transcriptional and post-translational regulation (Ahrens et al. 2017). Furthermore, small ORFs at large have been hypothesized to contribute greatly to phenotypic variability through mechanisms such as regulation of protein levels, function, recruitment (Orr et al. 2021), signaling or formation of protein complexes (Andrews and Rothnagel 2014; Brar and Weissman 2015; Schlesinger and Elsässer 2022; Chen et al. 2024; Middendorf et al. 2024).

An interesting perspective to our results would be to assess whether translated *de novo* ORFs tend to accumulate within specific regions of the genome. Indeed, proximity to conserved genes is one of the factors invoked as a mechanism by which *de novo* ORFs could benefit from increased transcription and translation, rendering them visible to natural selection (Andreatta et al. 2015; Kleppe and Bornberg-Bauer 2018; Grandchamp et al. 2022; Iyengar et al. 2024; Lebherz et al. 2024). This mechanism has been referred to as “transcriptional/translational readthrough”, and involves the emergence of newly coding, highly disordered sequences at the C-termini of highly expressed proteins, which provide conformational flexibility without disrupting protein function (Kleppe and Bornberg-Bauer 2018). *De novo* genes have been consistently described as being shorter than other genes, and the “grow slow and moult” hypothesis suggests that one of the ways *de novo* genes could progressively become longer as they age is through the acquisition of new domains at their C-termini (Bornberg-Bauer et al. 2015; Kleppe and Bornberg-Bauer 2018).

Overall, our findings suggest that pervasive translation takes place within intergenic regions in *Arabidopsis* species, providing a vast reservoir of transiently, lowly expressed, neutrally evolving short ORFs. Among those, a select few may ultimately give rise to functional genes, by progressively becoming longer, more highly expressed, and changing in GC content. Future work combining ribosome profiling, mass spectrometry, and functional assays across tissues and developmental stages will be essential to validate this repertoire of *de novo* ORFs and to determine their physiological relevance, particularly how they integrate into, and potentially disrupt, cellular interaction networks.

Material and methods

Plant Materials and Growth Conditions

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

Lysates preparation

Three individuals from each species were chosen for subsequent data production; based on the amount of material that had been gathered. Roots and leaves from the same single individual (at the exception of A. thaliana, where 9 individuals were merged into 3 groups of 3) were gathered together regardless of their sampling date. Tissue types were separately ground into a powder, using liquid nitrogen, a mortar and pestle. The obtained powder amounted to an approximate total of 1 g of tissue, per individual. Approximately 0.25 g of each tissue type was kept for ribosome footprint production, per individual. The powders were then suspended in 4 mL of lysis buffer, as described in (Wu and Hsu 2022). The obtained lysates were vortexed twice; distributed in 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 then transferred to refrigerated 1.5 mL tubes. Aliquots of 200 µL were produced and 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 seemed to contain on average much less material than leaf ones; root and leaf aliquots were pooled per individual; so as to obtain 75% of material coming from leaves, and 25% of material coming from roots.

Ribosome footprints

The lysates were treated with 50 units RNase I Ambion (Invitrogen, AM2295) per 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. An RNA extraction test was first performed using the Zymo RNA clean and concentrator kit (Zymo Research; R1017) on aliquots of all three replicates of A. lyrata, and on one aliquot of the second replicate of A. halleri. Standard protocol procedure was followed in order to extract RNAs between 17 and 200 nucleotides long. However, after subsequent quantification of those RNA extracts, it became apparent that they were too lowly concentrated in order to undergo gel migration (from less than 1 to 7 ng/µL; when we expected 10-100 ng/µL); and this extraction protocol was subsequently not used anymore. Instead, for the remaining aliquots, total RNA was extracted using a standard hot phenol procedure as described in (Baudin-Baillieu et al. 2016). Then, the totality of extracted RNAs from each aliquot were charged on 17% polyacrylamide; 7M TAE-urea gels, which were pre-migrated for 30 min at 100V. The TAE 1x migration of samples was performed at 100 V for approximately 5 h; with size markers at 28 and 30 nucleotides. After migration, gels were stained with SYBR-GOLD and gel slices between the 28 and 30 nt markers were excised. The excised gel band was pulverized and 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 um Corning column (Dutscher; 2515806); and centrifuged at maximal speed for 5mn. Per tube, 2uL 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 is adjusted with an equal, 100% ethanol volume; and 2 volumes (180 µL) of this adjusted buffer are added to the rRNA-depleted RPF samples (90 µL). Then, is added to this mix 1 volume of 100% ethanol (270 µL) to favor the retrieval of small RNAs between 17 and 200 nucleotides (ribosome footprints are typically approximately 30 nucleotides long). The total 540 µL are then deposited on a kit's column, and centrifuged for 30s at 16,000 x g at room temperature. Then, 400 µL of RNA Prep Buffer are added to the column, which is again centrifuged in identical conditions; followed by the addition of 700 µL of RNA Wash Buffer and a third identical centrifugation. Finally, 400 µL of RNA Wash Buffer are added; and the column is centrifuged for 1 min at 16,000xg at room temperature. The column is put on a 2mL tube and the ribosome footprints are eluded on the column with 15 µL of distilled, nuclease-free water; and then centrifuged for 1 min at 16,000xg at room temperature again. Again, the final microRNA concentration is measured by fluorescence with the micro-RNA Quant-iT kit (Thermofisher Scientific; Q32882).

Library preparation and sequencing

The 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.

Ribo-seq Data Analysis

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

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

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

Syntenic region annotation and alignment

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

PID in the ORF region within the pairwise alignment was ≥ 60%. ORFs that passed those criteria were used for Multiple Sequence Alignment (MSA); between the extended ORF sequence and every correctly aligned, extended syntenic region using MAFFT (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 50% in at least 3 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 if the gene's frame was the same as the reference ORF's one, then the ORF was considered as conserved.

Intrinsic disorder and structure prediction

Alphafold2 (v.2.3.0: Database cut-off data: 2022-12-19) (Jumper et al. 2021) predictions were performed on HPC cluster Palma II. Predictions with highest confidence (ranked_0) were used for analysis. Secondary elements were annotated using Stride (Frishman and Argos 1995). Disorder was predicted locally using AIUpred (no smoothing) (Erdős and Dosztányi 2024). 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*, the following five high-quality genome accessions were gathered: the North italian BN_1480, and the romanian CX_1858, DD_1911, DB_1905 and DO_1882 (unpublished assemblies, corresponding reads available at ENA project: PRJEB70880 (Batista et al. 2024)). For *A. lyrata*, the following five high-quality genome accessions were gathered: the North American PTP (unpublished assembly), INDP15 and TSS14.18 (ENA project: PRJEB80457 (Padilla-García et al. 2025)), the austrian GCA_026151145 (GenBank assembly accession: GCA_026151145.1) and the siberian NT1 (GenBank assembly accession:

GCA_963992905.1 (Scott et al. 2025)). The reference annotations from the reference genomes of *A. halleri* (AUBY1) and *A. lyrata* (MN47) were lifted off to the other accessions using the liftoff tool (v1.6.3) (Shumate and Salzberg 2021). Using the same parameters as previously, all potential ORF candidates were bioinformatically annotated with RibORF (Ji 2018), on the basis of their sequence only, in each accession. The coordinates of the syntenic region from the reference genome were transposed to each accession, and conserved syntenic regions in different accessions were aligned using MAFFT (Katoh 2002). The conservation of the ORF sequence was then assessed from these multiple alignments, using the same procedure as previously. To test whether translated *de novo* ORFs or untranslated intergenic ORFs were more conserved across accessions, we created a "conservation score" that is, per ORF, the number of accessions within which each ORF is found. A Mann-Whitney U test was performed across the 2 ORF categories for the observed conservation patterns; and "*de novo"* and "untranslated IGORF" labels were then randomly permuted 1,000 times, with new conservation scores and a new Mann-Whitney U test being computed for each iteration. The observed test score was then compared to the distribution of test scores under the null hypothesis (*de novo* and untranslated IGORFs have similar conservation pattern), and a corresponding p-value was assessed.

Polymorphism analysis

Genomic data from a total of 37 *A. halleri* individuals from central Europe (**Table S7**) (Suda et al. 2024; Pavan et al. 2025) and from 39 *A. lyrata* individuals from North America and central Europe (**Table S7**) (Mattila et al. 2017; Hämälä et al. 2018; Marburger et al. 2019; Takou et al. 2021; Pavan et al. 2025) were collected. Paired-end reads were mapped to the AUBY1 (*A. halleri*) or MN47 (*A. lyrata*) reference genome using bowtie 2 (v2.4.1) (Langmead and Salzberg 2012); and duplicates were marked and removed using Picard Tools (v2.21.4) (Anon 2019). Variant Calling and subsequent genotype joining were performed using GATK (v4.1.9.0) (Van der Auwera and O'Connor 2020). Variant calling files were converted to .fasta using custom bash scripts, with one file per chromosome. Using feature annotations, the corresponding sequences were then retrieved in each allele of each individual for each annotation. For intergenic *de novo* ORFs and untranslated IGORFs, in order to obtain enough polymorphic sites for pN/pS ratio computation, sequences were concatenated until a length equivalent to that of the shortest CDS was reached. The resulting sequences were aligned using MAFFT (Katoh

2002). The pN/pS ratios were computed from these alignments using dNdSpiNpiS_1.0 from the popPhyl project (Gayral et al. 2013) with parameters *-gapN_site=10*, *-gapN_seq=0.6*, *-remove_frameshift=yes* for CDS and *-remove_frameshift=no* for intergenic ORFs, and *-allow_internal_bc=yes*.

Plots were produced using R (v 4.2.2), with packages ggplot2 (Wickham 2009), ggsignif (Constantin and Patil 2021), ggpubr (Kassambara 2023), patchwork (Pedersen 2024), ComplexUpset (Michał Krassowski et al. 2022) and ComplexHeatmap (Gu et al. 2016), and alignment visualizations with Jalview (Waterhouse et al. 2009).

Acknowledgements

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). We thank the high performance computing service and Bilille at the University of Lille for providing computing resources. 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. Structure predictions for this publication were performed on the HPC cluster PALMA II of the University of Münster, subsidized by the Deutsche Forschungsgemeinschaft (INST 211/667-1).

Bibliography

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.

Ahrens JB, Nunez-Castilla J, Siltberg-Liberles J. 2017. Evolution of intrinsic disorder in eukaryotic proteins. *Cell. Mol. Life Sci.* 74:3163–3174.

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.

Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15:193–204.

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

Aubel M, Buchel F, Heames B, Jones A, Honc O, Bornberg-Bauer E, Hlouchova K. 2024. High-throughput Selection of Human *de novo* -emerged sORFs with High Folding Potential.Carreño CA, editor. *Genome Biol. Evol.* 16:evae069.

Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, Jentoft S. 2018. De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. *Mol. Biol. Evol.* 35:593–606.

Bartholomäus A, Del Campo C, Ignatova Z. 2016. Mapping the non-standardized biases of ribosome profiling. *Biol. Chem.* 397:23–35.

Basile W, Sachenkova O, Light S, Elofsson A. 2017. High GC content causes orphan proteins to be intrinsically disordered.Dunbrack RL, editor. *PLOS Comput. Biol.* 13:e1005375.

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.

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. *Genetics* 172:1675–1681.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. *Nucleic Acids Res.* 41:D36–D42.

Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nat. Commun.* 12:1–13.

Bornberg-Bauer E, Hlouchova K, Lange A. 2021. Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* 68:175–183.

Bornberg-Bauer E, Schmitz J, Heberlein M. 2015. Emergence of de novo proteins from “dark genomic matter” by “grow slow and moult.” *Biochem. Soc. Trans.* 43:867–873.

Brar GA, Weissman JS. 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* 16:651–664.

Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics* 179:487–496.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

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.

Carels N, Bernardi G. 2000. Two Classes of Genes in Plants. *Genetics* 154:1819–1825.

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, 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.

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.

Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. 2019. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* 29:932–943.

Erdős G, Dosztányi Z. 2024. AIUPred: combining energy estimation with deep learning for the enhanced prediction of protein disorder. *Nucleic Acids Res.* 52:W176–W181.

Frishman D, Argos P. 1995. Knowledge‐based protein secondary structure assignment. *Proteins Struct. Funct. Bioinforma.* 23:566–579.

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, Berk K, Dohmen E, Bornberg-Bauer E. 2022. New Genomic Signals Underlying the Emergence of Human Proto-Genes. *Genes* 13:284.

Grandchamp A, Czuppon P, Bornberg-Bauer E. 2024. Quantification and modeling of turnover dynamics of *de novo* transcripts in *Drosophila melanogaster*. *Nucleic Acids Res.* 52:274–287.

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

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.

Heames B, Schmitz J, Bornberg-Bauer E. 2020. A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila. *J. Mol. Evol.* 88:382–398.

Holehouse AS, Kragelund BB. 2023. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* [Internet]. Available from: https://www.nature.com/articles/s41580-023-00673-0

Hsu PY, Calviello L, Wu HYL, Li FW, Rothfels CJ, Ohler U, Benfey PN. 2016. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 113:E7126–E7135.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43:476–481.

Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* 8:1365–1379.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Resolution Using Ribosome Profiling. *Science* 324:218–223.

Ingolia NT, Hussmann JA, Weissman JS. 2019. Ribosome profiling: Global views of translation. *Cold Spring Harb. Perspect. Biol.* 11:1–20.

Iyengar BR, Bornberg-Bauer E. 2023. Neutral Models of *De Novo* Gene Emergence Suggest that Gene Evolution has a Preferred Trajectory.Harris K, editor. *Mol. Biol. Evol.* 40:msad079.

Iyengar BR, Grandchamp A, Bornberg-Bauer E. 2024. How antisense transcripts can evolve to encode novel proteins. *Nat. Commun.* 15:6187.

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.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* 40:1202–1210.

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

Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 25:1841–1842.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges.Prlic A, editor. *PLoS Comput. Biol.* 9:e1003118.

Lebherz MK, Fouks B, Schmidt J, Bornberg-Bauer E, Grandchamp A. 2024. DNA Transposons Favor De Novo Transcript Emergence Through Enrichment of Transcription Factor Binding Motifs.Luria V, editor. *Genome Biol. Evol.* 16:evae134.

Lei T, Chang Y, Yao C, Zhang H. 2022. A systematic evaluation revealed that detecting translated non-canonical ORFs from ribosome profiling data remains challenging.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* 103:9935–9939.

Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant J.* 58:485–498.

Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, Ge S, Guo YL. 2016. On the origin of de novo genes in arabidopsis thaliana populations. *Genome Biol. Evol.* 8:2190–2202.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* 4:865–875.

Madeira F, Madhusoodanan N, Lee J, Eusebi A, Niewielska A, Tivey ARN, Lopez R, Butcher S. 2024. The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res.* 52:W521–W525.

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, Eicholt LA. 2024. Random, de novo, and conserved proteins: How structure and disorder predictors perform differently. *Proteins Struct. Funct. Bioinforma.* 92:757–767.

Middendorf L, Iyengar BR, Eicholt LA. 2024. Sequence, Structure and Functional space of *Drosophila de novo* proteins. Bioinformatics Available from: http://biorxiv.org/lookup/doi/10.1101/2024.01.30.577933

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. 2014. Evolution: Dynamics of de novo gene emergence. *Curr. Biol.* 24:R238–R240.

Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife* 5:1–20.

Nielly-Thibault L, Landry CR. 2019. Differences between the raw material and the products of de novo gene birth can result from mutational biases. *Genetics* 212:1353–1366.

Ohno S. 1970. Evolution by Gene Duplication. Berlin, Heidelberg: Springer Berlin Heidelberg Available from: http://link.springer.com/10.1007/978-3-642-86659-3

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

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.

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 ? :734–763.

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

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

Roginski P, Grandchamp A, Quignot C, Lopes A. 2024. *De 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.

Schlesinger D, Elsässer SJ. 2022. Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. *FEBS J.* 289:53–74.

Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* 2:1626–1632.

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.

Singh R, Ming R, Yu Q. 2016. Comparative Analysis of GC Content Variations in Plant Genomes. *Trop. Plant Biol.* 9:136–149.

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* 45:831–835.

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.

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.

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.

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

Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet.* 15:1–23.

Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, Carvunis A-R. 2023. A vast evolutionarily transient 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.

Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer New York Available from: https://link.springer.com/10.1007/978-0-387-98141-3

Willis S, Masel J. 2018. Gene birth contributes to structural disorder encoded by overlapping genes. *Genetics* 210:303–313.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* 1.

Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* 3:1245–1252.

Wu HYL, Hsu PY. 2022. A custom library construction method for super-resolution ribosome profiling in Arabidopsis. *Plant Methods* 18:1–8.

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.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res.* 18:1446–1455.

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

Supplementary material

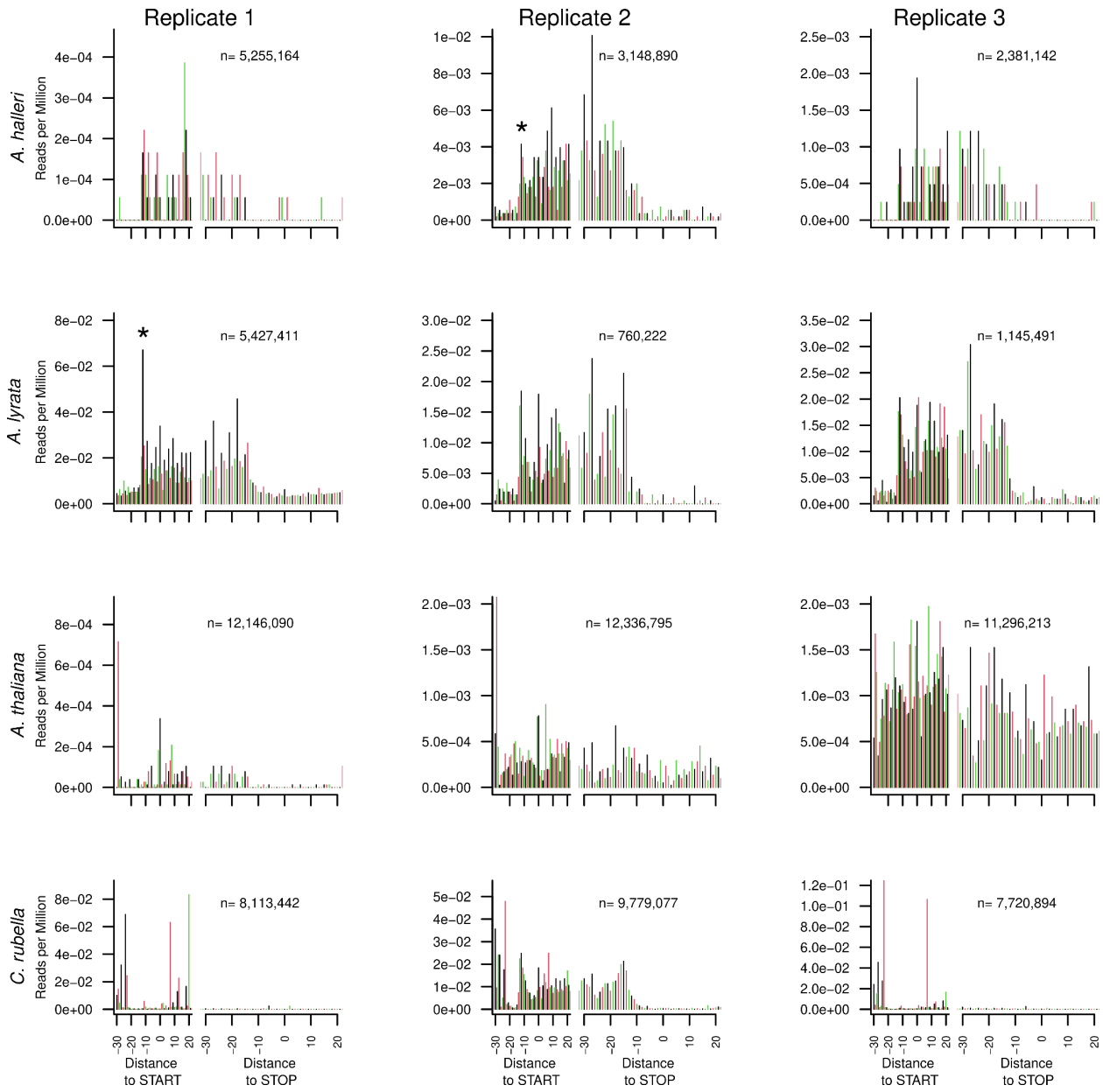


Figure S1: Metagene plots of 28-kmers from all ribosome profiling libraries produced in this study, before P-site shifting, where each line corresponds to a species, and each column to a replicate. Produced using data from all samples described in Table S1. Read accumulation is colored based on its frame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

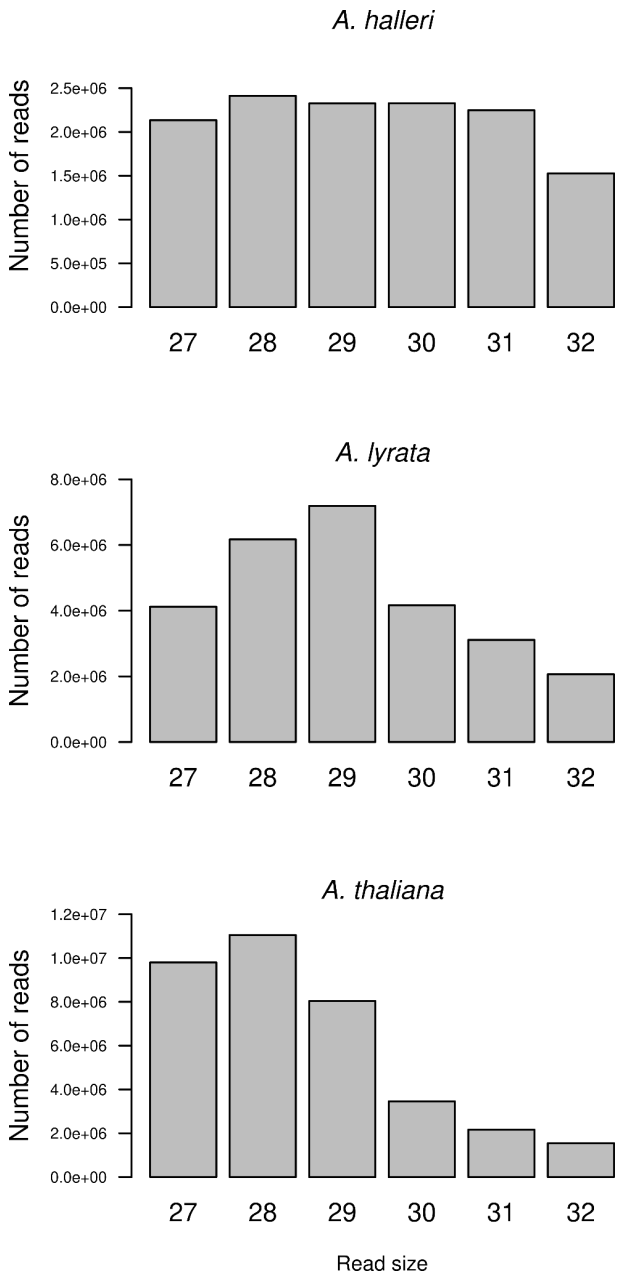


Figure S2: RPF k-mer distributions in the same libraries as those from Table S2.

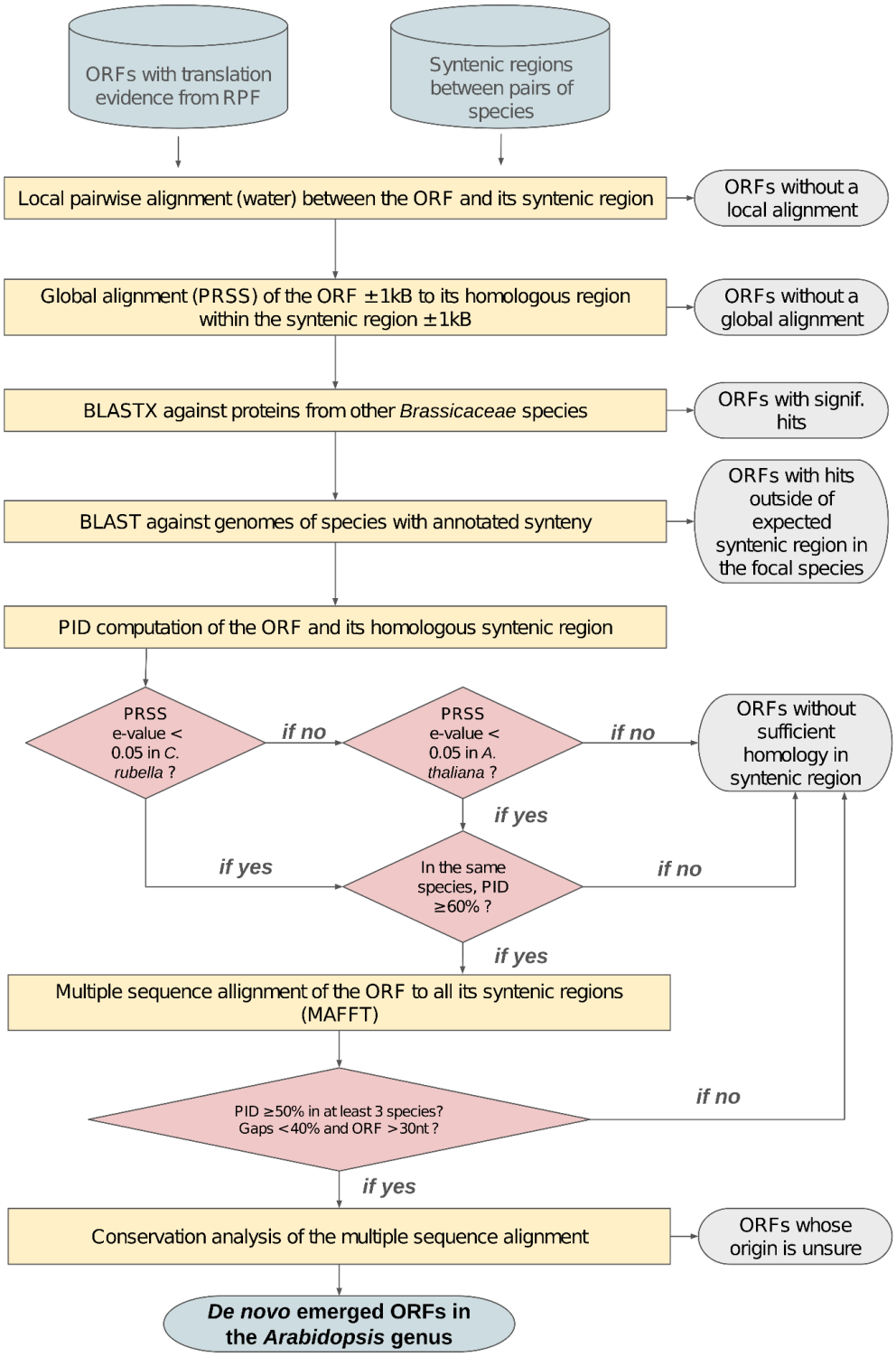


Figure S3: Flowchart of the bioinformatic pipeline used for the annotation of translated de novo ORFs. "ORFs with translation evidence from ribosome profiling" include both previously annotated CDS, and IGORFs.

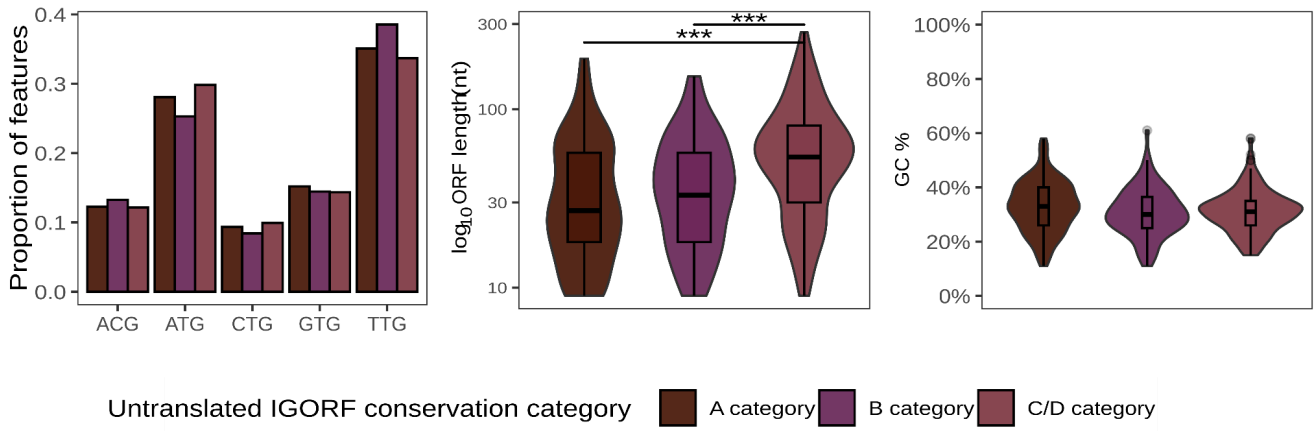


Figure S4: Sequences properties of untranslated IGORFs (START codon, length and GC content) depending on their conservation class. For length and GC content, all statistically significant differences were represented (Mann-Whitney U-test, * = p < 0.05, ** = p < 0.01, *** = p< 0.001).

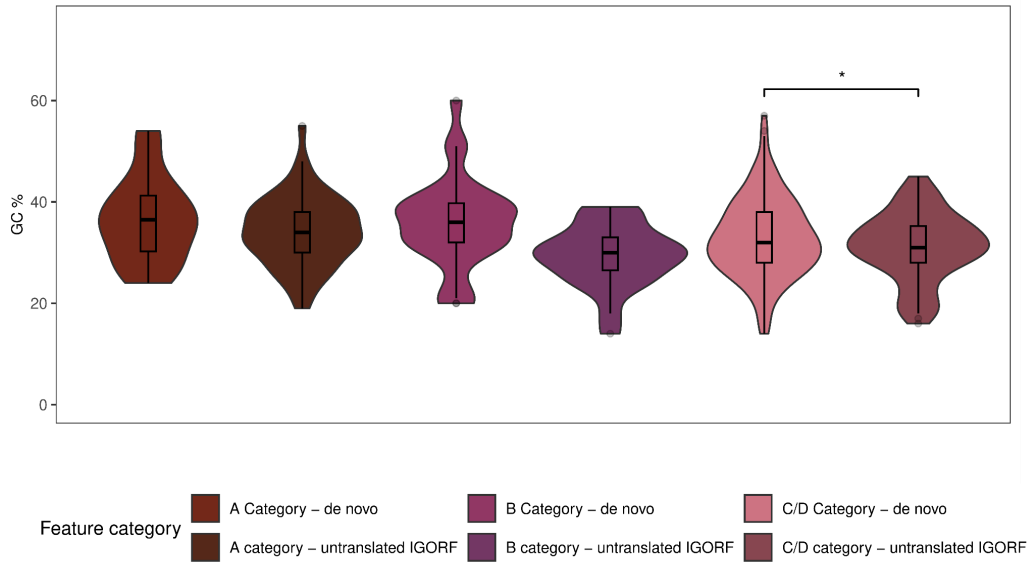


Figure S5: GC content of untranslated IGORs and *de novo* ORFs depending on their conservation class. All statistically significant differences were represented (Mann-Whitney U-test, * = p < 0.05, ** = p < 0.01, *** = p< 0.001).

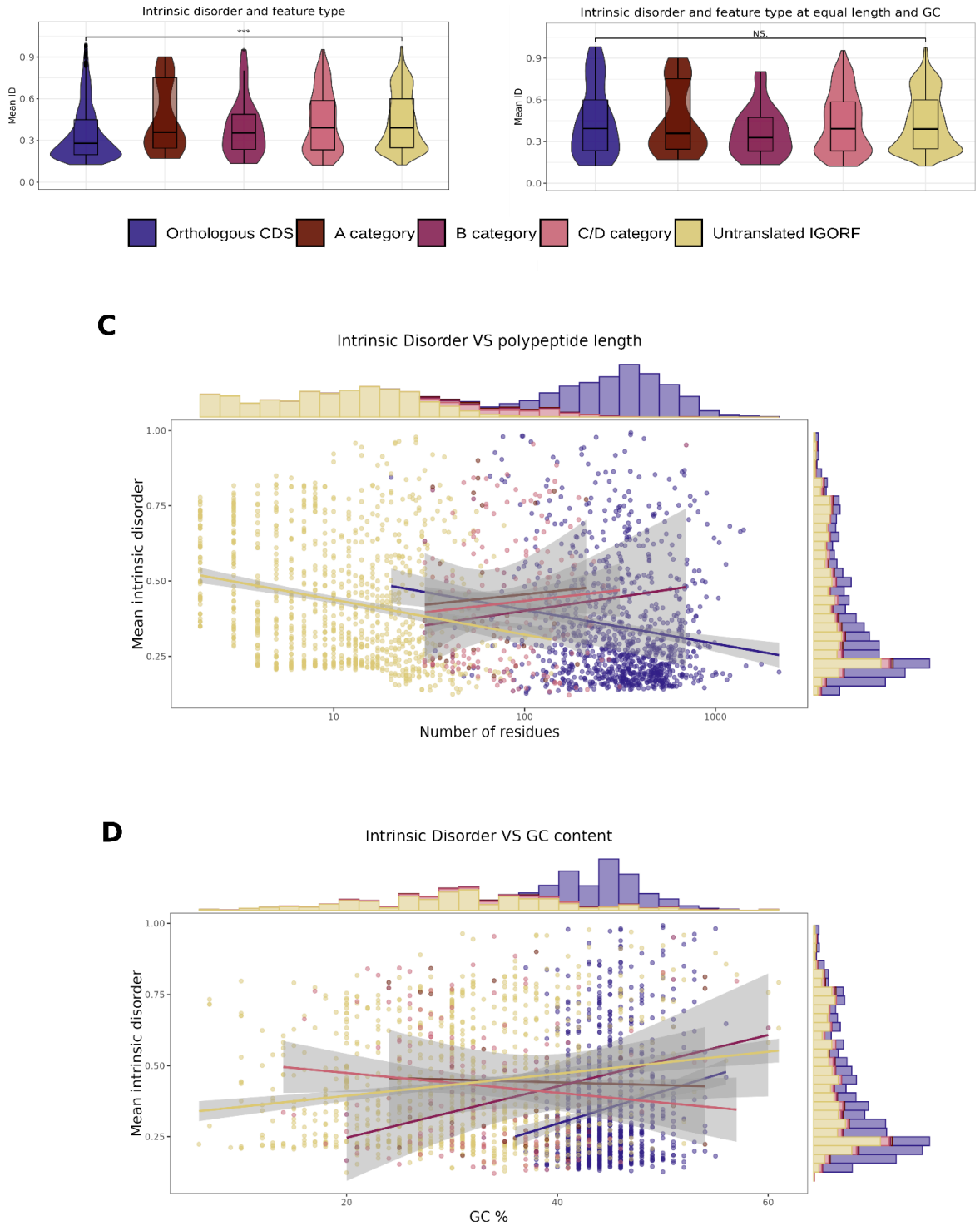


Figure S6: A) Measures of ISD in the translation products of conserved canonical CDS, different conservation classes of *de novo* ORFs and untranslated IGORFs. The only statistically significant difference (between orthologous CDS and untranslated IGORFs) is represented (Mann-Whitney U-test, *** = p< 0.001). B) Measures of ISD

in the translation products of conserved canonical CDS, different conservation classes of *de novo* ORFs and untranslated IGORFs; but where CDS were selected to match other categories in length and GC content. There is no longer any statistical difference. C) Linear regression of Mean Intrinsic Disorder in the translation products of the same previous categories, against the number of amino acid residues. Across categories, a significant negative association was found between the 2 variables (*adjusted* R^2 *= 0.03532, p-value < 2.2e-16*). D) Linear regression of Mean Intrinsic Disorder in the translation products of the same previous categories against the level of GC content in the DNA sequences encoding them. A significant positive association was found in orthologous CDS and IGORFs (in conserved CDS: *adjusted* R^2*=0.03359, p-value = 5.374e-09*).

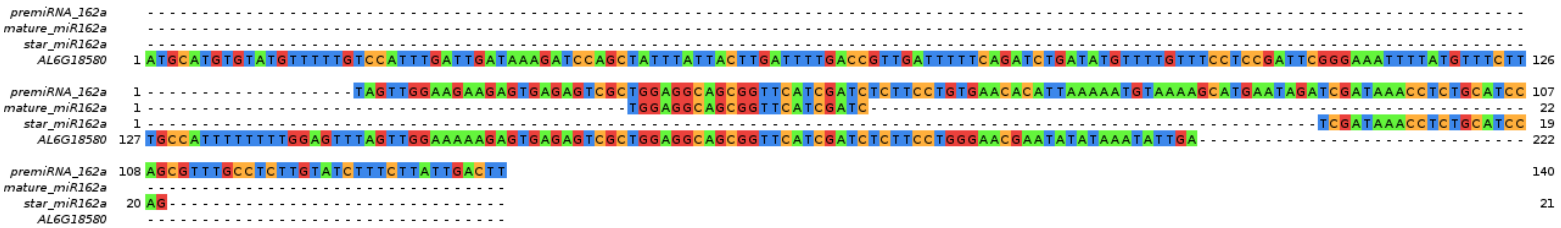


Figure S7: Alignment of the *A. lyrata* AL6G18580 gene to the *A. thaliana* miR162a precursor, mature miRNA and miRNA* (star miRNA).

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

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

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

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

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

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

	Species pair	**Translated ORFs found within syntenic regions**	**Without significant hits in protein coding genes of other *Brassicaceae* species**	**Without significant hits in genomic regions outside of synteny in focal species**	**With an extended alignment in *C. rubella* or *A. thaliana* of E-value < 0.05 and a PID > 60%**	**With a MSA of at least 50% PID in at least 3 sequences, and maximum of 40% gaps**	**With a *de novo* emergence**
A. halleri	*A. lyrata*	2,880	1,416	864	302	192	42
	A. thaliana	3,334					
	C. rubella	3,100					
A. lyrata	*A. halleri*	7,152	5,454	3,520	1,282	697	163
	A. thaliana	8,614					
	C. rubella	8,504					
A. thaliana	*A. halleri*	2,192	376	247	39	31	14
	A. lyrata	2,362					
	C. rubella	2,888					

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

Translated in:	Conservation category	Already annotated as CDS*	Newly identified as IGORF
A. halleri	Category A	6	2
	Category B	0	4
	Category D	3	27
	TOTAL:	**9**	**33**
A. lyrata	Category A	9	10
	Category B	6	20
	Category C	6	112
	TOTAL:	**21**	**142**
A. thaliana	Category A	4	3
	A. thaliana - specific	0	7
	TOTAL:	**4**	**10**
	ALL SPECIES TOTAL:	**34**	**185**
	SUM:	**219**	
	(without redundance betweens species)	**211**	

* in at least one species for category A and B

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

ORF name	Hits description	Query Cover	E value	Per. ident
GENE_732871	A. lyrata uncharacterized LOC110227695 mRNA, A. thaliana uncharacterized protein AT5G12043, A. thaliana genome	99%, 100%, 100%	3E-33, 2E-25, 2E-25	87.94, 84.51, 84.51
GENE_826985	A. lyrata probable apyrase 7 LOC9304022 mRNA, A. thaliana genome, A. thaliana GDA1/CD39 nucleoside phosphatase family protein AT4G19180 mRNA	100%, 62%, 62%	7E-40, 6.00E-31, 3E-29	87.8, 95.6, 94.51
GENE_829626	NO_HITS	NO_HITS	NO_HITS	NO_HITS
GENE_AL1G18050	NO_HITS	NO_HITS	NO_HITS	NO_HITS
GENE_AL3G13390	NO_HITS	NO_HITS	NO_HITS	NO_HITS
GENE_AL3G23310	A. lyrata uncharacterized LOC9318812 mRNA, A. thaliana genome A. thaliana uncharacterized protein AT3G11745 mRNA	100%, 100%, 100%	0.0000006, 4E-26, 4E-26	100, 88.68, 88.68
GENE_AL3G25260	A. thaliana genome, A. thaliana uncharacterized misc_RNA (AT3G00320)	73%, 73%	0.00000007, 0.00000007	100, 100
GENE_AL4G44240	A. lyrata uncharacterized LOC9318064 mRNA	100%	2.00E-49	100
GENE_AL4G45690	A. thaliana genome	52%	4.00E-17	92.86
GENE_AL5G37070	A. lyrata glycine-rich protein 3 short isoform LOC9312403 mRNA	100%	5.00E-92	100
GENE_AL6G11140	A. thaliana genome, A. thaliana uncharacterized misc_RNA AT3G00320	86%, 86%	5E-30, 5E-30	92.86, 92.86
GENE_AL6G18580	A. lyrata uncharacterized LOC110227896 ncRNA, A. lyrata microRNA aly-MIR162a precursor, A. thaliana miR162a primary transcript gene, Arabidopsis thaliana hypothetical protein MIR162a miRNA	100%, 54%, 78%, 27%	3E-110, 1E-53, 2E-42, 7E-17	100, 86.26, 86.26, 96.67
GENE_AL8G35520	A. lyrata extensin-3 LOC9302389 mRNA, A. thaliana Proline-rich extensin-like family protein AT5G59170 mRNA, A. thaliana genome	97%, 88%, 88%	0, 8E-86, 6E-14	100, 87.76, 96.23
IGORF_1	A. thaliana genome	100%	2.00E-60	100
IGORF_2	A. thaliana genome	100%	3.00E-95	100
IGORF_3	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_4	A. thaliana genome	100%	6.00E-26	100
IGORF_5	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_6	A. thaliana genome, B. rapa genome	100%, 42%	6E-44, 3.00E-07	100, 95.35
IGORF_7	A. thaliana genome	100%	2.00E-32	100
IGORF_8	A. thaliana genome	100%	1.00E-103	100
IGORF_9	A. thaliana genome	100%	2.00E-32	100
IGORF_10	A. thaliana genome	100%	2.00E-19	100
IGORF_12	A. thaliana genome, A. lyrata uncharacterized LOC9310279 misc_RNA	100%, 100%	2E-42, 3.00E-17	100, 86
IGORF_14	A. thaliana genome	100%	3.00E-16	100
IGORF_15	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_16	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_17	A. thaliana genome	98%	7.00E-56	89.84
IGORF_18	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_19	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_20	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_22	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_23	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_26	A. thaliana genome	100%	2.00E-14	96.3
IGORF_27	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_28	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_29	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_30	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_32	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_33	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_34	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_35	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_36	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_37	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_38	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_40	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_42	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_43	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_44	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_45	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_46	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_47	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_48	Sisymbrium irio genome	30%	0.0002	100

IGORF_49	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_50	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_52	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_53	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_55	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_56	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_57	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_58	Brassica oleracea genome	30%	3.00E-08	97.56
IGORF_61	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_63	A. thaliana genome	100%	3.00E-25	95.06
IGORF_65	A. lyrata uncharacterized LOC9325856 mRNA	100%	2.00E-08	100
IGORF_66	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_67	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_68	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_69	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_70	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_71	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_72	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_73	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_74	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_75	A. thaliana genome	86%	3.00E-18	84.96
IGORF_76	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_77	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_78	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_79	A. thaliana genome	65%	1.00E-14	80
IGORF_80	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_81	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_82	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_83	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_84	Brassica oleracea genome	88%	1.00E-08	100
IGORF_85	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_86	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_87	A. thaliana genome	100%	5.00E-34	96.81
IGORF_88	A. thaliana genome, A. thaliana other RNA AT1G51645 lncRNA	96%, 96%	1E-48, 1E-48	87.36 87.36
IGORF_89	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_90	A. thaliana genome	72%	1.00E-27	91.84
IGORF_92	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_93	A. thaliana genome	51%	1.00E-39	95.5
IGORF_94	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_96	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_97	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_98	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_100	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_102	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_103	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_104	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_105	Raphanus sativus U5 spliceosomal RNA LOC130497136 ncRNA	77%	0.0007	88.64
IGORF_106	A. thaliana genome	84%	3.00E-44	91.61
IGORF_107	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_108	A. thaliana genome	96%	8.00E-55	93.46
IGORF_109	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_110	A. thaliana genome	65%	2.00E-25	92.31
IGORF_111	A. thaliana genome	97%	9.00E-66	88.65
IGORF_112	A. thaliana genome	74%	2.00E-37	87.18
IGORF_113	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_114	A lyrata uncharacterized LOC110230800 ncRNA	100%	2.00E-11	100
IGORF_115	A. thaliana genome	100%	4.00E-12	90.62
IGORF_116	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_117	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_118	A. lyrata probable inactive leucine-rich repeat receptor-like protein kinase At3g03770 mRNA	53%	3.00E-23	100
IGORF_120	A. thaliana genome	83%	4.00E-11	92.86
IGORF_122	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_123	NO_HITS	NO_HITS	NO_HITS	NO_HITS

IGORF_124	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_125	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_126	A. thaliana miR172c primary transcript gene, Arabidopsis lyrata microRNA aly-MIR172c precursor, Brassica rapa genome	70%, 47%, 44%	2E-34, 4.00E-22, 4.00E-17	96.84, 100, 96.67
IGORF_129	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_130	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_131	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_132	A. thaliana genome	80%	6.00E-52	92.16
IGORF_133	A. thaliana genome	98%	7.00E-38	97.96
IGORF_134	A. thaliana genome, Brassica oleracea HDEM genome	70%, 58%	4E-26, 2.00E-13	97.37, 92.06
IGORF_138	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_140	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_141	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_143	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_144	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_145	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_148	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_149	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_150	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_151	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_152	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_155	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_156	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_157	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_158	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_159	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_161	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_162	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_163	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_164	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_165	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_167	A. thaliana genome, Arabidopsis thaliana FASCICLIN-like arabinogalactan protein 8 FLA8 mRNA, A. lyrata fasciclin-like arabinogalactan protein 8 LOC9318081 mRNA	99%, 71%, 66%	2E-33, 8.00E-22, 5.00E-19	98.84, 100, 100
IGORF_168	A. thaliana genome	90%	1.00E-30	91.74
IGORF_169	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_170	Camelina sativa uncharacterized LOC109132690 ncRNA	94%	0.0003	96.88
IGORF_171	A. thaliana genome	78%	5.00E-31	93.88
IGORF_172	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_176	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_178	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_177	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_179	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_180	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_181	A. thaliana genome	39%	4.00E-11	100
IGORF_183	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_185	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_186	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_189	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_190	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_191	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_192	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_193	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_194	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_195	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_196	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_198	A. lyrata nascent polypeptide-associated complex subunit alpha-like protein 3 LOC9309699 mRNA	73%	2.00E-66	99.32
IGORF_199	A. lyrata peroxidase 56 LOC9309780 mRNA	58%	1.00E-14	100
IGORF_200	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_201	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_202	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_203	NO_HITS	NO_HITS	NO_HITS	NO_HITS

IGORF_204	A. thaliana genome, A. thaliana uncharacterized misc_RNA AT5G00435 miscRNA	52%, 52%	0.000000002, 0.000000002	91.07, 91.07
IGORF_205	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_206	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_207	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_208	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_210	A. thaliana genome	100%	9.00E-33	91.23
IGORF_211	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_212	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_214	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_216	A. thaliana genome, Arabidopsis thaliana ascorbate peroxidase 3 APX3 mRNA	92%, 75%	2E-15, 2.00E-10	95, 95.83
IGORF_218	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_219	A. thaliana genome	65%	3.00E-27	97.44
IGORF_222	A. lyrata uncharacterized LOC9305728 mRNA	100%	7.00E-10	100
IGORF_223	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_225	A. thaliana genome	100%	4.00E-48	88.57
IGORF_226	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_227	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_228	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_229	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_230	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_231	Arabidopsis thaliana Ribosomal protein L35 AT5G45590 mRNA	98%	6.00E-10	78.63
IGORF_232	A. lyrata ankyrin repeat-containing protein ITN1 LOC110225629 mRNA	100%	2.00E-27	100
IGORF_233	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_235	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_236	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_238	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_241	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_242	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_244	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_246	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_247	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_248	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_252	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_251	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_253	NO_HITS	NO_HITS	NO_HITS	NO_HITS
IGORF_254	A. lyrata uncharacterized LOC9301140 mRNA, A. thaliana genome	63%, 32%	6E-46, 1.00E-12	100, 96.15

Table S6: Supplementary Megablast results for each translated *de novo* ORF (core nucleotide database nt, update date: 2025/07/31), with default search parameters

Species	Population	GPS coordinates	Sample accession	Reference
A. halleri	Harz Moutains (Germany)	51.6655555555556, 11.1584444444444	SAMEA115474513	*Suda et al., 2024*
A. halleri	Western Carpathians (Slovakia)	48.8563611111111, 20.9202777777778	SAMEA115474514	*Suda et al., 2024*
A. halleri	Western Carpathians (Slovakia)	48.7658333333333, 21.12975	SAMEA115474515	*Suda et al., 2024*
A. halleri	Harz Moutains (Germany)	51.8625833333333, 10.3001111111111	SAMEA115474516	*Suda et al., 2024*
A. halleri	Siegerland (Germany)	51.0054, 8.006606	SAMEA115474517	*Suda et al., 2024*
A. halleri	Kosice-okolie District (Slovakia)	48.7522777777778, 20.8963888888889	SAMEA115474518	*Suda et al., 2024*
A. halleri	Miasteczko Śląskie (Poland)	50.5035277777778, 18.93575	SAMEA115474519	*Suda et al., 2024*
A. halleri	Miasteczko Śląskie (Poland)	50.5035277777778, 18.93575	SAMEA115474520	*Suda et al., 2024*
A. halleri	Bayern (Germany)	50.41137, 11.55532	SAMEA115474521	*Suda et al., 2024*
A. halleri	Sauerland (Germany)	51.3063833333333, 8.4852	SAMEA115474522	*Suda et al., 2024*
A. halleri	Tatra Mountains (Poland)	49.27925, 19.9660833333333	SAMEA115474523	*Suda et al., 2024*
A. halleri	Tatra Mountains (Poland)	49.27925, 19.9660833333333	SAMEA115474524	*Suda et al., 2024*
A. halleri	Maramures County (Romania)	47.8084722222222, 23.6171388888889	SAMEA115474525	*Suda et al., 2024*
A. halleri	Suceava County (Romania)	47.5276944444444, 25.4276111111111	SAMEA115474526	*Suda et al., 2024*
A. halleri	Maramures County (Romania)	47.6947222222222, 23.7769722222222	SAMEA115474527	*Suda et al., 2024*
A. halleri	Cluj County (Romania)	46.8631666666667, 22.8076944444444	SAMEA115474528	*Suda et al., 2024*
A. halleri	Lana di Dentro (Italy)	46.49308, 10.88953	SAMN43174266	*Pavan et al. 2025*
A. halleri	Ponte Nossa (Italy)	45.85931, 9.87669	SAMN42903595	*Pavan et al. 2025*
A. halleri	Paisco Loveno (Italy)	46.05556, 10.24297	SAMN42903597	*Pavan et al. 2025*
A. halleri	Mortagne-du-Nord (France)	50.49377, 3.45634	SAMN43174278	*Pavan et al. 2025*
A. halleri	Wallenfels (Germany)	50.41137, 11.55532	SAMN42903598	*Pavan et al. 2025*

A. halleri	Wallenfels (Germany)	50.41137, 11.55532	SAMN42903599	*Pavan et al. 2025*
A. halleri	Falkenstein (Germany)	51.66556, 11.15844	SAMN43174267	*Pavan et al. 2025*
A. halleri	Langelsheim (Germany)	51.94870, 10.34884	SAMN42903593	*Pavan et al. 2025*
A. halleri	Croglio (Switzerland)	45.98750, 8.83951	SAMN43174284	*Pavan et al. 2025*
A. halleri	Zocca-Vicosoprano (Switzerland)	46.37158, 9.66014	SAMN43174279	*Pavan et al. 2025*
A. halleri	Kowari (Poland)	50.75864, 15.84672	SAMN42903592	*Pavan et al. 2025*
A. halleri	Zakopane (Poland)	49.27925, 19.96608	SAMN42914077	*Pavan et al. 2025*
A. halleri	Miasteczko Śląskie (Poland)	50.50353, 18.93575	SAMN42903594	*Pavan et al. 2025*
A. halleri	Blidari (Romania)	47.80847, 23.61714	SAMN43174271	*Pavan et al. 2025*
A. halleri	Munții Gutâi (Romania)	47.69472, 23.77697	SAMN43174272	*Pavan et al. 2025*
A. halleri	Tranisu (Romania)	46.86317, 22.80769	SAMN43174277	*Pavan et al. 2025*
A. halleri	Fundu Moldovei (Romania)	47.52769, 25.42761	SAMN43174273	*Pavan et al. 2025*
A. halleri	Galmeikogel - Annarotte (Austria)	47.84847, 15.39017	SAMN43174275	*Pavan et al. 2025*
A. halleri	Bad Gastein (Austria)	47.13403, 13.12675	SAMN43174276	*Pavan et al. 2025*
A. halleri	Castle Rabenstein (Austria)	47.24958, 15.30894	SAMN43174268	*Pavan et al. 2025*
A. halleri	Steinbachrotte (Austria)	47.93833, 15.43397	SAMN43174287	*Pavan et al. 2025*
A. lyrata	Pinery Provincial Park (North America)	43.26404, -81.83971	SAMN43781242	*Pavan et al. 2025*
A. lyrata	Pinery Provincial Park (North America)	43.26404, -81.83971	SAMN43781218	*Pavan et al. 2025*
A. lyrata	Pinery Provincial Park (North America)	43.26404, -81.83971	SAMN43781219	*Pavan et al. 2025*
A. lyrata	Rondeau Provincial Park (North America)	42.26138, -81.84638	SAMN43781240	*Pavan et al. 2025*
A. lyrata	Rondeau Provincial Park (North America)	42.26138, -81.84638	SAMN43781237	*Pavan et al. 2025*
A. lyrata	Rondeau Provincial Park (North America)	42.26138, -81.84638	SAMN43781234	*Pavan et al. 2025*

A. lyrata	Tobermory georgian bay (North America)	45.24166, -81.5175	SAMN43781213	*Pavan et al. 2025*
A. lyrata	Tobermory georgian bay (North America)	45.24166, -81.5175	SAMN43781216	*Pavan et al. 2025*
A. lyrata	Tobermory georgian bay (North America)	45.24166, -81.5175	SAMN43781214	*Pavan et al. 2025*
A. lyrata	Tobermory singing sands (North America)	45.1925, -81.58388	SAMN43781247	*Pavan et al. 2025*
A. lyrata	Tobermory singing sands (North America)	45.1925, -81.58388	SAMN43781227	*Pavan et al. 2025*
A. lyrata	Tobermory singing sands (North America)	45.1925, -81.58388	SAMN43781228	*Pavan et al. 2025*
A. lyrata	Indiana Dune National Lakeshore (North America)	41.66897, -87.04221	SAMN43781217	*Pavan et al. 2025*
A. lyrata	Indiana Dune National Lakeshore (North America)	41.66897, -87.04221	SAMN43781221	*Pavan et al. 2025*
A. lyrata	Indiana Dune National Lakeshore (North America)	41.66897, -87.04221	SAMN43781222	*Pavan et al. 2025*
A. lyrata	Storr (UK)	57.50, -6.16	SAMN06141186	*Mattila et al. 2017*
A. lyrata	Storr (UK)	57.50, -6.16	SAMN06141185	*Mattila et al. 2017*
A. lyrata	Stubbsand (Sweden)	63.22, 18.95	SAMN06141184	*Mattila et al. 2017*
A. lyrata	Stubbsand (Sweden)	63.22, 18.95	SAMN06141183	*Mattila et al. 2017*
A. lyrata	Stubbsand (Sweden)	63.22, 18.95	SAMN06141182	*Mattila et al. 2017*
A. lyrata	Stubbsand (Sweden)	63.22, 18.95	SAMN06141181	*Mattila et al. 2017*
A. lyrata	Plech (Germany)	49.63, 11.51	SAMN06141187	*Mattila et al. 2017*
A. lyrata	Plech (Germany)	49.63, 11.51	SAMN06141188	*Mattila et al. 2017*
A. lyrata	Plech (Germany)	49.63, 11.51	SAMN06141189	*Mattila et al. 2017*
A. lyrata	Mayodan (USA)	36.42, -79.97	SAMN06141193	*Mattila et al. 2017*
A. lyrata	Mayodan (USA)	36.42, -79.97	SAMN06141194	*Mattila et al. 2017*
A. lyrata	Mayodan (USA)	36.42, -79.97	SAMN06141195	*Mattila et al. 2017*
A. lyrata	Spiterstulen (Norway)	61.63, 8.4	SAMN06141173	*Mattila et al. 2017*

A. lyrata	Spiterstulen (Norway)	61.63, 8.4	SAMN06141174	*Mattila et al. 2017*
A. lyrata	Spiterstulen (Norway)	61.63, 8.4	SAMN06141175	*Mattila et al. 2017*
A. lyrata	Austria	47.90154, 15.96864	SAMEA5947868	*Marburger et al. 2019*
A. lyrata	Austria	47.90154, 15.96864	SAMEA5947869	*Marburger et al. 2019*
A. lyrata	Austria	47.92251, 15.98176	SAMEA5947870	*Marburger et al. 2019*
A. lyrata	Austria	47.92251, 15.98176	SAMEA5947871	*Marburger et al. 2019*
A. lyrata	Plech (Germany)	49.65, 11.45	SAMEA6943357	*Takou et al. 2021*
A. lyrata	Plech (Germany)	49.65, 11.45	SAMEA6943358	*Takou et al. 2021*
A. lyrata	Plech (Germany)	49.65, 11.45	SAMEA6943359	*Takou et al. 2021*
A. lyrata	Spiterstulen (Norway)	61.62, 8.40	SAMN09069416	*Hämälä et al. 2018*
A. lyrata	Spiterstulen (Norway)	61.62, 8.40	SAMN09069417	*Hämälä et al. 2018*
A. lyrata	Spiterstulen (Norway)	61.62, 8.40	SAMN09069418	*Hämälä et al. 2018*

Table S7: Summary of the population whole-genome sequencing data