The evolutionary history and functional specialization of microRNA genes in *Arabidopsis halleri* and *A. lyrata*

Flavia Pavan[1], Jacinthe Azevedo Favory[2], Eléanore Lacoste[3], Chloé Beaumont[1], Firas Louis[1], Christelle Blassiau[1], Corinne Cruaud[4], Karine Labadie[4], Sophie Gallina[1], Mathieu Genete[1], Vinod Kumar[5], Ute Kramer[5], Rita A. Batista[1], Claire Patiou[1], Laurence Debacker[1], Chloé Ponitzki[1], Esther Houzé[1], Eléonore Durand[1], Jean-Marc Aury[3], Vincent Castric[1], Sylvain Legrand[1].

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

[2] Laboratoire Génome et Développement des Plantes, UMR5096 CNRS/UPVD, Perpignan, France

[3] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

[4] Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

[5] Faculty of Biology and Biotechnology, Ruhr University Bochum, D-44801 Bochum, Germany

Author for correspondence : vincent.castric@univ-lille.fr

Abstract

MicroRNAs (miRNAs) are a class of small non-coding RNAs that play important regulatory roles in plant genomes. While some miRNA genes are deeply conserved, the majority appear to be species-specific, raising the question of how they emerge and integrate into cellular regulatory networks. To better understand this, we first performed a detailed annotation of miRNA genes in the closely related plants *Arabidopsis halleri* and *A. lyrata* and evaluated their phylogenetic conservation across 87 plant species. We then characterized the process by which newly emerged miRNA genes progressively acquire the properties of "canonical" miRNA genes, in terms of size and stability of the hairpin precursor, loading of their cleavage products into Argonaute proteins, and potential to regulate downstream target genes. Nucleotide polymorphism was lower in the mature miRNA sequence than in the other parts of the hairpin (stem, terminal loop), and the regions of coding sequences targeted by miRNAs also had reduced diversity as compared to their neighboring regions along the genes. These patterns were less pronounced for recently emerged than for evolutionarily conserved miRNA genes, suggesting a weaker selective constraint on the most recent miRNA genes. Our results illustrate the rapid birth-and-death of miRNA genes in plant genomes, and provide a detailed picture of the evolutionary processes by which a small fraction of them eventually integrate into "core" biological processes.

Key words: Arabidopsis, evolution, microRNA, polymorphism, species-specific genes

43 **1. Introduction**

44 The origins of evolutionary novelties have been a topic of considerable
45 interest in biology (Wagner, 2011). Following Francois Jacob's (1977) seminal
46 concept of molecular tinkering (Jacob, 1977), the emergence of novel biological
47 functions "from scratch" has long been considered an unlikely evolutionary event.
48 Instead, evolution was believed to proceed through the modification of existing
49 structures (such as protein-coding genes) following various forms of rearrangements
50 such as small or large-scale duplications, fusions or fissions. Recently, however, it
51 has become apparent that new genes do actually arise relatively readily as a result
52 of a variety of processes including pervasive transcription throughout the genome,
53 and the field has moved from "whether" new genes can arise to "how" they arise
54 (Van Oss and Carvunis, 2019). A particularly debated question is whether the newly
55 emerged "proto-genes" become gradually optimized by natural selection and
56 eventually acquire the "canonical" gene-like characteristics (as per the "continuum
57 model", Carvunis et al., 2012), or whether they result from the immediate apparition
58 of rare "hopeful monsters", *i.e.* DNA sequences that are already pre-adapted and
59 immediately exhibit gene-like characteristics with essentially no further optimization
60 (as per the "preadaptation model", Wilson et al., 2017; McLysaght and Guerzoni,
61 2015). This process has been mostly studied in the particular case of protein-coding
62 genes, and requires the broad-scale comparison of genes of different evolutionary
63 ages that were formed at different times in the past along a phylogeny. However, not
64 all genes are coding for proteins, and the study of the other sorts of genetic elements
65 populating the genome is necessary for a comprehensive understanding of
66 phenotypic evolution.

67 Regulatory RNAs are an important class of regulators of gene expression,
68 and among them microRNAs (miRNAs) are key post-transcriptional regulators of
69 gene expression in plants, animals, fungi and some viruses (Dexheimer and
70 Cochella, 2020; Nanbo et al., 2021). miRNAs are expressed from genes that do not
71 encode for proteins but are transcribed by RNA polymerase II into primary miRNAs
72 (pri-miRNAs). These pri-miRNAs adopt a hairpin-like structure recognized by
73 DICER-LIKE (DCL) proteins in plants. DCL proteins cleave the pri-miRNA once to
74 generate the pre-miRNA, and a second time to further release the miRNA/miRNA*
75 duplex. The mature miRNA, typically a single 21 nucleotides-long RNA, is loaded
76 into ARGONAUTE1 (AGO1) proteins, forming the RNA-induced silencing complex

77 (RISC) in association with other proteins. The RISC recognizes messenger RNA
78 (mRNA) targets through near-complete sequence complementarity with the mature
79 miRNA, leading to negative regulation through mRNA degradation or translation
80 inhibition (Reviewed in Zhan and Meyers, 2023; Ding and Zhang, 2023).

81 The recent availability of genome assemblies together with massive small
82 RNA sequencing data has enabled broad-scale comparisons of the repertoire of
83 miRNA genes across a growing number of plant and animal species, albeit with a
84 strong bias toward model organisms. These comparisons revealed striking
85 quantitative variation, with the total number of annotated miRNA genes ranging from
86 just a few dozens to hundreds of miRNA genes per genome (miRBase v22,
87 Kozomara et al., 2019). However, properly interpreting these variations has
88 remained challenging because annotation of miRNA genes in plant and animal
89 genomes is notoriously difficult due to their small size and the abundance of short
90 inverted repeats, the heterogeneity of annotation methods, of the quality of genome
91 assemblies, of molecular methodologies employed for small RNA sequencing, and of
92 tissue types being compared. In spite of these caveats, the data available clearly
93 indicate that, while some miRNAs are deeply conserved, lineage-specific miRNAs
94 are also common, even between closely related species, suggesting a rapid
95 evolutionary dynamics (Fahlgren et al., 2007; Cuperus et al., 2011; Nozawa et al.,
96 2012; Chávez Montes et al., 2014).

97 The current model posits that “proto-miRNA” genes originate from a variety of
98 sources, including the inverted duplication of protein-coding genes, transposable
99 element-related sequences, or regions of the genome that happened to contain
100 inverted repeats and acquired the ability to be transcribed (reviewed in Cui et al.,
101 2017). However, the abundance of proto-miRNAs relative to canonical miRNAs and
102 the processes by which proto-miRNAs transition into canonical miRNAs have rarely
103 been characterized in detail. Previous studies suggested that the process initially
104 starts from stem-loops exhibiting near perfect complementarity that are the preferred
105 substrate for DCL2, DCL3 or DCL4 proteins, imprecisely generating multiple
106 duplexes of 24-nt-long small interfering RNAs (siRNAs) that are then loaded into
107 AGO4 proteins. As the hairpin structure accumulates mutations over evolutionary
108 time its complementarity is progressively disrupted, facilitating recognition by DCL1

and leading to the precise production of a single 21-nt-long mature miRNA preferentially loaded in AGO1, hence acquiring features of “canonical” miRNA genes (Allen et al., 2004; Voinnet et al., 2009; Baldrich et al., 2018; Pegler et al., 2023). Recent miRNA genes were also suggested to have weaker and more limited spatio-temporal expression territories than the more conserved miRNA genes, and that they also tend to be processed less precisely by DCL proteins leading to the production of a more diverse population of mature miRNAs (Fahlgren et al., 2007, 2010; Ma et al., 2010; Chávez Montes et al., 2014). Young miRNA genes tend to target genes associated with adaptation to local environments (Wen et al., 2016; Bradley et al., 2017), while highly conserved miRNA genes more often target genes involved in crucial cellular processes in plant development and stress responses (Dong et al., 2022). Two models have been proposed for the evolution of miRNA-target interactions. In the "decay model" (Chen and Rajewsky, 2007; Roux et al., 2012), new miRNAs initially have many targets, most of which are deleterious, while only a few are beneficial. Over time, deleterious interactions are removed by natural selection and advantageous interactions are retained. In the "growth model" in contrast, the number of miRNA targets instead increases over the course of evolution (Nozawa et al., 2016). In this model, new miRNAs initially possess few targets, most of which are neutral with only a few being beneficial. This allows the level of expression of the miRNA gene to eventually increase and gradually acquire new targets over the course of evolution. While the target acquisition model has received some support in humans and mice (Chen and Rajewsky, 2007; Roux et al., 2012), the "growth model" has been favored in Drosophila (Nozawa et al., 2016). Hence, the relevance of these models, and the overall evolutionary significance of the newly acquired miRNA genes and their potential regulatory across genomes, has not been widely tested . Finally, young miRNA genes also seem to diverge more rapidly between related species, suggesting weaker functional constraints than that applying to older miRNA genes (Fahlgren et al., 2010; Ma et al., 2010). In *Arabidopsis thaliana,* the binding site within the genes targeted by miRNAs exhibits low polymorphism, indicating strong purifying selection (Ehrenreich and Purugganan, 2008; Smith et al., 2015). However, little is known about the microevolution of the binding site of the genes targeted by recently evolved miRNA genes.

141 In the genus Arabidopsis, a total of 221 miRNA genes have been annotated in
142 the plant model *A. thaliana* (PmiREN 2.0, Guo et al., 2022b), and the companion
143 papers by Ma et al., (2010) and Fahlgren et al., (2010) identified 154 and 164 miRNA
144 genes, respectively in *A. lyrata*, from which *A. thaliana* diverged about 5 Myrs ago
145 (Koch et al., 2000; Ossowski et al., 2010). These comparisons revealed a series of
146 miRNA genes specific to either species, but given the rapid evolutionary dynamics of
147 miRNA genes such broad-scale phylogenetic comparisons are inherently limited,
148 and the comparison of even more closely related species is needed, as they
149 represent a powerful way to reveal the most recently formed miRNA genes. *A. halleri*
150 diverged from *A. lyrata* only one million years ago (Roux et al., 2011) and is a
151 promising model, but the genome assemblies published for this species are highly
152 fragmented (Briskine et al., 2017; Legrand et al., 2019), and the repertoire of
153 annotated miRNAs is very incomplete (only 18 miRNAs have been deposited in the
154 PmiREN 2.0 database, Guo et al., 2022b).

155 In this study, we explored the recent evolutionary dynamics of miRNA genes
156 by focusing on *A. halleri* and *A. lyrata*. We first obtained a high-quality chromosome-
157 level reference genome assembly for *A. halleri* and used sRNA-seq data from a
158 variety of accessions to provide the first comprehensive annotation of miRNA genes
159 for this species and followed the same approach to compare them to those in the
160 closely related *A. lyrata* genome. Immunoprecipitation of AGO1 and AGO4 proteins
161 confirmed the validity of the majority of our miRNA gene predictions, including a
162 substantial fraction of those specific to either *A. halleri* or *A. lyrata*, and analysis of
163 the conservation patterns across the Viridiplantae provided a detailed picture of their
164 evolutionary progression along the proto-miRNA - canonical miRNA continuum.
165 Finally, we analyzed whole-genome resequencing data from natural *A. halleri* and *A.*
166 *lyrata* accessions and showed that the functional constraint varied along the miRNA
167 sequence in a manner that differed according to the evolutionary age of miRNA
168 genes.

169

170 ## 2. Results

171 **Reference-level assembly of a *A. halleri* genome**

172 We first produced a chromosome-level reference genome assembly for an
173 individual from Northern France (Auby-1, from the Auby population, 50.40624°N,
174 3.08265°E) based on a combination of long Oxford Nanopore Technology reads,
175 short Illumina reads and Hi-C data. Briefly, high molecular weight DNA from leaf
176 tissue was extracted and a total of 29 Gb of sequence were obtained using a
177 PromethION (Oxford Nanopore Technology). The 3.32 million reads had a N50 of
178 18.9 kbp (Supplemental Table S1). The high quality long reads were assembled
179 using NECAT (Chen et al., 2021) and then polished first using all long reads with
180 Racon (Vaser et al., 2017) and Medaka (https://github.com/nanoporetech/medaka)
181 and then using Illumina short-reads with Hapo-G (Aury and Istace, 2021)
182 (Supplemental Table S2). The resulting assembly was composed of 175 contigs and
183 had a cumulative size of 227 Mbp with an N50 of 25.9Mb (Supplemental Table S2).
184 The eight largest contigs covered 90% of the total length and had a size compatible
185 with complete chromosomes (ranging from 22.2 to 31.7 Mbp). The remaining
186 unanchored scaffolds represented only 8.4% of the assembly (Supplemental Figure
187 S1). Hi-C (omni-C) sequencing data were generated to facilitate the chromosome-
188 level assembly and were used to further orientate and anchor contigs to scaffolds
189 (Supplemental Figure S1). We assessed the completeness of the reference genome
190 using BUSCO and found 99.1% complete universal single-copy orthologs, 0.2%
191 fragmented universal single-copy orthologs and 0.7% missing universal single-copy
192 orthologs from the Brassica dataset odb10 (Supplemental Table S2). Overall, the
193 resulting assembly has a sharply higher contiguity than the one published by
194 Legrand et al., (2019) with 18-times less scaffolds and a 93-times higher N50
195 (Supplemental Table S3).
196
197 We used two approaches to annotate protein-coding genes in the genome.
198 First, we aligned the protein sequences of *A. lyrata* and *A. thaliana* against the
199 genome assembly using GeneWise (Birney et al., 2004) to search for homologs.
200 Second, RNA-sequencing data were mapped to the reference genome using Hisat2
201 (Kim et al., 2019) and assembled by Stringtie (Shumate et al., 2022). Finally, we
202 used Gmove (Dubarry et al., 2016) to combine these two sets of predictions. Overall,

203 a total of 34,721 protein-coding genes were predicted. We used OrthoFinder (Emms
204 and Kelly, 2019) to analyze orthology relationships between the predicted genes of
205 *A. halleri*, *A. lyrata* and *A. thaliana*. After removing orthogroups containing paralogs,
206 we identify 20,306 orthologous genes between *A. halleri* and *A. lyrata*, 13,082
207 orthologous genes between *A. lyrata* and *A. thaliana* and 13,977 orthologous genes
208 between *A. halleri* and *A. thaliana*.

209 **Annotation of the miRNA genes in the *A. halleri* Auby1 individual**

210 To obtain a comprehensive set of miRNA genes in the *A. halleri* reference
211 genome, we first generated ultra-deep small RNA sequencing (sRNA-seq) data from
212 two tissues (leaves and a mix of flower buds at different stages of development) of
213 the accession used to produce the reference genome (Auby-1). We obtained a total
214 of 206 and 159 million Illumina reads for the two sRNA-seq libraries (leaves and
215 buds, respectively) (Supplemental Table S4). To enhance our ability to annotate
216 miRNA genes, we combined predictions from miRkwood (Guigon et al., 2019) and
217 Shortstack (Johnson et al., 2016), two algorithms that are adapted for plant
218 genomes. While Shorstack is more conservative and predicts fewer miRNAs,
219 miRkwood includes less reliable miRNA predictions but still with a majority of the
220 miRNAs predicted in *A. thaliana* loaded in AGO1 or AGO4, which is considered high-
221 level evidence for their regulatory potential (Guigon et al., 2019). Overall, after
222 merging the predictions from the two tissues, we obtained a total of 332 predicted
223 miRNA genes in the *A. halleri* reference genome (Supplemental Table S4).

224 To investigate whether sequencing depth could be a limiting factor for the
225 discovery of miRNA genes, we randomly sub-sampled sequencing reads from the
226 library with the highest number of reads (the one obtained from leaves, comprising
227 206 million reads), and newly predicted the miRNA genes using the exact same
228 procedure in ten independent replicates for each sample size. We observed that a
229 depth of 165 million reads is required to predict 90% of the total set of miRNA genes
230 (Figure 1a), and observed no clear plateau of the number of predicted miRNA genes,
231 indicating that even such a high sequencing depth remains a limiting factor, and that
232 more miRNA genes with low abundance remain to be discovered. However, we note
233 a clear inflection of the saturation curve once the first 86 miRNA genes have been
234 discovered, suggesting that a limited set of miRNAs with relatively high abundance

235 can already be revealed with a lower sequencing depth (around 20 million reads, as
236 is classically done in many sRNA sequencing experiments).

237 **Core and accessory miRNA genes in the *A. halleri* and *A. lyrata* reference**
238 **genomes**

239 To evaluate the variation of the repertoire of miRNA genes, we then aligned
240 sRNA-seq data that we either generated ourselves (*n* = 6 libraries) or retrieved from
241 the Sequence Read Archive (SRA) at the NCBI (*n* = 13 libraries) onto the *A. halleri*
242 reference genome. For *A. lyrata*, we used the recently updated reference genome
243 (accession MN47, Kolesnikova et al., 2023) and aligned reads from *n* = 3 sRNA-seq
244 libraries that we generated and *n* = 10 sRNA-seq publicly available libraries. These
245 data originate from a diversity of geographical accessions, plant tissues (leaves,
246 buds and roots), developmental stages, sample preparation (such as True-seq,
247 Nextflex Small RNA-Seq, SOLiD Total RNA-Seq, ION total RNA-seq), sequencing
248 methods (SOLID, PROTON, Illumina) and sequencing depths (from two to 206
249 million reads) (Supplemental Table S4). Given this heterogeneity, the results are
250 expected to buffer the inherent technical biases associated with individual sRNA
251 sequencing experiments (Wright et al., 2019). Our analysis predicted between 46
252 and 267 miRNA genes per sample (Supplemental Table S4). After merging the
253 predictions across samples, we identified a total of 463 and 276 miRNA genes in *A.*
254 *halleri* and *A. lyrata*, respectively (Supplemental Data Set S1). The higher number
255 detected in *A. halleri* is expected because of the larger number of sequencing
256 datasets analyzed. Because a given miRNA precursor could produce different
257 mature miRNA molecules in different accessions, these miRNA genes together
258 resulted in a total of 678 and 521 mature miRNAs in *A. halleri* and *A. lyrata*,
259 respectively (*i.e.* on average, a miRNA gene produced 1.5 and 1.9 mature miRNAs
260 across all accessions in *A. halleri* and *A. lyrata*) (Supplemental Data Set S1;
261 Supplemental Data Set S2). About a third of these miRNA genes were predicted by
262 both softwares (287 in *A. halleri* and 87 in *A. lyrata*), while 145 and 176 genes were
263 unique to miRkwood and 31 and 13 were unique to Shortstack in *A. halleri* and *A.*
264 *lyrata,* respectively. The higher number of predictions made by miRkwood is in line
265 with Li et al., (2021), who showed that miRkwood is able to predict substantially
266 more miRNAs than other plant miRNA prediction tools.

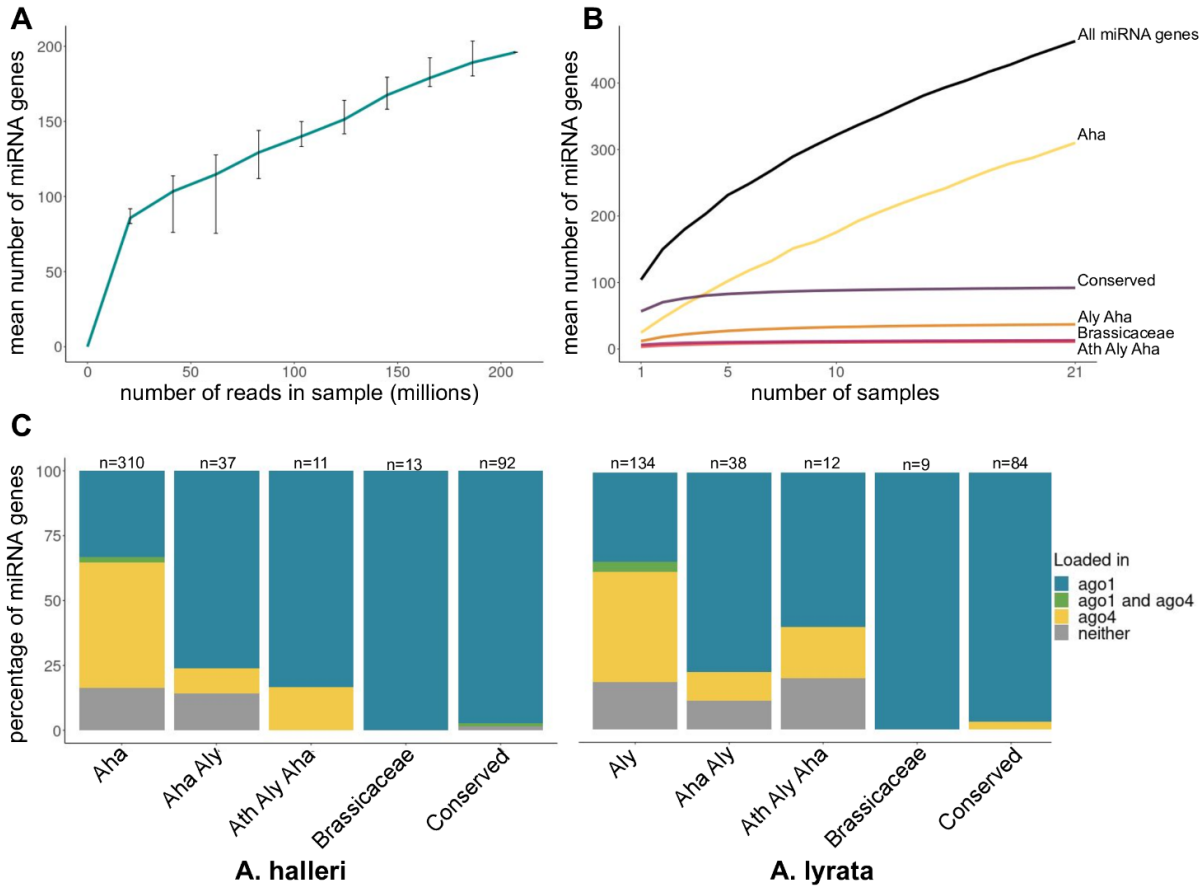
267 **Completeness of the repertoires**

268 While some miRNA genes were broadly shared and predicted in at least 80%
269 of the samples (8.6% and 9.8% in *A. halleri* and *A. lyrata*), a large proportion was
270 private to a single sample (52.3% and 42.8% in *A. halleri* and *A. lyrata*) (Figure 2a,
271 b). Hence, the number of "core" miRNA genes was relatively limited as compared to
272 the accessory miRNome, noting that different individuals from the same species can
273 carry or express different miRNA genes because of genetic or environmental
274 variation. To evaluate the completeness of the set of miRNA genes we predicted in
275 the *A. halleri* and *A. lyrata* genomes, we performed a saturation analysis by
276 randomly subsampling within the 21 and 13 individual samples from each species.
277 We performed 1,000 replicates for each sample size and evaluated how the number
278 of miRNA predictions increased with the number of samples upon which they are
279 based. For *A. halleri,* we found that 18 of the 21 samples were needed to reach 90%
280 of the total number of predictions (Figure 1b). Similarly, in *A. lyrata* 8 of the 11
281 samples needed to be included to reach 90% of the total number of predictions
282 (Supplemental Figure S2). Hence, it is clear that the repertoire of miRNA genes in
283 these two species was not saturated and was limited by the number of accessions
284 that have been sequenced so far. In particular, our results show that analyses based
285 on a single sequencing experiment in a single reference accession (as commonly
286 performed) are likely underestimating the number of miRNA genes in a species by at
287 least an order of magnitude. Altogether, our results suggest that our ability to
288 discover miRNA genes remains limited both by the number of accessions and the
289 sequencing depth.

290 **A majority of miRNA predictions are validated by AGO-IP**

291 We then performed immunoprecipitation of AGO1 and AGO4 proteins to
292 provide formal experimental validation of our predictions. To broaden the set of
293 miRNA genes we could discover, we analyzed three tissues (leaves, buds and roots)
294 from a pool of six *A. halleri* or *A. lyrata* individuals per populations (Auby, France and
295 I9, Italy for *A. halleri* and Plech, Germany for *A. lyrata*), and sequenced the small
296 RNAs associated with these proteins as well as the input material (total cellular
297 fraction). Out of the total set of miRNA genes predicted above, 314 and 147 were
298 present in the *A. halleri* and *A. lyrata* input samples. A large majority of these

299 predicted miRNA genes (88.2% and 83.4%, respectively) produced mature miRNAs
300 associated with either AGO1 or AGO4 proteins (Figure 1c). Consistent with previous
301 findings (Mi et al., 2008), the sRNAs loaded in AGO1 were predominantly 21-
302 nucleotides-long with a 5' uridine (38.6% in *A. halleri* and 33.05% in *A. lyrata*), while
303 the sRNAs loaded in AGO4 were predominantly 24-nucleotides-long with a 5'
304 adenosine (56.2% in *A. halleri* and 60.6% in *A. lyrata*) (Supplemental Figure S3).
305 Therefore, our bioinformatic annotation strategy identifies *bona fide* miRNAs with a
306 substantial number of canonical miRNA genes, including a large number of those
307 that are accession-specific.



308

309 **Figure 1. Annotation of miRNA gene repertoires.** (a) Sequencing depth saturation
310 curve for the most deeply sequenced accession (*A. halleri* Auby-1 accession,
311 leaves). Bars represent the 95% confidence interval. (b) Sample saturation curve for
312 miRNA gene repertoires in *A. halleri* according to their conservation. (c) Percentage
313 of miRNA genes loaded in AGO1 and/or AGO4 proteins out of 314 and 147 miRNA
314 genes expressed in the input fractions in *A. halleri* and *A. lyrata*.

315 **A minority of miRNA genes are conserved at a large phylogenetic scale**

316 To evaluate the evolutionary age of miRNA genes, we combined three
317 different strategies at increasingly divergent phylogenetic scales. First, we aimed at a
318 fine-scale comparison between *A. halleri* and *A. lyrata.* To do this, we considered
319 miRNA genes as syntenic if their hairpin sequences were reciprocal best-hits and if
320 they were flanked by syntenic genes. Second, we took advantage of the availability
321 of assembled genomes and sRNA sequencing experiments in eleven Brassicaceae
322 species to apply the exact same discovery pipeline we used in *A. halleri* and *A.*
323 *lyrata.* Details of the genome assemblies and sRNA-seq experiments included are
324 reported in Supplemental Table S5. Note that because of divergent genome
325 structures and variable quality of the genome assemblies we did not attempt to
326 recover synteny relationships for this phylogenetic level. Finally, we extended the
327 analysis to the broad set of Viridiplantae species included in the PmiREN 1.0
328 (Supplemental Table S6). database, which was constructed by uniformly processing
329 sRNA sequencing datasets and uses a recent set of criteria to identify miRNA genes
330 (Guo et al., 2020). Because precursor sequences can diverge rapidly, we aligned the
331 mature miRNAs predicted in *A. halleri* and *A. lyrata* (rather than full precursor
332 sequences) to the mature miRNAs predicted in the 87 species of the database
333 (Figure 2a), and considered mature miRNAs as homologous if they shared $\geqslant$ 85%
334 sequence similarity. Combining the results of the three analyses, we observed very
335 similar patterns of conservation in both species (Figure 2b, c). We defined five
336 groups of conservation for which we associated an age based on the divergence
337 time in the phylogeny: 1) deeply conserved miRNAs shared by very distant species.
338 This category represents 20% (n = 92) and 30% (n = 84) of the predicted miRNAs
339 genes in *A. halleri* and *A. lyrata*, respectively (Figure 2b, c), and includes well-
340 studied miRNA families such as miR156/miR157, miR166/miR161, miR169 and
341 miR395. 2) miRNAs shared across the Brassicaceae family. This category
342 represents 3% (n = 13) in *A. halleri* and 3% (n = 9) in *A. lyrata* (Figure 2b, c), and
343 also includes some well-studied miRNA families such as miR158, miR845, miR400.
344 3) miRNAs shared between *A. thaliana*, *A. halleri* and *A. lyrata*. This category
345 represents 2% (n = 11) in *A. halleri* and 4% (n = 12) in *A. lyrata* (Figure 2b, c), and
346 also includes some well-studied miRNA families such as miR822, miR823 and
347 miR842. 4) miRNAs shared only between *A. halleri* and *A. lyrata*. This category

348 represents 8% (*n* = 37) in *A. halleri* and 14% (*n* = 38) in *A. lyrata*, including
349 previously annotated families such as miR3433 and miR3443. 5) the *A. halleri*-
350 specific miRNAs represent 67% of the *A. halleri* repertoire (*n* = 310) and the *A.*
351 *lyrata*-specific miRNAs represents 49% of the *A. lyrata* repertoire (*n* = 134) (Figure
352 2b, c). Based on the divergence time between these two closely related species
353 (Roux et al., 2011), we estimate that this last category of miRNAs appeared at most
354 one million years ago. Overall, in both species we found that the vast majority of
355 annotated miRNAs were either broadly conserved or species-specific, with only a
356 small fraction of miRNAs showing intermediate levels of phylogenetic conservation.
357

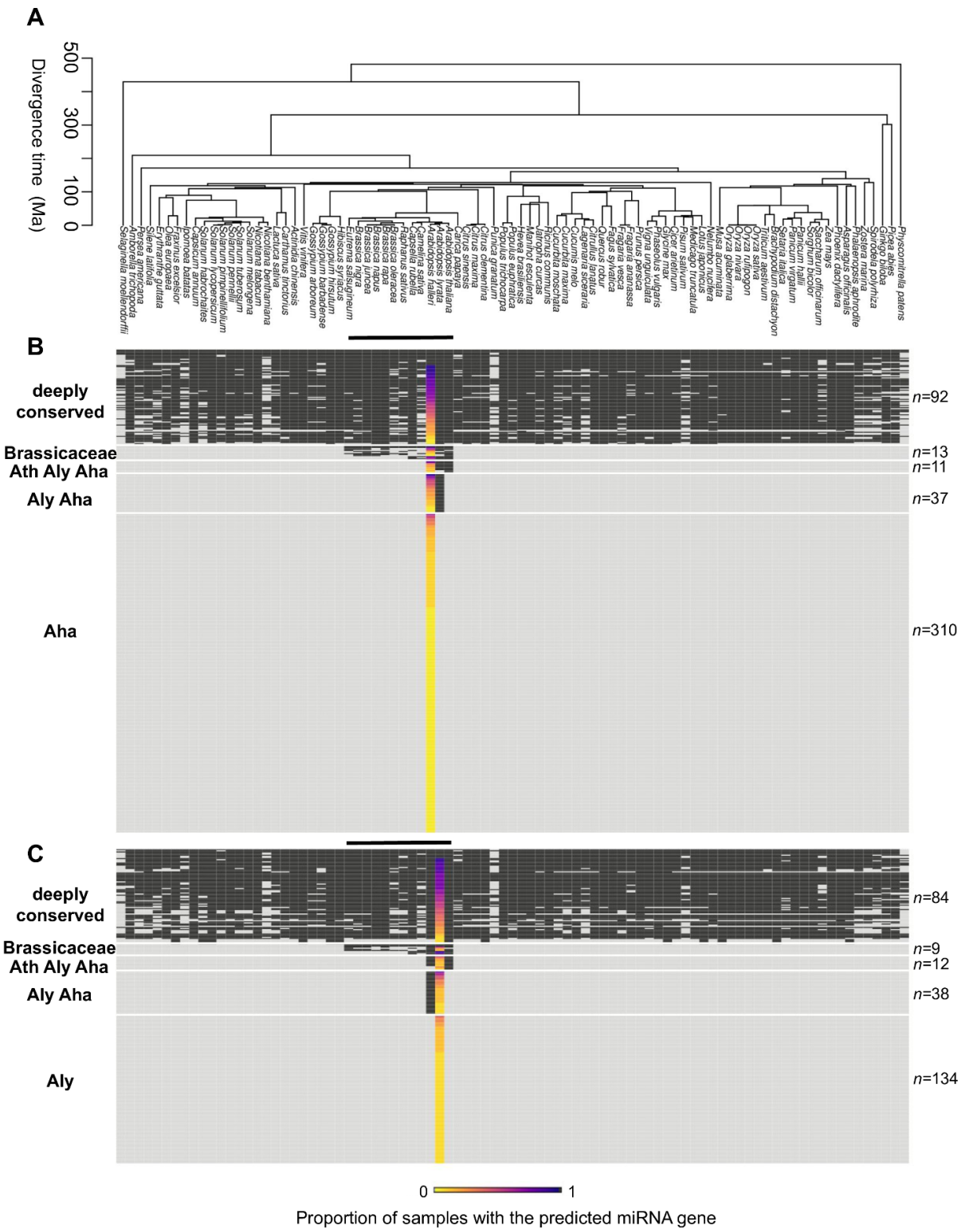


Figure 2: The majority of miRNA genes is either deeply conserved or species-specific. (a) Phylogenetic tree based on TimeTree v.5 of 87 Viridiplantae species present in the PmiREN database. Phylogenetic families are shown in supplementary table S6. The black bars indicate the Brassicaceae family. Overview of the miRNA gene conservation (b) in *A. halleri* and (c) in *A. lyrata*. Each line corresponds to one miRNA gene, and species are represented in columns. Black squares indicate the presence of an homolog/ortholog, and gray squares its absence in the corresponding

366 species. The number of miRNA genes in the five groups of conservation are
367 indicated on the left part of the figure (1: deeply conserved, 2: shared with the
368 Brassicaceae family, 3: shared with the Arabidopsis family, 4: shared between *A.*
369 *halleri* and *A. lyrata,* 5: species-specific). The proportion of accessions in which the
370 miRNA gene was predicted is indicated by the colored bars from yellow (unique
371 sample) to black (all samples).

372 **Natural variation of the repertoire of deeply conserved and species-specific**
373 **miRNAs**

374 We then determined how the set of miRNA genes in each group of
375 conservation varied with the number of samples included in the analysis. The
376 species-specific miRNA genes tended to be detected in a smaller number of
377 samples (8.2% and 10.8% of the samples in *A. halleri* and *A. lyrata,* respectively)
378 than the deeply conserved genes (detected in 61.8% and 59.6% of the sample in *A.*
379 *halleri* and *A. lyrata,* Figure 2b,c). Specifically, in *A. halleri*, 90% of the total number
380 of predictions of the most deeply conserved miRNA genes were already annotated
381 with only five of the 21 samples. Similarly, only six samples were needed to annotate
382 90% of the total number of miRNA genes shared within the Brassicaceae family,
383 respectively. In contrast, up to nine and 14 samples were needed to annotate miRNA
384 genes shared with *A. lyrata* and the *A. halleri*-specific genes (Figure 1b). Similarly, in
385 *A. lyrata*, only five and three individuals were needed to identify 90% of the deeply
386 conserved and the miRNA genes shared with the Brassicaceae family, while up to
387 eight and nine individuals were needed for the miRNA genes shared with *A. halleri*
388 and the *A. lyrata*-specific miRNA genes (Supplemental Figure S2). Overall, these
389 results indicate that our analysis of multiple samples probably represents a
390 comprehensive set of the deeply conserved miRNA genes, while the repertoire of
391 species-specific miRNA genes is not saturated even with a large number of samples.
392 Hence, including more samples would probably mostly increase the number of
393 species-specific miRNA genes.

394 **How young miRNAs become canonical miRNAs**

395 Based on our evaluation of the evolutionary age of miRNA genes, we then
396 sought to characterize how the more recent miRNAs genes differ from the more
397 ancient ones. For this analysis, we merged the orthologous miRNA genes between
398 *A. halleri* and *A. lyrata,* for which we took the average of each character value, *i.e*

399 reads abundance, length, stability, processing precision. Our dataset was composed
400 of 97 deeply conserved miRNA genes, 14 genes shared within the Brassicaceae
401 family, 14 genes shared between *A. thaliana*, *A. halleri* and *A. lyrata,* 38 genes
402 shared between *A. halleri* and *A. lyrata*, and 444 *A. halleri*- or *A. lyrata*-specific
403 miRNA genes. We used linear regression models to evaluate how a series of
404 molecular properties evolved with age of the miRNA genes. The mean normalized
405 level of expression of the miRNA genes increased from 74.4 reads per kilobase
406 million mapped reads (RPKM) for the species-specific miRNA genes to 278.9 RPKM
407 for the most deeply conserved (adjusted R^2=0.28; p-value=2.66e-55) (Figure 3a;
408 Supplemental Figure S5). Similarly, expression of the mature miRNA increased from
409 2.8 to 35.9 RPM (adjusted R^2=0.34; p-value=1.58e-70) (Supplemental Figure S4).
410 These results are consistent with previous studies that showed that conserved
411 miRNA genes tend to be expressed broadly at higher levels than the more recent
412 miRNA genes (Cuperus et al., 2011). We note that in spite of this general trend,
413 there is a strong variance within categories (as evidenced by the low adjusted R^2),
414 and some of the most recent miRNA genes could still be expressed quite
415 substantially, at levels comparable to those of some of the most conserved miRNAs.
416 Second, we analyzed the evolution of the size, stability and processing
417 precision of the hairpins produced by the miRNA genes. We found that the average
418 hairpin length tended to decrease over the course of evolution, with relatively long
419 hairpins for species-specific and Arabidopsis-specific miRNA genes (mean size of
420 213 nt and 261 nt, respectively), but a shorter mean size of only 155 nt for the deeply
421 conserved miRNAs (adjusted R^2= 0.05; p-value=1.57e-10) (Figure 3b; Supplemental
422 Figure S5). We then estimated the minimal free energy index (MFEI) of each hairpin
423 as an indicator of stability, for which a score far from zero indicates high stability. The
424 average MFEI increased from the species-specific (-1.21) to the deeply conserved
425 miRNA genes (-0.96; adjusted R^2=0.07; p-value=3.86e-13) (Figure 3c; Supplemental
426 Figure S5), corresponding to a decrease of the stability of the hairpin structure as
427 miRNAs became more ancient. The hairpin structure, in particular the presence of
428 bulges, is an important factor for cleavage by DCL proteins (Bologna et al., 2013).
429 Following Ma et al., (2010), we defined the DCL processing precision of each miRNA
430 gene as the abundance of mature miRNA sequences divided by the abundance of all
431 the reads mapping to the hairpin. A score close to one indicates a high processing
432 precision by DCL, while a score close to zero indicates an imprecise processing. The

433 average processing precision increased from the species-specific miRNA genes
434 (0.24) to the deeply conserved miRNA genes (0.67; adjusted R^2=0.38; p-
435 value=4.24e-78) (Figure 3d; Supplemental Figure S5). Altogether, our results show
436 that over the course of evolution, the hairpin produced by miRNA gene decreases in
437 length, becomes more unstable and is processed more precisely by DCL proteins.
438 Third, we examined the size and 5' nucleotide of miRNAs, as these features are
439 known to be important for miRNA biogenesis and functions (Mi et al., 2008). The
440 proportion of 21-nucleotides miRNAs with a uridine as the first 5' nucleotide
441 increased from the species-specific miRNAs (15%) to the deeply conserved (37%),
442 while the proportion of 24-nucleotides miRNAs with an adenosine as the first 5'
443 nucleotide decreased from 32% (species-specifics) to 0.3% (deeply conserved)
444 (Figure 3e). AGO1 proteins select mainly 21-nucleotides miRNAs with a 5' uridine,
445 while AGO4 proteins select mainly 24-nucleotides miRNAs with a 5' adenosine (Mi et
446 al., 2008), and accordingly we found that the vast majority of the conserved miRNAs
447 were loaded in AGO1. This was especially true for the most conserved miRNAs
448 (71/72, 99% and 66/68, 97% in *A. halleri* and *A. lyrata* respectively), but also for the
449 miRNAs shared across the Brassicaceae family (100% for both species, 8/8 and 7/7
450 in *A. halleri* and *A. lyrata* respectively), those shared across the Arabidopsis genus
451 (5/6, 83% and 3/5, 60% in *A. halleri* and *A. lyrata*) and those shared by *A. halleri* and
452 *A. lyrata* miRNAs (16/21, 89% and 14/18, 78% in *A. lyrata*). A substantial proportion
453 of the *A. halleri*- and of the *A. lyrata*-specific miRNAs (68/207, 33% and 17/49, 35%
454 respectively) were also loaded in AGO1 (Figure 1c). Loading into AGO4 followed the
455 opposite trend, as 99 of the 207 (48%) *A. halleri*-specific and 21 of the 49 (43%) *A.*
456 *lyrata*-specific miRNAs were found in the AGO4 fraction. This proportion decreased
457 rapidly as miRNA genes became older, with only 2/21 (9%) and 2/18 (11%) for
458 miRNA shared by *A. halleri* and *A. lyrata,* respectively, 1/6 (17%) and 1/5 (20%) for
459 miRNAs shared across the three Arabidopsis species, in *A. halleri* and in *A. lyrata*,
460 respectively. None of the deeply conserved miRNAs in *A. halleri* and only two of the
461 68 deeply conserved miRNAs in *A. lyrata* (3%) were associated with AGO4 (Figure
462 1c). Finally, dual loading in both AGO1 and AGO4 was relatively rare, with only
463 7/207 of the *A. halleri*-specific, 2/49 of the *A. lyrata*-specific and 2/72 of the deeply
464 conserved miRNAs in *A. halleri* being almost equally loaded in AGO1 and AGO4
465 (Figure 1c). Hence, our results provide a clear picture, where miRNAs produced by
466 nearly all ancient miRNA genes are almost exclusively loaded in AGO1, while

miRNAs produced by the very young miRNA genes are mainly loaded in AGO4 and a substantial proportion in AGO1. Thus, in spite of their limited conservation, a substantial proportion of these species-specific miRNAs may already have some regulatory potential.

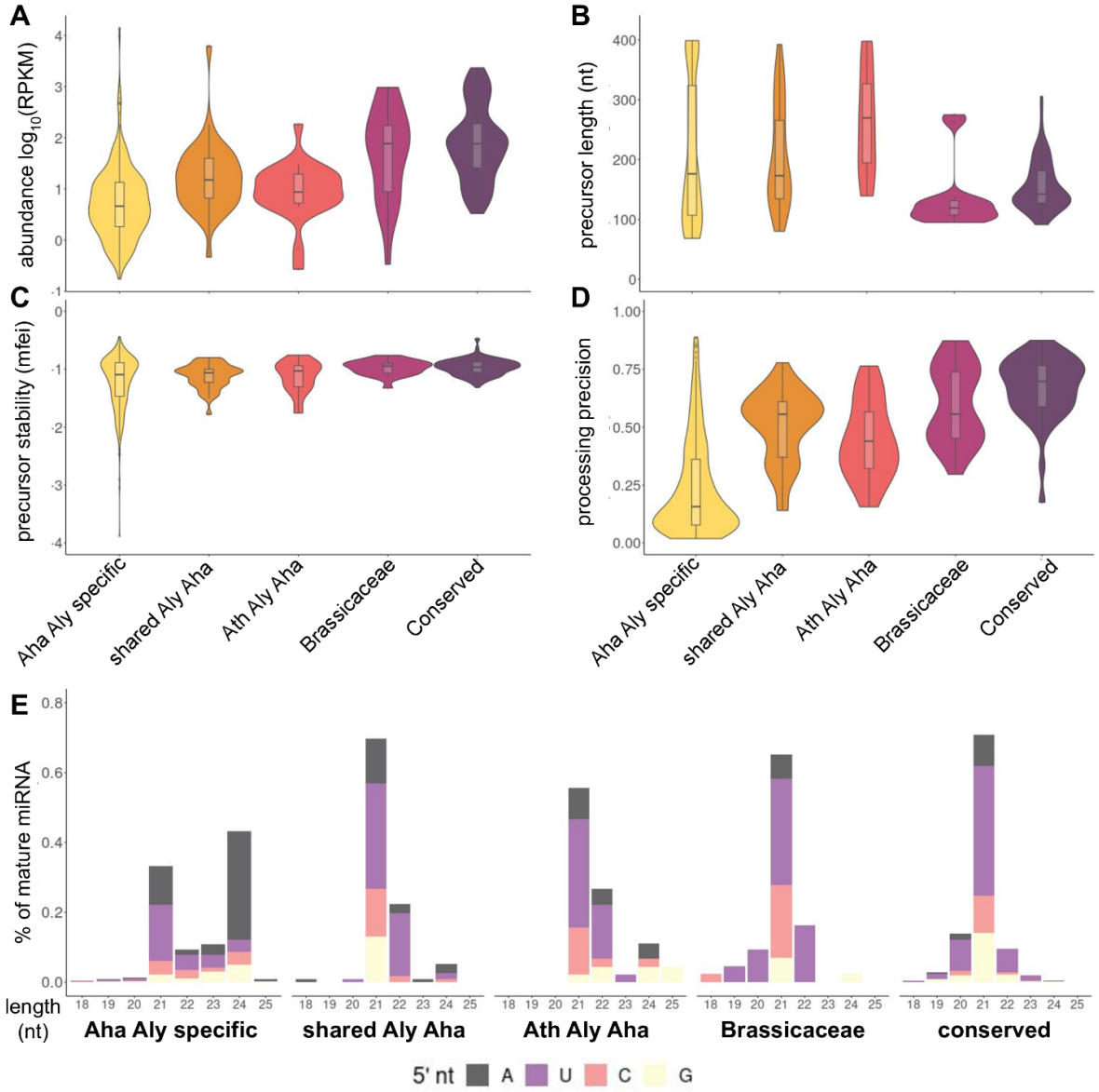
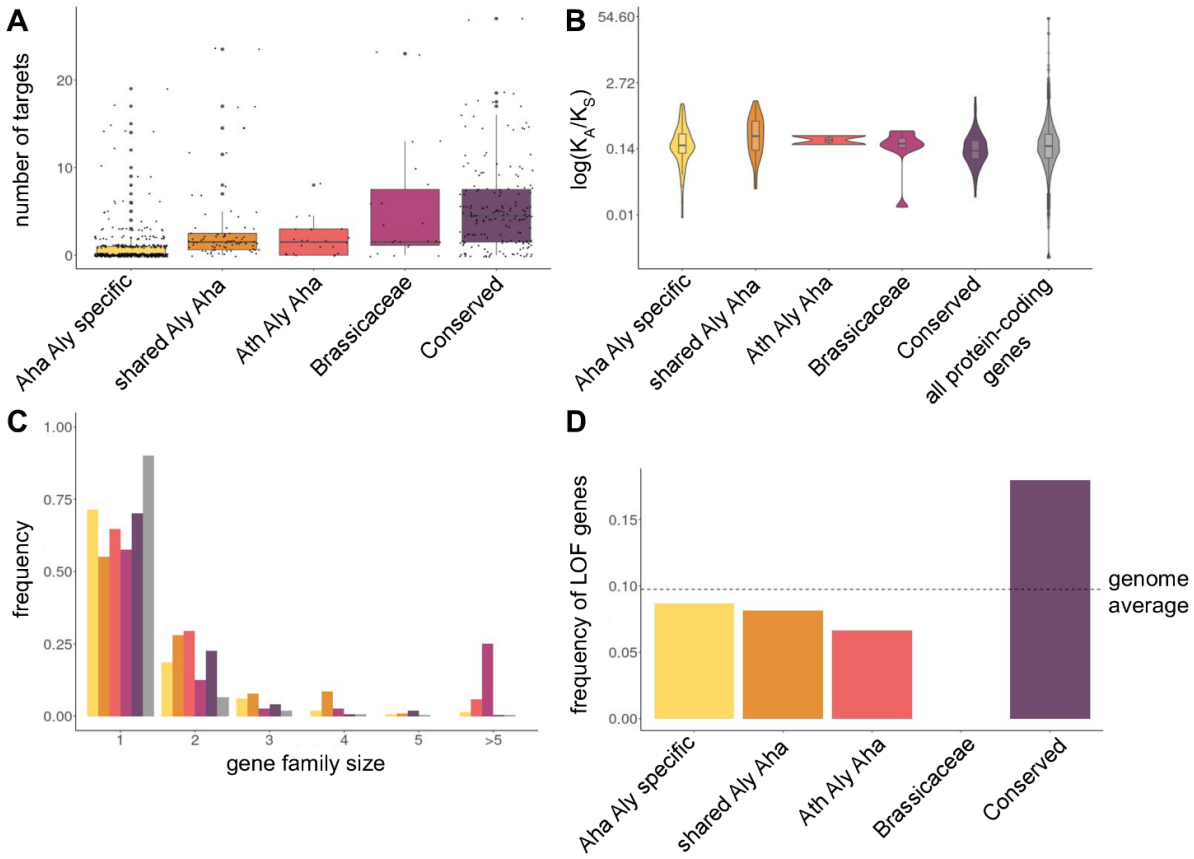


Figure 3: The characteristics of miRNA genes evolve slightly in the course of evolution. The orthologous genes between *A. halleri* and *A. lyrata* have been merged and categorized in five groups of conservation: the deeply conserved miRNA genes, those shared with the Brassicaceae family, those shared among *A. thaliana*, *A. halleri* and *A. lyrata*, those shared exclusively between *A. halleri* and *A. lyrata*, and the species-specific miRNA genes. (a) Expression level of the miRNA genes. (b) Length of the hairpin produced by the miRNA gene. (c) Hairpin stability, estimated using Minimum Free Energy Index (MFEI) calculation (d) DCL processing precision. (e) Mature miRNA size distribution and nature of the first 5' nucleotide.

481 **The number of essential targets increases over the course of evolution.**

482 To gain insight into the integration of miRNA genes in gene regulatory
483 networks, we predicted the targets in the coding sequences (CDS) across the
484 genome for each miRNA gene using Targetfinder (Bo and Wang, 2005), and first
485 evaluated the evolution of their number according to the age of the miRNA gene.
486 The number of predicted targeted genes per miRNA gene was positively correlated
487 with its age (adjusted R^2=0.17; p-value=3.09e-32) (Supplemental Figure S6),
488 increasing from species-specific miRNA genes (0.87 targets on average per miRNA)
489 to the most deeply conserved (5.42 targets on average per miRNA) (Figure 4a).
490 Second, we determined the essentiality of the genes targeted by miRNAs using
491 three proxies as in Legrand et al., (2019): 1) the size of the gene family (single-copy
492 genes are predicted to be more essential due to the lack of functional redundancy);
493 2) the k_A/k_S ratio calculated from orthologous genes between *A. halleri, A. lyrata* and
494 *A. thaliana*, for which lower values are expected for more essential genes; 3) the
495 presence of loss-of-function (LOF) phenotype in *A. thaliana* mutant*s* (Lloyd and
496 Meinke, 2012). After merging the orthologous miRNA genes, our dataset was
497 composed of 262 genes targeted by the most deeply conserved miRNA genes, 40
498 by the miRNA genes shared across the Brassicaceae family, 17 by the miRNA
499 genes shared between *A. thaliana*, *A. halleri* and *A. lyrata,* 129 by the miRNA genes
500 shared between *A. halleri* and *A. lyrata* and 150 by the species-specific miRNA
501 genes. The k_A/k_S ratios calculated from *A. halleri, A. lyrata* and *A. thaliana*
502 divergence were negatively correlated with age of the miRNA gene (adjusted
503 R^2=0.02; p-value=0.03), with a mean k_A/k_S of 0.22 and 0.33 for the genes targeted by
504 species-specific miRNAs and miRNAs shared between *A. halleri* and *A. lyrata*
505 respectively, and a lower value (k_A/k_S=0.18) for the genes targeted by the deeply
506 conserved miRNAs (Figure 4b; Supplemental Figure S6). The average gene family
507 size of the genes targeted was negatively correlated with age of the miRNA genes
508 (adjusted R^2=0.01; p-value=0.003), decreasing from 1.47 and 5.01 for the genes
509 targeted by species-specific and Brassicaceae-specific miRNAs to 1.38 for those
510 targeted by deeply conserved miRNAs (Figure 4c; Supplemental Figure S6). The
511 frequency of target genes with a LOF phenotype was correlated with age of the
512 miRNA gene (p-value=0.009). However, the frequency of LOF genes initially
513 decreased slightly (from 0.087 for the genes targeted by the species-specific genes,

514 close to the genomic average, to 0.066 for the genes targeted by miRNAs shared by
515 *A. halleri* and *A. lyrata*), but then increased sharply to 0.179 for those targeted by
516 deeply conserved miRNAs (Figure 4d). Overall, our results indicate that the number
517 of miRNA-target interactions increases over the course of evolution, along with the
518 proportion of interactions involving essential genes.
519

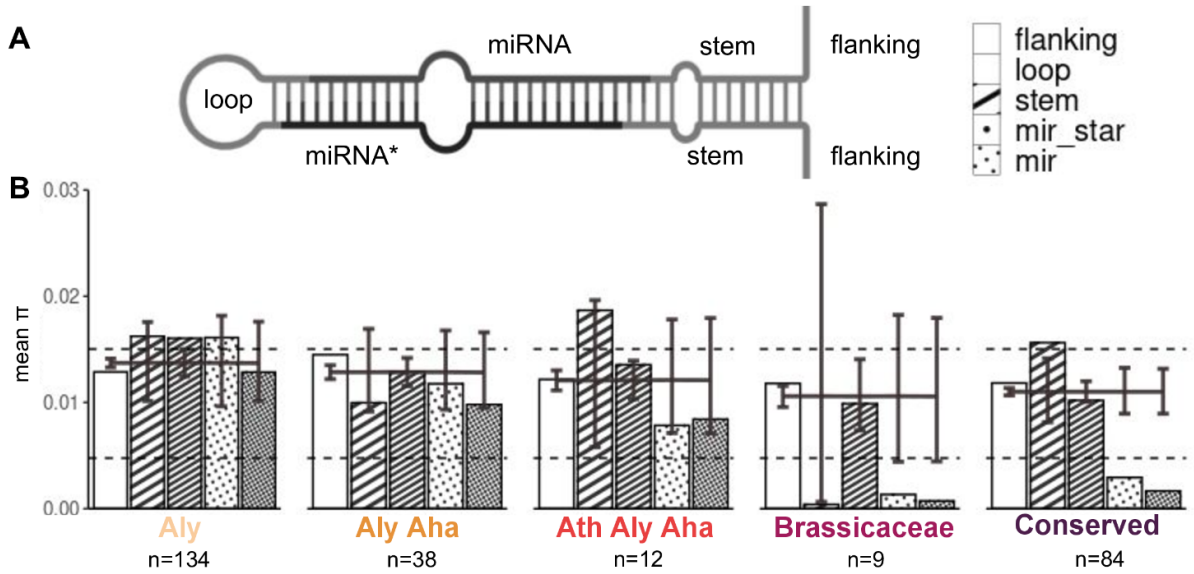


520
521 **Figure 4: The old miRNA genes have more essential targets than the young**
522 **ones.** (a) Number of targets per miRNA gene according to its age. (b) k_A/k_S ratios of
523 targeted genes calculated from *A. halleri, A. lyrata* and *A. thaliana* divergence. (c)
524 Frequency of the targeted gene family size. (d) Frequency of LOF phenotype genes
525 in the miRNA genes targets. The frequency of LOF phenotype genes in all the genes
526 present in the species is indicated by the dashed line.

527 **Functional constraint on the miRNA/miRNA* duplex over the course of**
528 **evolution**

529 To determine whether certain parts of the hairpin were more constrained by
530 natural selection than others, we investigated nucleotide polymorphism of the 276 *A.*
531 *lyrata* miRNA genes in 100 *A. lyrata* individuals from natural populations that we
532 either newly sequenced or retrieved from published datasets. We determined the
533 level of nucleotide polymorphism (π) for each part of the miRNA hairpins, including

534 the miRNA, the miRNA*, the rest of the stem, the loop, as well as 200 bp of
535 upstream and downstream flanking sequences (Figure 5a). Polymorphism of the
536 miRNA/miRNA* duplex showed a decrease of about 53% in *A. lyrata* compared to
537 the rest of the precursor (π=0.0062 vs. 0.0134), suggesting high selective constraint
538 (Supplemental Figure S7). Strikingly, polymorphism of the duplex in the deeply
539 conserved miRNA genes (mean π of 0.0062) was equivalent to the polymorphism of
540 the 0-fold degenerate positions of protein-coding genes (mean π of 0.0047 for both
541 species), suggesting that this part of the precursor evolves under considerable
542 selective constraint (Figure 5b). The overall level of polymorphism of the hairpin
543 decreased from the species-specific (mean π of 0.0153) to the deeply conserved
544 miRNA genes (0.0076) (Figure 5b). Polymorphism of the species-specific miRNA
545 genes was similar to the polymorphism of the 4-fold degenerate positions across the
546 genome (mean π of 0.0150). Thus, our results suggest that, collectively, the
547 youngest miRNA genes tend to evolve close to neutrality, although we note that this
548 conclusion does not preclude the possibility that some of them may be involved in
549 the control of important biological functions. In contrast, the selective constraint on
550 the more deeply conserved miRNAs is considerable, with levels of polymorphism of
551 the miRNA/miRNA* duplex even lower than those of the most strongly constrained
552 sites of protein-coding sequences.

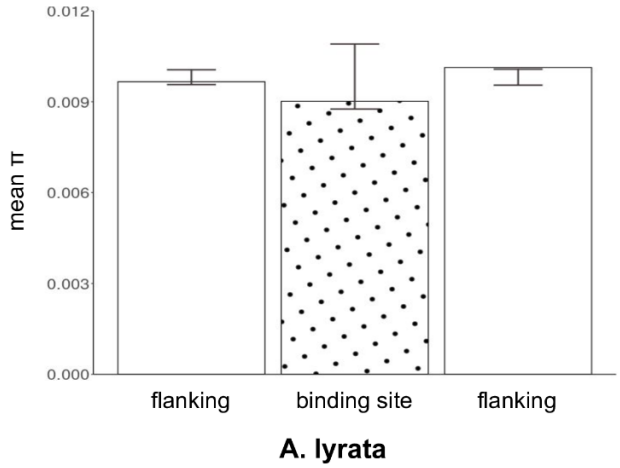


553
554 **Figure 5: Selective constraints increase over the course of the evolution of**
555 **miRNA genes.** (a) Description of the miRNA hairpin regions with the upstream and
556 downstream flanking regions (200 bp each). (b) *A. lyrata* average nucleotide diversity
557 in the different parts of the miRNA hairpins. The dashed lines represent the mean π

558 values for the 0 fold (lower line) and 4 fold (upper line) degenerate positions of all
559 protein-coding genes of the genome. The bars represent the 95% confidence interval
560 obtained by random permutation of nucleotides for 1,000 simulations.

561 **Natural selection on the miRNA binding sites**

562 We then asked whether the targeting by miRNAs could represent a detectable
563 functional constraint along the coding sequence of their target genes. To test this
564 hypothesis, we compared the polymorphism of 1,042 predicted binding sites in *A.*
565 *lyrata* with that of their 300 bp upstream and downstream flanking regions along the
566 target mRNAs. We observed slightly lower polymorphism of the binding site (average
567 π 0.0090 in *A. lyrata*) as compared to the flanking regions (average π 0.0098 in *A.*
568 *lyrata*), *i.e.* a 8.8% reduction, suggesting that the presence of the miRNA binding site
569 represents a detectable selective constraint on the CDS in addition to the original
570 constraint of coding for a specific set of amino-acids (Figure 6, Figure S8).



571

572 **Figure 6: miRNA binding sites** (dotted bar) **have lower** average nucleotide
573 diversity in *A. lyrata* than their upstream and downstream flanking regions (300 bp
574 each, white bars). The whiskers represent the 95% confidence interval under the
575 assumption of a random distribution of polymorphisms along the concatenated
576 sequence and were obtained by random permutation of nucleotides for 1,000
577 simulations.
578

579 ## 3. Discussion

580 **Challenges in the identification of miRNAs in plant genomes**

581 Proto-miRNAs have been proposed to emerge relatively readily, but studying
582 their emergence and evolution has remained challenging because this requires the
583 comparison of well-assembled and well-annotated closely related genomes, and

584 high-quality deep sRNA sequencing data. Our deep miRNA annotation of the closely
585 related *A. halleri* and *A. lyrata* genomes revealed both the long-term conservation
586 and important evolutionary lability of these genetic elements. Identifying the
587 complete set of miRNA genes in a species remains challenging for at least two
588 reasons. First, in line with Ma et al., (2010) and Cuperus et al., (2011), we find that
589 evolutionarily young miRNA genes tend to be expressed at low levels, and Chávez
590 Montes et al., (2014) suggested that their expression territories might be limited in
591 space and in time. Our results show that in spite of our extensive sequencing of a
592 diversity of samples from diverse environmental conditions, tissues, or accessions of
593 origin, the discovery of new miRNA genes is not yet exhausted in *A. halleri* and *A.*
594 *lyrata*. Specifically, we observed a relatively limited "core" miRNAome, and the
595 majority of miRNA genes belong to the "accessory" miRNAome, found in a single or
596 a few samples only. To achieve an even more comprehensive annotation, it will now
597 be necessary to take into account the genomic variation among accessions by
598 moving away from the alignment of sRNA-seq reads onto a single reference
599 genome, and assembling individual genomes across a diverse set of natural
600 accessions. While we took advantage of published data from a diversity of sources
601 to maximize the number of accessions and environmental conditions, an important
602 limitation is that we did not control for these factors. For a more detailed analysis it
603 will also be necessary to compare miRNA annotations across accessions cultivated
604 under a common garden environment. The second challenge is that annotation of
605 miRNA genes relies on a set of criteria that have remained debated (reviewed in
606 Axtell and Meyers, 2018). Here, we show that even though young miRNA genes
607 tend to exhibit non-canonical features and can thus be hard to distinguish from false
608 positives, in line with Guiguon et al., (2019) we observed that a vast majority of the
609 predicted miRNA genes were experimentally validated by AGO1- and AGO4-IP,
610 including a substantial fraction of the most evolutionarily recent ones. We conclude
611 that the criteria we used for miRNA annotation were relatively stringent, and that the
612 regulatory potential conferred by loading into AGO1 and/or AGO4 seems to be
613 acquired rapidly, at least for some of them.

614 Fahlgren et al., (2010) compared miRNAs in *A. lyrata* and *A. thaliana* and
615 estimated that 18% of them were *A. lyrata*-specific and 22% were *A. thaliana*
616 specific. Here, with our deeper annotation, we found an even higher proportion, with

617 up to 67% *A. halleri*- and 49% *A. lyrata*-specific miRNA genes. This difference
618 illustrates the increased sequencing power over the last decade, and the effect of
619 our strategy to multiply the number of accessions. In addition, Fahlgren et al., (2010)
620 estimated that 134 miRNA genes were shared exclusively between *A. lyrata* and *A.*
621 *thaliana*. Here, by including a much larger set of outgroup species (87 species in the
622 Pmiren database), we restricted this set to only 11 or 12 miRNA genes specific to the
623 Arabidopsis genus (depending on whether they were seen specifically in *A. halleri* or
624 *A. lyrata*). This further illustrates that the vast majority of miRNA genes are either
625 deeply conserved or species-specific, with very few showing intermediate levels of
626 conservation. This small number might still be an overestimation, since the set of
627 mature miRNA sequences for some species in the PmiREN database is probably
628 incomplete (*e.g. Nicotiana benthamiana n*=73, *Punica granatum n*=33 or *Pisum*
629 *sativum n*=51), possibly explaining the lack of homologs detected in the deeply
630 conserved group for some species.

631 **The evolutionary history of miRNA genes**

632 A large proportion of the miRNA genes we identified are species-specific and have
633 emerged recently, providing unprecedented power to explore the early steps of their
634 evolution. Collectively, our results largely support the verbal model of emergence of
635 miRNA genes proposed earlier (Allen et al., 2004; Voinnet et al., 2009; Baldrich et
636 al., 2018; Pegler et al., 2023), whereby young miRNA genes start from near-perfect
637 and relatively long hairpins, whose length and stability decrease by an accumulation
638 of mutations creating bulges over the course of evolution together with a decrease of
639 the diversity in size of the mature miRNA population, while the overall expression
640 level and processing precision of the hairpin increases. Our findings parallel the
641 observations made in the context of the *de novo* birth of protein-coding genes
642 (Carnuvis et al., 2012; Wilson et al., 2017). Just like a large number of ORFs can be
643 identified in a genome, we also identified a very large number of potential candidates
644 being tested by natural selection, with possibly neutral or deleterious effects initially.
645 Then only a very small fraction are retained over the long run, and eventually control
646 essential cellular functions. The question of whether the miRNA genes that
647 eventually become fixed have been slowly optimized by natural selection from
648 imperfect progenitors, or rather represent “hopeful monsters” that were immediately
649 beneficial when they arose is difficult to address directly. Yet, we note that the

650 variance of molecular features among the group of the most recent miRNA genes is
651 very large, so their distribution largely overlaps with that of the canonical miRNA
652 genes. Hence, our results are consistent with the idea that at least some of them
653 may already exhibit features allowing them to function as efficiently as the highly
654 conserved canonical miRNA genes.

655 **Origin of young miRNA genes**

656 Previous studies in rice (Zhang et al. 2011), wheat (Poretti et al., 2019; Crescente et
657 al. 2022) and more generally in angiosperms (Guo et al. 2022a) suggested that
658 transposons are an important source of origin of new miRNA genes. In contrast,
659 Fahlgren et al. (2010) and Nozawa et al. (2012) suggested that a large proportion of
660 miRNA genes in *A. thaliana* emerged from protein-coding genes. These differences
661 observed between plant species could be due to variation of the genomic content of
662 transposable elements in these species, or related to the age of the miRNA genes
663 used in these studies. In our study, we observe a high proportion of young miRNA
664 genes that produce 24 nt sRNAs and are loaded into AGO4, indicating their possible
665 integration into RNA-directed DNA Methylation (RdDM) pathway. One of the main
666 roles of the RdDM pathway is the stable silencing of transposons (Erdmann et al.
667 2020), thus it is tempting to speculate that the young miRNA genes whose products
668 are loaded in AGO4 could originate from transposons and may participate in their
669 regulation, as was documented e.g. by Borges et al. (2018). An interesting next step
670 will be to formally evaluate the contribution of the different possible progenitor
671 sequences that could have given rise to the vast repertoire of new miRNA genes we
672 uncovered.

673 **Integration of young miRNA genes in the regulatory network**

674 Although there are examples of young miRNA genes having important functional
675 roles (Wen et al., 2016; Bradley et al., 2017), our results suggest most of them are
676 unlikely to have essential biological functions, and are rapidly lost by genetic drift,
677 mutation or natural selection (Fahlgren et al., 2010; Ma et al., 2010; Nozawa et al.,
678 2012; Smith et al., 2015). Here, we found that the expression level of young miRNA
679 genes was low and their miRNA/miRNA* duplex evolved largely neutrally,
680 suggesting that these genes may not have a significant effect on the cell or the

individual. On the other hand, some miRNA genes are deeply conserved and the question of how a new miRNA integrates the functional regulatory network without impairing the fitness of the individual is still debated. Chen and Rajewsky, (2007) argued that in humans young miRNA genes have many targets that appear at random in the genome, few of which are neutral or advantageous and many of which are slightly deleterious and will be lost. In contrast, Nozawa et al., (2016) argued that young miRNA genes have only few targets, most of which are neutral, and only a small fraction of which are beneficial. Neutral miRNA-targets interactions are rapidly lost through drift mutations, while beneficial ones are conserved under purifying selection. During this period, the expression level of the miRNA can increase to enable efficient suppression of its important targets and the miRNA may also acquire new targets because the chances of forming pairs with mRNAs is higher when it is more highly expressed itself (Nozawa et al., 2016). We observed that the young miRNA genes have fewer essential targets than older ones, supporting the “growth model”. Nonetheless, we found that the proportion of these interactions involving essential genes decreased before increasing again in the deeply conserved genes. This trend could result from natural selection initially removing deleterious interactions as the expression level of the miRNA gene increases.

A striking result of our analysis is the reduced nucleotide diversity of the miRNA binding site along the mRNA sequence. However, the extent of the reduction we observed is a lot weaker than that observed in *A. thaliana* by Ehrenreich and Purugganan, (2008). This study focused on miRNA binding sites that were validated by experimental data, so are probably enriched for the interactions with the strongest magnitude of regulatory effect. In addition, the annotation of miRNAs in this study relied on more limited data, and so were also enriched for the “low hanging”, most highly expressed miRNA genes that are easier to detect. It would be interesting to extend our analysis to evaluate the effect of the choice of miRNA-target interactions on the magnitude of the reduced diversity within the target sites.

Evolutionary significance of new miRNA genes

It is clear from our results that not all miRNA genes in a genome have the same evolutionary age. Some have been present for extended periods of time, while others emerged very recently. While it is clear that the most conserved miRNA

genes fulfill essential biological functions, the evolutionary significance of the species-specific miRNA genes is harder to establish. This difficulty parallels that encountered for other genomic elements or cellular features. For instance, the evolution of long non-coding RNAs has been hotly debated. While key roles have been documented for some, (e.g. Statello et al., 2021), overall they seem to have little to no actual evolutionary importance, and most of them are largely dispensable (Goudarzi et al., 2019). Similarly, while alternative splicing is now recognized as a widespread phenomenon, the fraction of alternative splicing events with actual adaptive role is possibly low, and the variation of this feature among species is best explained by the drift-barrier hypothesis (Benitière et al. 2023; Lynch 2007). Here, even though the species-specific miRNA are not conserved, we cannot exclude that some have important biological functions. One example of non-conserved miRNA genes obviously fulfilling an important biological function is given by the sRNA precursors controlling dominance interactions between self-incompatibility alleles in Arabidopsis (Durand et al., 2014). Similarly, sRNAs determining the patterns of adaptation to the local environments encountered by specific accessions would also not be expected to show strong conservation. Given the large number of new miRNA being tested by natural selection at a given time, it is possible that non-conserved miRNA may play an important role in the rapid adaptation of plants to changing environments. At the same time, it is also possible that the majority of species-specific miRNA genes may in fact be neutral, as suggested by their low number of predicted targets and the fact that the proportion of LOF genes among their predicted target genes closely matches that of a random draw across the genome.

To achieve a better understanding of the origin of new miRNA genes, it will now be necessary to investigate the molecular nature of their potential progenitors across the genome. In addition, their actual regulatory impact is currently hard to measure, and speculation can only be made on the basis of very indirect evidence. Designing experiments to determine whether at least some of them actually have the capacity to regulate their predicted target genes will be a challenging, yet fascinating next step.

743
744 ## 4. Materials and methods

745 **Plant material**

746 *A. halleri* and *A. lyrata* plants were gathered from natural populations (see
747 Supplementary table S6) and subsequently cultivated in standard greenhouse
748 conditions. This cultivation aimed to produce leaves and buds for DNA and RNA
749 extractions. For argonaute immunoprecipitation experiments, cuttings from six *A.*
750 *halleri* Auby, ten *A. halleri* I9 and six *A. lyrata* Plech individuals were cultivated in
751 hydroponic conditions in a growth medium composed of 1 mM $Ca(NO_3)_2$, 0.5 mM
752 $MgSO_4$, 3 mM KNO_3, 0.5 mM $NH_4H_2PO_4$, 0.1 µM $CuSO_4$, 0.1 µM NaCl, 1 µM KCl, 2
753 µM $MnSO_4$, 25 µM H_3BO_3, 0.1 µM $(NH_4)_6Mo7O_{24}$, 20 µM FeEDDHA, and 1 µM (*A.*
754 *lyrata*) or 10 µM (*A. halleri*) $ZnSO_4$. The pH of the solution was maintained at 5.0
755 using MES acid buffer (2 mM). Roots were collected after six weeks.

756 ***A. halleri* reference genome**

757 **High-molecular-weight DNA extraction, PromethION library preparation**
758 **and sequencing**

759 Two grams of fresh leaves were collected and flash-frozen. High molecular weight
760 genomic DNA was extracted as described in (Belser et al., 2018 and Vacherie et al.,
761 2022). For Nanopore library preparation, the smallest genomic DNA fragments were
762 first eliminated using the Short Read Eliminator Kit (Pacific Biosciences, Menlo Park,
763 CA, USA). Starting with 1 µg of genomic DNA, libraries were then prepared
764 according to the protocol « 1D Native barcoding genomic DNA (with EXP-NBD104
765 and SQK-LSK109) » provided by Oxford Nanopore Technologies (Oxford Nanopore
766 Technologies Ltd, Oxford, UK), with some minor exceptions (increased incubation
767 times for enzymatic steps and purification on beads). All the DNA recovered after the
768 ligation of the barcode step was pooled with the DNA of three other samples before
769 the final adaptor ligation. Each library, containing a total of four barcoded samples,
770 was loaded on a R9.4.1 PromethION flow cell. In order to maintain the translocation
771 speed, flow cells were refueled with 250µl Flush Buffer when necessary. Reads were
772 basecalled using Guppy version 5.0.16 or 5.1.13. The Nanopore long reads were not
773 cleaned and raw reads were used for genome assembly.

774 **Illumina library preparation and sequencing**

775 A PCR free library was prepared following the Kapa Hyper Prep Kit procedure
776 provided by Roche (Roche, Basel, Switzerland). The library was sequenced on a
777 paired-end mode with an Illumina NovaSeq 6000 instrument (Illumina, San Diego,
778 CA, USA) using a 151 base-length read chemistry.

779 **Hi-C library preparation and sequencing**

780 Two Omni-C libraries were prepared using the Dovetail Omni-C Kit (Dovetail
781 Genomics, Scotts Valley, CA, USA) following the Mammalian Cell Protocol for
782 Sample Preparation version 1.4 after plant nuclei isolation. Briefly, flash-frozen
783 young leaves (400 mg to 1 g) were cryoground in liquid nitrogen and pure nuclei
784 were first isolated following the “High Molecular Weight DNA Extraction from
785 Recalcitrant Plant Species" protocol described by Workman et al.
786 (https://doi.org/10.1038/protex.2018.059). Once the nuclei had been isolated, the
787 pellets were treated as mammalian cells and were fixed with DSG and
788 formaldehyde; chromatin was randomly digested with DNase I and then extracted.
789 Chromatin ends were repaired and ligated to a biotinylated bridge adapter, followed
790 by proximity-ligation of adapter-containing ends. After proximity ligation, crosslinks
791 were reversed and DNA was purified from proteins. Purified DNA was treated to
792 remove biotin that was not internal to ligated fragments, and a sequencing library
793 was generated using NEBNext Ultra enzymes and Illumina-compatible adapters.
794 Biotin-containing fragments were isolated using streptavidin beads before PCR
795 enrichment of the library. The final libraries were sequenced on an Illumina
796 NovaSeq6000 instrument (Illumina, San Diego, CA, USA) with 2x 150 read length,
797 and 83 million Omni-C reads were generated.

798 **RNA library preparation and sequencing**

799 Starting with 100 ng of total RNA, rRNA were first depleted from RNA samples
800 extracted from leaves using the QIAseq FastSelect –rRNA Plant Kit (Qiagen, Hilde,
801 Germany). A library was then prepared following the NEBNext Ultra II Directional
802 RNA Library Prep for Illumina Kit procedure (New England Biolabs, Ipswich, MA,
803 USA). A library was directly prepared from 100 ng of total RNA extracted from floral
804 buds using the NEBNext Ultra II Directional RNA Library Prep for Illumina Kit.
805 Libraries were sequenced with an Illumina NovaSeq 6000 instrument (Illumina) using
806 a paired-end 151 base-length read chemistry.

807 **Assembly of the *Arabidopsis halleri* reference genome**

808 We generated three sets of read samples: the complete set of reads, 30X coverage
809 of the longest reads, and 30X coverage of the filtlong
810 (https://github.com/rrwick/Filtlong) highest-score reads (Supplemental Table S1). We
811 then launched three different assemblers, Smartdenovo (Liu et al., 2021), Flye
812 (Kolmogorov et al., 2019), and NECAT (Chen et al., 2021) on these three subsets of
813 reads with the exception that NECAT was specifically run on the entire set of reads
814 due to the implementation of a downsampling algorithm in its pipeline. Smartdenovo
815 was launched with the parameters -k 17, as advised by the developers in case of
816 larger genomes and -c 1 to generate a consensus sequence. Flye was launched with
817 an estimated genome size of 240 Mbp and the -nano-raw option. NECAT was
818 launched with a genome size of 240 Mbp and all other parameters set to their default
819 values. Out of the 7 different assemblies obtained (Supplemental Table S8), we
820 selected the Necat output for its higher contiguity (N50 > 1Mb) to continue our
821 workflow. The Necat output was polished one time using Racon (Vaser et al., 2017)
822 with Nanopore reads, then one time with Medaka
823 (https://github.com/nanoporetech/medaka) (model r941_prom_hac_g507) and
824 Nanopore reads, and two times with Hapo-G (1.3.4) (Aury and Istace, 2021) and
825 Illumina short reads. We obtained an assembly of 518 contigs.

826 However the cumulative size of the assembly was higher than expected due to the
827 high heterozygosity rate (320 Mb vs 240 Mb), and suggesting that the assembly size
828 was currently inflated by the presence of allelic duplications. As indicated by BUSCO

829 (Waterhouse et al., 2018) and KAT (Mapleson et al., 2017) (Supplemental Table S2
830 and Figure S9A), we observed the two alleles for many genes and a significant
831 proportion of homozygous kmers were present twice in the assembly. We used
832 HaploMerger2 (Huang et al., 2017) with default parameters and generated a haploid
833 version of the assembly (Batch A twice to remove major misjoin and one run of
834 Batch B). Haplomerger2 detected allelic duplications through all-against-all
835 alignments and chose for each alignment the longest genomic regions (parameter -
836 selectLongHaplotype), which may generate haplotype switches but ensure to
837 maximize the gene content. We obtained two haplotypes: a reference version
838 composed of the longer haplotype (when two haplotypes are available for a genomic
839 locus) and a second version, named alternative, with the corresponding other allele
840 of each duplicated genomic locus. At the end of the process, *A. halleri* haploid
841 assembly has a cumulative size of 225 Mb, closer to the expected one, and KAT
842 analysis showed a reduction of allelic duplications (Figure S9B). Additionally, the
843 contig N50 benefited greatly from the separation and combination of the two
844 haplotypes, rising to 3.3 Mb (Supplemental Table S2). Final assembly was polished
845 one last time with Hapo-G and Illumina short reads to ensure that no allelic regions
846 present twice in the diploid assembly have remained unpolished.

847 Chromosome-scale assembly was achieved using Hi-C data (Supplemental Table
848 S1) and the 3D-DNA pipeline (version 180419) (https://github.com/aidenlab/3d-dna).
849 Hi-C raw reads were aligned against the assembly (-s none option) using Juicer
850 (Durand et al., 2016). The resulting merged_nodups.txt file and the assembly were
851 given to the run-asm-pipeline.sh script with the options "--editor-repeat-coverage 5 --
852 splitter-coarse-stringency 30 --editor-coarse-resolution 100,000". Contact maps were
853 visualized through the Juicebox tool (version 1.11.08)
854 (https://github.com/aidenlab/Juicebox) and edited to adjust the construction of
855 chromosomes or break misjoins (Supplemental Figure S1). After edition, the
856 new.assembly file was downloaded from the Juicebox interface, filtered and
857 converted into a fasta file using the juicebox_assembly_converter.py script. Finally,
858 Hapo-G was run one last time on the chromosome-scale haploid assembly.

859 **Genome annotation of the *Arabidopsis halleri* reference assembly**

860 The *A. halleri* reference genome was masked using RepeatMasker (v.4.1.0, default
861 parameters) (Smit AFA, Hubley R, Green P. RepeatMasker.
862 http://repeatmasker.org/) and a home-made library of transposable elements (based
863 on four *Arabidopsis* species) available on the *A. halleri* repository (see Data
864 availability section). Using this procedure, 48.6% of the input assembly was masked.
865
866 Gene prediction was done using as input homologous proteins and RNA-Seq data.
867 Proteins from *Arabidopsis thaliana* (TAIR10) and *Arabidopsis lyrata* (extracted from
868 uniprot database) were aligned against *A. halleri* masked genome assembly in two
869 steps. Firstly, BLAT (default parameters) (Kent, 2002) was used to quickly localize
870 corresponding putative genes of the proteins on the genome. The best match and
871 matches with a score ≥ 90% of the best match score were retained. Secondly, the
872 alignments were refined using Genewise (default parameters) (Birney et al., 2004),
873 which is more precise for intron/exon boundary detection. Alignments were kept if
874 more than 50% of the length of the protein was aligned to the genome.
875
876 To allow the detection of expressed and/or specific genes, we also used short-read
877 RNA-Seq data extracted from two tissues (leaves and flower buds) of the same *A.*
878 *halleri* individual. Short-reads were mapped on the genome assembly using HiSat2
879 (version 2.2.1 with default parameters) (Kim et al., 2019). Bam files were then sorted
880 and merged by tissue and Stringtie (version 2.2.1) (Shumate et al., 2022) was
881 launched on each tissue with the following parameters (--rf -p 16 -v -m 150). At each
882 genomic locus, we kept only the most expressed transcript.
883
884 Finally, we integrated the protein homologies and transcripts using a combiner called
885 Gmove (-m 10000 -e 3 -score) (Dubarry et al., 2016). This tool can find CDSs based
886 on genome located evidence without any calibration step. Briefly, putative exons and
887 introns, extracted from the alignments, were used to build a simplified graph by
888 removing redundancies. Then, Gmove extracted all paths from the graph and
889 searched for open reading frames (ORFs) consistent with the protein evidence.
890 Completeness of the gene catalogs was assessed using BUSCO version 4.0.2 with
891 the Brassica dataset odb10 and default parameters (Supplemental Table S2).

892 **Identification of miRNAs**

893 **sRNA extraction, library preparation and sequencing**

894 Total RNA from *A. halleri,* Auby1, PL22, I30 and *A. lyrata* CP99 and MN47 samples
895 were extracted with the miRNeasy minikit (Qiagen). For *A. halleri* PL22, I30 and *A.*
896 *lyrata* CP99 and MN47, 3 µg of total RNA were sent to LC Sciences for library
897 construction and sequencing. For *A. halleri* Auby1, total RNA (3µg) was purified with
898 RNA Clean and Concentrator-5 kit (Zymo Research, Irvine, CA, USA, Ref. ZR1016),
899 keeping only small RNAs fragments (17 - 200 nt) for the small RNAseq library
900 preparation. Libraries were prepared following the NEXTflex Small RNA-Seq Kit v3
901 protocol provided by Perkin Elmer (Perkin Elmer, Waltham, MA, USA), starting with
902 200 ng of treated RNA, and ending with gel-free size selection and clean-up. The
903 libraries were sequenced with an Illumina NovaSeq 6000 instrument (Illumina) using
904 a paired-end 151 base-length read chemistry.

905 **Additional data collection**

906 Alongside the sRNA sequencing data produced in this study, various sets of sRNA
907 sequencing data for *A. halleri*, *A. lyrata*, *A. thaliana, Camelina sativa, Capsella*
908 *rubella, Raphanus sativus, Brassica oleracea, B. rapa, B. napus, B. juncea, B. nigra*
909 and *Eutrema salsugineum* were obtained from the NCBI SRA database
910 (https://www-ncbi-nlm-nih-gov.inee.bib.cnrs.fr/sra) (detailed information can be found
911 in Supplemental Tables S4 and S7).

912 **Identification of putative miRNA genes**

913 The raw sRNA reads were processed according to miRkwood recommendations
914 (https://bioinfo.univ-lille.fr/mirkwood/smallRNAseq/BED_file.php) using Python
915 scripts performing adapter removal, trimming and quality filtering. Then, the sRNA
916 reads were aligned to the reference genome of the respective species using Bowtie1
917 (Langmead et al., 2009), allowing for zero mismatch for the sample from *A. halleri*
918 Auby1 and allowing for one mismatch for the other samples to be able detect isomirs
919 (miRNA variants). The reference genomes used were those of the Auby1 (this
920 present study) and MN47 accessions (Kolesnikova et al., 2023) for *A. halleri* and *A.*
921 *lyrata*, respectively. For the remaining species, genome assemblies were
922 downloaded from NCBI ASSEMBLY database (https://www-ncbi-nlm-nih-

923 gov.inee.bib.cnrs.fr/assembly/) (detailed information can be found in Supplemental
924 Table S5).
925 Our annotation strategy consisted of combining miRNAs predicted by miRkwood
926 (score >=5) (Guigon et al., 2019) and Shortstack 4.0.2 (Johnson et al., 2016).
927 miRkwood include a set of filters defined in Axtell and Meyers, (2018) such as a
928 threshold for the stability of the hairpin (MFEI < -0.8), for the reads mapping to each
929 arm of the hairpin (at least ten), the accuracy of precursor cleavage, the existence of
930 the mature miRNA (read frequency at least 33%), the presence of the
931 miRNA/miRNA* duplex and its stability (Guignon et al., 2019). Then, we merged the
932 common predictions between the different samples and removed the predictions that
933 fell into small chromosomal contigs to obtain a unique repertory for each species.
934 Finally, to gain higher confidence in these predictions we mapped our sRNA read
935 data onto predicted miRNA precursors using structVis v0.4 for manual observation
936 (https://github.com/MikeAxtell/strucVis) (Supplemental Data Set S2).

937 **Experimental validation of miRNA predictions**

938 **Deep-sequencing of Argonaute-associated small RNAs**

939 The Argonaute immunoprecipitations have been done according to Barre-Villeneuve
940 et al. (2024) with anti-AGO1 antibodies (AS09 527, Agrisera) and Anti-AGO4
941 antibodies (AS09 617, Agrisera). Inflorescence, leaf and root tissues from pooled
942 individuals of *A. halleri* (Auby and I9) and *A. lyrata* (Plech) were ground in liquid
943 nitrogen and were homogenized in extraction buffer EB (50 mM Tris-HCl at pH 7.5,
944 150 mM NaCl, 5 mM $MgCl_2$, 0.2% v/v NP40, 10% glycerol, 10 µM MG132)
945 containing the EDTA-free protease inhibitor cocktail (Roche). After 15 min of
946 incubation at 4°C, cell debris were removed by centrifugation at 21,000*g* for 30 min
947 at 4°C. The clarified lysate was incubated for 2 h at 4°C at 7-10 rpm, with AGO1 or
948 AGO4 antibodies (from agrisera), and then 1 h at 4°C at 7-10 rpm with dynabeads
949 protein A (Invitrogen), equilibrated with the EB . Beads were isolated using a
950 magnetic rack, and washed once with 1 mL of EB and 4 times with 1 mL of PBS
951 (Gibco). The sRNA were extracted from total/inputs and immunoprecipitated
952 fractions using respectively Trizol and Trizol-LS, according to supplier instructions
953 (Invitrogen). Subsequent sRNA libraries were performed and sequenced by the
954 POPS platform from the plants science institute of Paris-Saclay (IPS2).

955 **Bioinformatic analysis of AGO-IP libraries**

956 After removal of adaptors, trimming and quality filtering, sequences were aligned
957 onto the *A. halleri* and *A. lyrata* reference genomes with Bowtie1 allowing for one
958 mismatch. We searched for an exact match between mature miRNA and sRNA read
959 sequences and considered a miRNA loaded in AGO protein if more than 5 reads
960 were found in the immunoprecipitate data. For each sample, reads were normalized
961 per million total mapped reads (RPM). Enrichment with respect to the
962 immunoprecipitate was calculated as the ratio of reads in the immunoprecipitate to
963 reads in the input.

964 **Conservation analysis of miRNA genes**

965 **Synteny analysis of miRNA genes**

966 The orthology maps of genes between *A. halleri* vs. *A. lyrata*, *A. halleri* vs. *A.*
967 *thaliana* and *A. lyrata* vs. *A. thaliana* were constructed using protein sequences with
968 OrthoFinder v2.5.4 (Emms and Kelly, 2019) using default parameters. Only
969 orthogroups that contain one-to-one orthologues per species were kept for further
970 comparison. Orthologous miRNAs between *A. halleri*, *A. lyrata* and *A. thaliana* were
971 identified using the gene orthology maps described above. We selected miRNA
972 genes located between upstream and downstream orthologous genes and restricted
973 the size of the chromosomal fragment to 100 kb. The sequences of framed miRNA
974 genes were aligned using the best-hit approach, commonly used to establish
975 orthology relationships within genomes (Ward and Moreno-Hagelsieb, 2014). Two
976 miRNA genes were considered syntenic if they were a reciprocal best match.

977 **miRNA genes conservation across Viridiplantae**

978 The miRNA families and the conservation across Viridiplantae were assigned based
979 on similarity of mature miRNA sequences using the PmiREN 1.0 database (Guo et
980 al., 2020). This database is specialized for plant miRNAs and is based on a
981 standardized analysis of sRNA-seq data, which reduces the variability between
982 predictions that would be due to the use of different tools. We filtered the database
983 on mature miRNA sequence length requiring 18-nt to 25-nt sequences. In addition,
984 we enriched the database with the predicted miRNAs from ten Brassicaceae species
985 (*A. thaliana, Brassica juncea, B. napus, B. nigra, B. oleracea, B. rapa, Capsella*

986 *rubella, Eutrema salsugineum, Camelina sativa, Raphanus sativus*), allowing us to
987 be more precise about the conservation status of the miRNAs inside the
988 Brassicaceae family. Then, the sequences of the mature miRNAs were aligned using
989 Exonerate (Slater et al., 2005), allowing for three mismatch/gap/insertion. Alignments
990 with a unique distant species (outside the Brassicaceae family) were considered as
991 false positives.

992 **Characterization of features of miRNA genes**

993 We assessed the thermodynamic stability of the precursors using the Minimum Free
994 Energy Index (MFEI) according to the equation MFEI = [MFE / sequence length x
995 100] / (G+C%) (Guignon et al., 2019). We determined the secondary structure MFE
996 of the precursors using the RNAfold software (Lorenz et al., 2011) and used Python
997 scripts to calculate the GC content.
998 From secondary structure, we further defined the different parts of the miRNA
999 precursors (miRNA/miRNA* duplex, loop, stem and the flanking regions) using
1000 python scripts.
1001 We calculated the abundance miRNAs in each sample where they were predicted
1002 and took the average value. The abundance of precursors was defined as the reads
1003 mapping the precursor normalized per 1,000,000 total mapped reads and the
1004 precursor length (RPKM). The abundance of mature miRNAs was normalized per
1005 1,000,000 total mapped reads (RPM).
1006 The associations between miRNA features and their age were examined with
1007 regression linear models using R (v4.1.2; R Core Team 2023).

1008 **Target characterization**

1009 **Target prediction**

1010 We identified the potential miRNA targets in the CDS of *A. halleri* and *A. lyrata* using
1011 TargetFinder (Bo and Wang, 2005) with default parameters, which provide the best
1012 balance between specificity and sensitivity (Srivastava et al., 2014). We applied a
1013 cut-off penalty score of ≤ 3 as recommended in Fahlgren et al., (2007) for reliable
1014 miRNA-mRNA target interactions.

1015 **Proxies of essentiality of *A. halleri* and *A. lyrata* genes**

1016 Three proxies have been used as in Legrand et al., (2019) to assess gene
1017 essentiality. Briefly, the all-against-all Blast method was employed using the CDS to
1018 estimate the size of the gene family. The hits with a query coverage inferior to 50%
1019 and/or an e-value superior to 1e-30 were discarded. Ka/Ks was estimated using
1020 KaKs_Calculator2.0 (Wang et al., 2010) with the Goldman and Yang method
1021 (Goldman et al., 1994) from the alignments of pairs of orthologous CDS between *A.*
1022 *halleri* vs. *A. thaliana* and *A. lyrata* vs. *A. thaliana* obtained using Water from the
1023 EMBOSS package (Rice et al., 2000). Finally, loss of function genes were identified
1024 using a dataset composed of 2400 Arabidopsis genes with a loss-of-function mutant
1025 phenotype (Lloyd and Meinke, 2012).
1026 The associations between target gene features and miRNA gene ages were
1027 examined with regression linear models using R (v4.1.2; R Core Team 2023), except
1028 for loss-of-function genes proxy for which we used a Chi-squared test on all
1029 conservation groups.

1030 **Polymorphism analysis**

1031 **Data collection**

1032 To assess the genomic diversity, we analyzed 100 *A. lyrata* individuals from natural
1033 accessions. In addition to the genomic data produced, we downloaded WGS data
1034 obtained by Takou et al., (2021) and Mattila et al., (2017). The set was composed of
1035 39 individuals from Michigan, USA (this study); Spiterstulen, Norway (24 individuals)
1036 (Mattila et al., 2017; Takou et al., 2021); Stubbsand, Sweden (6 individuals) (Mattila
1037 et al., 2017); Plech, Germany (18 individuals) (Mattila et al., 2017; Takou et al.,
1038 2021); Austria (7 individuals) (Takou et al., 2021); Mayodan, USA (6 individuals)
1039 (Mattila et al., 2017).

1040 **Variant calling and pi calculation**

1041 After adapters removal, the reads were mapped to the reference genomes of *A.*
1042 *halleri* and *A. lyrata* using bowtie2 (Langmead et al., 2012) and PCR duplicated
1043 reads were removed with picard MarkDuplicates version 2.21.4 (available at
1044 http://broadinstitute.github.io/picard). GATK version 4.1.9.0 (McKenna et al., 2010)
1045 was used to call and annotate single nucleotide polymorphisms (SNPs) using

haplotypecaller. Individual GVCF files were subjected to joint genotyping to obtain a .vcf file with information on all sites, both variant and invariant. We extracted the precursors, targets and flanking regions and filtered the resulting .vcf files with VCFtools version 0.1.16 (Danecek et al., 2011). Because invariant sites do not have quality scores, we created individual .vcf files for variant and invariant sites. Invariant sites were identified by setting the minor allele frequency to zero (--max-maf), whereas variant sites have a minor allele count ≥1 (--mac). We filtered variant sites using the following options --remove-indels --min-alleles 1 --max-alleles 2 --max-missing 0.75 --minDP 5 --minQ 30. Subsequently, we indexed both .vcf files with tabix of SAMtools and combined them with BCFtools version 1.12 (Danecek et al., 2021). The average number of nucleotide differences between genotypes (π) was calculated using VCFtools version 0.1.16 (Danecek et al., 2011). The average number of nucleotide differences between genotypes (π) was calculated using VCFtools version 0.1.16 (Danecek et al., 2011). Additionally, we carried out permutation tests to assess the probability that the differences we observed could be due to our result being different from chance alone and thus determining its significance. Specifically, For example, each nucleotide associated with its nucleotide diversity value (π) was permuted within the hairpin, and then the average π of each region of the hairpin was calculated. This was repeated a large number of times (n=1000), allowing us to define a confidence interval. If the average π observed for the hairpin part was outside the confidence interval, this meant that the observed value was different from chance and therefore significant.

Acknowledgments

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. We thank Anamaria Nesculesca, Filipe Borges for taking part in FP’s PhD committee, and Blake Meyers, Noah Fahlgren, Patricia Baldrich and Xavier Vekemans for discussions.

Supplemental data

The following materials are available in the online version of this article.

1079 **Supplemental Figure S1.** Curated chromosome-scale assembly of a reference *A.*
1080 *halleri* accession (Auby-1).
1081 **Supplemental Figure S2.** Completeness of the miRNA gene repertories according
1082 to the numbers of individuals sampled in *A. lyrata*.
1083 **Supplemental Figure S3.** Size distribution and nature of the 5'nt of AGO1 and
1084 AGO4 associated miRNAs.
1085 **Supplemental Figure S4.** Mature miRNA expression according to their
1086 conservation.
1087 **Supplemental Figure S5.** Linear regression of the miRNA genes characteristics
1088 according to their age.
1089 **Supplemental Figure S6.** Linear regression of the miRNA genes target
1090 characteristics according to their age.
1091 **Supplemental Figure S7.** The miRNA/miRNA* duplex is strongly constrained by
1092 natural selection.
1093 **Supplemental Figure S8.** Average nucleotide diversity for the miRNA binding site
1094 and upstream and downstream flanking regions (300 bp each) in *A. lyrata* mRNA
1095 targets according to the conservation of the miRNA gene targeting them.
1096 **Supplemental Figure S9.** KAT plot.
1097 **Supplemental Data Set S1.** *Arabidopsis halleri* and *A. lyrata* predicted miRNA
1098 genes.
1099 **Supplemental Data Set S2.** *Arabidopsis halleri* and *A. lyrata* predicted miRNA
1100 genes structVis visualization.
1101 **Supplemental Table S1.** Comparison of nanopore readset statistics.
1102 **Supplemental Table S2.** Assembly statistics of the reference accession (Auby1)
1103 throughout the process.
1104 **Supplemental Table S3.** Comparison of *A. halleri* PL22 and Auby1 genome
1105 assemblies.
1106 **Supplemental Table S4.** sRNAseq datasets for miRNA predictions.
1107 **Supplemental Table S5.** sRNAseq Datasets for miRNA predictions in the
1108 Brassicaceae family.
1109 **Supplemental Table S6.** Phylogenetic families of the 87 plant species used to
1110 analyze the conservation of the miRNA genes.
1111 **Supplemental Table S7.** GPS coordinates of the plant material collected for this
1112 study.
1113 **Supplemental Table S8.** Comparison of assemblies statistics.
1114
1115 **Funding**
1116 This project was funded by Région Nord Pas de Calais (MICRO^2 project) to VC and
1117 SL, ERC (NOVEL project, grant #648321) to VC, ANR (project TE-MoMa, grant
1118 ANR-18-CE02-0020-01) to VC and SL.
1119

1120 **Availability of supporting data**

1121 The Illumina and Oxford Nanopore sequencing data of the *A. halleri* reference
1122 genome are available in the European Nucleotide Archive under the following project
1123 PRJEB70878. The small RNA sequencing data are available in the NCBI-SRA
1124 database under the following BioProject PRJNA1098478. All the previously
1125 undiscovered miRNA loci were deposited at miRBase (https://www.mirbase.org;
1126 Kozomara et al. 2019). The supplementary data sets are available in Figshare under
1127 the following accession numbers : https://doi.org/10.6084/m9.figshare.25746267.v1,
1128 https://doi.org/10.6084/m9.figshare.25737294.v3 and
1129 https://doi.org/10.6084/m9.figshare.25737288.v3

1130

1131 **Credit authorship contribution statement**

1132 Cultivated plants: CPo, EH, FP. Performed molecular biology experiments: FP, JAF,
1133 CB, CC, LD, RAB, VK. Analysed data : FP, SL, CPa, FL, SG, JMA, EL, RAB, MG.
1134 Contributed samples: VK, UK. Designed the project : VC, SL, JAF, ED. Wrote the
1135 manuscript : FP, VC, SL, JMA, EL.

1136

1137 **Declaration of Competing Interest**

1138 The authors declare that they have no known competing financial interests or
1139 personal relationships that could have appeared to influence the work reported in
1140 this paper.

1141

1142 **5. References**

1143

1144 **Allen, E., Xie, Z., Gustafson, A.M., Sung, G.-H., Spatafora, J.W., and Carrington,**
1145 **J.C.** (2004). Evolution of microRNA genes by inverted duplication of target gene
1146 sequences in Arabidopsis thaliana. Nat Genet **36**: 1282–1290.

1147 **Aury, J.-M. and Istace, B.** (2021). Hapo-G, haplotype-aware polishing of genome
1148 assemblies with accurate reads. NAR Genomics and Bioinformatics **3**: lqab034.

1149 **Axtell, M.J. and Meyers, B.C.** (2018). Revisiting Criteria for Plant MicroRNA
1150 Annotation in the Era of Big Data. Plant Cell **30**: 272–284.

1151 **Baldrich, P., Beric, A., and Meyers, B.C.** (2018). Despacito: the slow evolutionary
1152 changes in plant microRNAs. Current Opinion in Plant Biology **42**: 16–22.

1153 **Barre-Villeneuve, C., Laudié, M., Carpentier, M.-C., Kuhn, L., Lagrange, T., and**
1154 **Azevedo-Favory, J.** (2024). The unique dual targeting of AGO1 by two types of
1155 PRMT enzymes promotes phasiRNA loading in *Arabidopsis thaliana*. Nucleic
1156 Acids Research: gkae045.

1157 **Belser, C. et al.** (2018). Chromosome-scale assemblies of plant genomes using
1158 nanopore long reads and optical maps. Nature Plants **4**: 879–887.

1159 **Bénitìere, F., Necsulea, A., and Duret, L.** (2024). Random genetic drift sets an
1160 upper limit on mRNA splicing accuracy in metazoans.

1161 **Birney, E., Clamp, M., and Durbin, R.** (2004). GeneWise and Genomewise.
1162 Genome Res. **14**: 988–995.

1163 **Bo, X. and Wang, S.** (2005). TargetFinder: a software for antisense oligonucleotide
1164 target site selection based on MAST and secondary structures of target mRNA.
1165 Bioinformatics **21**: 1401–1402.

1166 **Bologna, N.G., Schapire, A.L., Zhai, J., Chorostecki, U., Boisbouvier, J.,**
1167 **Meyers, B.C., and Palatnik, J.F.** (2013). Multiple RNA recognition patterns
1168 during microRNA biogenesis in plants. Genome Res. **23**: 1675–1689.

1169 **Borges, F., Parent, J.-S., Van Ex, F., Wolff, P., Martínez, G., Köhler, C., and**
1170 **Martienssen, R.A.** (2018). Transposon-derived small RNAs triggered by miR845
1171 mediate genome dosage response in Arabidopsis. Nat Genet **50**: 186–192.

1172 **Bradley, D. et al.** (2017). Evolution of flower color pattern through selection on
1173 regulatory small RNAs. Science **358**: 925–928.

1174 **Briskine, R.V., Paape, T., Shimizu-Inatsugi, R., Nishiyama, T., Akama, S., Sese,**
1175 **J., and Shimizu, K.K.** (2017). Genome assembly and annotation of *Arabidopsis*
1176 *halleri* , a model for heavy metal hyperaccumulation and evolutionary ecology.
1177 Molecular Ecology Resources **17**: 1025–1036.

1178 **Carvunis, A.-R. et al.** (2012). Proto-genes and de novo gene birth. Nature **487**: 370–
1179 374.

1180 **Chávez Montes, R.A., Rosas-Cárdenas, D.F.F., De Paoli, E., Accerbi, M.,**
1181 **Rymarquis, L.A., Mahalingam, G., Marsch-Martínez, N., Meyers, B.C.,**
1182 **Green, P.J., and De Folter, S.** (2014). Sample sequencing of vascular plants

1183 demonstrates widespread conservation and divergence of microRNAs. Nat
1184 Commun **5**: 3722.

1185 **Chen, K. and Rajewsky, N.** (2007). The evolution of gene regulation by transcription
1186 factors and microRNAs. Nat Rev Genet **8**: 93–103.

1187 **Chen, Y. et al.** (2021). Efficient assembly of nanopore reads via highly accurate and
1188 intact error correction. Nat Commun **12**: 60.

1189 **Crescente, J.M., Zavallo, D., Del Vas, M., Asurmendi, S., Helguera, M.,**
1190 **Fernandez, E., and Vanzetti, L.S.** (2022). Genome-wide identification of MITE-
1191 derived microRNAs and their targets in bread wheat. BMC Genomics **23**: 154.

1192 **Cui, J., You, C., and Chen, X.** (2017). The evolution of microRNAs in plants. Current
1193 Opinion in Plant Biology **35**: 61–67.

1194 **Cuperus, J.T., Fahlgren, N., and Carrington, J.C.** (2011). Evolution and Functional
1195 Diversification of *MIRNA* Genes. Plant Cell **23**: 431–442.

1196 **Danecek, P., Auton A, Abecasis G, Albers CA, Banks E, DePristo MA,**
1197 **Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000**
1198 **Genomes Project Analysis Group** (2011). The variant call format and
1199 VCFtools. Bioinformatics **27**: 2156–2158.

1200 **Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O.,**
1201 **Whitwham A., Keane T., McCarthy S.A., Davies R.M., and Li H**. (2021)
1202 Twelve years of SAMtools and BCFtools. GigaScience, **10**, 2021, 1–4

1203 **Dexheimer, P.J. and Cochella, L.** (2020). MicroRNAs: From Mechanism to
1204 Organism. Front. Cell Dev. Biol. **8**: 409.

1205 **Ding, N. and Zhang, B.** (2023). microRNA production in Arabidopsis. Front. Plant
1206 Sci. **14**: 1096772.

1207 **Dong, Q., Hu, B., and Zhang, C.** (2022). microRNAs and Their Roles in Plant
1208 Development. Front. Plant Sci. **13**: 824240.

1209 **Dubarry, M. et al.**, *Gmove a Tool for Eukaryotic Gene Predictions Using Various*
1210 *Evidences* (F1000Research, 2016).

1211 **Durand, E., Méheust, R., Soucaze, M., Goubet, P.M., Gallina, S., Poux, C.,**
1212 **Fobis-Loisy, I., Guillon, E., Gaude, T., Sarazin, A., Figeac, M., Prat, E.,**
1213 **Marande, W., Bergès, H., Vekemans, X., Billiard, S., Castric, V.** (2014).

1214 Dominance hierarchy arising from the evolution of a complex small RNA
1215 regulatory network. Science **346**: 1200–1205.

1216 **Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S.,**
1217 **and Aiden, E.L.** (2016). Juicer Provides a One-Click System for Analyzing Loop-
1218 Resolution Hi-C Experiments. Cell Systems **3**: 95–98.

1219 **Erdmann R.M. and Picard C.L**. RNA-directed DNA Methylation. PLoS Genet.
1220 2020:**16**(10):e1009034. https://doi.org/10.1371/journal.pgen.1009034

1221 **Ehrenreich, I.M. and Purugganan, M.D.** (2008). Sequence Variation of MicroRNAs
1222 and Their Binding Sites in Arabidopsis. Plant Physiology **146**: 1974–1982.

1223 **Emms, D.M. and Kelly, S.** (2019). OrthoFinder: phylogenetic orthology inference for
1224 comparative genomics. Genome Biol **20**: 238.

1225 **Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M.,**
1226 **Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L., and**
1227 **Carrington, J.C.** (2007). High-Throughput Sequencing of Arabidopsis
1228 microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. PLoS
1229 ONE **2**: e219.

1230 **Fahlgren, N., Sullivan, C.M., Kasschau, K.D., Chapman, E.J., Cumbie, J.S.,**
1231 **Montgomery, T.A., Gilbert, S.D., Dasenko, M., Backman, T.W.H., Givan,**
1232 **S.A., and Carrington, J.C.** (2009). Computational and analytical framework for
1233 small RNA profiling by high-throughput sequencing. RNA **15**: 992–1002.

1234 **Fahlgren, N., Jogdeo, S., Kasschau, K.D., Sullivan, C.M., Chapman, E.J.,**
1235 **Laubinger, S., Smith, L.M., Dasenko, M., Givan, S.A., Weigel, D., and**
1236 **Carrington, J.C.** (2010). MicroRNA Gene Evolution in *Arabidopsis lyrata* and
1237 *Arabidopsis thaliana*. The Plant Cell **22**: 1074–1089.

1238 **François Jacob** (1977). Evolution and Tinkering. Science **196**: 1661–1666.

1239 **Goldman N. and Yang Z.** (1994) A codon-based model of nucleotide substitution for
1240 protein-coding DNA sequences. Molecular Biology and Evolution.

1241 **Goudarzi, M., Berg, K., Pieper, L.M., and Schier, A.F.** (2019). Individual long non-
1242 coding RNAs have no overt functions in zebrafish embryogenesis, viability and
1243 fertility. eLife **8**: e40815.

1244 **Guigon, I., Legrand, S., Berthelot, J.-F., Bini, S., Lanselle, D., Benmounah, M.,**
1245 **and Touzet, H.** (2019). miRkwood: a tool for the reliable identification of
1246 microRNAs in plant genomes. BMC Genomics **20**: 532.

1247 **Guo, Z. et al.** (2020). PmiREN: a comprehensive encyclopedia of plant miRNAs.
1248 Nucleic Acids Research **48**: D1114–D1121.

1249 **Guo, Z., Kuang, Z., Deng, Y., Li, L., and Yang, X.** (2022a). Identification of
1250 Species-Specific MicroRNAs Provides Insights into Dynamic Evolution of
1251 MicroRNAs in Plants. IJMS **23**: 14273.

1252 **Guo, Z., Kuang, Z., Zhao, Y., Deng, Y., He, H., Wan, M., Tao, Y., Wang, D., Wei,**
1253 **J., Li, L., and Yang, X.** (2022b). PmiREN2.0: from data annotation to functional
1254 exploration of plant microRNAs. Nucleic Acids Research **50**: D1475–D1482.

1255 **Huang, S., Kang, M., and Xu, A.** (2017). HaploMerger2: rebuilding both haploid sub-
1256 assemblies from high-heterozygosity diploid genome assembly. Bioinformatics
1257 **33**: 2577–2579.

1258 **Johnson, N.R., Yeoh, J.M., Coruh, C., and Axtell, M.J.** Improved Placement of
1259 Multi-mapping Small RNAs. *G3 Genes*|*Genomes*|*Genetics* **6:**2103–2111.

1260 **Kent, W.J. (**2002) BLAT—The BLAST-Like Alignment Tool.

1261 **Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L.** (2019). Graph-
1262 based genome alignment and genotyping with HISAT2 and HISAT-genotype.
1263 Nat Biotechnol **37**: 907–915.

1264 **Koch, M.A., Haubold, B., and Mitchell-Olds, T.** (2000). Comparative Evolutionary
1265 Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in Arabidopsis,
1266 Arabis, and Related Genera (Brassicaceae). Molecular Biology and Evolution **17**:
1267 1483–1498.

1268 **Kolesnikova, U.K., Scott, A.D., Van De Velde, J.D., Burns, R., Tikhomirov, N.P.,**
1269 **Pfordt, U., Clarke, A.C., Yant, L., Seregin, A.P., Vekemans, X., Laurent, S.,**
1270 **and Novikova, P.Y.** (2023). Transition to Self-compatibility Associated With
1271 Dominant *S* -allele in a Diploid Siberian Progenitor of Allotetraploid *Arabidopsis*
1272 *kamchatica* Revealed by *Arabidopsis lyrata* Genomes. Molecular Biology and
1273 Evolution **40**: msad122.

1274 **Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A.** (2019). Assembly of long,
1275 error-prone reads using repeat graphs. Nat Biotechnol **37**: 540–546.

1276 **Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S.** (2019). miRBase: from
1277 microRNA sequences to function. Nucleic Acids Research **47**: D155–D162.

1278 **Kubota, S., Iwasaki, T., Hanada, K., Nagano, A.J., Fujiyama, A., Toyoda, A.,**
1279 **Sugano, S., Suzuki, Y., Hikosaka, K., Ito, M., and Morinaga, S.-I.** (2015). A
1280 Genome Scan for Genes Underlying Microgeographic-Scale Local Adaptation in
1281 a Wild Arabidopsis Species. PLoS Genet **11**: e1005361.

1282 **Kumar, S., Suleski, M., Craig, J.M., Kasprowicz, A.E., Sanderford, M., Li, M.,**
1283 **Stecher, G., and Hedges, S.B.** (2022). TimeTree 5: An Expanded Resource for
1284 Species Divergence Times. Molecular Biology and Evolution **39**: msac174.

1285 **Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L.** (2009). Ultrafast and
1286 memory-efficient alignment of short DNA sequences to the human genome.
1287 Genome Biol 10: R25.

1288 **Langmead, B. and Salzberg, S.L.** (2012). Fast gapped-read alignment with Bowtie
1289 2. Nat Methods **9**: 357–359.

1290 **Legrand, S. et al.** (2019). Differential retention of transposable element-derived
1291 sequences in outcrossing Arabidopsis genomes. Mobile DNA **10**: 30.

1292 **Li, J., Reichel, M., Li, Y., and Millar, A.A.** (2014). The functional scope of plant
1293 microRNA-mediated silencing. Trends in Plant Science **19**: 750–756.

1294 **Li, Q., Liu, G., Bao, Y., Wu, Y., and You, Q.** (2021). Evaluation and application of
1295 tools for the identification of known microRNAs in plants. Appl Plant Sci **9**.

1296 **Liu, H., Wu, S., Li, A., and Ruan, J.** (2021). SMARTdenovo: a de novo assembler
1297 using long noisy reads. Gigabyte **2021**: 1–9.

1298 **Lloyd, J. and Meinke, D.** (2012). A Comprehensive Dataset of Genes with a Loss-
1299 of-Function Mutant Phenotype in Arabidopsis. Plant Physiology **158**: 1115–1129.

1300 **Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C.,**
1301 **Stadler, P.F., and Hofacker, I.L.** (2011). ViennaRNA Package 2.0. Algorithms
1302 Mol Biol **6**: 26.

1303 **Lynch, M.** (2007). The evolution of genetic networks by non-adaptive processes. Nat
1304 Rev Genet **8**: 803–813.

1305 **Ma, Z., Coruh, C., and Axtell, M.J.** (2010). *Arabidopsis lyrata* Small RNAs:
1306 Transient *MIRNA* and Small Interfering RNA Loci within the *Arabidopsis* Genus.
1307 Plant Cell **22**: 1090–1103.

1308 **Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo,**
1309 **B.J.** (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and
1310 genome assemblies. Bioinformatics **33**: 574–576.

1311 **Mattila, T.M., Tyrmi, J., Pyhäjärvi, T., and Savolainen, O.** (2017). Genome-Wide
1312 Analysis of Colonization History and Concomitant Selection in Arabidopsis lyrata.
1313 Molecular Biology and Evolution **34**: 2665–2677.

1314 **McLysaght, A. and Guerzoni, D.** (2015). New genes from non-coding sequence:
1315 the role of de novo protein-coding genes in eukaryotic evolutionary innovation.
1316 Phil. Trans. R. Soc. B **370**: 20140332.

1317 **Mi, S. et al.** (2008). Sorting of Small RNAs into Arabidopsis Argonaute Complexes Is
1318 Directed by the 5′ Terminal Nucleotide. Cell **133**: 116–127.

1319 **Nanbo, A., Furuyama, W., and Lin, Z.** (2021). RNA Virus-Encoded miRNAs:
1320 Current Insights and Future Challenges. Front. Microbiol. **12**: 679210.

1321 **Nozawa, M., Fujimi, M., Iwamoto, C., Onizuka, K., Fukuda, N., Ikeo, K., and**
1322 **Gojobori, T.** (2016). Evolutionary Transitions of MicroRNA-Target Pairs.
1323 Genome Biol Evol **8**: 1621–1633.

1324 **Nozawa, M., Miura, S., and Nei, M.** (2012). Origins and Evolution of MicroRNA
1325 Genes in Plant Species. Genome Biology and Evolution **4**: 230–239.

1326 **Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M.,**
1327 **Shaw, R.G., Weigel, D., and Lynch, M.** (2010). The Rate and Molecular
1328 Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. Science **327**: 92–
1329 94.

1330 **Pegler, J.L., Oultram, J.M.J., Mann, C.W.G., Carroll, B.J., Grof, C.P.L., and**
1331 **Eamens, A.L.** (2023). Miniature Inverted-Repeat Transposable Elements: Small
1332 DNA Transposons That Have Contributed to Plant MICRORNA Gene Evolution.
1333 Plants **12**: 1101.

1334 **Poretti, M., Praz, C.R., Meile, L., Kälin, C., Schaefer, L.K., Schläfli, M., Widrig, V.,**
1335 **Sanchez-Vallet, A., Wicker, T., and Bourras, S.** (2020). Domestication of High-
1336 Copy Transposons Underlays the Wheat Small RNA Response to an Obligate
1337 Pathogen. Molecular Biology and Evolution **37**: 839–848.

1338 **Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P.** (2006). A diverse and
1339 evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. Genes Dev. **20**:
1340 3407–3425.

1341 **Rice, P.** EMBOSS: The European Molecular Biology Open Software Suite.

1342 **Roux, C., Castric, V., Pauwels, M., Wright, S.I., Saumitou-Laprade, P., and**
1343 **Vekemans, X.** (2011). Does Speciation between Arabidopsis halleri and
1344 Arabidopsis lyrata Coincide with Major Changes in a Molecular Target of
1345 Adaptation? PLoS ONE **6**: e26872.

1346 **Roux, J., Gonzàlez-Porta, M., and Robinson-Rechavi, M.** (2012). Comparative
1347 analysis of human and mouse expression data illuminates tissue-specific
1348 evolutionary patterns of miRNAs. Nucleic Acids Research **40**: 5890–5900.

1349 **Shumate, A., Wong, B., Pertea, G., and Pertea, M.** (2022). Improved transcriptome
1350 assembly using a hybrid of long and short reads with StringTie. PLoS Comput
1351 Biol **18**: e1009730.

1352 **Slater, G. and Birney, E.** (2005). Automated generation of heuristics for biological
1353 sequence comparison. BMC Bioinformatics **6**: 31.

1354 **Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M.** (2021). Gene regulation by
1355 long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol **22**: 96–
1356 118.

1357 **Smith, L.M., Burbano, H.A., Wang, X., Fitz, J., Wang, G., Ural-Blimke, Y., and**
1358 **Weigel, D.** (2015). Rapid divergence and high diversity of miRNAs and miRNA
1359 targets in the Camelineae. Plant J **81**: 597–610.

1360 **Srivastava, P.K., Moturu, T.R., Pandey, P., Baldwin, I.T., and Pandey, S.P.**
1361 (2014). A comparison of performance of plant miRNA target prediction tools and
1362 the characterization of features for genome-wide target prediction. BMC
1363 Genomics **15**: 348.

1364 **Takou, M., Hämälä, T., Koch, E.M., Steige, K.A., Dittberner, H., Yant, L., Genete,**
1365 **M., Sunyaev, S., Castric, V., Vekemans, X., Savolainen, O., and Meaux, J.D.**
1366 (2021). Maintenance of Adaptive Dynamics and No Detectable Load in a Range-
1367 Edge Outcrossing Plant Population. Molecular Biology and Evolution **38**: 1820–
1368 1836.

1369 **Vacherie B**., **Labadie, K., Falentin, C.,** (2022). HMW DNA extraction for Long Read
1370 Sequencing using CTAB. protocols.io https://dx.doi.org/10.17504/protoc
1371 ols.io.bp2l694yzlqe/v1

1372 **Van Oss, S.B. and Carvunis, A.-R.** (2019). De novo gene birth. PLoS Genet **15**:
1373 e1008160.

1374 **Vaser, R., Sović, I., Nagarajan, N., and Šikić, M.** (2017). Fast and accurate de
1375 novo genome assembly from long uncorrected reads. Genome Res. **27**: 737–
1376 746.

1377 **Voinnet, O.** (2009). Origin, Biogenesis, and Activity of Plant MicroRNAs. Cell **136**:
1378 669–687.

1379 **Wagner, A.** (2011). The molecular origins of evolutionary innovations. Trends in
1380 Genetics **27**: 397–410.

1381 **Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J.** (2010). KaKs_Calculator 2.0: A
1382 Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies.
1383 Genomics, Proteomics & Bioinformatics **8**: 77–80.

1384 **Ward, N. and Moreno-Hagelsieb, G.** (2014). Quickly Finding Orthologs as
1385 Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss?
1386 PLoS ONE **9**: e101850.
1387

1388 **Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P.,**
1389 **Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M.** (2018). BUSCO
1390 Applications from Quality Assessments to Gene Prediction and Phylogenomics.
1391 Molecular Biology and Evolution **35**: 543–548.
1392

1393 **Wong, G.Y. and Millar, A.A.** (2023). Target Landscape of Conserved Plant
1394 MicroRNAs and the Complexities of Their Ancient MicroRNA-Binding Sites. Plant
1395 Cell Physiol **64**:604-621.
1396

1397 **Wen, M., Lin, X., Xie, M., Wang, Y., Shen, X., Liufu, Z., Wu, C.-I., Shi, S., and**
1398 **Tang, T.** (2016). Small RNA transcriptomes of mangroves evolve adaptively in
1399 extreme environments. Sci Rep **6**: 27551.

1400 **Wilson, B.A., Foy, S.G., Neme, R., and Masel, J.** (2017). Young genes are highly
1401 disordered as predicted by the preadaptation hypothesis of de novo gene birth.
1402 Nat Ecol Evol **1**: 0146.

1403 **Wright, C., Rajpurohit, A., Burke, E.E., Williams, C., Collado-Torres, L., Kimos,**
1404 **M., Brandon, N.J., Cross, A.J., Jaffe, A.E., Weinberger, D.R., and Shin, J.H.**
1405 (2019). Comprehensive assessment of multiple biases in small RNA sequencing
1406 reveals significant differences in the performance of widely used methods. BMC
1407 Genomics **20**: 513.

1408 **Zhan, J. and Meyers, B.C.** (2023). Plant Small RNAs: Their Biogenesis, Regulatory
1409 Roles, and Functions. Annu. Rev. Plant Biol. **74**: 21–51.

1410 **Zhang Y, Jiang W, and Gao L**. Evolution of MicroRNA Genes in Oryza sativa and
1411 Arabidopsis thaliana: An Update of the Inverted Duplication Model. PLoS ONE.
1412 2011:**6**(12):e28073.

1413