AN ABSTRACT OF THE THESIS OF

Sanjuro Jogdeo for the degree of Master of Science in Molecular and Cellular Biology presented on June 22, 2012.

Title: Rapid Evolution of Post-transcriptionally Regulated RESTORER OF FERTILITY-LIKE Genes in the Genus *Arabidopsis*

Abstract approved:

James C. Carrington

The Pentatricopeptide Repeat (*PPR*) gene family produces RNA-binding proteins that target organellar transcripts. The *PPR* family is expanded in land plants, with nearly 450 genes identified in *Arabidopsis thaliana.* In plants with a Cytoplasmic Male Sterility (CMS) phenotype, members of the *PPR* family can act as a *RESTORER OF FERTILITY (Rf)* and are part of a subset of genes called *RESTORER OF FERTILITY-LIKE* (*RFL*). Unlike other *PPR* transcripts, *RFL* transcripts are targets of both microRNA (miRNA) and trans-acting siRNA (tasiRNA) and produce secondary siRNA after initial miRNA- or tasiRNA-guided cleavage. We utilized the *A. lyrata* genome assembly and high-throughput sequencing of small RNA to examine the evolutionary dynamics of the *PPR* gene family and the pattern of small RNA targeting of *RFL* transcripts. We found an expanded set of 539 *PPR* genes in *A. lyrata*, 51 of which were in the *RFL* group, often in multiple collinear copies when compared to their *A. thaliana* orthologs. In-species *RFL* paralogs appear to be more related to one another than to their collinear orthologs, which is possible evidence of gene conversion or ectopic recombination. miRNA targeting of *RFL* transcripts is largely conserved with nearly two-thirds of all target sites maintained. TasiRNA targeting was less conserved with roughly one-third of comparable validated tasiRNA targets maintained in both species. However, when clusters of potential tasiRNA targets were considered, roughly two-thirds of target sites are conserved. Production of secondary siRNA from *A. lyrata PPR* transcripts is less well defined than in *A.*

thaliana, with strong signals coming from phases that are not concordant with the miRNA- or tasiRNA-guided cleavage sites.

Rapid Evolution of Post-transcriptionally Regulated RESTORER OF FERTILITY-LIKE Genes in the Genus *Arabidopsis*

by
Sanjuro Jogdeo

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented June 22, 2012
Commencement June 2013

Master of Science thesis of Sanjuro Jogdeo presented on June 22, 2012.

APPROVED:

Major Professor, representing the Molecular and Cellular Biology Program

Director of the Molecular and Cellular Biology Program

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Sanjuro Jogdeo, Author

ACKNOWLEDGEMENTS

I would like to thank Jim Carrington for taking me as a graduate student. Noah Fahlgren was of invaluable assistance in preparing this document; I am extremely grateful for his guidance and help. Chris Sullivan was instrumental in facilitating everything computational. Other members of the Carrington lab, including Josh Cuperus, Kristin Kasschau, Alberto Carbonell, Steen Hoyer, Hernan Garcia-Ruiz, Aron Curzon, Annie Bailey, and Atsushi Takeda provided stimulating discussions and encouragement. Small RNA sequencing libraries used in this work were prepared by Kristin Kasschau, Elisabeth Chapman, Tai Montgomery, and Lisa Smith. Many thanks to Ian Small and Sota Fujii who provided PPR-related HMM models and *RFL* transcript sequences. Joey Spatafora helped with phylogenetics. Dee Denver helped with molecular analysis and was on my committee, as were Todd Mockler, and Kari van Zee. The CGRB office staff, Rosa Hill, Gail Millimaki, Jeannine Cropley, and Goretti Nguyen, provided support throughout my program and a place to work, as did Stephanie Bollman and Felipe Arredondo in Brett Tyler's lab. Particular thanks go to my parents, who often had the good sense not to ask. Coffee and donuts were provided by Interzone and Nutcakes, respectively. The *A. lyrata* genome sequence, without which this work could not be done, was produced by the U.S. Department of Energy Joint Genome Institute under the Community Sequencing Program. The MCB program and the Botany and Plant Pathology Department provided teaching opportunities. Grant support for this work came from the National Science Foundation (MCB-1231726) and the National Institutes of Health (AI043288).

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICIES

Rapid Evolution of Post-transcriptionally Regulated RESTORER OF FERTILITY-LIKE Genes in the Genus *Arabidopsis*

INTRODUCTION

Small RNA Biogenesis and Function

Small RNA regulation of gene expression in the model plant *Arabidopsis thaliana* is accomplished through a diverse set of functional pathways. The core effector of regulation is the RNA-induced Silencing Complex (RISC), which is comprised of an ARGONAUTE (AGO) protein and a single strand of RNA that is typically between 21 and 24 nucleotides (nt) in length (Baumberger, 2005; Rivas et al., 2005; Hammond et al., 2000; Reinhart et al., 2002; Jones-Rhoades and Bartel, 2004; Kasschau et al., 2007) . The single-stranded AGO-associated RNA originates from a double-stranded RNA duplex with certain characteristics that are critical for RISC loading. Both strands of the duplex have two unpaired nucleotide residues at the 3' ends that are essential for interaction with AGO proteins (Elbashir et al., 2001). One strand, called the guide strand, preferentially associates with an AGO protein based primarily on the relative thermodynamic stability of each 5' end of the duplex and 5' terminal base composition (Schwarz et al., 2003; Mi et al., 2008; Montgomery, Howell, et al., 2008; Takeda et al., 2008). There are ten members of the *AGO* gene family in the *A. thaliana* genome and each is associated primarily with a specific silencing pathway, although the function of several AGO proteins is still unclear (reviewed by Mallory and Vaucheret, 2010). *AGO1* is most often associated with miRNA and shows a strong preference for small RNA with a 5' terminal Uracil (U) (Baumberger, 2005; Mi et al., 2008; Cuperus et al., 2010; Wu et al., 2009). Mature miRNA act as specificity factors for the RISC complex through small RNA complementarity to target transcripts, though the complementarity is often not perfect (Fire et al., 1998; Zamore et al., 2000). In plants, strong complementarity between target and guide RNAs in positions 2-13 of the guide is essential for efficient targeting (Allen et al., 2005; Schwab et al., 2005). The RISC complex binds to mRNA transcripts in a sequence-specific manner and inhibits translation through either catalytic slicing or translational repression (Hammond et al., 2000; Llave et al., 2002; Lanet et al., 2009; Brodersen et al., 2008). Slicing of the mRNA occurs between nucleotides paired to positions 10 and 11 from the 5' end of the small RNA (Elbashir et al., 2001).

Small RNA biogenesis pathways are diverse and can include both endogenous and exogenous factors. MicroRNA (*MIRNA*) genes are encoded in the genome and are transcribed by RNA Polymerase II as primary transcripts (pri-miRNA) (Kim et al., 2009; Cai et al., 2004). Rather than being routed to the translational machinery, pri-miRNA form self-complementary foldbacks with at least one hairpin-like stem-loop structure that includes the miRNA/miRNA-complementary region (miRNA*). In order to release the miRNA/miRNA* duplex, the loop-distal arms are cleaved by an RNAaseIII-like endonuclease, *DICER-LIKE 1 (DCL1)* in a fashion that leaves the essential 2nt 3' overhang and a second *DCL1* cleavage event removes the loop and loop-proximal nucleotides, again leaving a 2nt 3' overhang (Kurihara and Watanabe, 2004; Vaucheret, 2006; Addo-Quaye et al., 2009). The remaining double-stranded molecule is the mature miRNA/miRNA* duplex. Recent comparisons within the *Brassicaceae* family and between a broader selection of plant species have revealed several distinctive patterns of *MIRNA* evolution (reviewed by Cuperus et al., 2011). There is strong evidence that many *MIRNA* genes arise from inverted duplications of their targets (Allen et al., 2004; Fahlgren et al., 2010, 2007), as demonstrated by the relatively high degree of sequence similarity between *MIRNA* transcripts and target genes. Transcripts from young *MIRNA* genes tend to be less abundant than deeply conserved *MIRNA* genes, with correspondingly lower levels of mature miRNA. Conserved *MIRNA* are often found in multi-gene families, sometimes containing nearly 20 members, whereas young *MIRNA* tend to be single copy (Xie et al., 2005; Jones-Rhoades and Bartel, 2004; Rajagopalan et al., 2006). Processing precision of pri-miRNA foldbacks by DCL proteins can also vary. In deeply conserved *MIRNA* genes, processing of the pri-miRNA by DCL is often very precise, leading to a high ratio of miRNA to non-miRNA sequences that derive from *MIRNA* foldback structures. By contrast, young *MIRNA* genes tend to be processed more imprecisely, with similar abundances of miRNA and other small RNA aligning to the foldback arms (Ma et al., 2010). The foldback structures themselves also tend to be more branched and longer in young *MIRNA* than their conserved counterparts.

In the canonical miRNA silencing model, transcripts cleaved by RISC complexes are routed to the RNA degradation pathway. Alternatively, RISC targeting of a subset of transcripts can trigger the secondary production of short-interfering RNAs (siRNA). For instance, the non-protein coding trans-acting siRNA-generating (*TAS*) genes are initially targeted and cleaved by AGO-miRNA complexes that trigger the production of trans-acting siRNA (tasiRNA), which are specialized secondary siRNA that regulate additional target transcripts in trans. In *A. thaliana*, there are 4 *TAS* gene families. The *TAS1* and *TAS3* families each contain three genes while *TAS2* and *TAS4* are single gene families. Only *TAS3* is broadly conserved among land plants (Vazquez et al., 2004; Axtell et al., 2006). *TAS1*, *TAS2*, and *TAS4* have not been found outside of the *Brassicaceae* lineage. *TAS1* and *TAS2* are both targeted by the AGO1-miR173 complexes, *TAS3* is targeted by AGO7-miR390, and *TAS4* is targeted by AGO1-miR828 (Vazquez et al., 2004; Yoshikawa et al., 2005; Allen et al., 2005; Rajagopalan et al., 2006; Montgomery, Yoo, et al., 2008; Montgomery, Howell, et al., 2008; Axtell et al., 2006; Cuperus et al., 2010; Chen et al., 2010; Manavella et al., 2012). The mechanism that initiates tasiRNA production varies based on the *TAS* transcript involved. *TAS3* requires two cleavage and/or binding events by AGO7-miR390 for tasiRNA production to proceed (Axtell et al., 2006; Montgomery, Howell, et al., 2008), whereas a single cleavage event of the *TAS1* and *TAS2* transcripts is sufficient to initiate tasiRNA production. In all *TAS* transcripts, initiation is followed by RNA-DEPENDENT RNA POLYMERASE (RDR6) processing, which converts single-stranded RNA to double-stranded RNA. This double-stranded RNA is processively cut by DCL4 such that double-stranded small RNA duplexes are created in a 21nt phase from the initial miRNA-guided cleavage site (Allen et al., 2005; Yoshikawa et al., 2005; Montgomery, Yoo, et al., 2008; Vazquez et al., 2004). The small RNA duplexes formed in this way are predominantly 21nt in length and are able to form RISC complexes with AGO proteins. The structure of the small RNA duplex containing the guide strand responsible for the primary cleavage event is an important factor in determining whether secondary small RNA production will occur. Small RNA from duplexes containing asymmetric bulged nucleotide can reprogram RISC to trigger secondary small RNA production. (Manavella et al., 2012; Cuperus et al., 2010; Chen et al., 2010). Small RNA that are in phase with the primary cleavage site are often abundant, but phage slippage can occur and lead to abundant small RNA production at phase-forward sites (Howell et al., 2007). Regardless, a strong phasing signal can still be detected as far as 14 cycles away from the initial cleavage site. Many tasiRNA have a terminal 5' U and combine with AGO1 to form RISC complexes, and tasiRNA with a 5' Adenine (A) bind AGO2 and comprise 20% of the AGO2-bound small RNA (Mi et al., 2008).

The biological functions of tasiRNA targeting are only partially understood. *TAS3* tasiRNA target *AUXIN RESPONSE FACTOR* (*ARF*) transcripts *ARF3* and *ARF4* in *A. thaliana* (Chitwood and Timmermans, 2010)*.* *ARF3* and *ARF4* play an important role in plant development and tasiRNA regulation of these transcripts is critical to the timing of developmental phase transitions (Fahlgren et al., 2006; Chitwood and Timmermans, 2010; Chitwood et al., 2009; Garcia et al., 2006; Adenot et al., 2006). *TAS4*-siRNA81(-) targets three MYB transcription factors and regulates anthocyanin production through an autoregulatory feedback mechanism (Luo et al., 2011). While several *TAS1* and *TAS2* tasiRNA have been demonstrated to target several members of the Pentatricopeptide (*PPR*) family of genes, the biological function of this targeting has not yet been elucidated.

Pentatricopeptide Repeat Gene Family

The *PPR* gene family in *A. thaliana* is large with approximately 441 members. *PPR* genes encode RNA-binding proteins that are characterized by tandem or near-tandem repeats of a PPR 35 amino acid motif (Aubourg et al., 2000; Lurin et al., 2004; Small and Peeters, 2000). Approximately 80 percent of PPR proteins are predicted to have organelle signaling sequences (Lurin et al., 2004; Small and Peeters, 2000; Aubourg et al., 2000) and they have been implicated in RNA editing (Kotera et al., 2005; Okuda et al., 2007; Chateigner-Boutin and Small, 2007), transcript splicing (Prikryl et al., 2010; Koprivova et al., 2010; Schmitz-Linneweber et al., 2006; Nakamura et al., 2003), and regulation of expression (Hashimoto et al., 2003; Chi et al., 2010; Liu, Rodermel, et al., 2010; Hammani et al., 2011; Liu, Yu, et al., 2010) in chloroplasts and mitochondria (reviewed by Schmitz-Linneweber and

Small, 2008). Recent work has found that the repeat motif structure most likely allows PPR proteins to bind RNA in a sequence-specific manner, with a roughly one-to-one correspondence between the number of PPR motifs present in the protein and the number of nucleotides targeted (Zehrmann et al., 2011; Delannoy et al., 2007; Nakamura et al., 2003). PPR proteins are divided into two subfamilies based on the type of PPR domain present in the protein. In addition to the canonical P-type motif, there are several PPR-related motifs that are used to differentiate the subfamilies (Lurin et al., 2004). The PLS subfamily contains P-type motifs interspersed with Long (L) and Short (S) motifs, and can also contain other PPR-related C-terminal motifs (Lurin et al., 2004). The P subfamily (PPR-P) contains predominantly P-type motifs and does not typically contain any PPR-related C-terminal motifs.

The *PPR* family often includes a *RESTORER OF FERTILITY* (*Rf*) gene in plants with a cytoplasmic male sterility (CMS) phenotype. CMS is usually the result of a chimeric or defective mitochondrial gene, the product of which disrupts pollen development and renders the pollen unviable (reviewed by Schnable and Wise, 1998). Because of the highly deleterious nature of CMS in autogamous plants, naturally occurring CMS phenotypes are found in allogamous hermaphroditic species. The specific genes involved in CMS vary between species or even between two accessions of the same species, though the CMS phenotype is often associated with defects in mitochondrial gene products, especially those that encode subunits of ATP synthase or are in a nearby coding region (reviewed by Hanson and Bentolila, 2004). Populations with a CMS phenotype generally also have an *Rf* gene that restores viable pollen production. Of the *Rf* genes identified to date, most are *PPR* genes in the *P* subfamily. CMS is found in over 140 species, roughly half of which are in natural populations (Laser and Lersten, 1972; Frank, 1989), but the nuclear restorers for many species have not been identified. The first *PPR Rf* gene was found in petunia (Bentolila et al., 2002) and the genetic basis of CMS has been identified in several other species including *Oryza sativa* (rice), *Raphanus sativus* (radish), *Zea mays* (maize), *Sorghum bicolor*, *Helianthus annuus* (sunflower), and several species in the *Brassica* genus (reviewed by Hanson and Bentolila, 2004). The majority of nuclear *Rf* genes associated with CMS have either been unambiguously identified as *PPR* genes, or were mapped to *PPR* rich regions that contain both *Rf* and non-*Rf PPR* genes. In some species, including maize (Laughnan and Gabay-Laughnan, 1983) and radish (Ogura, 1968; Ikegaya, 1986), more than one chimeric gene was found to cause CMS. Radish has two well-studied CMS cytoplasms, dubbed Ogura and Kosena and a single PPR-P protein restores fertility for both types (Brown et al., 2003; Koizuka et al., 2003). The active radish *Rf* gene is in close proximity to two other *PPR* genes, one of which is a likely pseudogene (Uyttewaal et al., 2008). The Petunia *Rf* is a *PPR* gene and is proximal to another nearly identical *PPR-P* gene (Bentolila et al., 2002). An *Rf* gene in *Mimulus gutatus* has not been identified but there are *Rf*-associated loci that can independently restore fertility and which have a high density of *PPR* genes (Barr and Fishman, 2010). Although no CMS phenotype was found in wild accessions of the outcrossing *A. lyrata,* certain hybrids of between different regional accessions display a CMS phenotype (Leppälä and Savolainen, 2011). These and other examples highlight the role that certain *PPR-P* genes play in restoring pollen viability and the genomic context in which the *Rf* genes are found.

Although autogamous *A. thaliana* does not have a CMS phenotype, it does have a grouping of *Rf-like* (*RFL*) *PPR* genes that are homologous to *Rf* genes found in other species (Fujii et al., 2006; O'Toole et al., 2008; Howell et al., 2007). All *RFL* genes appear to have a common ancestor or group of ancestors and have undergone lineage-specific expansions (Fujii et al., 2011). The *A. thaliana* expansion can be identified in a phylogeny of *PPR* genes as a clade of 30 recently diverged genes amongst the larger set of *PPR-P* genes. Twenty-four of these genes are located in two clusters on Chromosome 1, a small cluster of four genes within a 200kb span starting at position 4,183,066 and a large cluster of twenty-two genes spanning approximately 1 Megabase (Mb) starting at position 23,176,930. Many genes in region surrounding the large cluster, both *PPR* and non-*PPR*, have paralogs in the small cluster, indicating that the two regions may have been formed by a segmental duplication (Geddy and Brown, 2007). Several of these clustered *RFL* genes are considered pseudogenes (Lurin et al., 2004) and one, *AT1G62860*, is identified as a pseudogene in The Arabidopsis Information Resource (TAIR) version 10 (TAIR10) (ftp://ftp.arabidopsis.org/home/tair/, on www.arabidopsis.org, February 8, 2011). In the *A. lyrata* hybrids which produce a CMS phenotype, QTL mapping has identified a restorer-related marker in close proximity to the large *RFL* cluster on *A. lyrata* chromosome 2 (Leppälä and Savolainen, 2011). Unfortunately, the nearest flanking markers were several million bases away and thus the restorer gene could not be definitively identified as a *PPR* gene.

Small RNA Targeting of *PPR* Genes

All 30 *RFL* genes in *A. thaliana* are predicted or validated targets of various small RNA. Three miRNA have been shown to target *RFL* genes in *A. thaliana*: miR400, miR161.1, and miR161.2 (Sunkar and Zhu, 2004; Vazquez et al., 2004; Addo-Quaye et al., 2008; German et al., 2008; Allen et al., 2004; Axtell et al., 2006; Howell et al., 2007). miR161.1 and miR161.2 are processed from the same pri-miRNA foldback but from different locations on the foldback arms (Allen et al., 2004). The two *MIRNA* genes show several characteristics of recent evolutionary origins. In *A. thaliana,* mature miR400 is expressed at low levels compared to other miRNA (Addo-Quaye et al., 2008; Fahlgren et al., 2007). In *A. lyrata*, the *MIR400* gene has been identified but mature miR400 has not been sequenced (Fahlgren et al., 2010; Ma et al., 2010). Mature miRNA from *MIR161* are highly expressed but *MIR161* has a high degree of sequence similarity to its targets, consistent with a recent origin by inverted duplication (Allen et al., 2004; Fahlgren et al., 2007). All three mature miRNA are 21nt in length.

A. thaliana RFL genes are validated targets of 8 tasiRNA: *TAS2* D6(-), *TAS2* D9(-), *TAS2* D11(-), *TAS2* D12(-), *TAS1a* D9(-), *TAS1b* D4(-), *TAS1c* D6(-), *TAS1c* D10(-) (Rhoades et al., 2002; Allen et al., 2004; Sunkar and Zhu, 2004; Yoshikawa et al., 2005; Howell et al., 2007). Allen et. al. (2005) specified a nomenclature for identifying tasiRNA in the following way: The *TAS* transcript name is followed by the 21nt cycle number and terminated with the strand of origin in parenthesis. Thus, *TAS2* D6(-) is a tasiRNA that originates from the 6th 21nt phasing cycle after the miR173-guided cleavage site on the *TAS2* transcript. Additionally, it comes from the negative strand of the *TAS2* transcript (which is possible because of RDR6 reverse transcription of the cleaved transcript). As can be seen from the tasiRNA names, the functional tasiRNA that have been validated to target *PPR* transcripts all come from the negative strand of the *TAS* transcript and come from 4 to 12 cycles away from the cleavage site. *TAS2* D6(-) is the only validated *PPR*-targeting tasiRNA that is in the 22nt size class, the remaining 7 tasiRNA are 21nt in length.

Along with *TAS* transcripts, *PPR* transcripts are a primary source of phased small RNA. Initial cleavage of *PPR* transcripts and subsequent processing by RDR6 and DCL4 result in the generation of predominantly 21nt small RNA (Vazquez et al., 2004; Allen et al., 2005; Howell et al., 2007; Axtell et al., 2006). Phased production of small RNA from *A. thaliana PPR* transcripts is initiated by miR161.1-, miR161.2-, and *TAS2* D6(-)-guided cleavage (Howell et al., 2007). The phasing signal from miRNA-guided cleavage is relatively weak, but cleavage guided by *TAS2* D6(-) is associated with a strong phasing signal (see below for description of phasing signal calculation), which generally occurs at the 5’ end of the transcript (Howell et al., 2007). Seventeen of the 30 *RFL* genes produce phased small RNA, and in most cases the phasing is coincident with the miR161, miR400, or *TAS2* D6(-) cleavage sites.

The recent release of the *Arabidopsis lyrata* genome sequence provides an excellent opportunity to study the evolution of small RNA targeting of the structurally dynamic *RFL* sub-family of genes. Lineage-specific expansions of both the *RFL* genes and the *TAS1* and *TAS2* gene families highlight the need to study the dynamics of targeting interactions in closely related species. Separated by approximately 10 million years (Koch et al., 2000; Wright et al., 2002; Ossowski et al., 2010), *A. thaliana* and *A. lyrata* are two of the most closely related plants with full genome sequences and are thus uniquely positioned for this analysis.

As a matter of convention, *A. thaliana* genes will be referred to by TAIR accession. Because extensive re-annotation of the *A. lyrata* genome was necessary, new gene names were used in lieu of those published by Hu et. al (2011). *A. lyrata* genes include coordinate information as part of the gene name. For example, *Al-RFL2_894* is an *RFL* gene on scaffold 2 of the genome assembly with a start position near

894kb. Genomic differences between *A. thaliana* and *A. lyrata* will be stated using *A. thaliana* as the reference, unless otherwise specified. For instance, a deletion in *A. lyrata* could also be an insertion in *A. thaliana*, but for convenience, it will be referred to as a deletion in *A. lyrata* (relative to *A. thaliana*). It is recognized that without polarizing data the differences cannot be attributed to one or the other species.

RESULTS

PPR Gene Annotation

In order to identify *A. lyrata PPR* genes, we created a six-frame translation of the *A. lyrata* genome (Hu et al., 2011; Fahlgren et al., 2010) and used the HMMER package (Eddy, 1998, http://hmmer.janelia.org/) and Pfam Hidden Markov Models (HMM) for the PPR motif (http://pfam.janelia.org/) to identify and map PPR motifs to the six-frame translation. We mapped the six-frame translation motif positions back to genomic coordinates and compared them to gene annotations compiled by the Joint Genome Institute (Hu et al., 2011), and in some cases, to new annotations generated by the gene-finding programs GENSCAN (Burge and Karlin, 1997), GeneMark (Lukashin and Borodovsky, 1998) and fgenesh (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind). Any gene model containing a cluster of at least two exonic PPR motifs was considered a *PPR* gene candidate. In order to isolate the *RFL* genes from the larger group of *PPR* genes, we sought to differentiate between *PPR* genes in the *P* and *PLS* subfamilies. To accomplish this, we again used the HMMER package but with HMMs specific to the PPR motif subtypes and against the candidate PPR amino acid sequences rather than the whole genome. Any gene containing at least 50% P subtype motifs was considered a *P* subfamily gene. Based on this analysis, we found 539 *PPR* genes in *A. lyrata* (315 *P* subfamily and 224 *PLS* subfamily).

To ensure our annotation process was consistent with prior efforts to characterize *PPR* genes, we used the same method as above to re-annotate the *A. thaliana* genome using the TAIR10 genome assembly. We identified 461 *PPR* genes (257 *P*-type and 204 *PLS*-type). A previous study (Lurin et al., 2004) found 441 *PPR* genes in *A. thaliana* (241 *P*-type and 200 *PLS*-type). The version of the Arabidopsis Genome Initiative (AGI) *A. thaliana* gene models available at the time included only 421 out of the 441 genes identified by Lurin et al. (2004), and we used this smaller number as a basis of comparison since it is unclear whether the remaining 20 genes were ever included in subsequent *A. thaliana* genome releases. Our re-annotation identified 413 of the 421 genes. For these 413 genes, we compared our assignments of *PPR* genes to *P* or *PLS* subfamilies to those described in Lurin et al. (2004) and found only 5 cases where the assignments differed. Several of these genes had low scoring PPR motif predictions and were borderline cases. Given the similarity of our findings with those of Lurin et al. (2004), we are confident that our annotation process is consistent with previous work. It is interesting to note that when the *PPR* transcripts are separated by subfamily and binned by number of predicted PPR motifs they contain, the pattern of expansion is different between the two subfamilies. The *P* subfamily is expanded in *A. lyrata* across most of the bins (Figure 1A), whereas the *PLS* subfamily bin counts are roughly equivalent between species (Figure 1B). Also, *A. lyrata* appears to have many more genes with five or fewer PPR motif predictions (Figure 1C).

Different methods of identifying PPR motifs are useful under different circumstances. To differentiate between these different methods, we will use the term “peptide PPRs” for PPR motifs that are identified on the predicted peptide sequence of a gene. The term “6-frame PPRs” will be used for PPR motifs identified on the 6-frame translation of the whole genome.

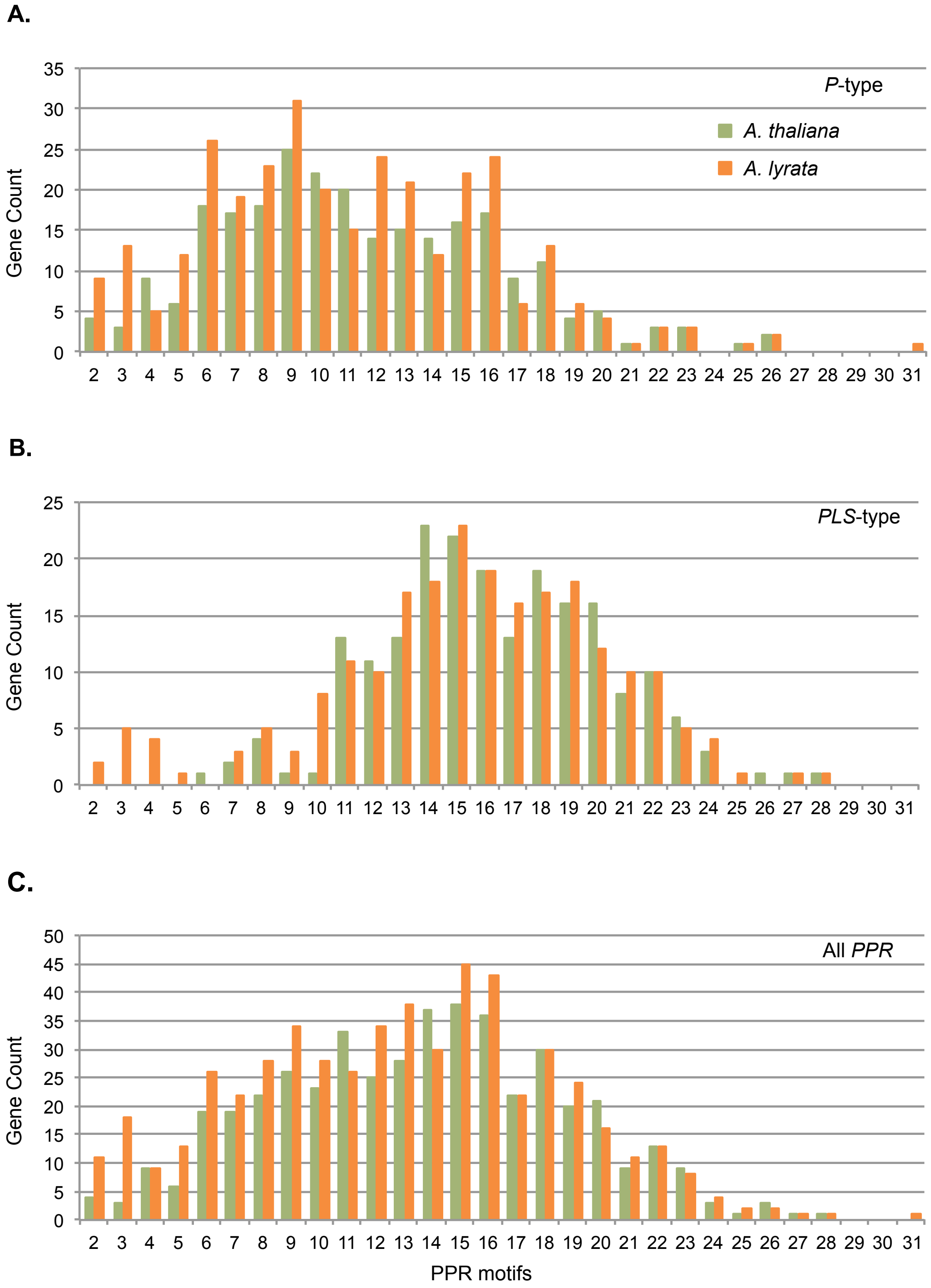


Figure 1. Comparison of *PPR* gene sub-types in *A. lyrata* and *A. thaliana*. *PPR* genes from each species were binned by the number of peptide PPR motifs found in each translated gene. **A.** *P*-type genes. **B.** *PLS*-type genes. **C.** All *PPR* genes. Color key shown in **A.** is the same for all three panels.

Identification of *RFL* Orthologs

Orthologs can be defined as "…genes derived from a single ancestral gene in the last common ancestor of the compared species" (Koonin, 2005). As we will show, *A. lyrata* often has multiple copies of *PPR* genes that are collinear with a single gene in *A. thaliana*. Although ancestry was not always determined in this study, we will use the term ortholog or co-ortholog to describe genes that are related by both sequence similarity and collinearity, or, in cases where a gene in one species has no collinear ortholog in the other species, the unpaired gene and its best match in the other species. In many cases we also sought to distinguish between orthologs that are found in collinear blocks and those that are not. We will refer to these as collinear orthologs and non-collinear orthologs, respectively.

Previous studies identified 28 *RFL* genes in *A. thaliana* (O'Toole et al., 2008; Fujii et al., 2011; Geddy and Brown, 2007). Subsequent updates to *A. thaliana* gene models from TAIR split *AT1G62910* and *AT1G63580*, adding *AT1G62914* and *AT1G64583*, increasing the total number *RFL* genes to 30. Lurin et al. (2004) marked several *RFL* genes as potential pseudogenes (*AT1G62860*, *AT1G63230*, *AT1G63320*, and *AT1G63630*), and one of these, *AT1G62860*, is annotated as a pseudogene in TAIR10. In order to identify *A. lyrata* genes orthologous to *A. thaliana RFL* genes, we used a whole-genome alignment from a previous study (Fahlgren 2010) to plot comparative alignments of *A. thaliana RFL* genes and their orthologous regions in *A. lyrata*. These plots often revealed extensive expansions of *PPR* genes in *A. lyrata* with as many as 6 collinear copies of a single *A. thaliana* gene. There was only one locus where a single *A. lyrata* gene was found to have multiple collinear *A. thaliana* orthologs, as is discussed below. In total, 41 *A. lyrata* collinear orthologs were found for 28 *A. thaliana* genes. To identify non-collinear *A. lyrata RFL* genes and to confirm the genes found through whole-genome alignment, we created four neighbor-joining trees of the *A. lyrata PPR-P* subfamily using different substitution models and gap penalties. As in *A. thaliana* (Howell et al., 2007; Fujii et al., 2011), the *RFL* genes in *A. lyrata* formed a distinctive clade of recently expanded genes and this grouping was largely consistent across all four trees (one of these is depicted in Figure 2). The *A. lyrata RFL* clade includes nine genes that were not found by the collinear search and four that were found by the collinear search, but at a sufficient distance from the collinear region to be considered non-collinear orthologs (Figure 2). Two of the nine non-collinear genes were found adjacent to one another and roughly midway between the collinear orthologs of *AT1G12620* and *AT1G12700*. A third gene, *Al-RFL1_5179*, was found at this locus but was not present in the *A. lyrata RFL* clade. Based on its proximity to, and sequence similarity with, other *RFL* genes, we considered it an *RFL* gene. Two of the four neighbor-joining trees contained two sequences in the *RFL* clade that we did not consider *RFL* genes (Figure 2). One predicted gene is just upstream of *Al-RFL2_1393* and has two exons separated by a ~4kb intron that consists almost entirely of unassembled sequence. The PPR motifs are divided between the two exons. While it seems likely that there is at least one additional *PPR* gene at this locus, the current assembly does not permit a high quality annotation. The other predicted gene contains five PPR motifs but was short (~900bp) and was not in close proximity to other *RFL* or *PPR* genes. The orthologous gene in *A. thaliana* is a *PPR* gene of ~2kb in size but is not part of the *RFL* clade. It is predicted to contain up to eight PPR motifs. We chose to exclude the *A. lyrata* gene from our analysis as it is likely a non-*RFL* pseudogene. For *RFL* genes in either species with no collinear ortholog, we determined the most appropriate non-collinear ortholog by BLAST (Altschul et al., 1997) alignment to other *RFL* protein sequences.

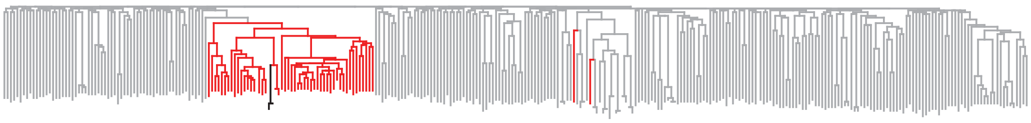


Figure 2. Neighbor-joining tree of *A. lyrata* P-type PPR peptide sequences. Red highlighted branches are RFL peptide sequences. Gray branches are non-RFL PPR-P peptide sequences. Two branches in the center of the RFL clade highlighted in black are not considered RFL sequences.

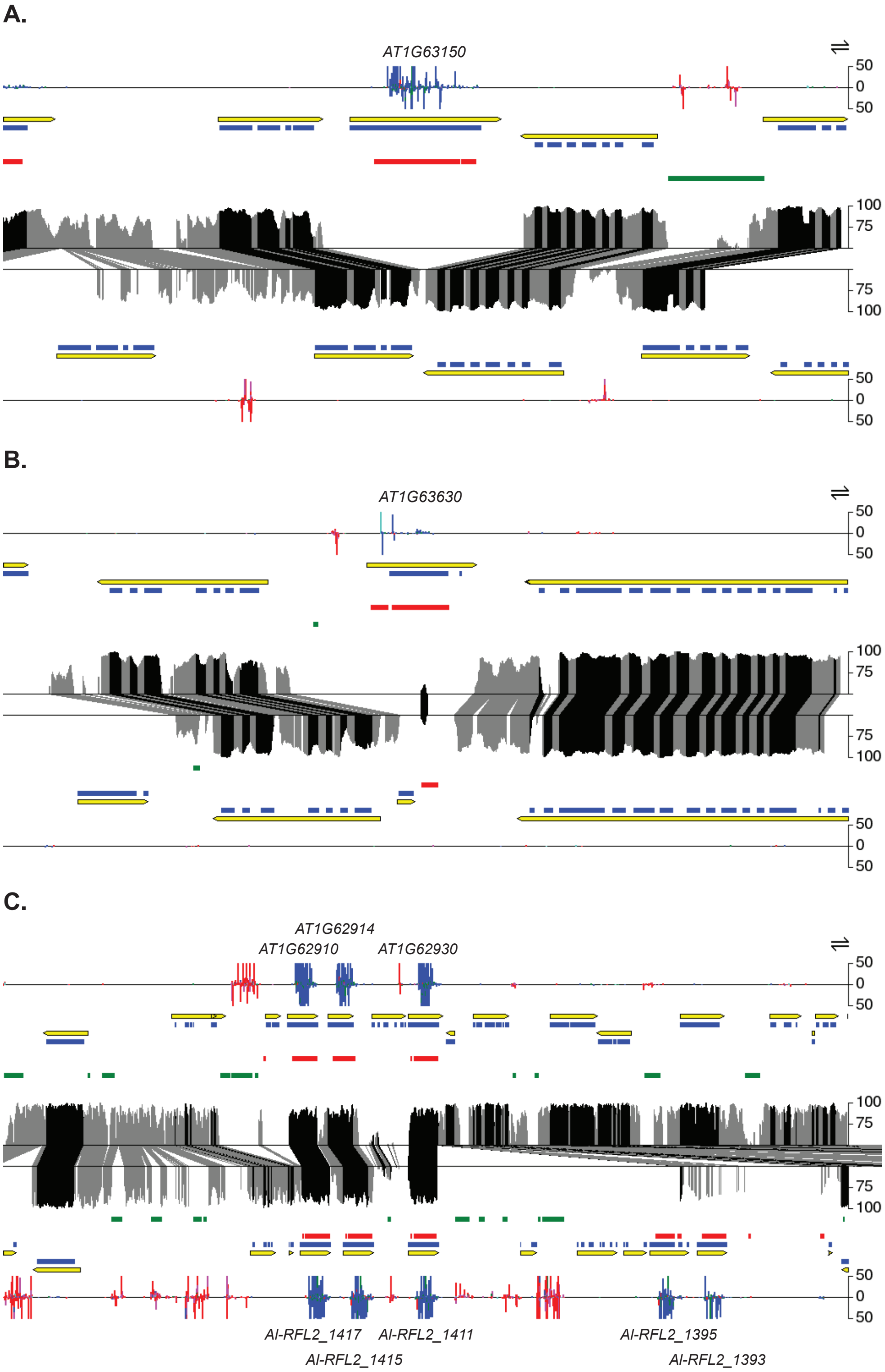
A. thaliana genes *AT1G63150* and *AT1G63630* had no detectable collinear orthologs in *A. lyrata*. For both genes, genomic alignments contain gaps in *A. lyrata* where collinear orthologs were expected, indicating likely insertions in *A. lyrata* or deletions in *A. thaliana* (Figure 3A). BLASTN alignments of *AT1G63150* against *A. lyrata RFL* genes yielded many similar hits, the strongest of which were to *Al-RFL2_629* and *Al-RFL2_894*. *Al-RFL_629* does not have a collinear ortholog in *A. thaliana*. *Al-RFL2_894* has a collinear ortholog but it is in a region that is highly diverged from *A. thaliana* (see additional discussion below). Given the stronger nucleotide alignment, *Al-RFL2_894* is considered the ortholog of *AT1G63150* for future comparisons. The *A. lyrata* genomic region that is orthologous to *AT1G63630* is predicted to have 2 PPR motifs in the 6-frame translation, but there is no predicted gene that overlaps those motifs (Figure 3B). Interestingly, Lurin et al. (2004) marked *AT1G63630* as a probable pseudogene in *A. thaliana*. As with *AT1G63150*, the best match of *AT1G63630* in *A. lyrata* is to a gene with no collinear ortholog, *Al-RFL2_700*. The strongest alignment to *Al-RFL2_700* in *A. thaliana* is to *AT1G63230*, but *Al-RFL2_700* is also a close match of *AT1G63150*. It is possible that there were either lineage-specific expansions of *Al-RFL2_894* and *Al-RFL2_700*, or that their ancestral genes experienced expansions prior to speciation, followed by a lineage-specific deletions.

There were two instances of paralogous gene clusters found in *A. lyrata* some distance away from their *A. thaliana* orthologs. One of these clusters was in the region orthologous to *A. thaliana* genes *AT1G62910*, *AT1G62914*, and *AT1G6290*, which contains collinear orthologs *Al-RFL2_1417*, *Al-RFL2_1415*, and *Al-RFL2_1411*. Two genes, *Al-RFL2_1393* and *Al-RFL2_1395*, can be found approximately 10-15kb upstream of the collinear orthologs and these are closely related to the three *A. thaliana* genes (Figure 3C). BLAST alignments of *Al-RFL2_1393* against the three *A. thaliana* transcripts yielded nearly identical results (see Appendix 1 for BLAST output). A peptide alignment shows a slightly better alignment to *AT1G62930* but over a shorter sequence length. *Al-RFL2_1393* also closely aligns to *AT1G62930* and *AT1G62910* but the *AT1G62930* alignment is slightly better in both cases. As noted above, the locus just upstream of Al-RFL_*1393* contains a gene prediction that is disrupted by ~4kb of unassembled sequence. This raises the possibility that there is a third gene to accompany *Al-RFL2_1393* and *Al-RFL2_1395*, and that the entire region collinear to the *A. thaliana* genes has been duplicated.

The second non-collinear cluster of *PPR* genes includes *Al-RFL1_5179* and *Al-RFL1_5181* (putative pseudogenes, see below) and *Al-RFL1_5184*. The cluster is located approximately 50kb and 25kb away from collinear orthologs of *AT1G12620* and *AT1G12700*, respectively. BLAST results indicate that all three genes are closely related to *AT1G12620*, *AT1G12300*, and *AT1G12775* and somewhat less to *AT1G12700*. *Al-RFL_5181* had better BLAST alignment to *AT1G12300* and those two genes were designated as non-collinear orthologs. *Al-RFL_5179* and *Al-RFL_5184* were designated as orthologs of *AT1G12620* because of their stronger alignment to that *A. thaliana* gene.

Figure 3. *A. thaliana RFL* genes without collinear orthologs in *A. lyrata*. Whole-genome alignment plots for the region surrounding **A.** *AT1G63630* and **B.** *AT1G63150* and **C.** *AT1G62910*, *AT1G62914*, and *AT1G62930*, and the orthologous regions in *A. lyrata*. Tracks from outermost to innermost are small RNA histogram, gene model (yellow), exons (blue), 6-frame PPR motifs (red), transposable elements (green), and aligned residues (black and grey lines). *PPR* genes are labeled above the relevant gene model. Small RNA histograms plot the number of reads in a scrolling window with a 1nt scroll and a 100nt window on both strands with upper and lower bars representing reads found on the Watson and Crick strands, respectively. The maximum number of reads plotted is 50. Size classes of 20-24nt are represented from smallest to largest by the colors turquoise, blue, green, fuchsia, red, and dark red. Aligned residues are connected by black lines if they lie within exonic sequence or grey lines if they are in intronic or intergenic sequence. Alignment height is proportional to similarity on a scrolling average, calculated over a 100nt window with a 1nt scroll. Only regions with an average sequence similarity between 50 and 100 percent are plotted. The arrow glyph in the upper right corner represents the directionality of the lower track. If the arrows point in different directions (as is the case in both plots), the lower tracks are reversed. If the arrows point in the same direction, the lower tracks are not reversed.

Figure 3.


A.
AT1G63150
⇌
50
0
50
100
75
75
100
50
0
50
B.
AT1G63630
⇌
50
0
50
100
75
75
100
50
0
50
C.
AT1G62914
AT1G62910
AT1G62930
⇌
50
0
50
100
75
75
100
50
0
50
Al-RFL2_1417
Al-RFL2_1411
Al-RFL2_1395
Al-RFL2_1415
Al-RFL2_1393

Annotation of *RFL* Pseudogenes

We examined *A. lyrata RFL* genes in greater detail to identify probable pseudogenes and to assess the nature of the local expansions. We identified seven probable pseudogenes in *A. lyrata* by using a combination of comparisons to *A. thaliana* orthologs, number of predicted peptide PPR motifs, gene structure, and 6-frame PPR overlaps with introns. Three of these genes, *Al-RFL2_1559*, *Al-RFL2_1211*, and *Al-RFL2_911* had fewer than seven predicted peptide PPR motifs and had transcript lengths of less than 700nt (Appendix 2). The other four genes, *Al-RFL1_5291*, *Al-RFL2_898*, *Al-RFL1_5179*, and *Al-RFL1_5181*, were all at least 1000nt long and were predicted to have at least 5 peptide PPR motifs. *Al-RFL1_5291* is about half the length of its collinear ortholog, *AT1G12775* and has half the number of predicted peptide PPR motifs. *Al-RFL1_5291* has a predicted intron at the 3' end but high sequence similarity of the intronic sequence to exonic *AT1G12775* sequence. The 6-frame PPR motifs occur continuously across the intron, indicating that the intron may be an attempt by the gene calling programs to avoid a frameshift mutation. There are two PPR motifs predicted on the 6-frame translation just beyond the end of *Al-RFL1_5291*, indicating that premature stop codons and possible deletions have disrupted the 3' end of the gene (Appendix 2).

The region that surrounds *Al-RFL2_911*, *Al-RFL2_894*, and *Al-RFL2_898* is collinear with the region surrounding *A. thaliana* genes *AT1G63320*, *AT1G63330*, and *AT1G63400*, but the whole-genome alignment shows extensive segmental change between the two *PPR*-rich regions, including an ~15kb deletion in *A. lyrata* centered on *Al-RFL2_894* (Figure 4A). The whole-genome alignment split the orthology assignment of *Al-RFL2_894* between *AT1G63330* and *AT1G63400*, although it seems more likely that an insertion or deletion event took place adjacent to *Al-RFL2_894* and that *Al-RFL2_894* is orthologous to only one of the *A. thaliana* genes (Figure 4A). Further analysis revealed that *AT1G63330* is closely related to another *A. thaliana RFL* gene, *AT1G62590*, and that the next gene downstream of *AT1G63330*, which is not a *PPR* gene, is related to the upstream neighbor of *AT1G62590*. Therefore, *AT1G63330* and *AT1G62590* and their respective neighbors were likely part of an inverted duplication, but one that did not include *AT1G63320*.

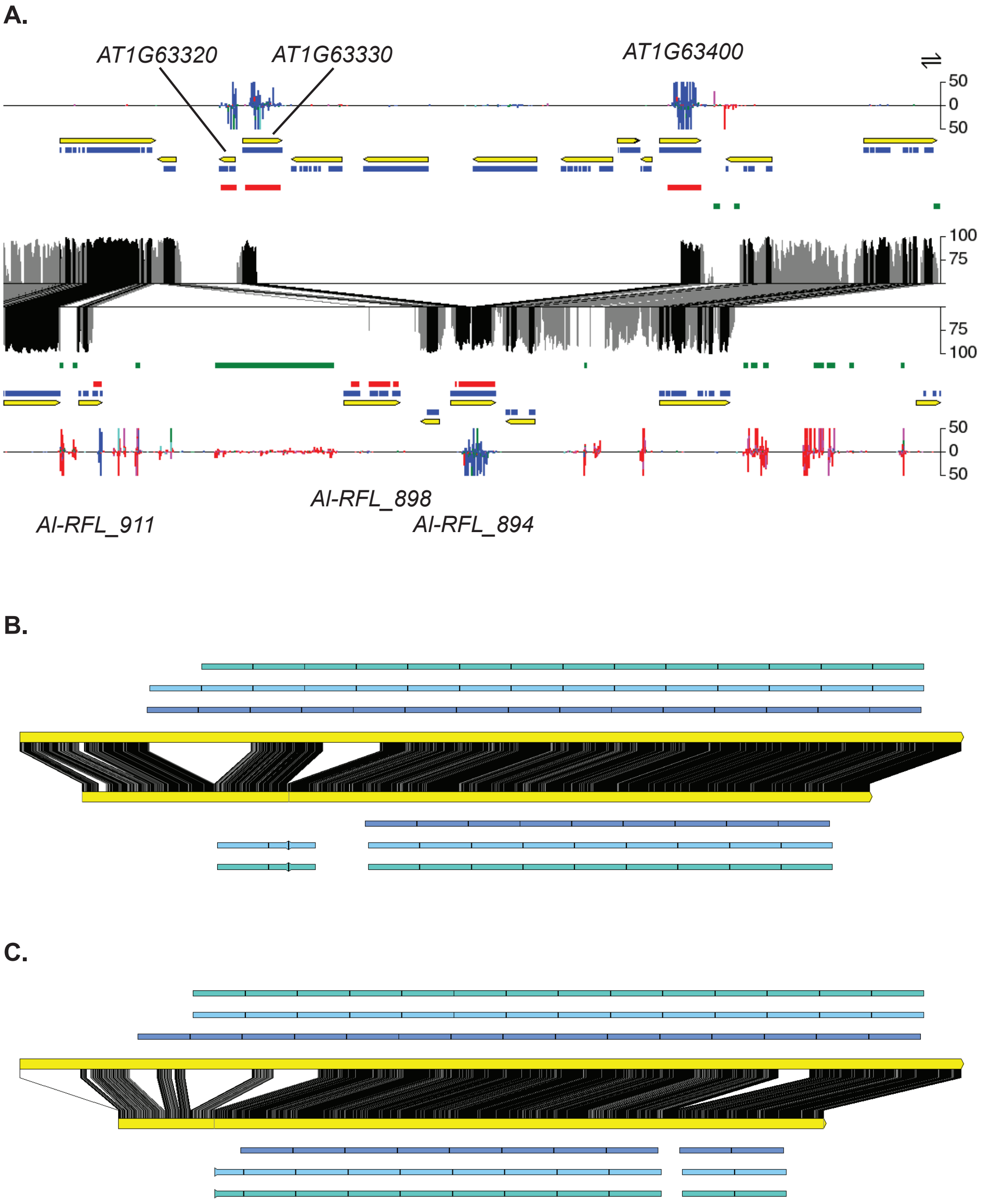


Figure 4. Evaluation of pseudogenes.
A. Whole genome alignment of the *A. thaliana* region between *AT1G63320* and *AT1G63400* with ~15kb flanking sequences on both sides. Features are as described in Figure 3. **B.** and **C.** depict aligned transcripts for *AT1G12300* / *Al-RFL1_5181* and *AT1G62670* / *Al-RFL2_1566*, respectively. Yellow bars represent spliced transcripts with the *A. thaliana* transcript in the upper position. A vertical black line within the yellow transcript box indicates intron splice junction position. Other tracks from outermost to innermost are 6-frame PPR_fs motifs (green), 6-frame PPR_ls motifs (light blue), and peptide PPR motifs (dark blue). 6-frame motifs are mapped to the transcript with exon-intron overlaps indicated by a slight bulge at the boundary. Aligned residues are joined by black (identical) or grey (mismatched) lines.

BLASTP alignments of the Al-RFL2_894 amino acid sequence to other *A. thaliana* PPR sequences did not provide definitive evidence for a particular orthology relationship. Al-RFL2_894 has a roughly 80 percent sequence similarity with nine *A. thaliana* RFL peptide sequences. AT1G63330 and AT1G63400 are marginally worse alignments than the other seven *A. thaliana* sequences, as well as several other non-collinear *A. lyrata RFL* sequences (see Appendix 3 for BLAST results). Given the inconclusive BLAST and genome-wide alignments, and the strong sequence homology between all three genes, we decided to categorize both *AT1G63330* and *AT1G63400* as collinear orthologs of *Al-RFL2_894*, despite the ~15kbp distance between them. It should be noted that *AT1G63150* was also categorized as a non-collinear ortholog of *Al-RFL2_894* due to BLAST similarity and the lack of a collinear ortholog for *AT1G63150*. It is possible that the progenitor of this gene group had multiple lineage-specific expansions or that it existed in high copy number in a common ancestor and experienced independent deletions in each *Arabidopsis* lineage. *Al-RFL2_898* appears to be part of an insertion in *A. lyrata,* and although it is adjacent to *Al-RFL_894*, it aligns most closely to the non-collinear *AT1G64100*. *Al-RFL2_898* has extensive evidence of pseudogenization. Three introns were predicted, likely due to a frameshift mutation and in-frame premature stop codons. A third intron prediction may be the result of the insertion of a low-complexity AT-rich segment. Relatively large gaps between peptide PPRs, which are not normally observed, and several 6-frame PPRs that overlap intron-exon junctions suggest that *Al-RFL2_898* is not a functional *PPR* gene. *AT1G63320* is adjacent to this highly diverged region and appears to have an additional insertion/deletion in the orthologous *A. lyrata* region. There is a pseudogenic *PPR* gene, Al-RFL2_911, at the border of the insertion that we assigned to *AT1G63320* as a collinear ortholog. As stated earlier, this pseudogene is the only collinear ortholog that did not appear in the *RFL* clade (Figure 2).

Neighboring genes *Al-RFL1_5179* and *Al-RFL1_5181* were both considered potential pseudogenes. *Al-RFL1_5179* has an ~1100nt transcript, which is somewhat shorter than other *PPR* transcripts, and only has five PPR motifs, suggesting that it is a pseudogene. *Al-RFL1_5181* appears to have a deletion that eliminated several PPR motifs when compared to its non-collinear ortholog, *AT1G12300* (Figure 4B). In addition, a predicted intron seems likely to be an effort to avoid a premature stop codon. While 6-frame PPR motifs are predicted to traverse this intron, peptide PPR motifs are not predicted to span it. Thus, although there are 12 PPR motifs predicted by the 6-frame translation, only nine are predicted to occur in the peptide sequence. *Al-RFL2_1566* is one of the six collinear orthologs of *AT1G62670* and has a pattern similar to *Al-RFL1_5181*; disrupted PPRs on the 5' end of the peptide sequence and an intron that seems to be predicted to bypass a stop codon (Figure 4C). However, other evidence suggests that *Al-RFL2_1566* may be functional. The lack of similarity between *Al-RFL2_1566* and *AT1G62670* at the 5' end could reflect an incorrectly predicted start codon. A start codon further upstream would eliminate the intron but would still leave a truncated 6-frame PPR motif. In *Al-RFL2_1566*, only the 6-frame PPR at the furthest 5' end of the gene is disrupted, whereas in *Al-RFL1_5181* the disruption is in the second and third 6-frame PPRs from the 5' end. Thus, although the structural characteristics of *Al-RFL1_5181* and *Al-RFL2_1566* are similar, evidence suggested that *Al-RFL1_5181*, but not *Al-RFL2_1566*, was a pseudogene.

In total, we identified 51 *RFL* genes in *A. lyrata*, but because two *A. lyrata* genes had multiple orthologs in *A. thaliana*, 54 ortholog pairs were identified (Figure 5 and Table 1). Only genes in the *A. thaliana* large and small *RFL* clusters on Chromosome 1 were represented by multiple co-orthologs in *A. lyrata*; the *A. thaliana* genes outside of the two clusters had only one ortholog each. The four genes in the small *A. thaliana RFL* cluster had a total of 12 co-orthologs in *A. lyrata* with a minimum of two co-orthologs per *A. thaliana* gene and a maximum of 7 co-orthologs for *AT1G12620*. Within the large *RFL* cluster of 20 *A. thaliana* genes, two had no *A. lyrata* orthologs. The remaining 18 *A. thaliana* genes had 36 orthologs in total, of which 26 were collinear and 10 were non-collinear. Three of the seven single-ortholog genes in the large genomic cluster are at the downstream edge of the cluster. In general, non-clustered *RFL* genes and *RFL* genes at the edge of the clusters were less likely to be duplicated.

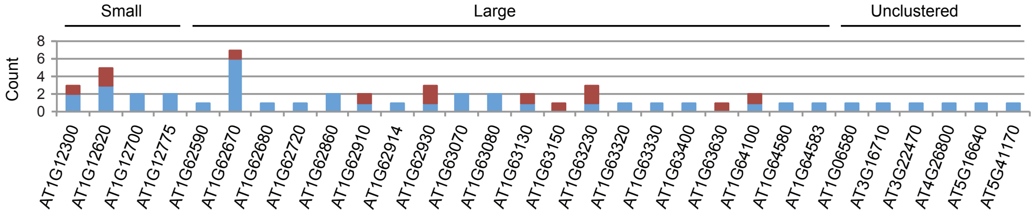


Figure 5. *A. thaliana RFL* co-orthologs found in *A lyrata*.

Number of *A. lyrata* orthologs found for each *A. thaliana RFL* gene. Collinear and non-collinear orthologs are depicted by blue and red bars, respectively. *A. thaliana* genes are grouped by cluster and ordered by position on the genome within each cluster.

Table 1. ***A. lyrata RFL*** **genes.**
RFL genes found in *A. lyrata* are listed in the first column with their *A. thaliana* ortholog(s) in the second column. Column three indicates whether they are collinear orthologs. Column four lists the transcript length. Column five lists which loci are considered pseudogenes. *A. lyrata* gene names in the first column include information on chromosome and start position.

Model Name	*A.t.* Ortholog	Collinear	CDS Length (nt)	Pseudo
Al-RFL1_2385	*AT1G06580*	yes	1,503	no
Al-RFL1_4995	*AT1G12300*	yes	1,911	no
Al-RFL1_4999	*AT1G12300*	yes	2,307	no
Al-RFL1_5181	*AT1G12300*	no	1,602	yes
Al-RFL1_5122	*AT1G12620*	yes	1,866	no
Al-RFL1_5125	*AT1G12620*	yes	1,578	no
Al-RFL1_5127	*AT1G12620*	yes	1,866	no
Al-RFL1_5179	*AT1G12620*	no	1,098	yes
Al-RFL1_5184	*AT1G12620*	no	1,866	no
Al-RFL1_5213	*AT1G12700*	yes	1,920	no
Al-RFL1_5216	*AT1G12700*	yes	1,776	no
Al-RFL1_5291	*AT1G12775*	yes	1,062	yes
Al-RFL1_5296	*AT1G12775*	yes	1,701	no
Al-RFL2_1611	*AT1G62590*	yes	1,902	no
Al-RFL2_1557	*AT1G62670*	yes	1,863	no
Al-RFL2_1559	*AT1G62670*	yes	681	yes
Al-RFL2_1561	*AT1G62670*	yes	1,884	no
Al-RFL2_1566	*AT1G62670*	yes	1,419	no
Al-RFL2_1571	*AT1G62670*	yes	1,857	no
Al-RFL2_1575	*AT1G62670*	yes	1,881	no
Al-RFL2_5036	*AT1G62670*	no	2,172	no
Al-RFL2_1553	*AT1G62680*	yes	1,653	no
Al-RFL2_1521	*AT1G62720*	yes	1,476	no
Al-RFL2_1453	*AT1G62860*	yes	1,974	no
Al-RFL2_1458	*AT1G62860*	yes	2,070	no
Al-RFL2_629	*AT1G62910*	no	1,908	no
Al-RFL2_1417	*AT1G62910*	yes	1,890	no
Al-RFL2_1415	*AT1G62914*	yes	1,884	no
Al-RFL2_1393	*AT1G62930*	no	1,824	no
Al-RFL2_1395	*AT1G62930*	no	1,815	no
Al-RFL2_1411	*AT1G62930*	yes	1,860	no
Al-RFL2_1211	*AT1G63070*	yes	636	yes
Al-RFL2_1215	*AT1G63070*	yes	2,385	no
Al-RFL2_1201	*AT1G63080*	yes	1,872	no
Al-RFL2_1205	*AT1G63080*	yes	1,899	no
Al-RFL2_729	*AT1G63130*	no	1,857	no
Al-RFL2_1159	*AT1G63130*	yes	1,686	no
Al-RFL2_894	*AT1G63330* *AT1G63400* *AT1G63150*	yes yes no	1,887	no
Al-RFL2_700	*AT1G63230* *AT1G63630*	no no	1,509	no
Al-RFL2_1054	*AT1G63230*	yes	1,653	no
Al-RFL2_1082	*AT1G63230*	no	1,662	no
Al-RFL2_911	*AT1G63320*	yes	561	yes
Al-RFL2_464	*AT1G64100*	yes	1,656	no
Al-RFL2_898	*AT1G64100*	no	1,626	yes
Al-RFL2_112	*AT1G64580*	yes	1,365	no
Al-RFL2_109	*AT1G64583*	yes	1,812	no
Al-RFL3_7093	*AT3G16710*	yes	1,513	no
Al-RFL3_10057	*AT3G22470*	yes	1,857	no
Al-RFL7_6500	*AT4G26800*	yes	1,311	no
Al-RFL6_6831	*AT5G16640*	yes	1,515	no
Al-RFL7_21693	*AT5G41170*	yes	1,581	no

Comparison of Structural Changes in *RFL* Genes

Comparisons of gene structure between *A. thaliana* and *A. lyrata* orthologs yielded several interesting results. *AT1G63230*, which Lurin et al. (2004) identified as a probable pseudogene, has one collinear ortholog, *Al-RFL2_1054*, and two non-collinear orthologs, *Al-RFL2_700* and *Al-RFL2_1082*. All three *A. lyrata* orthologs are between 1500 and 1675 bp in length and contain 12 PPR motifs. They also have similar structure, with a conserved 3' intron that intersects the final 3' PPR motif. By contrast, *AT1G63230* is only 972 bp in length with an ~500 bp 5' UTR. The 6-frame PPR analysis placed 2 PPR motifs within the long 5' UTR. It is likely that *AT1G63230* is shortened version of its *A. lyrata* orthologs and that mutations have created several nonsense codons in the 5' end of the gene. These results are consistent with the proposal that *AT1G63230* is a pseudogene.

AT1G62860 was also identified by Lurin et al. (2004) as a probable pseudogene and it has two collinear orthologs in *A. lyrata*. One ortholog, *Al-RFL2_1453*, has a predicted intron that may be the result of a frameshift mutation, but it still retains 15 peptide PPR motifs that are nearly continuous across the peptide sequence, so there was insufficient evidence to consider it a pseudogene. The other ortholog, *Al-RFL2_1458*, has 17 PPR motifs and does not have the 5' intron found in *AT1G62860*. These data suggest that *Al-RFL2_1458* is a functional relative of AT1G62860 and further supports that *AT1G62860* is non-functional.

The structural differences between *AT1G64100* and its single collinear ortholog, *Al-RFL2_464*, are an interesting case. *Al-RFL2_464* has two fewer peptide PPR motifs, primarily from a truncation of the 3' end of the gene and a frameshift mutation resulting in an ~100 bp intron. However, it also has a deletion of approximately 45 bp that eliminates a sequence gap between two PPR motifs in *AT1G64100*. Thus, in spite of the addition of an intron and a small deletion, PPR motifs are continuous along the Al-RFL2_464 peptide sequence (Figure 6A), which is expected for a canonical PPR protein. It is possible that the small deletion was a favorable loss that improved the functionality of the *A. lyrata* gene. It is, of course, also possible that the change occurred in *A. thaliana* and that the gene can be functional without a strict tandem arrangement of the PPR motifs on the translated sequence.

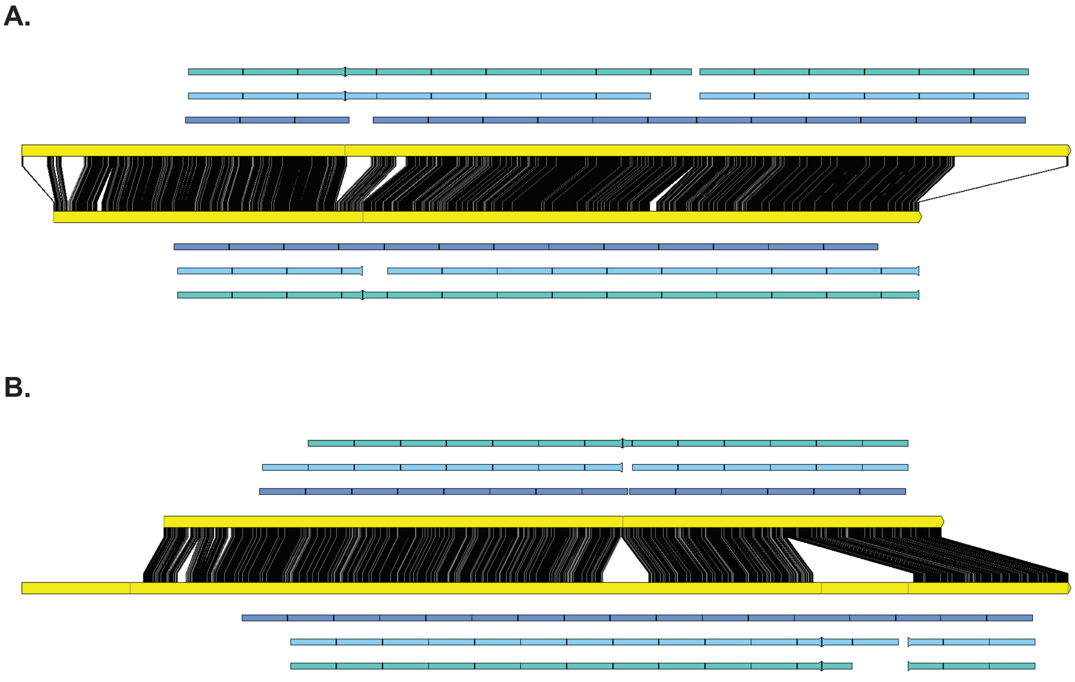


Figure 6. Structural and genomic changes in *PPR*-rich regions.
Features as described in Figure 4. **A.** Alignment plot of *AT1G64100* (top) and *Al-RFL2_464* (bottom). **B.** Alignment plot of *AT1G63070* (top) and *Al-RFL2_1215* (bottom).

AT1G63070 and its collinear ortholog *Al-RFL2_1215* have significant structural differences but both have continuous PPR motifs across their predicted peptide sequence (Figure 6B). *AT1G63070* has a single predicted intron in a non-orthologous location to the two predicted introns in *Al-RFL2_1215* that fall within the PPR encoding region. The intron farthest 3' within *Al-RFL2_1215* is ~2kbp long, which is atypical of *A. thaliana* genes, although it may not be for *A.* lyrata genes. In spite of the structural differences, both genes are otherwise typical *PPR* genes.

Al-RFL2_1415 is predicted to be longer and to have three additional PPR motifs when compared to its ortholog in *A. thaliana, AT1G62914*. The farthest 3' 6-frame

PPR motif in *AT1G62914* overlaps the end of the gene, although there are no predicted motifs beyond the end of the gene. This may indicate that a premature stop codon has pseudogenized *AT1G62914* and that *Al-RFL2_1415* is a functional relative.

Al-RFL2_1395 has three fewer PPR motifs than its ortholog *AT1G62910*, two additional predicted introns, and intron-overlapping 6-frame PPR motifs. One of the introns is the result of a 170bp gap in the *A. lyrata* genome assembly. It is possible that the sequences surrounding the gap may be less accurate, making it difficult to determine the exact nature of the changes that took place at this locus.

Putative *A. lyrata* co-orthologs of *A. thaliana* clade *PPR* genes were identified by both genome position and sequence similarity, but many *A. lyrata* co-orthologs had significant structural variation relative to their *A. thaliana* counterparts. Structural variation included the number predicted introns, the number of predicted peptide PPRs, gene model truncation, and 6-frame PPR motifs that overlap exon-intron boundaries. As we examined the ortholog sets in *A. lyrata*, we found that *A. thaliana* genes with single orthologs were more likely to have conserved structure than genes with multiple co-orthologs.

Comparing the number of introns between orthologs can illustrate structural differences but it also suffers from certain biases. The *A. thaliana* genome is much more thoroughly researched and its annotations have gone through many rounds of fine-tuning. The *A. lyrata* genome is largely machine annotated and we observed that the machine annotations produced long genes with many introns, a pattern that is atypical of most *A. thaliana* genes. It is more likely that these were cases of separate genes concatenated in error, which could inflate the number of introns in *A. lyrata* genes relative to *A. thaliana*. Gene-calling algorithms can add introns to gene models in order to avoid an in-frame stop codon or frameshift insertions/deletions, which also inflates the number of introns, or results in the prediction of a coding region where none exists. To minimize these biases we manually annotated certain genes, and sought to minimize the length of the gene models to maintain consistency with canonical Arabidopsis genes, but this could have the effect of undercounting introns. As an alternate measure of structural change, we counted the number of PPR HMM calls from the 6-frame translation that overlapped non-coding regions of a gene. PPR motifs that overlap introns or UTR regions may be indicators that the machine annotations is attempting to bypass internal stop codons. We compared three measures for each ortholog pair: intron count, the number of PPR motifs found in the translated peptide sequence, and the number of PPRs from the 6-frame genome translation that overlap non-coding sequence.

Thirteen single-ortholog pairs were found whereas 41 pairs are part of multi-gene ortholog groups. The majority of the multi-ortholog groups contain a single *A. thaliana* gene and multiple *A. lyrata* genes, but two *A. lyrata* genes, *Al-RFL2_700* and *Al-RFL2_894*, have multiple co-orthologs in *A. thaliana*. In multi-ortholog pairs, the number of introns found in *A. lyrata* relative to *A. thaliana* ranged from three greater to three fewer, the number of peptide PPR motifs ranged five greater to eleven fewer, and the number of 6-frame PPR motifs overlapping non-coding DNA (introns and UTR regions) ranged from four greater to two fewer (Figure 7A). For single-ortholog pairs, there was far less dispersion in the structural variation metrics (Figure 7A). Single-ortholog pairs had between two greater and one fewer introns, three greater and two fewer peptide PPR motifs, and one greater and one fewer 6-frame PPR motifs overlapping non-coding DNA regions in *A. lyrata* relative to *A. thaliana* (Figure 7A). The 54 ortholog pairs examined include pseudogenes and non-collinear orthologs, both of which are more likely to be diverged from one another. We therefore calculated the same metrics but excluded non-collinear orthologs and any ortholog pair where either gene was considered a pseudogene (hereafter referred to as pseudogenic pairs). The more conservative analysis included 12 single-ortholog pairs and 22 multi-ortholog pairs, and measurements of structural changes in the multi-ortholog group were similar to those in the single-ortholog group (Figure 7B). *A. lyrata* genes in the multi-ortholog group had intron, peptide PPR, and 6-frame PPR overlaps differences ranging from two greater to three fewer, three greater to five fewer, and four greater to one fewer, respectively (Figure 7B). For the single-ortholog group, *A. lyrata* had intron, peptide PPR, and 6-frame PPR overlaps ranging from one greater to one fewer, three greater to one fewer, and one greater to one fewer, respectively (Figure 7B).

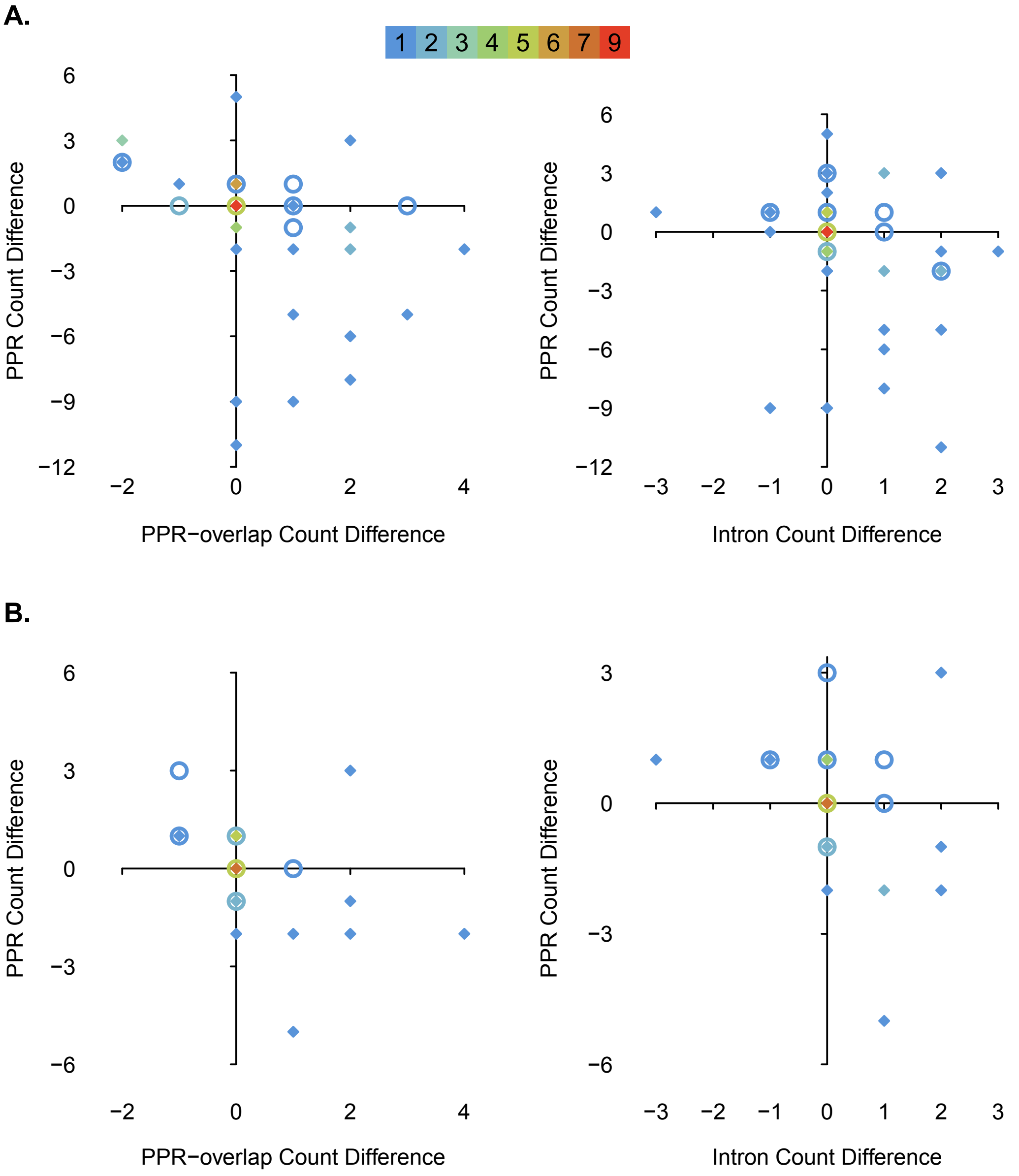


Figure 7. Comparison of gene structure in ortholog pairs.
Comparison of the number of peptide PPR motifs with number of introns (left panels) and number of 6-frame PPR motifs that map to introns or UTR sequence (right panels). Circles and diamonds represent the existence of an ortholog pair with the specified structural differences. The number of ortholog pairs represented by each symbol is color coded according to the key at the top of the figure. Diamond symbols represent pairs from the multi-ortholog group and circles represent pairs from the single-ortholog group. Scales for both vertical and horizontal axes represent the number of structural elements found in *A. lyrata* minus the number found in *A. thaliana*. Thus, positive numbers represent a greater number of elements in *A. lyrata* than *A. thaliana*. **A.** Comparison of all 54 ortholog pairs. **B.** Comparison of only collinear non-pseudogenic pairs.

Physically Distant *RFL* Paralogs Clade Together

Although orthology relationships were determined by collinearity, there were several *A. lyrata RFL* genes whose sequences were more similar to *A. lyrata RFL* genes in other collinear ortholog groups than to their collinear orthologs in *A. thaliana*. These conflicting results were observed in BLAST alignments as well as on certain branches of the gene phylogeny.

As noted above, *A. thaliana RFL* genes are primarily grouped in two genomic clusters on Chromosome 1. Within these genome-scale clusters, individual *RFL* genes are sometimes found adjacent to one another but individual genes, or groups of genes, are separated by 17–200kb of sequence. *A. lyrata RFL* genes are also distributed in two gene clusters but with a higher average gene density at each locus. The most parsimonious explanation for this pattern is that the duplication events that created the widely spaced *RFL* genes took place before the *A. thaliana* and *A. lyrata* lineages diverged. After the lineage split, local duplications or deletions created the observed density differences where single *A. thaliana RFL* genes have multiple *A. lyrata* co-orthologs. If this were the case, one would expect to find collinear orthologs more related to each other than to *RFL* genes in other collinear groups because there was less time to diverge. Yet in several cases we observed that *A. lyrata* genes were more similar to distant paralogs than to their *A. thaliana* orthologs, even though the paralog was a collinear ortholog to a different *A. thaliana* gene. In order to examine this pattern in greater detail, we prepared a phylogeny of the combined non-pseudogenic collinear orthologs from both species (Figure 8). We observed two branches where *A. lyrata* genes tended to be more similar to each other than to their respective collinear orthologs. Within the large genomic cluster of *A. lyrata RFL* genes, there were two clades of *A. lyrata* genes that are separate from a single group of their *A. thaliana* orthologs. One of these *A. lyrata* groups includes five co-orthologs to *AT1G62670* (Figure 8A1), which might be expected if the expansion in *A. lyrata* took place after the two species diverged. However, this same group includes orthologs of *AT1G62590*, *AT1G62910*, and *AT1G62914* that are 36kb, 136kb, and 138kb away, respectively, from the closest ortholog of *AT1G62670* or *AT1G62680*. The co-orthologs of *AT1G62910* and *AT1G62914* might also be explained by lineage-specific expansions, however, these eight genes span ~200kb of sequence and individually have collinear orthologs with which we would expect to them to form phylogenetic groups. The second set of 4 *A. lyrata* genes from the large genomic cluster (Figure 8A2) includes three genes that are adjacent to each other and one gene that is ~300kb away. The corresponding *A. thaliana* genes from the large cluster also form their own distinct clade (Figure 8B). A similar pattern was observed for the small genomic cluster with two clades of *A. lyrata* genes (Figure 8C1 and 8C2) separated from a single clade of four *A. thaliana* genes (Figure 8D). The *A. thaliana* genes span 171kb and the *A. lyrata* genes span 301kb. The separate *A. thaliana* and *A. lyrata* clades are surprising because collinear orthologs are expected to branch together. In contrast, *RFL* ortholog pairs found outside or at the edge of the gene clusters paired together (Figure 8, lower portion). Although recent expansions in the *A. lyrata RFL* genes may confound the use of phylogenies for this purpose, the grouping of paralogs into clades was unexpected.

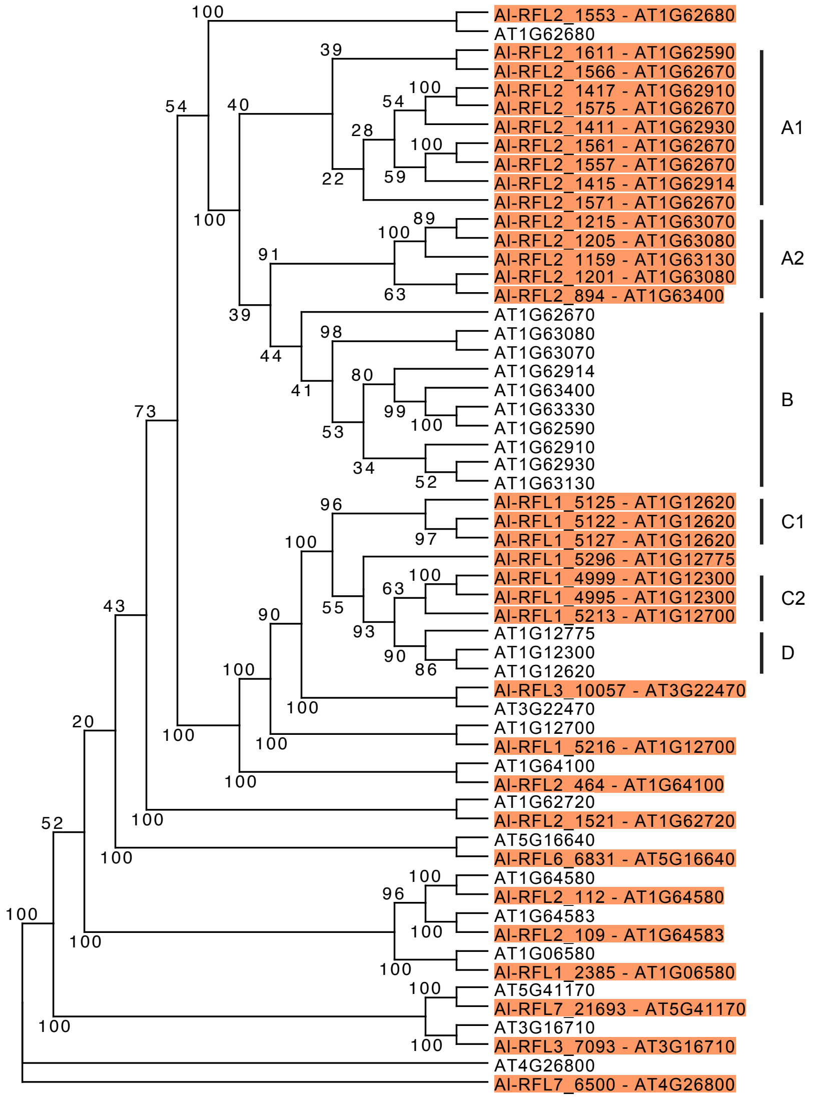


Figure 8. Phylogeny of collinear non-pseudogenic *RFL* transcript sequences. Maximum likelihood tree of collinear non-pseudogenic orthologs. *A. lyrata* genes are highlighted and labeled with the *A. lyrata* gene name followed by the *A. thaliana* ortholog. Bootstrap values appear adjacent to branch points. **A1** and **A2.** Clades of *A. lyrata* genes from the large genomic cluster. **B**. Clade of *A. thaliana* genes from the large genomic cluster. **C1** and **C2**. Clades of A. lyrata genes from the small genomic cluster. **D.** Clade of *A. thaliana* genes from the small genomic cluster.

Presence of a Collinear Ortholog to the Radish *Rf* Gene

A single *PPR* gene (*Rfo*) was shown to rescue two different types of CMS in radish (Brown et al., 2003; Koizuka et al., 2003). The *Rfo* gene is in a cluster of three tandemly arranged *PPR* genes with no collinear orthologs in *A. thaliana*, in spite of the extensive collinearity in the surrounding region (Brown et al., 2003). We sought to determine whether the *A. lyrata* genomic region collinear with the Radish *Rfo* gene contained a *PPR*-encoding gene. The genomic region in *A. thaliana* that is collinear with the Radish *Rfo* locus is between *AT1G63670* and *AT1G63720*, which is found on Chromosome 1 between positions 23,617,767 and 23,635,856 in TAIR10 (Brown et al., 2003). In *A. lyrata*, this *A. thaliana* region maps to an inverted region between positions 714,410 and 752,591 on Scaffold_1, which contains a single *RFL* gene, *Al-RFL2_729*. A BLASTP alignment of the Radish *Rfo* peptide sequence against a combination of *A. thaliana* and *A. lyrata RFL* peptide sequences revealed that Radish *Rfo* is most similar to the *A. thaliana* peptide sequences encoded by *AT1G62860*, *AT1G64100*, and their respective orthologs in *A. lyrata*, with sequence identities between 54% and 66%. Surprisingly, Al-RFL2_729 was not one of the strongest matches with only 46% sequence identity to Radish *Rfo*. Brown et. al (2003) found that the Radish *Rfo* gene is most closely related to three other *RFL* genes in *A. thaliana*, including *AT1G63630* that is 40kb upstream of where the collinear ortholog was expected, which was inconsistent with our BLAST results. This difference was reconciled when the BLAST results were sorted by percent identity rather than by bit score. When sorted by percent identity, AT1G63630 is one of the three *A. thaliana* genes with the highest level of sequence identity to Radish *Rfo*, but over only 257 residues. The peptide sequences discussed above have alignment lengths of at least 500 amino acids (see Appendix 4 for a summary of BLAST results). Although a collinear ortholog to *Rfo* is present in *A. lyrata*, the sequences of the two gene products are highly diverged.

Conservation of *TAS* Genes and Functional tasiRNA

As discussed earlier, small RNA derived from *TAS1* and *TAS2* genes have been found to target *RFL* transcripts, and in some cases, to trigger the production of *PPR*-

derived phased small RNA. Genome-wide alignments revealed that *TAS1b*, *TAS1c*, and *TAS2* are conserved in *A. lyrata* but that *TAS1a* is absent (Figure 9). The regions that flank the *aly-TAS1b* gene and the region orthologous to the *ath-TAS1a* gene are rich in transposable elements and produce large quantities of 24nt small RNA, which are generally associated with heterochromatin (Figure 9A-B). The region surrounding *aly-TAS1c* and *aly-TAS2* is less diverged (they are adjacent to one another) than the flanking regions around *aly-TAS1*b and *aly-TAS1*c (Figure 9C).

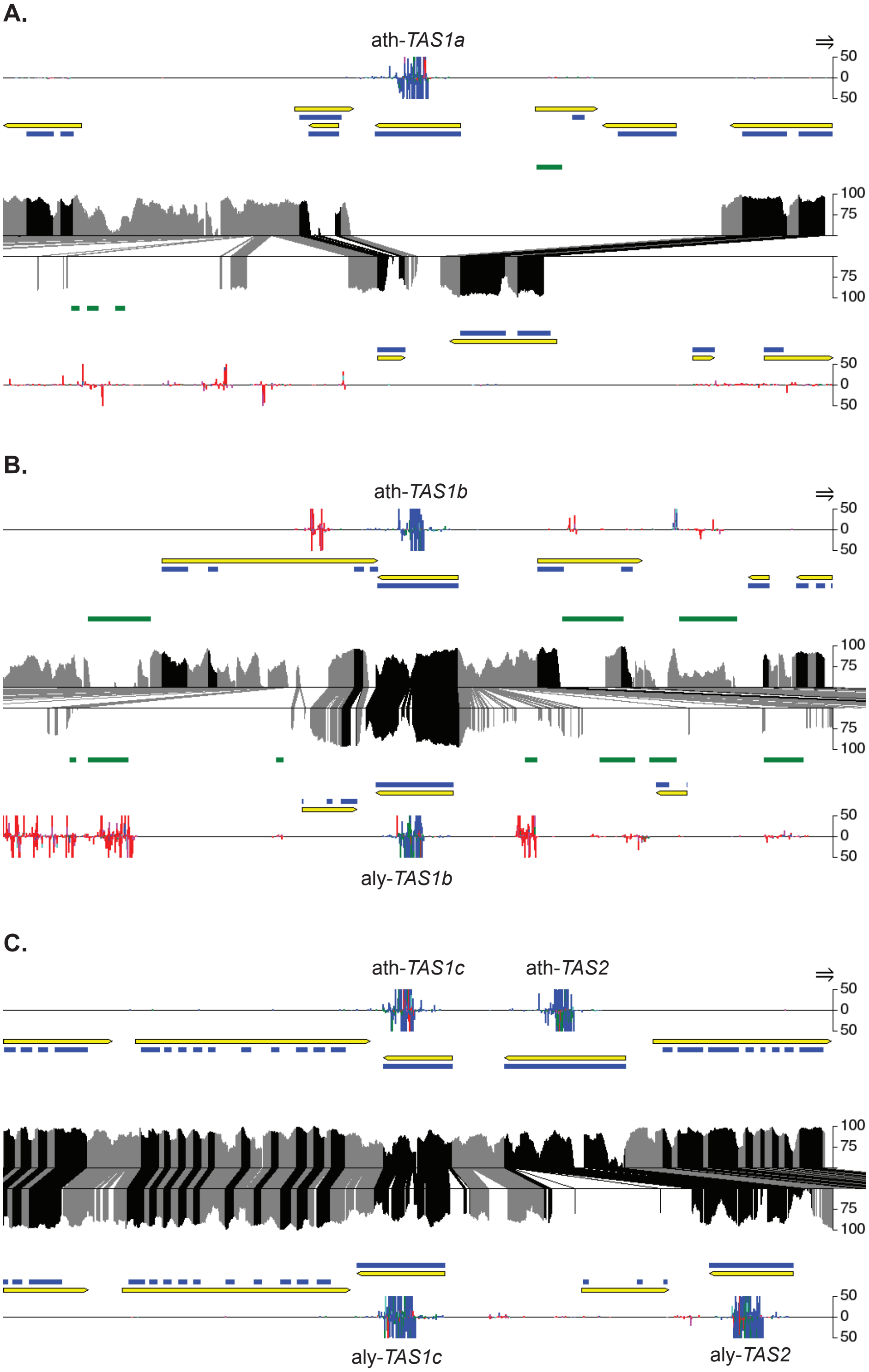


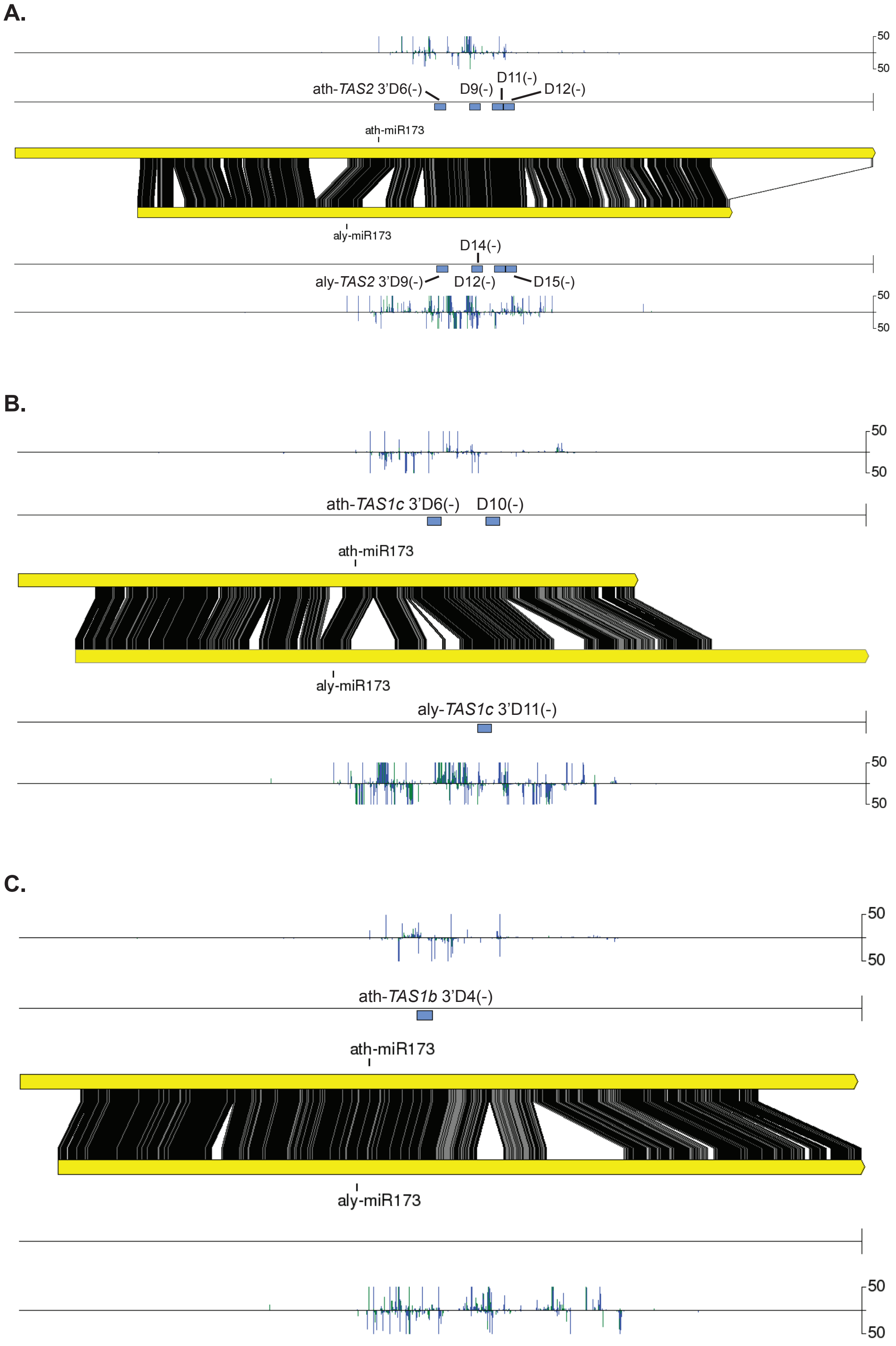
Figure 9. Conservation of *TAS* genes.
Whole-genome alignments are presented as in Figure 3. **A.** *TAS1a*. **B.** *TAS1b*. **C.** *TAS1c* and *TAS2*

TAS1 and *TAS2* transcripts produce tasiRNA (ath-*TAS1a* 3'D9(-), ath-*TAS1b* 3'D4(-), ath-*TAS1c* 3'D6(-), ath-*TAS1c* 3'D10(-), ath-*TAS2* 3'D6(-), ath-*TAS2* 3'D9(-), ath-*TAS2* 3'D11(-), and ath-*TAS2* 3'D12(-)) that are either predicted or validated to target *PPR* transcripts in *A. thaliana*. We searched for similar tasiRNA sequences in *A. lyrata TAS* transcript sequences and in previously sequenced small RNA libraries (Fahlgren et al., 2010) and found only five out of eight small RNA that matched perfectly. All four *PPR*-targeting tasiRNA sequences from *A. thaliana TAS2* are perfectly conserved in *A. lyrata* (Figure 10A). There are two insertions in *A. lyrata* between the miR173-guided cleavage site and the *TAS2* 3'D6(-) ortholog, but these insertions only increase the number of cycles that are traversed before the *TAS2* 3'D6(-) ortholog is reached. Thus, the ath-*TAS2* 3'D6(-) sequence is found in the ninth cycle in *A. lyrata* and the other three tasiRNA sequences are also shifted three cycles forward. These insertions have not disrupted the phase position of the sequences. ath-*TAS2* 3'D6(-) is still in phase with the miR173-guided cleavage site while other three sequences are processed 2nt phase forward, just as they are in *A. thaliana*.

One out of five *TAS1* tasiRNA that target *PPR* transcripts in *A. thaliana* were found perfectly conserved in *A. lyrata*. The ath-*TAS1c* 3'D6(-) sequence is conserved on the aly-*TAS1c* transcript but in cycle 11 rather than cycle 6 because of three insertions between the miR173-guided cleavage site and the *A. lyrata* locus orthologous to ath-*TAS1c* 3'D6(-) (Figure 10B). Ath-*TAS1c* 3'D10(-) can be aligned to an orthologous locus in aly-*TAS1c*, but there are three nucleotide substitutions when the sequences are compared. It is interesting to note that there are three insertions in the aly-*TAS1c* transcript sequence between the miR173-guided cleavage site and the ath-*TAS1c* 3'D10(-) ortholog and that the total number of inserted residues is 105, which is an exact multiple of 21. This is similar to the insertions in *TAS2* and the result is conservation of phase position, even if the sequence itself is not conserved. A closer sequence match to ath-*TAS1c* 3'D10(-) is found on the aly-*TAS1b* transcript at positions 661 to 681 and has only one nucleotide substitution. Although tasiRNA reads corresponding to this exact locus were not found in the sequencing libraries, tasiRNA from adjacent positions also have the potential to target ath-*TAS1c* 3'D10(-) orthologous sites and are found in the sequencing libraries, as will be covered in greater detail below. The closest match to ath-*TAS1b* 3'D4(-) is also found at cycle 4 in aly-*TAS1b* (Figure 10C), but the sequence has three nucleotide substitutions, one of which changes the 5' terminal residue from a U to an A. aly-*TAS1b* 3'D4(-) was sequenced 137 times in four libraries. The closest match to ath-*TAS1a* 3'D9(-) was aly-*TAS1b* 3'D9(-) but with two nucleotide substitutions (not at the 5' start site).

Figure 10. Transcript alignments for conserved *TAS* genes.
Nucleotide alignments of *TAS* gene transcripts. A. *TAS2*. B. *TAS1c*. C. *TAS1b*. Aligned transcripts (horizontal yellow bars joined by black and grey lines) are shown as in Figure 6. miR173-guided cleavage sites are labeled. Validated *PPR*-targeting tasiRNA are shown as blue boxes in a separate track. Outer tracks display the number of 21nt and 22nt small RNA reads sequenced from each position of the transcript, to a maximum of 50. Small RNA sequenced from the positive strand are shown above the horizontal axis while those sequenced from the negative strand are shown below the axis.

Figure 10.


A.
50
50
ath-TAS2 3'D6(-)
D9(-)
D11(-)
D12(-)
ath-miR173
aly-miR173
D14(-)
aly-TAS2 3'D9(-)
D12(-)
D15(-)
50
50
B.
50
50
ath-TAS1c 3'D6(-)
D10(-)
ath-miR173
aly-miR173
aly-TAS1c 3'D11(-)
50
50
C.
50
50
ath-TAS1b 3'D4(-)
ath-miR173
aly-miR173
50
50

Phased tasiRNA Production in *A. lyrata*

In order to compare tasiRNA production in *A. thaliana* and *A. lyrata*, we aligned 21nt small RNA reads derived from wild-type *A. thaliana* and *A. lyrata* to *TAS1* and *TAS2* transcripts, recapitulated the *A. thaliana* phase score measurements, and compared them to phase scores in *A. lyrata*. The use of phase scores to identify phased small RNA was formalized using small RNA libraries from wild-type *A. thaliana* and RDR2-defective mutants (Howell et al., 2007). Each position on a transcript is given a phase score based on the number of reads that align to downstream in-phase positions that are within a particular calculation window. The general formula for calculating the phasing score P is

$$P = \ln\left[\left(1 + \sum_{i=1}^{c} k_i\right)^{n-m}\right], P > 0$$

where c is the number of 21nt cycles in a calculation window, k is the number of reads whose alignment starts at the transcript position specified by the i^{th} cycle, n is the number of occupied in-phase cycles in the scanning window, and m is a minimum threshold for read occupancy, which can be adjusted by the user to make the measure more or less stringent. Reads that align to the negative strand are shifted downstream by two positions to account for the 2nt 3' overhang left by DCL processing. It can be seen that larger values of *c* could lead to greater values of *P* as reads from additional downstream cycles are added. Using larger values of *m* will require a greater number of occupied cycles (*n*) in order for *P* to be greater than 0. For this study we used the values c = 8 and m = 2. Thus, each phase score was calculated using the cumulative number of reads at the starting position and seven downstream positions, each separated by 21nt. Because *m* was set to two, read occupancy at three cycles was required for the value of *P* to be greater than 0. In addition, we required the presence of at least three reads at any position to be included in the occupied cycles count. Phasing peaks that are in the same 21nt register as the original miR173-guided cleavage site will be referred to as phase 0 peaks. If they occur in a cycle that is derived from x positions forward from the miRNA-guided cleavage site, they will be referred to as phase x peaks. Similarly, tasiRNA derived from cleavage-concordant cycles may be referred to as being in phase 0.

In ath-*TAS1b*, raw read peaks in concordance with the miR173-guided cleavage site were the most prominent, although there were also several peaks from non-concordant phases (Figure 11A, *A. thaliana* upper panel). It is interesting to note that Howell et al. (2007) observed only one major raw read peak, which was in phase 0, cycle 7. In the current study, there are two relatively high phase 0 peaks in cycles 6 and 9 as well as a phase 17 peak in cycle 5 that is the maximum for this transcript (Figure 11A, *A. thaliana* upper panel). The phase plots reflect this strong non-concordant phase signal (Figure 11A, *A. thaliana* lower panel). Phase 0 had the highest phase score of 33 but was closely followed by small RNA in phase 17 with a maximum phase score of 32. When small RNA reads from these two phases were summed across all cycles, they accounted for 76 percent of all small RNA reads aligned to ath-*TAS1b*. The highest phase score in aly-*TAS1b* was 38 and occurred in phase 0, however, after cycle 7, secondary peaks in phase 1 were higher than those of phase 0 (Figure 11A, *A. lyrata* lower panel). There were also a large number of reads from phase 2, primarily from the 3' side of the transcript and phase scores from phase 2 reached a maximum of 13. Reads from phases 0, 1, and 2 comprised 65 percent of the total number of reads from aly-*TAS1b*.

In ath-*TAS1c*, the miR173 cleavage-concordant phase signal was initially the strongest, with a maximum peak score of 26 (Figure 11B, *A. thaliana* lower panel). However, phase 0 peaks were quickly eclipsed by phase 1 peaks, which reached a maximum score of 38. The amplitude of the phase shift was similar to that found by Howell et. al (2007), in which the phase 1 to phase 0 RPM-normalized read ratio for cycles 1-5 was 1:5. In cycles 6-10, the same ratio was 22:1. In this study, the ratio of phase 1 to phase 0 RPM-normalized reads was 1:26 in cycles 1-4, before the phase 1 reads exceeded those of the cleavage-concordant peaks. In the next 4 cycles, the ratio was 6:1. While the amplitude of the shift was similar to that found by Howell et al. (2007), it occurred one cycle forward in our study and the higher ratio was found in phase 0 reads rather than phase 1 reads. Summing reads in cycles 1-5 and 6-10, we found that the ratios were 1:7 and 13:1 respectively. In *A. lyrata,* cleavage-concordant phase scores rose to a maximum of 54 and were never exceeded by phase 1 scores, which achieved a maximum of 42 (Figure 11B, *A. lyrata* lower panel). Cycles 1, 4, 5, and 9 from phase 0 had RPM-normalized read counts of greater than 1,000 RPM and the total number of phase 0 reads was 9,095 RPM. The sum of phase1 reads across all cycles was 4,506 RPM, with cycles 10 and 15 having RPM-normalized read counts over 1,000 RPM. Between cycles 1 and 9, the ratio of phase 1 to phase 0 reads was 1:34 while in the subsequent 9 cycles the ratio was 12:1. While phase 2 did not have any cycles with read counts over 1,000 RPM, it did have two phases with read counts of over 500 RPM and total reads for this phase were 2,113 RPM. Reads from phase 0, phase1, and phase 2 accounted for 48 percent of all reads from aly-*TAS1c*.

A. thaliana TAS2 phasing in this study was dominated by cleavage-concordant phase peaks. The two dominant RPM-normalized read peaks were in cycles 3 and 9 containing 5,068 and 3,029 RPM, respectively. Phase 1 reads were somewhat abundant with a total of 1,527 RPM, primarily from phase 9, which produced 839 RPM. No other phase achieved a total read count of more than 300 RPM. The relative abundance of phase 0 reads ensured that phase scores from phase 0 were highest in nearly every cycle, with a maximum phase score of 54 (Figure 11C, *A. thaliana* lower panel). Reads from phase 0 of *A. lyrata TAS2* were also in high abundance but were not as dominant relative to out-of-phase reads as in *A. thaliana*. In *A. lyrata*, cycles 4 and 5 in phase 0 produced 5,386 and 7,797 RPM, respectively. Reads in phase 1 were present in greater numbers starting in cycle 9, with a peak at cycle 10 of 2,756 RPM. The ratio of phase 1 to phase 0 reads in cycles 1-8 was 106:1 and between cycles 9-16 was 4:1. Phase scores for phase 0 peaked at 48 and phase 1 phase scores peaked at 41 (Figure 11C, *A. lyrata* lower panel). The overall portion of reads in phase 0 and phase 1 was 73 percent.

Figure 11. Phased small RNA from *TAS* transcripts.
TAS1b, TAS1c and *TAS2* transcripts are depicted with two plots for each transcript. Horizontal axes represent the length of the entire transcript. The top plot displays the number of reads found at each position of the transcript. Reads are normalized by library size (reads per million reads in the library). The lower plot depicts phase scores for each position on the transcript, plotted at the first position of the eight-cycle window. miR173 target sites are shown in red. Reads and phase scores that are in the same phase as the cleavage site are colored red for 15 cycles after the cleavage site. A 21nt scale bar is in the upper right corner of each pair of plots. **A.** Phased small RNA from *TAS1b*. **B.** Phased small RNA from *TAS1c*. **C.** Phased small RNA from *TAS2*.

Figure 11.

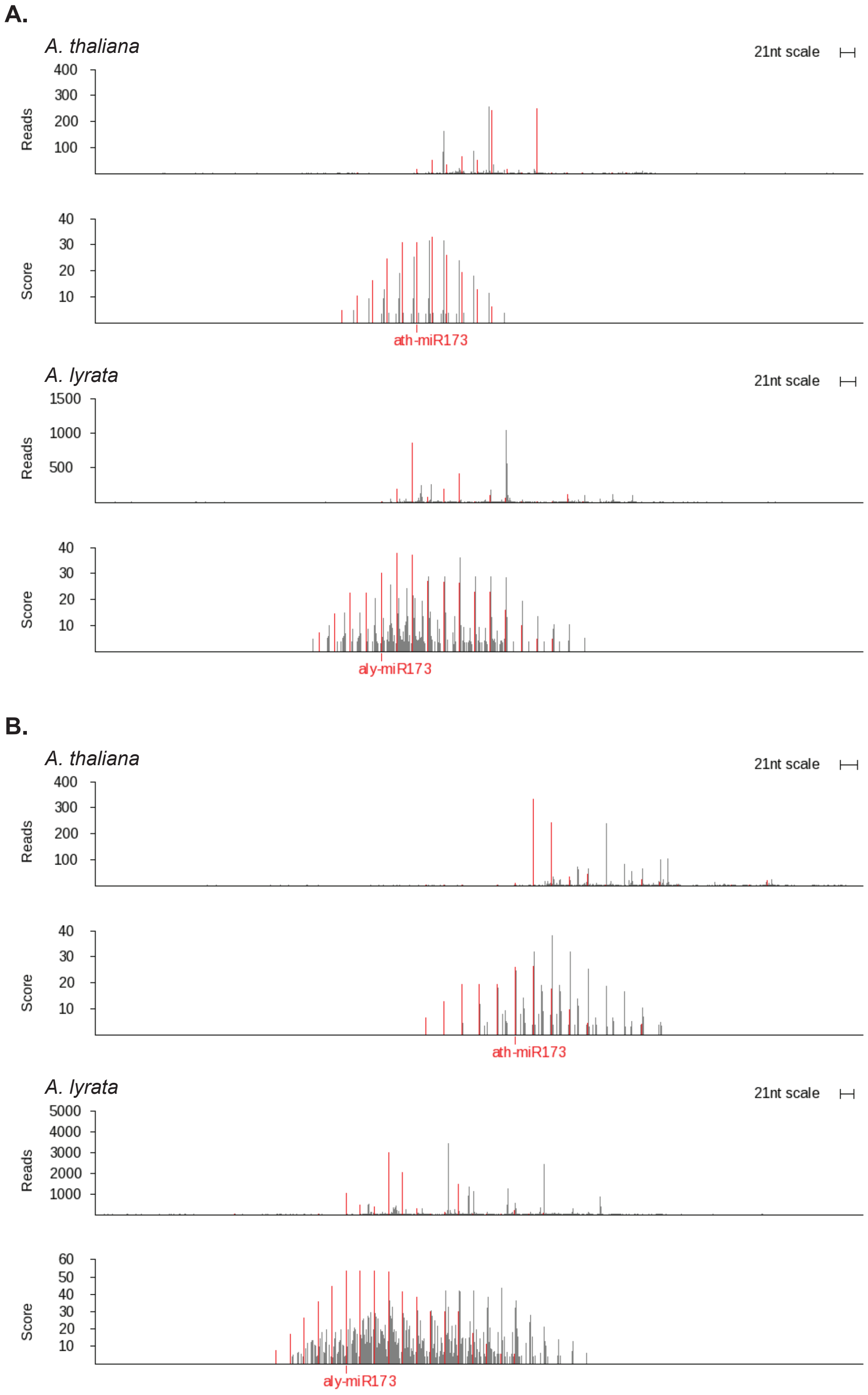
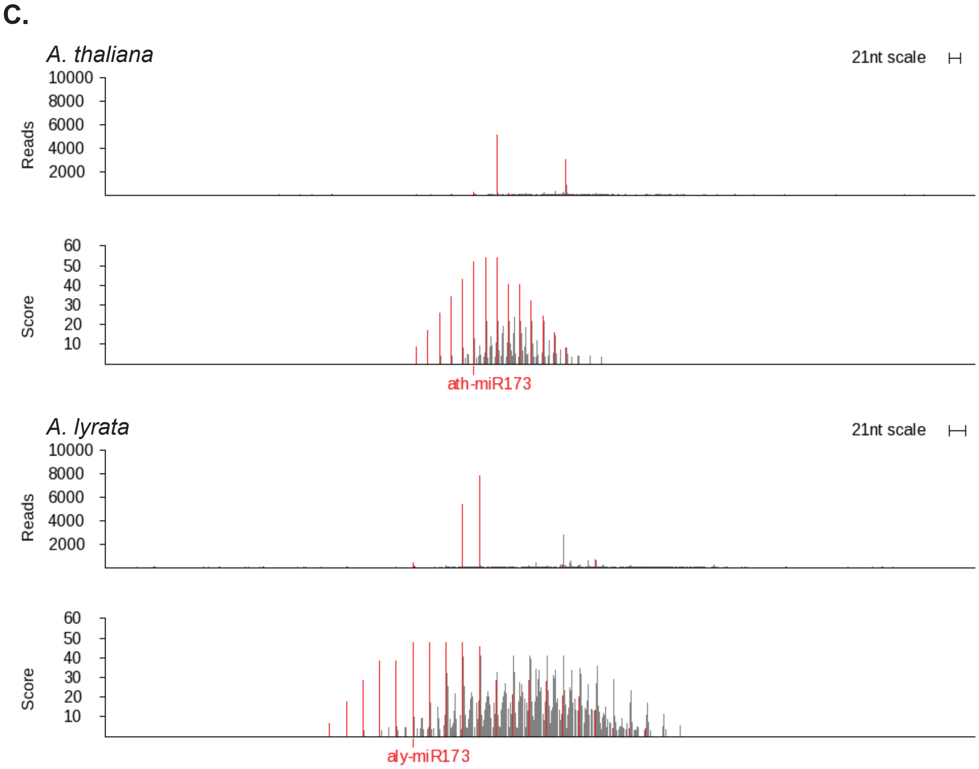

A.
A. thaliana
21nt scale
Reads
400
300
200
100
Score
40
30
20
10
ath-miR173
A. lyrata
21nt scale
Reads
1500
1000
500
Score
40
30
20
10
aly-miR173
B.
A. thaliana
21nt scale
Reads
400
300
200
100
Score
40
30
20
10
ath-miR173
A. lyrata
21nt scale
Reads
5000
4000
3000
2000
1000
Score
60
50
40
30
20
10
aly-miR173

Figure 11 (Continued).



When all phase plots for *A. thaliana* and *A. lyrata* in Figure 11 are compared, the *A. lyrata* phasing signals appear more dispersed in both occupancy and amplitude than the *A. thaliana* phasing signals. This observation is borne out when occupancy and amplitude are examined more closely. The phase signals from ath-*TAS1b* were particularly discrete, with 12 out of 21 phases having a score of 0 (Figure 11A). In contrast, every phase in aly-*TAS1b* had a score, ranging between 3 and 38 (Figure 11A). Similarly, *TAS1c* in *A. thaliana* has 11 phases with no score while in *A. lyrata*, every phase has a score of at least 10 (Figure 11B). In this respect, aly-*TAS2* and ath-*TAS2* are the most similar with 4 and 2 phases having a score of 0, respectively. Not only are a greater number of phases occupied in *A. lyrata*, the non-concordant phase score amplitudes comprise a larger portion of the total phase scores. To quantify this observation, we used Shannon's entropy as a measure of phase score dispersion according to the following formula:

$$-\sum_{i=0}^{20} p_i * log_{21}(p_i)$$

where *p* is the cumulative phase score across all cycles for a particular phase divided by the sum of all phase scores, and *i* is the phase number. Entropy values approaching 1 indicate a highly dispersed set of phase scores while values approaching 0 indicate that phase scores are concentrated in a smaller number of phases. In *A. thaliana*, the entropy values for *TAS1b*, *TAS1c*, and *TAS2* are 0.43, 0.59, and 0.69 respectively. In *A. lyrata*, the values are 0.82, 0.92, and 0.90. The higher entropy values found in *A. lyrata* highlight the noisier tasiRNA processing from *TAS* transcripts.

miRNA Targeting of *RFL* Transcripts

In *A. thaliana*, miR161.1, miR161.2, and miR400 have been experimentally shown to guide cleavage of *RFL* transcripts and in some cases miRNA from more than one *MIRNA* gene will target the same transcript. (Allen et al., 2004; Howell et al., 2007; German et al., 2008; Addo-Quaye et al., 2008). We sought to determine how many of the miRNA-target relationships were conserved between *A. lyrata* and *A. thaliana* and the frequency of target formation and loss. We used the TargetFinder program (Fahlgren et al., 2007; available at http://carringtonlab.org/resources/targetfinder) to evaluate the ability of the miRNA sequences to target *PPR* transcripts. TargetFinder aligns small RNA to potential target transcripts and assigns a target score to the alignment based on factors that are known to impact targeting. A score of zero indicates a perfect alignment while scores of greater than four indicate an unlikely, though not impossible, target relationship. For this study, a predicted target is one with a score of four or less. For each *RFL* ortholog pair, we compared the presence or absence of predicted miRNA targeting against the transcript sequences and compared the TargetFinder scores when both orthologs were predicted targets.

We analyzed each pair of orthologs individually, separating any pairs that included a pseudogene, hereafter referred to as pseudogenic pairs. We first looked for complete gain or loss of miRNA targeting in either species. In *A. thaliana*, none of the four *RFL* genes in the small genomic cluster on Chromosome 1 are targeted by miRNA. In *A. lyrata*, 7 out of 9 small cluster orthologs are targeted by at least one of the three miRNA with a score of 3 or 4. Only *Al-RFL1_5213* and *Al-RFL1_5296*, orthologs of *AT1G12700* and *AT1G12775*, respectively, are not predicted miRNA targets. *Al-RFL1_4999*, a collinear co-ortholog of *AT1G12300*, and *Al-RFL1_5184*, a non-collinear co-ortholog of *AT1G12620* were each predicted targets of two different miRNA unlike their orthologs, which are not predicted miRNA targets (Figure 12A and 12B). *AT1G64100* and its ortholog *Al-RFL2_464* (Figure 12C) were the only other case where an *A. lyrata RFL* transcript was predicted to be a miRNA target when its *A. thaliana* ortholog was not. Conversely, *Al-RFL2_5036* and *Al-RFL2_109* are not predicted miRNA targets while their *A. thaliana* orthologs are. In both of these cases, the target scores in *A. thaliana* were three or greater.

Because *RFL* transcripts are often targeted by more than one miRNA, there were many individual miRNA targeting gain/loss events between the two species that did not result in a complete loss of miRNA targeting. There were 18 instances of *A. thaliana* miRNA-target pairs with no corresponding target in *A. lyrata*, and only two of these had target scores of less than 3. The first of these is the ortholog pair *AT1G63080* and *Al-RFL2_1201* where *AT1G63080*, is predicted to be targeted by miR400 and miR161.2 with scores of 4 and miR161.1 with a score of 2.5 (Figure 12D). Only miR400 is predicted to target *Al-RFL2_1201* but with a strong score of 1 (Figure 12D). It is interesting that both orthologs have at least one strong target score but from different miRNA. The second ortholog pair is *AT1G62590* and *Al-RFL2_1611* (Figure 12E). Both are predicted targets of miR400 (scores of 2 and 1, respectively) but are predicted targets of different miR161 variants. The target prediction score of miR161.2 on *AT1G62590* is 1 and score for miR161.1 on *Al-RFL2_1611* is 2.5.

Figure 12. miRNA-guided cleavage sites on *RFL* transcripts.
Pairwise ortholog transcript alignments for **A.** *AT1G12300* / *Al-RFL1_4999*. **B.** *AT1G12620* / *Al-RFL1_5184.* **C.** *AT1G64100* / *Al-RFL2_464*. **D.** *AT1G63080* / *Al-RFL2_1201*. **E.** *AT1G62590* / *Al-RFL2_1611*. Alignment and PPR motif features shown as in Figure 4. Three additional feature tracks are shown. The outermost track for each transcript displays a per-position histogram of 21nt and 22nt small RNA reads with perfect alignments to the transcript. Histogram bars above the horizontal axis indicate alignments to the sense strand of the transcript while bars below the axis indicate alignment to the antisense strand. The maximum number of reads shown is 50. Innermost two tracks show small RNA-guided cleavage sites. The innermost track indicates miRNA- and tasiRNA-guided cleavage positions with labels and vertical lines, but only for those small RNA that have been validated to target *RFL* transcripts in *A. thaliana*. The outer of the two tracks depicts each predicted tasiRNA-guided cleavage site as a vertical line.

Figure 12.

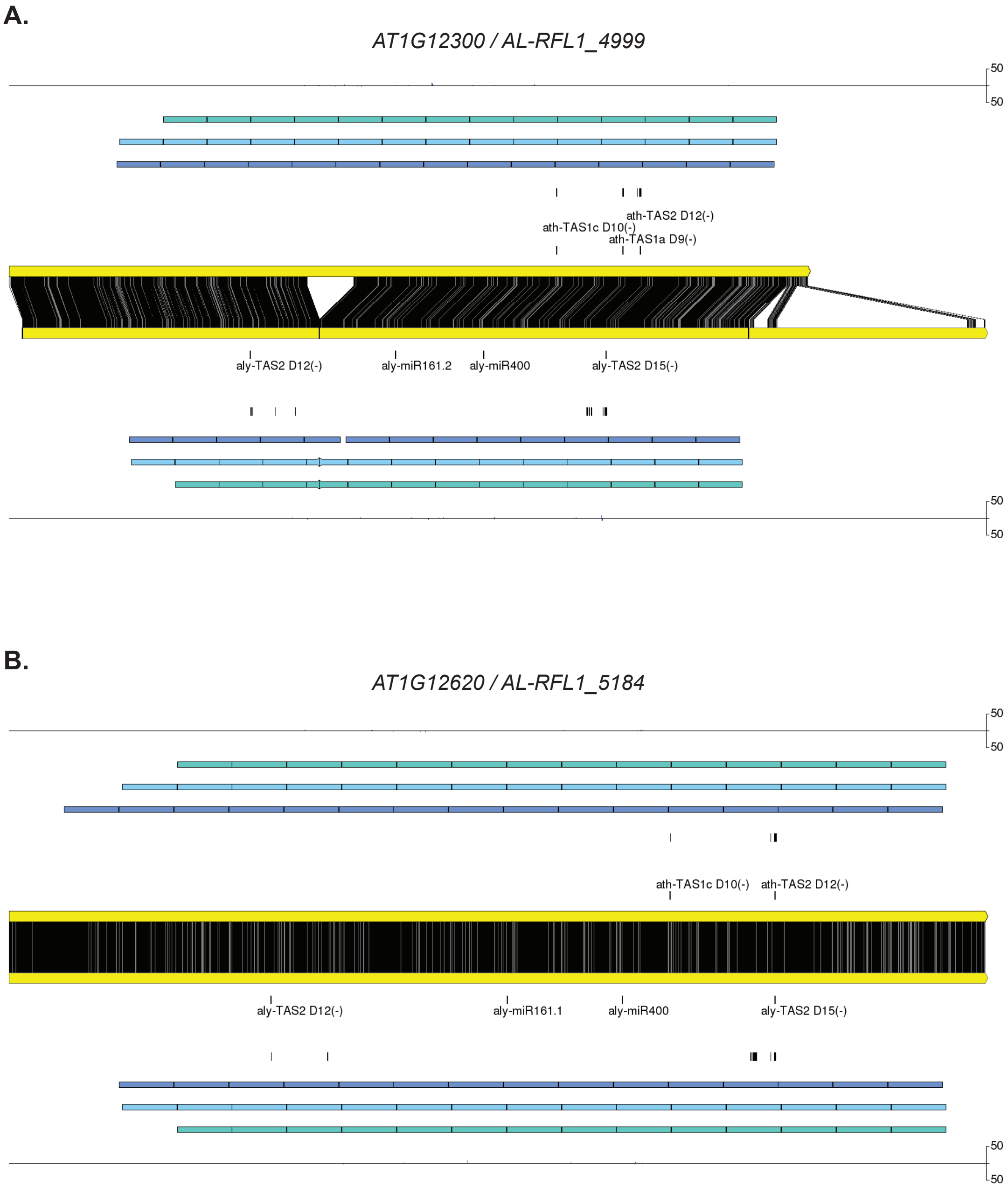

A.
AT1G12300 / AL-RFL1_4999
50
50
ath-TAS2 D12(-)
ath-TAS1c D10(-)
ath-TAS1a D9(-)
aly-TAS2 D12(-)
aly-miR161.2
aly-miR400
aly-TAS2 D15(-)
50
50
B.
AT1G12620 / AL-RFL1_5184
50
50
ath-TAS1c D10(-)
ath-TAS2 D12(-)
aly-TAS2 D12(-)
aly-miR161.1
aly-miR400
aly-TAS2 D15(-)
50
50

Figure 12 (Continued).

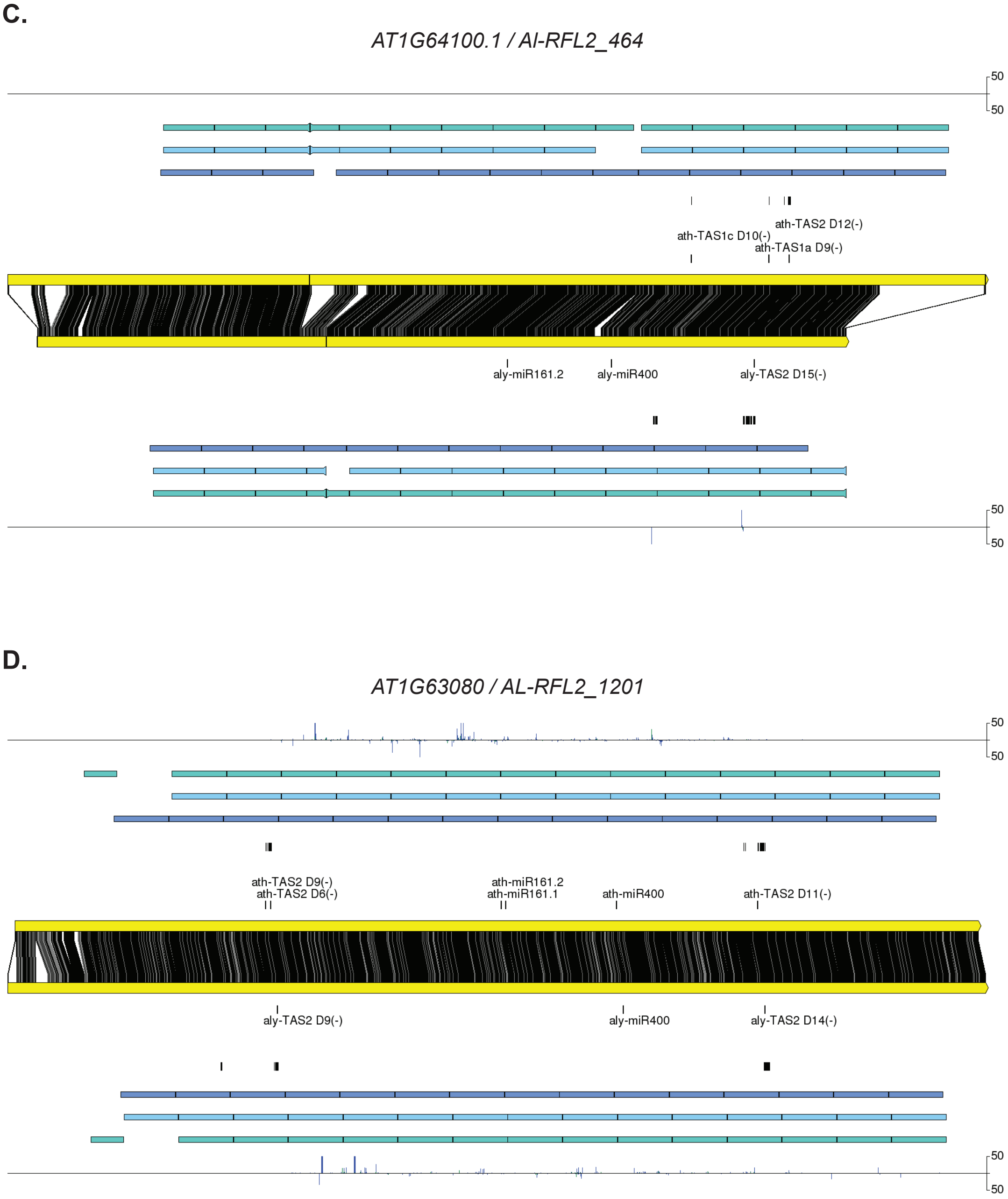
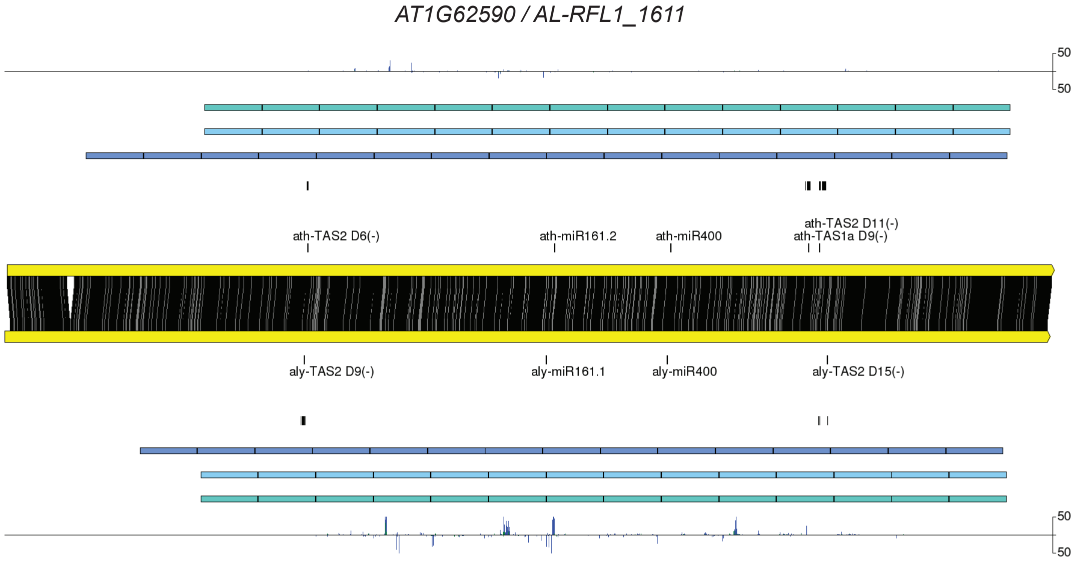


Figure 12 (Continued).

E.



We found 15 *A. lyrata* transcripts targeted by miRNA where the corresponding targeting was absent in *A. thaliana*, five of which had a target score of less than three. Two ortholog pairs, *AT1G62590* / *Al-RFL2_1611* and *AT1G64100* / *Al-RFL2_464*, were discussed above. Two cases involved the same gene, *Al-RFL2_894*, which is collinear with *AT1G63330* and *AT1G63400*, and is the non-collinear ortholog of *AT1G63150*. *Al-RFL2_894* is targeted by all three miRNA with scores of 2 or 2.5. *AT1G63150* is not a predicted target of miR400 but is targeted by miR161.1 and miR161.2, with target scores of 4 and 1.5, respectively. Similarly, *AT1G63330* is not targeted by miR161.1 but is targeted by miR400 and miR161.2 with target scores of 3 and 1, respectively. *AT1G63400* is targeted by all three miRNA. The fifth *A. lyrata* target is *Al-RFL2_1215* is a predicted target of miR161.1 with a target score of 2.5, while its ortholog *AT1G63070* is not a predicted miR161.1 target.

It is interesting that *RFL* transcripts from the six non-clustered genes had few changes in targeting. Four of the six had identical targets and target scores and one had a 0.5 shift in target score. Only *AT1G06580* / *Al-RFL1_2385* had significant differences, with the loss of miR161.2 targeting in *A. lyrata* and a slight change in the miR400 target score from 0 in *A. thaliana* to a 1 in *A. lyrata*. In total, only 4 of the 12 predicted target scores changed for the non-clustered ortholog pairs. Out of a total of 91 miRNA target loci, 33 had a gain or loss in predicted miRNA targeting. Out of 58 cases of conserved miRNA targeting, 37 target scores changed, though the vast majority were small changes in value. Target scores changed by a value of more than 2 in only three ortholog pairs. One of these ortholog pairs, *AT1G63080* / *Al-RFL2_1201* was discussed above. The second ortholog pair, *Al-RFL2_1575* is targeted by miR161.2 with a score of 0 while its ortholog, *AT1G62670*, is targeted with a score of 4. The third ortholog pair had a nearly identical pattern; *Al-RFL2_1417* has a miR161.2 score of 0 and *AT1G62910* has a score of 4.

Ortholog pairs that included putative pseudogenes were also analyzed for target gain/loss events. There were several gains / losses of miRNA targeting but when targeting was conserved, the change in target score was never greater than 0.5. *A. lyrata* genes in the small genomic *RFL* cluster that are putative pseudogenes were generally not predicted targets of miRNA, All three *A. lyrata* genes appear to be truncated copies of their *A. thaliana* counterparts, which may explain their lack of targets. Only miR400 was predicted to target *Al-RFL1_5181* in the small cluster, but with a high score of 4. From the large genomic cluster, miRNA targeting was completely conserved in thee out of nine ortholog pairs: *AT1G62670* / *Al-RFL2_1559*, *AT1G63070* / *Al-RFL2_1211* and *AT1G63230* / *Al-RFL2_1054*. The pseudogene *AT1G62860* has two collinear co-orthologs, *Al-RFL2_1453* and *Al-RFL2_1458*. Both *A. lyrata* transcripts contain miR161.2 targets whereas *AT1G62860* contains only a miR161.1 target site. Three non-collinear pseudogenic ortholog pairs had a loss of miR161.1 targeting in *A. lyrata* while the remaining two pseudogenic pairs have no predicted miRNA target sites. Overall, within the non-pseudogenic group, 64% of target sites were conserved (58 out of 91 target sites) whereas among the pseudogenic pairs, only 44% of target sites were conserved (8 out of 18). Figure 13 displays conservation of miRNA targeting for each ortholog pair.

A.

A. thaliana gene	*A. lyrata* gene	miR400		miR161.1		miR161.2	
		At	*Al*	*At*	*Al*	*At*	*Al*
AT1G06580	*Al-RFL1_2385*	0	1	3.5	3.5	4	
AT1G12300	*Al-RFL1_4995*						3
AT1G12300	*Al-RFL1_4999*		3				4
AT1G12620	*Al-RFL1_5122*				4		
AT1G12620	*Al-RFL1_5125*		4				
AT1G12620	*Al-RFL1_5127*		4				
AT1G12700	*Al-RFL1_5213*						
AT1G12700	*Al-RFL1_5216*						3
AT1G12775	*Al-RFL1_5296*						
AT1G62590	*Al-RFL2_1611*	2	1		2.5	1	
AT1G62670	*Al-RFL2_1557*	3	2	3	2.5	4	4
AT1G62670	*Al-RFL2_1561*	3	2	3	2.5	4	4
AT1G62670	*Al-RFL2_1566*	3	4	3	4	4	4
AT1G62670	*Al-RFL2_1571*	3		3	2.5	4	2
AT1G62670	*Al-RFL2_1575*	3	3	3	2.5	4	0
AT1G62680	*Al-RFL2_1553*						
AT1G62720	*Al-RFL2_1521*	1	1				
AT1G62910	*Al-RFL2_1417*	4	3	3	2.5	3	0
AT1G62914	*Al-RFL2_1415*	4	2	4	2.5	3	2
AT1G62930	*Al-RFL2_1411*	3	3	3	2.5	3	
AT1G63070	*Al-RFL2_1215*	3.5	3		2.5	3.5	
AT1G63080	*Al-RFL2_1201*	4	1	2.5		4	
AT1G63080	*Al-RFL2_1205*	4	3	2.5	2	4	3.5
AT1G63130	*Al-RFL2_1159*	3	3	3	3	3	3.5
AT1G63330	*Al-RFL2_894*	3	2		2.5	1	2
AT1G63400	*Al-RFL2_894*	4	2	2.5	2.5	2	2
AT1G64100	*Al-RFL2_464*		3				2
AT1G64580	*Al-RFL2_112*	3	3				
AT1G64583	*Al-RFL2_109*	4		3.5		4	
AT3G16710	*Al-RFL3_7093*	2.5	2			4	
AT3G22470	*Al-RFL3_10057*	2	2				
AT4G26800	*Al-RFL7_6500*					4	4
AT5G16640	*Al-RFL6_6831*	4	4	3.5	3.5	2.5	2.5
AT5G41170	*Al-RFL7_21693*			3.5	3.5	1	1
AT1G12620	*Al-RFL1_5184*		4		4		
AT1G62670	*Al-RFL2_5036*	3		3		4	
AT1G62910	*Al-RFL2_629*	4	3	3		3	1.5
AT1G62930	*Al-RFL2_1393*	3		3	4	3	
AT1G62930	*Al-RFL2_1395*	3	3	3	2.5	3	3.5
AT1G63130	*Al-RFL2_729*	3	3	3		3	1.5
AT1G63150	*Al-RFL2_894*		2	4	2.5	1.5	2
AT1G12775	*Al-RFL1_5291*						
AT1G62670	*Al-RFL2_1559*	3	3	3	2.5	4	3.5
AT1G62860	*Al-RFL2_1453*			4			2
AT1G62860	*Al-RFL2_1458*			4			2
AT1G63070	*Al-RFL2_1211*	3.5				3.5	
AT1G63230	*Al-RFL2_1054*			3	3.5	3	3
AT1G63320	*Al-RFL2_911*						
AT1G64100	*Al-RFL2_898*						
AT1G12300	*Al-RFL1_5181*		4				
AT1G12620	*Al-RFL1_5179*						
AT1G63230	*Al-RFL2_1082*			3		3	3
AT1G63230	*Al-RFL2_700*			3		3	3
AT1G63630	*Al-RFL2_700*			3		2.5	3

B.

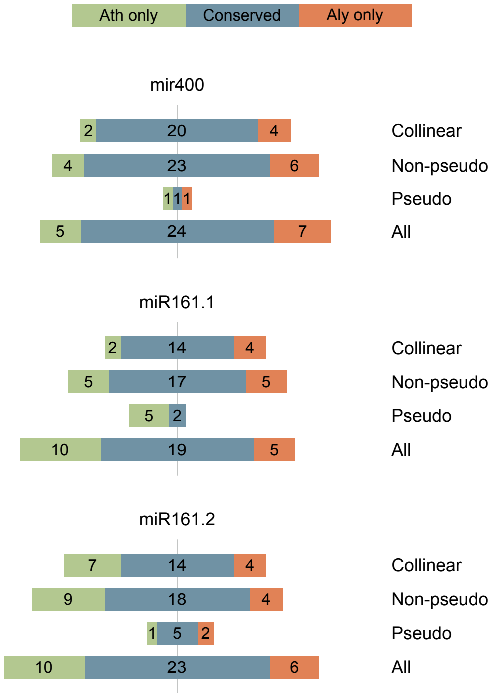


Figure 13. miRNA target scores and conservation.
A. The first and second columns are *A. thaliana* and *A. lyrata* gene models, respectively. Collinear orthologs are shown with red text, non-collinear with black text. Non-pseudogenes are shaded grey. miRNA are listed at the top with targets scores in sub-columns *At* for *A. thaliana* and *Al* for *A. lyrata*. Scores are shown in the body of the table. If more than one target exists on a transcript, the best score is listed. **B.** Number of conserved and unique miRNA target sties from pairwise comparisons of gene orthologs. Counts are summarized at 4 levels: collinear pairs, non-collinear and collinear pairs (i.e. non-pseudogenic pairs), pseudogenic pairs, and all pairwise comparisons. These are specified as Collinear, Non-pseudo, Pseudo, and All, respectively.

TasiRNA Targeting of *RFL* Transcripts

As a first step in examining tasiRNA targeting of *PPR* transcripts, we identified small RNA sequences that aligned perfectly to *TAS* transcripts separately for each species and retained only those that were 21nt or 22nt in length and were represented by at least two reads. We used the TargetFinder program to identify potential *PPR* targets of these selected tasiRNA. In doing so, we observed that tasiRNA target predictions appear in clusters along the target transcript. Prior studies that validated *PPR* targets found evidence for more than one cleavage site on certain transcripts (Yoshikawa et al., 2005; Allen et al., 2004). Because tasiRNA are processed imperfectly by DCL4, targeting of an *RFL* transcript by one tasiRNA often appears to be accompanied by potential, possibly weak, targeting by other tasiRNA that originate from adjacent start positions on the same *TAS* transcript. In examining the conservation of tasiRNA targeting, we evaluated not only the eight tasiRNA that were previously validated to target *PPR* transcripts (hereafter referred to as validated tasiRNA), but also the presence or absence of tasiRNA target clusters, which could indicate maintenance of targeting even in the absence of validated tasiRNA targets. Because many *PPR* transcripts are predicted targets of the same tasiRNA at more than one locus, each target locus was analyzed separately. The term orthologous target locus (OTL) will be used to indicate a tasiRNA target cluster and its orthologous locus on an orthologous transcript. For consistency, a single tasiRNA target site is also considered a cluster. For example, the first cluster of tasiRNA targets on *AT1G62910* has a single ath-*TAS2* 3'D6(-) target site. *AT1G62910* has two co-orthologs in *A. lyrata*, but only one of these has a cluster of tasiRNA targets at the orthologous locus that contains an aly-*TAS2* 3'D9(-) target site (aly-*TAS2* 3'D9(-) is the ath-*TAS2* 3'D6(-) orthologous RNA). In this example there are two OTLs (one for each ortholog pair), but targeting is conserved in only one. When non-validated tasiRNAs are discussed, the tasiRNA sequence will be identified in relation to the *TAS* locus from which it is derived and will include species (ath and aly for *A. thaliana* and *A. lyrata*, respectively), *TAS* transcript name, start position when aligned to the *TAS* transcript, strand, and length. For instance, ath-*TAS2*_776_-_22 is a 22nt tasiRNA derived from the negative strand of *TAS2* starting at position 776. Read counts provided in this section are raw read counts.

We first examined targets of ath-*TAS2* 3'D6(-), which is the primary tasiRNA involved in triggering secondary siRNA production from *RFL* transcripts, and its ortholog aly-*TAS2* 3'D9(-). There are 24 predicted OTLs between the two species, only two of which are associated with pseudogenes. Within the non-pseudogenic pairs, 11 OTLs have conserved targeting, three OTLs have ath-*TAS2* 3'D6(-)-specific targeting, and 8 OTLs have aly-*TAS2* 3'D9(-)-specific targeting. In the 11 OTLs with conserved targeting, ath-*TAS2* 3'D6(-) / aly-*TAS2* 3'D9(-) is often the tasiRNA with the lowest score and highest abundance within the cluster. Only *AT1G63130* is targeted by a more abundant tasiRNA with a lower targeting score, and this was another validated tasiRNA, ath-*TAS2* 3'D9(-). Of the five OTLs where ath-*TAS2* 3'D6(-) targeting is unique to *A. thaliana*, two ortholog pairs, *AT1G12775 / Al-RFL1_5291* (a pseudogene) and *AT1G12775 / Al-RFL1_5296* (Figure 14A), have corresponding tasiRNA target clusters but no aly-*TAS2* 3'D9(-) target. Their lowest target scores of 3.5 and 4 are high but similar to the score of 3 that is found in *A. thaliana.* Ath-*TAS2* 3'D6(-) and ath-*TAS2* 3'D9(-) target within the same cluster on *AT1G12775* and are represented by 129 and 90 reads, respectively. Although ath-*TAS2* 3'D9(-) has the lowest target score (3) within its target cluster, it is not the tasiRNA with the highest read count. ath-*TAS2*_842_-_22 is processed one nucleotide upstream of ath-*TAS2* 3'D9(-) and is the most abundant tasiRNA targeting this cluster with 801 reads. Because it is 22nt in length, it is predicted to cleave *AT1G12775* at the same site as ath-*TAS2* 3'D9(-). The orthologous tasiRNA cluster in *A. lyrata* contains only two target sites, one of which is targeted by aly-*TAS2* 3'D12(-), which has both the highest read count (37) and lowest score (4) within this small cluster of targets. A summary of ath-*TAS2* 3'D6(-) / aly-*TAS2* 3'D9(-) targeting can be found in Figures 15A and 15B.

Ath-*TAS2* 3'D9(-) / aly-*TAS2* 3'D12(-) targeting is present in 27 OTLs, 9 of which are on pseudogenic pairs, a much higher proportion than ath-*TAS2* 3'D6(-) / aly-*TAS2* 3'D9(-) (Figures 15A and 15B). Only 2 of the18 OTLs on non-pseudogenic pairs contained conserved targets in both *RFL* orthologs. *AT1G12775 / Al-RFL1_5296* targeting is one of the two conserved loci and was discussed above. Targeting was also conserved in *AT3G22470* / *Al-RFL3_10057*. ath-*TAS2* 3'D9(-) has the lowest target score (2) within the target cluster on *AT3G22470*, but ath-*TAS2*_842_-_22 is also predicted to target this cluster with a higher score of 2.5. On *Al-RFL3_10057*, aly-*TAS2* 3'D12(-) has the lowest target score of 2. aly-*TAS1c*_458_-_21 also targets this cluster and has the highest read count (1132) but a high target score of 4. Five of 18 non-pseudogenic OTLs contained targets specific to *A. thaliana.* Four of these five contained tasiRNA target clusters in *A. lyrata*. Only one out of eleven *A. lyrata* specific OTLs contain corresponding tasiRNA target clusters in *A. thaliana*, a lower proportion than the *A. thaliana*-specific clusters. This lone OTL on *AT5G16640* (Figure 14B) is targeted only by ath-*TAS2* 3'D6(-), whereas the *A. lyrata* target cluster contains both an aly-*TAS2* 3'D9(-) target and an aly-*TAS2* 3'D12(-) target. The remaining 10 OTLs with *A. lyrata*-specific targeting are all in the small genomic cluster of *RFL* genes and represent four *A. thaliana* genes and eight *A. lyrata* genes. One *A. lyrata* transcript, *Al-RFL1_5127*, contains three non-conserved aly-*TAS2* 3'D12(-) target sites in three separate OTLs. *Al-RFL1_5296* has two aly-*TAS2* 3'D12(-) target sites within a single OTL, whereas its ortholog *AT1G12775* has only one (Figure 14A). The only non-pseudogenic small cluster transcript in *A. lyrata* that lacks an aly-*TAS2* 3'D12(-) target site is *Al-RFL1_5216* (Appendix 3). There are nine OTLs in pseudogenic pairs, although this represents only six *A. thaliana* genes, two of which are pseudogenes with multiple co-orthologs. Six of nine OTLs have conserved ath-*TAS2* 3'D9(-) / aly-*TAS2* 3'D12 targeting, two ath-*TAS2* 3'D9(-) targets have no corresponding tasiRNA target cluster in *A. lyrata*, and one aly-*TAS2* 3'D12(-) target has no corresponding tasiRNA target cluster in *A. thaliana*.

The targeting and conservation pattern of ath-*TAS2* 3'D11(-) is distinctive in several ways. Within the 14 non-pseudogenic OTLs with *A. thaliana*-specific ath-*TAS2* 3'D11(-) targets, ath-*TAS2* 3'D11(-) never has the lowest target score or the highest number of reads in the tasiRNA target cluster that surrounds it. ath-*TAS2* 3'D11(-) is 21nt in length, is excised 2 phases forward from the cleavage-concordant phase, and has a read count of 67. The 21nt tasiRNA generated 1nt downstream in phase 3, ath-*TAS2*_886_-_21, has a read count of 114 and targets with an equal or lower target score in every *A. thaliana PPR* transcript targeted by ath-*TAS2* 3'D11(-).

These two tasiRNA generally do not have the lowest target scores in their target cluster, however lower scoring tasiRNA had read counts no higher than 11 in all of the targeted transcripts. *A. lyrata* targeting displays a pattern similar to that of *A. thaliana*. In the 8 conserved OTLs, aly-*TAS2* 3'D14(-), with a read count of 189, has neither the lowest scores nor the highest read count within the target clusters. Another tasiRNA, aly-*TAS2*_662_-_21, which is orthologous to ath-*TAS2*_886_-_21 (it is processed from the orthologous position 1nt downstream from aly-*TAS2* 3'D14(-) and has the same sequence as ath-*TAS2*_886_-_21) is represented by 301 reads, which is the highest read count in all 8 OTLs, and in three cases is tied for the best target score. Other tasiRNA with targets in these clusters are represented by a maximum of 32 reads. Thirteen of the 14 OTLs with *A. thaliana*-specific targeting contain tasiRNA target clusters in *A. lyrata,* and 6 of these contain predicted aly-*TAS2*_662_-_21 targets. There are only two OTLs on pseudogenic pairs of which one is conserved and one has *A. thaliana*-specific targeting (Figures 15A and 15B).

Ath-*TAS2* 3'D12(-) / aly-*TAS2* 3'D15(-) targets are present in 22 OTLs on non-pseudogenic pairs, 11 of which are conserved targets in both species (Figures 15A and 15B). Only 1 OTL has *A. thaliana*-specific targeting while 10 have *A. lyrata*-specific targeting. In *A. thaliana*, there are several tasiRNA that are at least as abundant as ath-*TAS2* 3'D12(-), which is represented by 71 reads and is processed in phase 2. Ath-*TAS2*_904_-_21, ath-*TAS2*_905_-_21, and ath-*TAS2*_907_-_21 are processed from phases 0, 1, and 3 of the same cycle and are represented by 75, 68, and 113 reads, respectively. Ath-*TAS2* 3'D12(-) is tied with at least one of these other three tasiRNA for the lowest target score in 9 out of 11 OTLs, and the lowest score in the other two OTLs is produced by ath-*TAS2*_904_-_21. All four tasiRNA have the lowest target score in the second cluster on *AT1G64100*. In *A. lyrata*, aly-*TAS2* 3'D15(-) is tied for the lowest target score in 9 out of the 11 conserved OTLs, though with only 18 reads, it is never the most abundant tasiRNA in any OTL. In all 8 OTLs, aly-*TAS2*_682_-_21, with 24 reads processed from phase 1 in cycle 15, is slightly more abundant than aly-*TAS2* 3'D15(-) and shares its low score. Out of 10 OTLs with *A. lyrata*-specific targeting, 8 contain *A. thaliana* tasiRNA target clusters. For these OTLs, the tasiRNA with the lowest target scores in *A. thaliana* are predominantly derived from phases 6, 7, and 8 in cycle 14, but all are of relatively low abundance with read counts of no greater than 11. The most abundant *A. thaliana* tasiRNA with targets in the 8 OTLs is from phase 3 of cycle 11 and is represented by 114 reads, but its target score of 3.5 is high compared to the best scoring tasiRNA, which have target scores of 0.5. In *A. lyrata*, aly-*TAS2* 3'D15(-) targets have the lowest score in four of ten OTLs but three of these four OTLs map to the same *A. lyrata* gene, *Al-RFL2_894*, which has three non-collinear co-orthologs in *A. thaliana* (Figures 14C-E). All three *A. thaliana* transcripts have tasiRNA target clusters orthologous to aly-*TAS2* 3'D15(-) but none are predicted to include ath-*TAS2* 3'D12(-) targets. Similar to the *A. thaliana* targeting, tasiRNA sequences derived from phases 5, 6, or 7 were responsible for the lowest target score in four of the remaining 10 OTLs. *A. lyrata* target scores were generally higher in OTLs with *A. lyrata*-specific targeting, with an average target score of 2.3, than in OTLs with conserved targeting, which had an average of 1.3. It should be noted that within the two OTLs with no tasiRNA target cluster in *A. thaliana,* the *A. lyrata* clusters are comprised of a single target less than 15nt away from the next cluster in *A. lyrata*. However, ath-*TAS2* 3'D12(-) does not target either neighboring cluster so these do represent potential gain or loss of targeting.

Although ath-*TAS1c* 3'D6(-) is conserved in *A. lyrata* as aly-*TAS1c* 3'D11(-), it has very few targets in either species. The only conserved targeting is in two separate OTLs on *AT5G41170* and *Al-RFL7_21693*, where target scores of 2.5 and 3.5 were observed. Aly-*TAS1c* 3'D11(-) has 3 targets in *A. lyrata* that are not conserved in *A. thaliana*, although all *A. thaliana* loci have tasiRNA target clusters at the orthologous locus. There is only one ortholog pair in which an ath-*TAS1c* 3'D6(-) target was present in *A. thaliana* but absent in *A. lyrata*. Although an orthologous tasiRNA target cluster is present on the *A. lyrata* transcript, there are only 5 unique tasiRNA sequences that target the *A. lyrata* transcript, all of which have target scores of 4 and are of low abundance with no more than 28 reads. There are no pseudogenic pairs targeted by either ath-*TAS1c* 3'D6(-) or aly-*TAS1c* 3'D11(-).

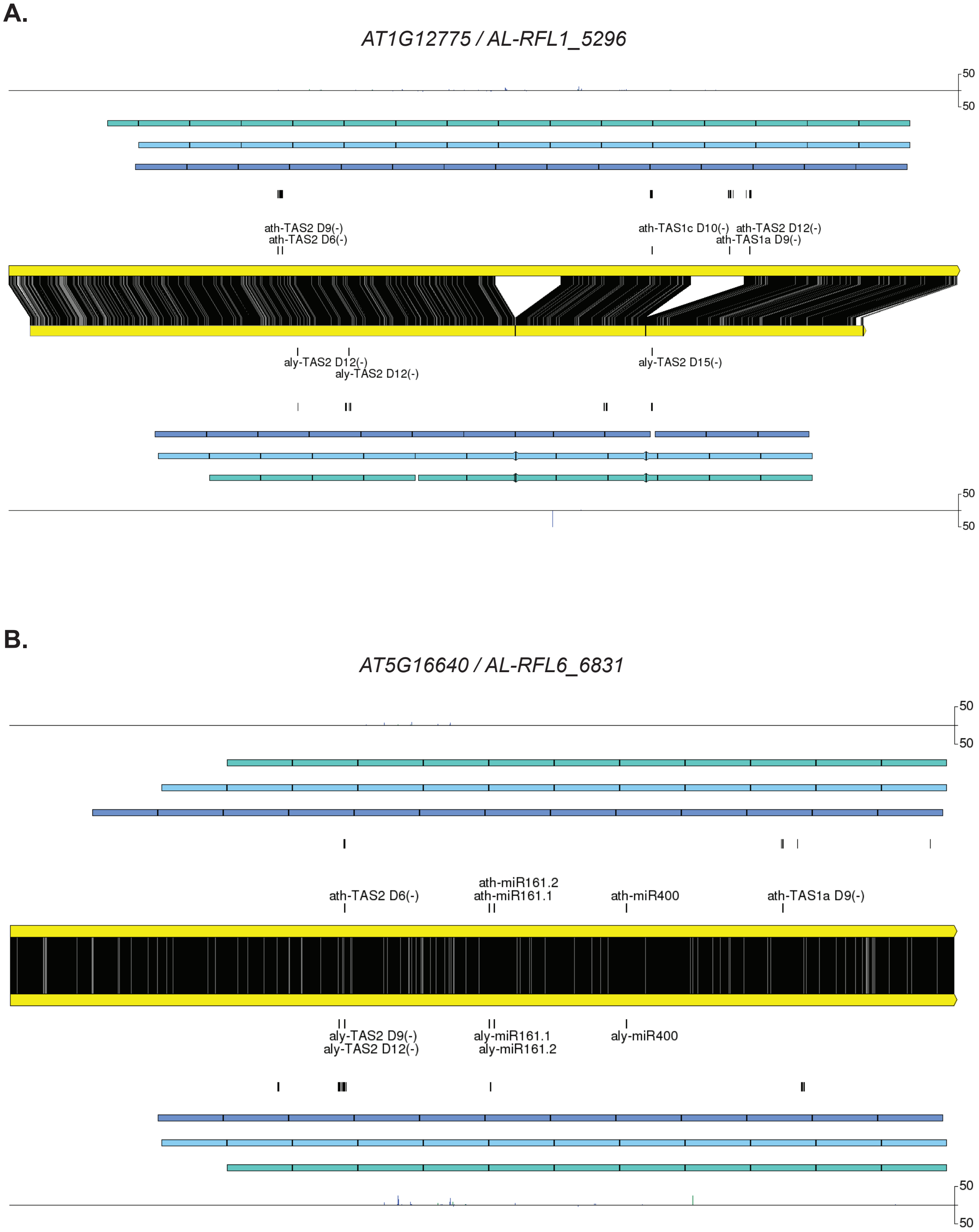
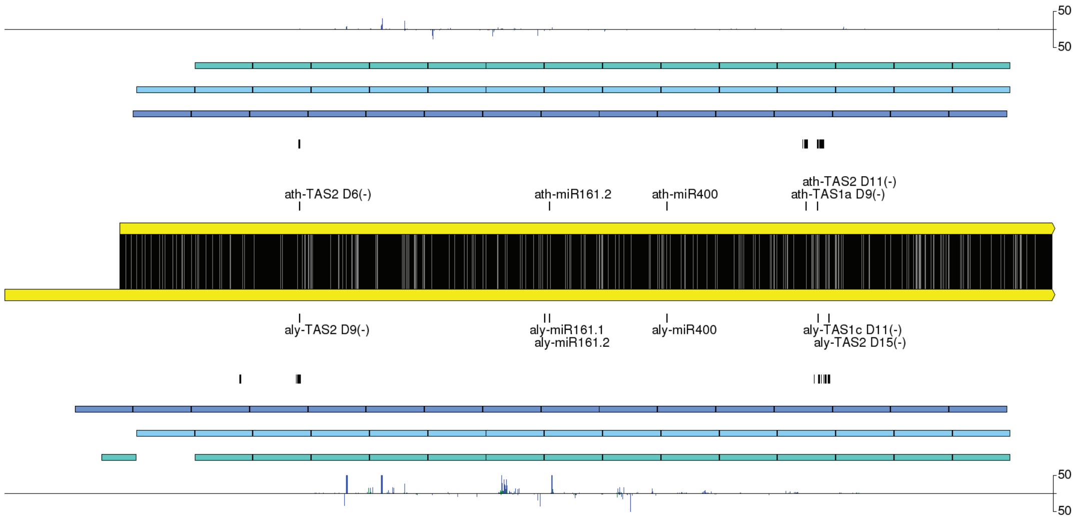


Figure 14. TasiRNA-guided cleavage sites on *RFL* transcripts.
All panels are pairwise ortholog alignments with multiple feature tracks, as described in Figure 12. **A.** *AT1G12775 / Al-RFL1_5296.* **B.** *AT5G16640 / Al-RFL6_6831*. **C.** *AT1G63330 / Al-RFL2_894*. **D.** *AT1G63400 / Al-RFL2_894.* **E.** *AT1G63150 / Al-RFL2_894*.

Figure 14 (Continued).

C.

AT1G63330 / AL-RFL2_894



D.

AT1G63400 / AL-RFL2_894

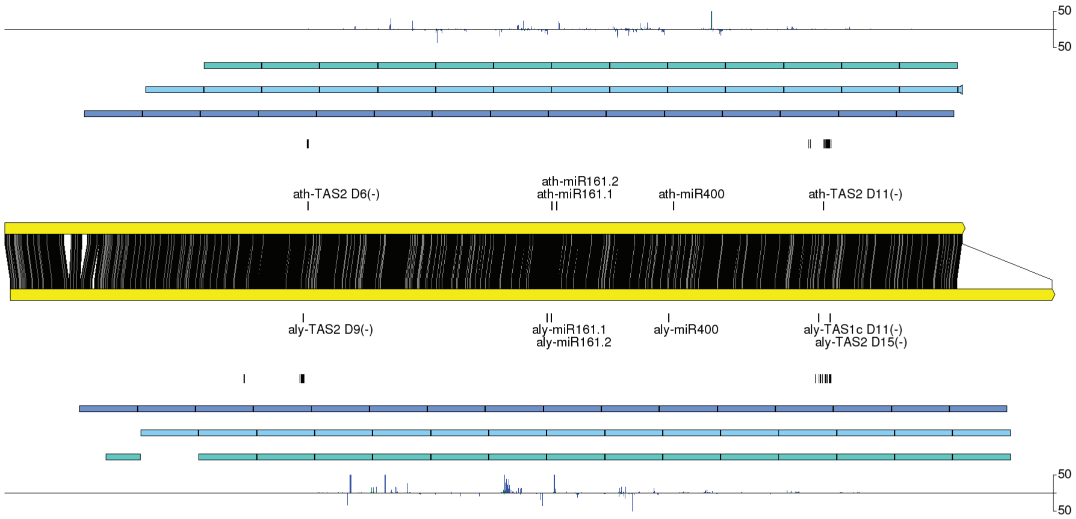
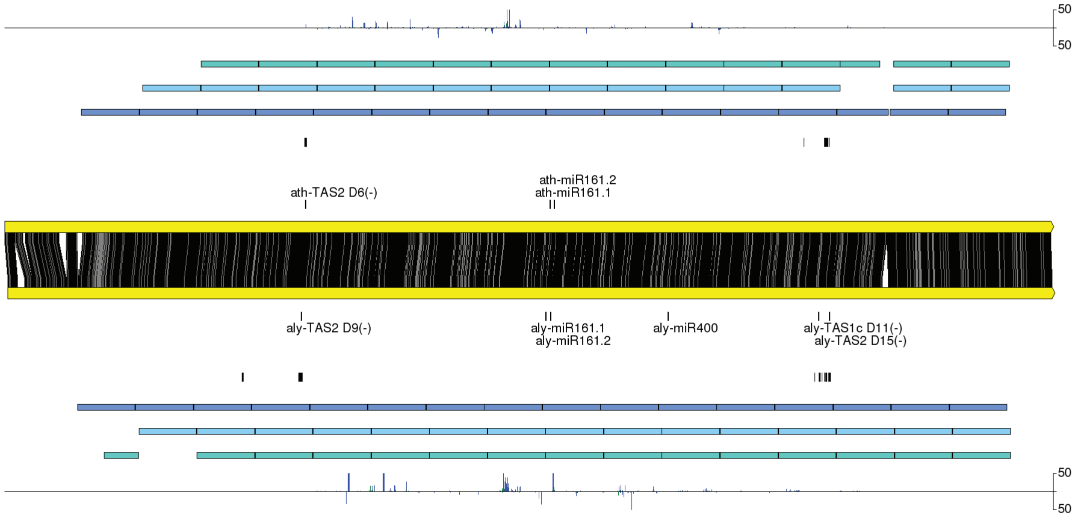


Figure 14 (Continued).

E.

AT1G63150 / AL-RFL2_894



The remaining validated tasiRNA sequences, ath-*TAS1a* 3'D9(-), ath-*TAS1b* 3'D4(-), and ath-*TAS1c* 3'D10(-), were not found in *A. lyrata* sequencing libraries or transcript sequences. Ath-*TAS1a* 3'D9(-) targeting is found in 14 OTLs, 10 of which have no corresponding target clusters in *A. lyrata*. The four OTLs with tasiRNA target clusters were associated with two genes in *A. thaliana*, *AT1G12300* and *AT1G412700*. The relevant OTL on *AT1G12300* is targeted by only four tasiRNA whereas the corresponding clusters on co-orthologs *Al-RFL1_4995* and *Al-RFL1_4999*, which have identical sequences at the target locus, have predicted targets for 17 tasiRNA from aly-*TAS1b* and aly-*TAS1c*. The aly-*TAS1b* locus that produces these tasiRNA is between positions 554 and 580 on the aly-*TAS1b* transcript sequence, a region that overlaps a 34bp insertion/deletion when compared to *A. thaliana* (Figure 16A). This 34bp sequence is similar to a sequence on the aly-*TAS1c* transcript between positions 552 and 585, which also overlaps an insertion / deletion when compared to ath-*TAS1c* (Figure 16B), and which is the source of aly-*TAS1c*-derived tasiRNA that target *Al-RFL1_4995* and *Al-RFL1_4999*. Although there are sequence similarities between these two tasiRNA-producing regions on aly-*TAS1b* and aly-*TAS1c*, the aly-*TAS1b* tasiRNA are produced primarily from a region that is absent in *A. thaliana*, whereas the aly-*TAS1c* tasiRNA are produced from a region that is present in *A. thaliana*. In spite of the presence of an orthologous ath-*TAS1c* locus, only ath-*TAS1c*_580_-_21 is predicted to target an *RFL* transcript, *AT1G12775*. The paucity of *RFL*-targeting tasiRNA from this *A. thaliana* locus is attributable to three changes in the *TAS1c* nucleotide sequence between *A. thaliana* and *A. lyrata*, two of which improve the alignment to the target loci in *A. lyrata* and one of which is neutral. These are complemented by four changes in the target sequence, all of which enhance the complementarity of *A. lyrata* tasiRNA to the target locus (Figure 16C). *AT1G12700* and its co-orthologs have a similar pattern. A small cluster of four tasiRNA target sites on *AT1G12700* is orthologous to larger clusters of targets sites on *Al-RFL1_5216* and *Al-RFL1_5213* (Figure 16D). The tasiRNA that target these regions are derived from the same locus as those that target the co-orthologs of *AT1G12300*, although there is more variation among the tasiRNA that are predicted to target *Al-RFL1_5216* and *Al-RFL1_5213* because their target sequences are not

Figure 15. Pairwise comparison of tasiRNA targeting.
A. Comparison of validated tasiRNA targeting for ortholog pairs. Gene models are shown as in Figure 13. Validated tasiRNA are listed at top. For each tasiRNA, the number of species-specific OTLs found in each species is listed in the first two sub-columns, *At* for the *A. thaliana* and *Al* for *A. lyrata*. Numbers in bold italics indicates the presence of an orthologous tasiRNA target cluster in the other species. The third sub-column (Co) displays the number of OTLs where the validated tasiRNA targets both species. Backgrounds for each number are shaded as a heat map with blue corresponding to zero and dark orange corresponding to four. The Novel Clusters column depicts target clusters without validated tasiRNA targeting. **B.** Summary of the number of conserved and unique tasiRNA target sties from the pairwise comparisons of gene orthologs presented in Figure 15A. Data presented as in Figure 13B. Dark bars above either the *A. thaliana* or *A. lyrata* blocks represent the proportion of OTLs that have corresponding tasiRNA target clusters. Thus, if the green *A. thaliana* block is labeled with a 12 and has a dark bar extending across one quarter of its length, there are 12 OTLs in which the validated tasiRNA targets only *A. thaliana*, of which 3 have tasiRNA target clusters in the orthologous position in *A. lyrata*.

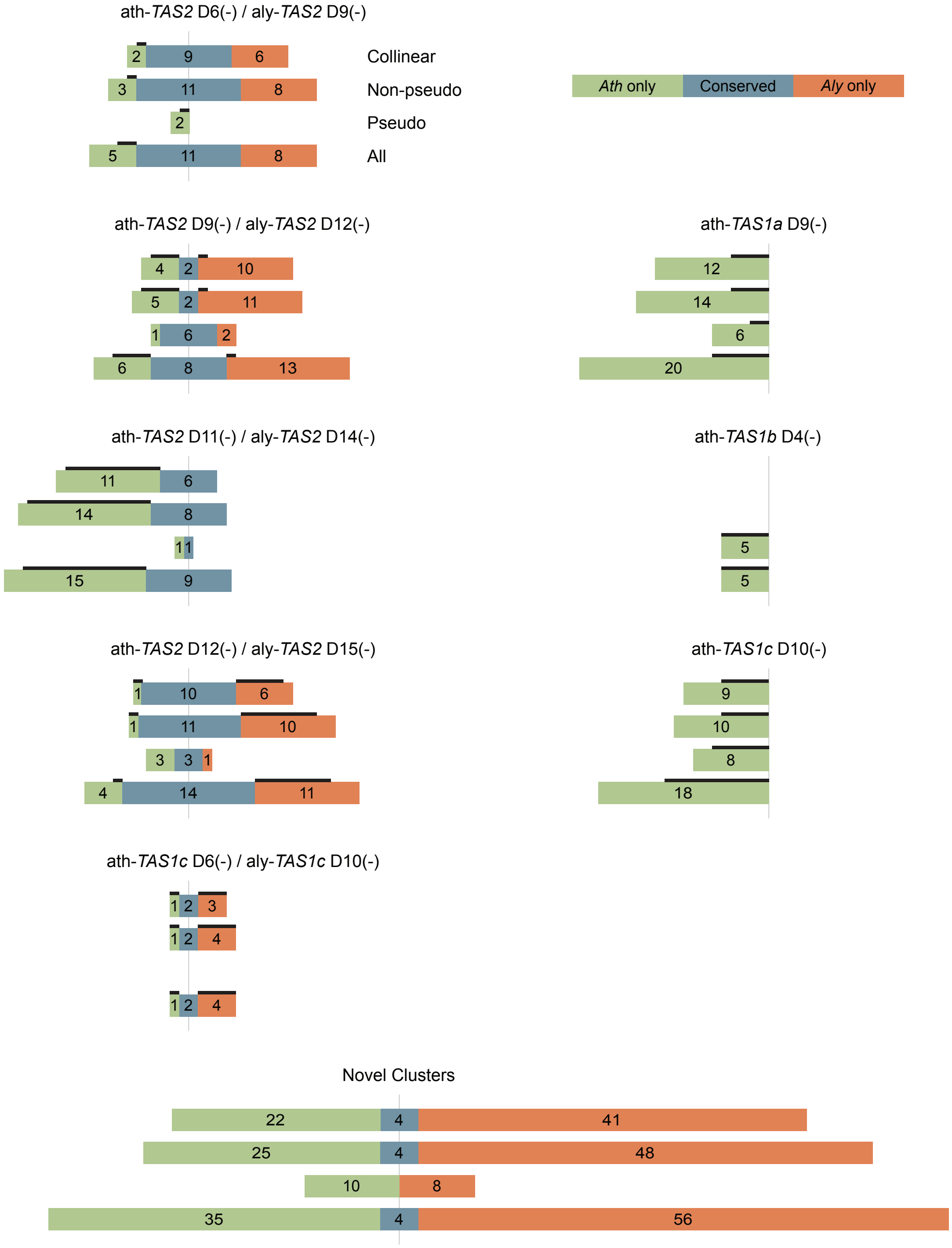
Figure 15.

A.

		ath-TAS2 **D6(-)**			***ath-TAS2*** **D9(-)**			***ath-TAS2*** **D11(-)**			***ath-TAS2*** **D12(-)**			***ath-TAS1c*** **D6(-)**			***ath-TAS1a*** **D9(-)**			***ath-TAS1b*** **D4(-)**			***ath-TAS1c*** **D10(-)**			**Novel Clust**		
A. thaliana **gene**	***A. lyrata*** **gene**	***At***	***Al***	**Cl**	***At***	***Al***	**Cl**	***At***	***Al***	**Cl**	***At***	***Al***	**Cl**	***At***	***Al***	**Cl**	***At***	***Al***	**Cl**	***At***	***Al***	**Cl**	***At***	***Al***	**Cl**	***At***	***Al***	**Cl**
AT1G06580	*Al-RFL1_2385*																1	0	0							1	1	0
AT1G12300	*Al-RFL1_4995*	0	1	0	0	1	0				1	1	1				***1***	0	0				1	0	0			
AT1G12300	*Al-RFL1_4999*				0	1	0				1	1	1				***1***	0	0				1	0	0	0	2	0
AT1G12620	*Al-RFL1_5122*				0	1	0				1	1	1										***1***	0	0	0	2	0
AT1G12620	*Al-RFL1_5125*				0	1	0				1	1	1										1	0	0	0	2	0
AT1G12620	*Al-RFL1_5127*				0	3	0				1	1	1										1	0	0	0	1	0
AT1G12700	*Al-RFL1_5213*				0	1	0				1	1	1				***1***	0	0				***1***	0	0	0	1	0
AT1G12700	*Al-RFL1_5216*										1	1	1				***1***	0	0				***1***	0	0			
AT1G12775	*Al-RFL1_5296*	***1***	0	0	1	2	1				1	1	1				1	0	0				***1***	0	0			
AT1G62590	*Al-RFL2_1611*	1	1	1				***1***	0	0	0	1	0				1	0	0									
AT1G62670	*Al-RFL2_1557*	0	1	0				***1***	0	0																2	2	0
AT1G62670	*Al-RFL2_1561*	0	1	0				***1***	0	0																2	2	0
AT1G62670	*Al-RFL2_1566*							***1***	0	0																2	0	0
AT1G62670	*Al-RFL2_1571*	0	1	0				1	0	0																2	1	0
AT1G62670	*Al-RFL2_1575*	0	1	0				***1***	0	0																2	0	0
AT1G62680	*Al-RFL2_1553*										1	1	1													3	3	1
AT1G62720	*Al-RFL2_1521*							***1***	0	0	0	***1***	0				1	0	0							0	1	0
AT1G62910	*Al-RFL2_1417*	1	1	1				1	1	1	0	***1***	0															
AT1G62914	*Al-RFL2_1415*	1	1	1				***1***	0	0																1	2	0
AT1G62930	*Al-RFL2_1411*	0	1	0				1	1	1	0	***1***	0				1	0	0							0	2	0
AT1G63070	*Al-RFL2_1215*	1	1	1	***1***	0	0	1	1	1																0	1	0
AT1G63080	*Al-RFL2_1201*	1	1	1	***1***	0	0	1	1	1																1	1	0
AT1G63080	*Al-RFL2_1205*	1	1	1	***1***	0	0	***1***	0	0				0	***1***	0										1	1	0
AT1G63130	*Al-RFL2_1159*	1	0	0	1	0	0	1	1	1																0	1	0
AT1G63330	*Al-RFL2_894*	1	1	1				***1***	0	0	0	***1***	0	0	***1***	0	1	0	0							0	1	0
AT1G63400	*Al-RFL2_894*	1	1	1				***1***	0	0	0	***1***	0	0	***1***	0										1	1	0
AT1G64100	*Al-RFL2_464*										1	1	1				1	0	0				***1***	0	0			
AT1G64580	*Al-RFL2_112*										***1***	0	0															
AT1G64583	*Al-RFL2_109*																									0	4	0
AT3G16710	*Al-RFL3_7093*																									2	2	1
AT3G22470	*Al-RFL3_10057*				1	1	1																			2	4	2
AT4G26800	*Al-RFL7_6500*													***1***	0	0										1	1	0
AT5G16640	*Al-RFL6_6831*	1	1	1	0	***1***	0										1	0	0							2	3	0
AT5G41170	*Al-RFL7_21693*							1	1	1				2	2	2										1	3	0
AT1G12620	*Al-RFL1_5184*				0	1	0				1	1	1										1	0	0	0	2	0
AT1G62670	*Al-RFL2_5036*	0	1	0				***1***	0	0																2	2	0
AT1G62910	*Al-RFL2_629*	1	0	0				1	1	1	0	***1***	0													0	1	0
AT1G62930	*Al-RFL2_1393*							1	1	1	0	1	0				1	0	0									
AT1G62930	*Al-RFL2_1395*	0	1	0				***1***	0	0							1	0	0							0	1	0
AT1G63130	*Al-RFL2_729*	1	1	1	***1***	0	0	***1***	0	0	0	***1***	0															
AT1G63150	*Al-RFL2_894*	1	1	1							0	***1***	0	0	***1***	0										1	1	0
AT1G12775	*Al-RFL1_5291*	***1***	0	0	1	1	1				1	0	0				1	0	0				1	0	0	0	1	0
AT1G62670	*Al-RFL2_1559*							1	0	0																2	0	0
AT1G62860	*Al-RFL2_1453*				1	1	1				1	1	1				1	0	0	***1***	0	0				2	2	0
AT1G62860	*Al-RFL2_1458*				1	1	1				1	1	1				***1***	0	0	***1***	0	0				2	2	0
AT1G63070	*Al-RFL2_1211*	1	0	0	1	0	0	1	1	1																		
AT1G63230	*Al-RFL2_1054*				1	1	1													***1***	0	0	***1***	0	0	1	0	0
AT1G63320	*Al-RFL2_911*																1	0	0									
AT1G64100	*Al-RFL2_898*										1	1	0				1	0	0				***1***	0	0			
AT1G12300	*Al-RFL1_5181*										1	1	1				***1***	0	0				***1***	0	0	0	2	0
AT1G12620	*Al-RFL1_5179*				0	1	0				1	0	0										1	0	0	0	1	0
AT1G63230	*Al-RFL2_1082*				1	1	1													***1***	0	0	***1***	0	0	1	0	0
AT1G63230	*Al-RFL2_700*				1	1	1													***1***	0	0	***1***	0	0	1	0	0
AT1G63630	*Al-RFL2_700*				0	1	0																***1***	0	0	1	0	0

Figure 15 (Continued).

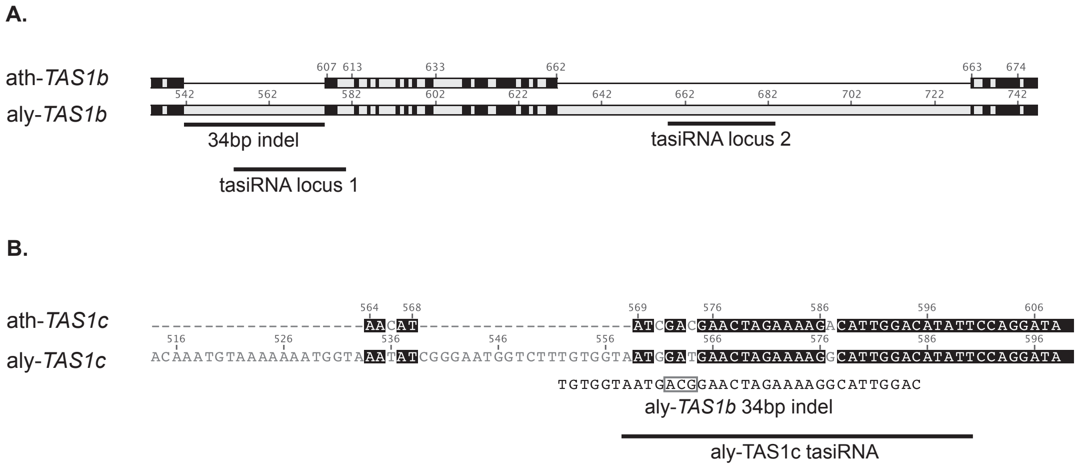
B.



identical to one another. The target sequence on *Al-RFL1_5216* is identical to that on *AT1G12700*, whereas *Al-RFL1_5213* contains several nucleotide substitutions in the target region. Of the 10 OTLs with no tasiRNA cluster in *A. lyrata*, three are co-orthologs of *AT1G63930*, two of which are non-collinear. An additional six OTLs with *A. thaliana*-specific targeting were found in pseudogenic pairs, two of which included tasiRNA clusters in *A. lyrata*. *AT1G62860* is considered a pseudogene in *A. thaliana* but has two co-orthologs in *A. lyrata*, *Al-RFL2_1453*, which we have categorized as a pseudogene, and *Al-RFL2_1458*. Although *Al-RFL2_1453* did not have a target cluster strictly orthologous to the ath-*TAS1a* 3'D9(-) target site, there is a target cluster less than 10 nucleotides away. *Al-RFL2_1458* has an unambiguous orthologous target cluster. Another pseudogene, *Al-RFL1_5181*, is a co-ortholog of *AT1G12300*. TasiRNA from the same locus that target the other *AT1G12300* co-orthologs and the *AT1G12700* co-orthologs also target these three *A. lyrata* transcripts at positions orthologous to the ath-*TAS1a* 3'D9(-) targets.

Figure 16. TasiRNA biogenesis and targeting.
A. Partial alignment of *A. thaliana* and *A. lyrata TAS1b* transcript sequences with original positions periodically indicated above the transcript line. Matching regions are shown in black, mismatched regions in grey. TasiRNA locus 1 and tasiRNA locus 2 produce tasiRNA that target loci orthologous to ath-*TAS1a* 3'D9(-) and ath-*TAS1c* 3'D10(-) orthologous sites, respectively. **B.** Partial alignment of *TAS1c* with identical residues highlighted in black. The sequence immediately below the alignment labeled "aly-*TAS1b* 34bp indel" is sequence identified in panel A**.** The locus labeled "aly-*TAS1c* tasiRNA" produces tasiRNA that target loci orthologous to ath-*TAS1a* 3'D9(-) orthologous sites. **C.** Alignment of two orthologous target sites and their associated tasiRNA. Top two sequences are aligned targets with nucleotide substitutions shown on the top sequence as lower case letters. Alignment of complementary *Al-RFL1_4995* and aly-*TAS1c*_559_-_21 are shown by a series of colons and periods and can be used to calculate target score. Colons indicate a perfect match, periods indicate a wobble match, and a blank space indicates non-matching residues. Details of target scoring can be found in the Methods section. Alignment of orthologous tasiRNA is shown in the lower two sequences. **D.** Alignment of an orthologous multi-target loci and tasiRNA producing regions of *TAS1c* transcripts. Alignment features shown as in panel C. Numbers in first and last rows indicate target scores for tasiRNA whose alignment starts at aligned position on the *TAS* transcript. Numbers in bold italics indicate that the target score is 0.5 greater than the integer indicated (e.g. 4.5 rather than 4). Underlined numbers highlight tasiRNA sequences found in the sequencing library.

Figure 16.



C.

```
AT1G12300              5' AUGGg-GAAucAGAaAAAGCg 3'
Al-RFL1_4995           5' AUGGA-GAACUAGAGAAAGCA 3'
                          ::::: ::::::::.:: :::
aly-TAS1c_559_-_21     3' UACCUACUUGAUCUUUUCCGU 5'
ath-TAS1c_569_-_21     3' UAgCUgCUUGAUCUUUUCuGU 5'
```

D.

```
22nt target score                55554565 6
21nt target score                6544445655
AT1G12700 (1354-1383)         5' GGGaAgCuuGAAAAgGCAUUGGAAAUuUUU 3'
                                 :.  :.:: ::::::.::::::: :: ::
ath-TAS1c (572-591, comp)     3' CUgCUUGAUCUUUUCuGUAACCUGUAUAAC 5'
aly-TAS1c (562-581, comp)     3' CUaCUUGAUCUUUUCcGUAACCUGUAUAAC 5'
                                 :. :::: :::::: :::::::: :::::
Al-RFL1_5213 (1465-1494)      5' GGGgAaCcaGAAAAaGCAUUGGAAAUaUUU 3'
21nt target score                4444555434
22nt target score                445555444
```

On non-pseudogenic pairs, ath-*TAS1c* 3'D10(-) targets are present in 10 OTLs, 5 of which contain orthologous tasiRNA target clusters in *A. lyrata* (Figures 15A and 15B). The 10 OTLs are related to 5 transcripts in *A. thaliana* with *AT1G12620* alone represented in four OTLs because of the high number of co-orthologs present in *A. lyrata.* As discussed above, a comparison of ath-*TAS1c* 3'D10(-) and its orthologous sequence in *A. lyrata*, aly-*TAS1c*_670_-_21, revealed the presence of three nucleotide substitutions. Four OTLs in *A. lyrata* with target clusters orthologous to ath-*TAS1c* 3'D10(-) are targeted by aly-*TAS1c*_670_-_21 and other tasiRNA derived from adjacent positions on the *TAS1c* transcript, including aly-*TAS1c*_669_-_21, which is in phase 1 rather than phase 2 and has both the lowest target score and the greatest number of reads in each cluster. In the fifth conserved OTL, *Al-RFL1_5216* is targeted by three tasiRNA derived from aly-*TAS1b* rather than aly-*TAS1c*. These aly-*TAS1b* tasiRNA are derived from positions 659 through 683 and tasiRNA from this region, along with tasiRNA from the region around aly-*TAS1c*_669_-_21, also target ath-*TAS1c* 3'D10(-) orthologous sites on pseudogenic pairs, where six out of eight OTLs contain tasiRNA target clusters. Aly-*TAS1c*_670_-_21 is the most abundant tasiRNA and has the lowest target score in five out of six OTLs on pseudogenic pairs, while aly-*TAS1c*_669_-_21 has that status in the sixth. Although both *TAS* gene and target sequences have diverged over time, it seems likely that tasiRNA from the ath-*TAS1c* 3'D10(-) orthologous locus in *A. lyrata* continue to target these transcripts, along with tasiRNA from aly-*TAS1b*.

The five ath-*TAS1b* 3'D4(-) associated OTLs are limited to two pseudogenic *A. thaliana* transcripts, *AT1G62860* and *AT1G63230*, and their five co-orthologs in *A. lyrata* (Figures 15A and 15B). Although there are three nucleotide substitutions between ath-*TAS1b* 3'D4(-) and its ortholog aly-*TAS1b*_448_-_21, tasiRNA from the aly-*TAS1b*_448_-_21 locus continue to target *RFL* transcripts. For these five OTLs, ath-*TAS1b* 3'D4(-) targeting invariably occurs in the same cluster as ath-*TAS2* 3'D9(-) targeting, which is conserved in all five *A. lyrata* orthologs. On both *A. thaliana* genes, the ath-*TAS1b* 3'D4(-) / ath-*TAS2* 3'D9(-) clusters are targeted only by tasiRNA derived from ath-*TAS2* and ath-*TAS1b*. In contrast, *A. lyrata* target clusters also have aly-*TAS1c*-derived tasiRNA targets sites. The locus on aly-*TAS1c* that produces these tasiRNA overlaps a 63bp sequence between positions 402 and 464 on the aly-*TAS1c* transcript that is absent in *A. thaliana*. A re-examination of the ath-*TAS2* 3'D9(-) / aly-*TAS2* 3'D12(-) targeting revealed that out of 17 OTLs where an ath-*TAS2* 3'D9(-) orthologous cluster is present in *A. lyrata*, 13 contain aly-*TAS1c* targets derived from the 63bp insertion, although 8 of these are targeted by only 1 tasiRNA, aly-*TAS1c*_456_-_21. None of the *A. thaliana* ath-*TAS2* 3'D9(-) target clusters in the 17 OTLs contained ath-*TAS1c*-derived targets. Thus, it appears that *A. thaliana* has a somewhat homogenized targeting profile at these target loci when compared to *A. lyrata*.

When looked at more broadly, there are several changes in targeting by perfectly conserved tasiRNA that are worth noting. When only target clusters are considered, ath-*TAS2* 3'D6(-) and ath-*TAS2* 3'D9(-) targets appeared to have less stable targeting than ath-*TAS2* 3'D11(-) and ath-*TAS2* 3'D12(-) (Figure 15B). Clusters in which ath-*TAS2* 3'D6(-) and ath-*TAS2* 3'D9(-) targets are present were conserved in 12 of 22 and 7 of 18 OTLs, respectively. This contrasts with ath-*TAS2* 3'D11(-) and ath-*TAS2* 3'D12(-) clusters in which 21 of 22 and 20 of 22 OTL clusters were conserved. When the physical clusters of *RFL* genes on the genome are considered, another set of differences appear. Ath-*TAS2* 3'D9(-) targets only one transcript from the small genomic *RFL* gene cluster in *A. thaliana*, whereas its ortholog aly-*TAS2* 3'D12(-) targets 10 out of the 12 transcript from the small cluster genomic cluster in *A. lyrata*.

There were many tasiRNA target clusters that were not associated with previously validated tasiRNA in *A. thaliana* or their orthologous RNAs in *A. lyrata* (hereafter referred to as novel clusters or novel OTLs). A total of 95 novel clusters were identified, the majority of which are lineage-specific. Thirty-five clusters are specific to *A. thaliana*, 56 are specific to *A. lyrata*, and 4 are conserved (Figures 15A and 15B). Two of the four conserved clusters are less than 50nt apart on *AT3G22470* and *Al-RFL3_10057*. The 5' cluster on *AT3G22470* consists of a single tasiRNA target with a score of 4. The tasiRNA, ath-*TAS1a*_602_-_21, is processed from a start position 2nt downstream from the ath-*TAS1a* 3'D9(-) start position and the two sequences largely overlap. The orthologous *A. lyrata* cluster is potentially targeted by 12 tasiRNA, all with a target score of 4, and all originating from the same loci in aly-*TAS1b* and aly-*TAS1c* that were identified as targeting ath-*TAS1a* orthologous sites. The cluster sizes are reversed in the 3' clusters with nine *A. thaliana* tasiRNA targets compared to only one in *A. lyrata*. All *A. thaliana* tasiRNA that target this locus are derived from *TAS2* in phases 10-13 and are of relatively low abundance, with between 2 and 34 reads. *AT1G62680* and *Al-RFL2_1553* also have a conserved target cluster of just one target site. The tasiRNA that target these sites are from the same locus as ath-*TAS2* 3'D6(-) / aly-*TAS2a* 3'D9(-) but are 21nt in length. The fourth conserved pair of novel clusters is on *AT3G16710* and *Al-RFL3_7093* and involves only one target site with a high target score and targeting tasiRNA of low abundance (6 and 8, respectively).

The 91 remaining novel OTLs represent 22 unique clusters on 15 *A. thaliana RFL* transcripts and 58 unique clusters on 33 *A. lyrata RFL* transcripts. The *A. lyrata* clusters are, on average, targeted by a greater number of tasiRNA. There were 43 predicted tasiRNA target sites in the 22 *A. lyrata* target clusters as compared to 206 predicted target sites in the 58 *A. lyrata* clusters, although 40 of these targets were present on just one transcript, *Al-RFL1_5184*. Targets on *A. thaliana* transcripts generally had high target scores. There were 31 targets with a score of 4, 9 targets with a score of 3.5, 3 targets with a score of 3, and no targets with scores below 3. Although the majority of *A. lyrata* target loci also had high target scores, there were 24 target loci that had a score of less than 3. The majority of tasiRNA that are targeting the novel clusters are derived from regions that we have examined as part of the validated tasiRNA targeting analysis. The regions surrounding the seven validated tasiRNA account for 30 out of 43 total targets in *A. thaliana*. The three regions on aly-*TAS1b*, two regions on aly-*TAS1c*, and four regions on aly-*TAS2* account for 141 of the 206 targets in *A. lyrata*.

The various ways in which tasiRNA derived from aly-*TAS1b* and aly-*TAS1c* compensated for the missing *TAS1a* ortholog is reflected in the number of unique tasiRNA sequenced. In *A. thaliana*, there were ten unique tasiRNA derived from ath-*TAS1b* and the ratio of ath-*TAS1b* tasiRNA to ath-*TAS1c* tasiRNA was 10:50 (Figure 17A). In contrast, the ratio of ath-*TAS1a* tasiRNA to ath-*TAS1c* tasiRNA is 40:50. In *A. lyrata*, the ratio of aly-*TAS1b* tasiRNA to aly-*TAS1c* tasiRNA is 59:101, which is lower than that of ath-*TAS1a*:ath-*TAS1c* but higher than that of ath-*TAS1b*:ath-*TAS1c*. Another difference between the tasiRNA populations and their targets is the ratio between tasiRNA that are predicted to target *RFL* transcripts and those that are predicted to target any *PPR* transcript. In *A. thaliana*, between 40 and 45 percent of *PPR*-targeting ath-*TAS1*-derived tasiRNA are predicted to target *RFL* transcripts. This contrasts with tasiRNA derived from ath-*TAS2*, where 85 percent of *PPR*-targeting transcripts are predicted to target *RFL* transcripts. *A. lyrata* has a much more consistent targeting profile with between 63 to 74 percent of *PPR*-targeting tasiRNA from all *A. lyrata TAS1/TAS2* sources targeting *RFL* transcripts (Figure 17A).

Because the 5' nucleotide is a key factor in determining AGO pairing (Mi et al., 2008; Montgomery, Howell, et al., 2008; Takeda et al., 2008), a pervasive shift in 5' terminal nucleotide composition might indicate a change in the prominence of AGO1 (5'U preference) or AGO2 (5'A preference) in effecting tasiRNA targeting. In examining the conservation of validated tasiRNA, we observed several examples of nucleotide substitutions in tasiRNA sequences that changed the 5' residue from a U in *A. thaliana* to an A in *A. lyrata*. Ath-*TAS1b* 3'D4(-) and ath-*TAS1c* 3'D10(-) both have a 5' U whereas their orthologs, aly-*TAS1b*_448_-_21 and aly-*TAS1c*_670_-_21, respectively, both have a 5' A. Six additional tasiRNA are produced from positions adjacent to the source of ath-*TAS1c* 3'D10(-). Two of these have a 5' U and three have a 5' A. By contrast, nine additional tasiRNA are produced from positions adjacent to aly-*TAS1c*_670_-_21, six of which have a 5' A and three of which have a 5' U. To see if these specific U to A substitutions in aly-*TAS1b* and ath-*TAS1c* are part of an overall pattern, we sorted tasiRNA that potentially target *RFL* transcripts by *TAS* source and compared their 5' terminal nucleotide composition frequencies (Figure 17B). We first performed pairwise comparisons of 5' terminal nucleotide ratios between the three orthologous *TAS1/TAS2* transcripts, none of which were significantly different. We subsequently performed pairwise comparisons of 5' terminal nucleotide ratios between all possible pairings of *TAS1/TAS2* transcripts, which was 21 comparisons in total. Significant differences resulted from comparisons of aly-*TAS1b* to three other transcripts: aly-*TAS2*, ath-*TAS1c*, and ath-*TAS2* (p-values of 0.0002522, 0.0021, 0.0004857, respectively, by Fisher's Exact Test using a Bonferroni correction for multiple comparisons). In these comparisons, the 5' nucleotide ratios of ath-*TAS1b* were not significantly different than those of ath-*TAS2* (p=0.7313). Although the 5' nucleotide ratio of aly-*TAS1b* is not significantly different from that of ath-*TAS1b*, it is possible that tasiRNA 5' nucleotide composition is changing between members of the *TAS* families, even if it not changing significantly between orthologs.

A.

Unique tasiRNA

PPR	*A. thaliana*	*A. lyrata*
TAS1a	40	0
TAS1b	10	59
TAS1c	50	101
TAS2	60	135

RFL	*A. thaliana*	*A. lyrata*
TAS1a	17	0
TAS1b	4	37
TAS1c	23	75
TAS2	51	94

B.

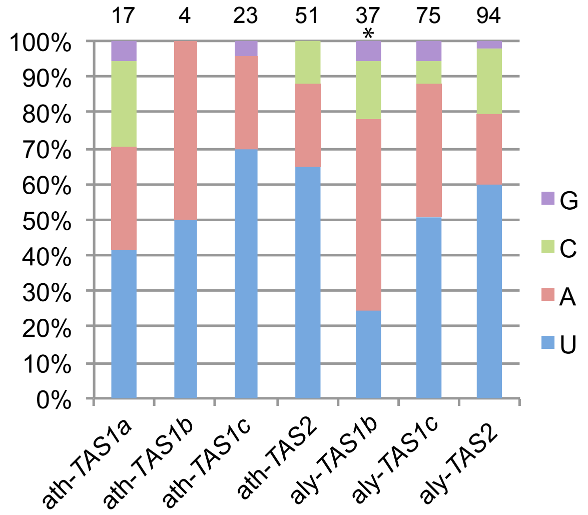


Figure 17. *PPR*-targeting tasiRNA and 5' nucleotide composition.
A. Number of unique *PPR*-targeting 21 and 22nt tasiRNA found in sequencing libraries. Upper panel shows the number of tasiRNA that are predicted to target *PPR* transcripts. Lower panel represents the subset of tasiRNA sequences that are predicted to specifically target *RFL* transcripts. **B.** The proportion of Guanine (G), Cytosine (C), Adenine (A), and Uracil (U) residues found at the 5' terminus of each *RFL*-targeting tasiRNA is shown. Colors representing each residue are shown to the right. TasiRNA are grouped by source and marked on the x-axis. Total number of reads from each source is displayed at the top. An asterisk indicates a significant difference in 5' nucleotide proportions.

Phased Small RNA Production from *PPR* transcripts

Prior studies in *A. thaliana* found three small RNA sequences that initiate production of phased small RNA from *RFL* transcripts: Ath-*TAS2* 3'D6(-), miR161.1, and miR161.2 (Montgomery, Yoo, et al., 2008; Allen et al., 2005; Howell et al., 2007). In several cases, phasing peaks were identified that did not correspond to a known phasing initiator, which we will hereafter refer to as cryptic peaks or cryptic phasing. We examined the phasing patterns of small RNA associated with *A. lyrata RFL* transcripts to identify conserved phasing associated with the three known small RNA initiators as well as the presence of other phasing signals. We first compared the *A. thaliana* phasing patterns observed in this study to results from a prior study with phasing scored as described previously (Howell et al., 2007). Howell et al. (2007) used small RNA derived from wild type and *rdr2* mutants and a read count threshold of one while we used wild type small RNA libraries and a read count threshold of three for this study. With a few exceptions, phasing was observed in the same transcripts with the same cleavage initiation sites. In two cases, *AT1G62914* (shown in Figure 18A, identified by Howell et al. (2007) as the 3' end of *AT1G62910*) and *AT1G63150*, miR161.1-directed phasing was observed by Howell et al. (2007) but not found in this study. The increase in the read count threshold may explain this difference; when a lower threshold was used, phasing peaks coincident with a miR161.1-guided cleavage site was found. We observed two cases of miR161-initiated phasing and two cases of ath-*TAS2* 3'D6(-)-initiated phasing that were not observed by Howell et al. (2007). miR161.2-initiated phasing was found in *AT1G63150* and a single phasing peak consistent with miR161.2 initiation was found in *AT1G63400* (Appendix 5). In both of these cases, the phasing scores were never greater than five, indicating a relatively low read count and a low tasiRNA cycle occupancy rate. Ath-*TAS2* 3'D6(-)-initiated phasing was found in *AT1G63330* and *AT1G63400* (Appendix 5). The *AT1G63330* signal was relatively weak with only 3 peaks and peak phase scores of 4. The *AT1G63400* signal was slightly better with six peaks, one of which had a score of 8. In an unusual case, miR161.1-initiated phasing peaks in *AT1G63400* were to the 3' side of the cleavage site rather than the 5' side of the cleavage site, as they were in Howell et al. (2007). Additionally, phasing peaks coincident with miR161.1-guided cleavage can be found on both sides of the cleavage site in *AT1G63130* (Figure 18B), *AT1G63910* (Appendix 5), and *AT1G63930* (Appendix 5). In all cases, the peaks are interspersed among cryptic peaks, raising the possibility that these peaks are not actually associated with the miR161.1-guided cleavage site but with another cleavage event.

When we compared *RFL*-derived small RNA phasing in *A. thaliana* to *A. lyrata*, several broad patterns emerged. Surprisingly, phasing consistent with aly-miR400-guided cleavage was found in eight transcripts, six of which are collinear orthologs and two of which are non-collinear (see Figure 18C for an example and Figure 19 for a summary). No ath-miR400-initiated phasing was found in *A. thaliana* by Howell et al. (2007), however, when the read count threshold is lowered to a single read for the *A. thaliana* small RNA libraries used in this study, ath-miR400-initiated phasing can be found in six transcripts. When the read count threshold is lowered to one read in *A. lyrata*, 13 transcripts have at least one phasing peak coincident with the miR400-guided cleavage site. In nearly every case of miR400-initiated phasing, the phase score is less than 3. miR400-directed phasing was also found to occur on both sides of the cleavage site. ath-*TAS2* 3'D6(-) / *aly-TAS2a 3'D*9(-) phasing occurs exclusively on the 3' side of the cleavage site and the miR161-directed phasing identified by Howell et al. (2007) occurs only on the 5' side of the cleavage site. The low scores and unusual phasing patterns displayed by miR400-initiated phasing may indicate that there is no real association with the miR400-guided cleavage site but that some other small RNA is targeting the *RFL* transcripts in the same phase, possibly small RNA derived from the same transcript or closely related paralogs. Another potentially new source of phasing was found on the *A. lyrata* transcript of *Al-RFL2_700*, coincident with an *aly-TAS2a 3'D*12(-)-guided cleavage site (Figure 18D). This gene has no collinear ortholog in *A. thaliana* but its closest BLAST hit is to *AT1G63230*. It is also the closest BLAST hit to *AT1G63630*, which also has no collinear ortholog. There are only four phase peaks consistent with *aly-TAS2a 3'D*12(-)-guided cleavage on *Al-RFL2_700* and these have scores of less than five.

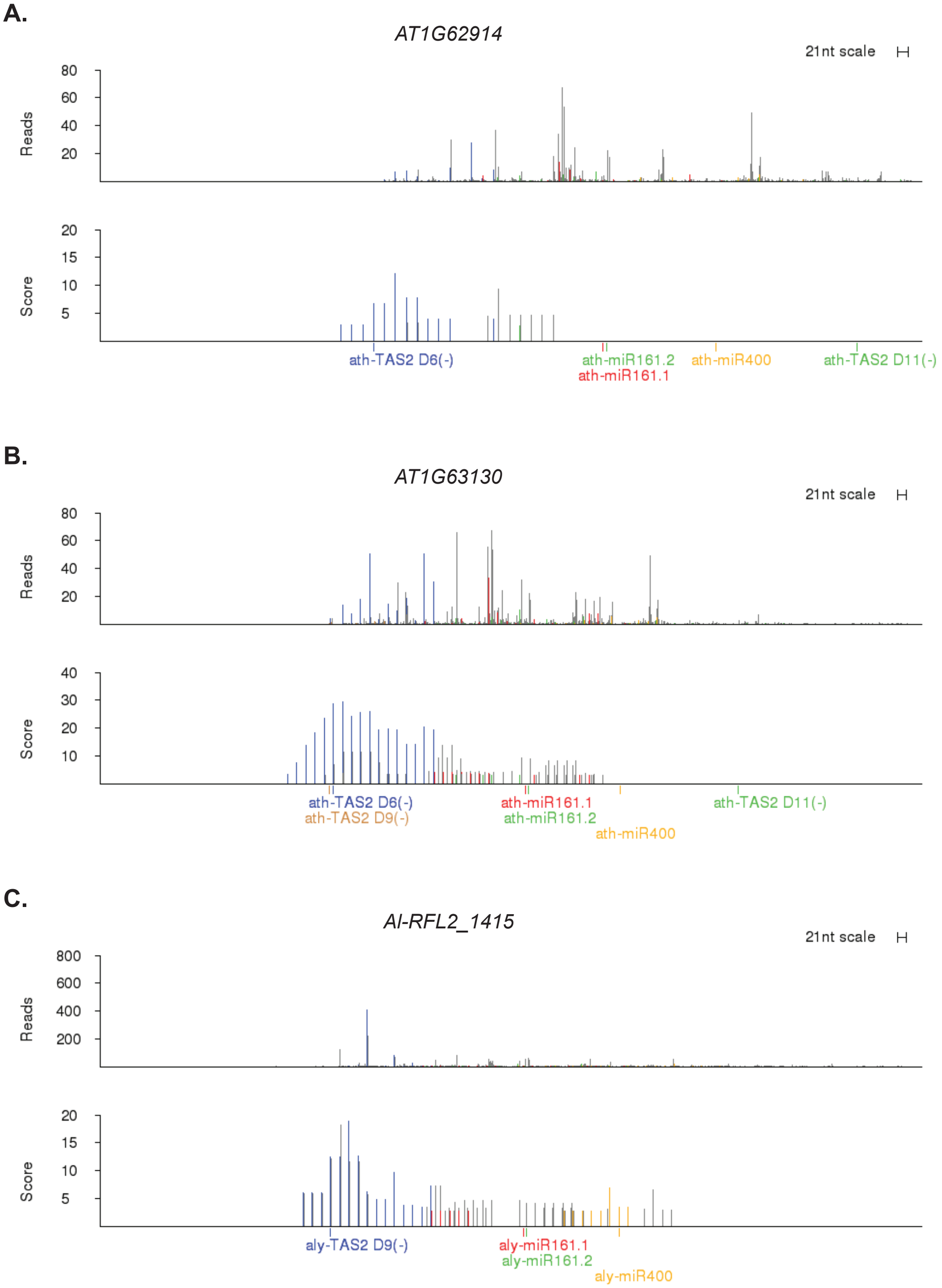
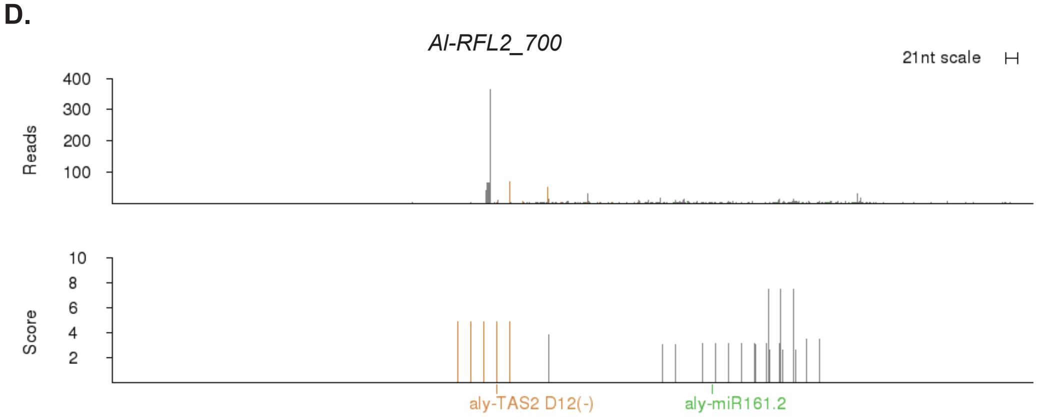


Figure 18. ***PPR*-derived small RNA.**
Phasing plots are depicted as in Figure 11. Small RNA-guided cleavage sites are shown in color with in-phase peaks and aligned small RNA read counts shown in the same color. Cleavage sites from different targets but in the same phase are shown in a single color. **A.** *AT1G62914*. **B.** *AT1G63130*. **C.** *Al-RFL2_1415*. **D.** *Al-RFL2_700*.

Figure18 (Continued).



Among non-pseudogenic ortholog pairs, ath-*TAS2* 3'D6(-) / aly-*TAS2* 3'D9(-)-initiated phasing was conserved more often than gained or lost (Figure 19A and Figure 19B). Of the eight pairwise gains in *A. lyrata* phasing relative to *A. thaliana*, four were in collinear co-orthologs of *AT1G62670* and were the result of a aly-*TAS2* 3'D9(-) target site not present in *A. thaliana* (a fifth collinear co-ortholog is not targeted by aly-*TAS2* 3'D9(-)) (Figure 19A). It seems likely that targeting and phasing arose in these paralogs either prior to the divergence of the two species and was subsequently lost in *A. thaliana*, or that it arose soon after the species divergence but before an *A. lyrata*-specific duplication. Phasing was conserved in ten pairwise comparisons and was found to be *A. thaliana*-specific in two comparisons. Within pseudogenic ortholog pairs, phased small RNAs are found from the *A. thaliana* transcript of *AT1G63070*, but not from its pseudogenic collinear co-ortholog *Al-RFL2_1211*. Excluding pseudogenic pairs, ath-*TAS2* 3'D6(-) / *aly-TAS2a 3'D*9(-)-initiated phasing was conserved, gained, or lost in *A. lyrata* in 10, 8, and 2 ortholog pairs, respectively (Figure 19B). These numbers are distorted somewhat by the tendency for *A. lyrata* to have paralogs within each ortholog group, which means a single change in the *A. lyrata* lineage can lead to multiple gains or losses in pairwise comparisons. In order to reduce this distortion, we adjusted the numbers to count gains and losses only once within each ortholog group. For instance, the four *A. lyrata* co-orthologs to *AT1G62670* that all exhibited *aly-TAS2a 3'D*9(-) initiated phasing were counted only once as a gain of phasing in that group. After these adjustments, conservation was still predominant with nine ortholog groups displaying conserved phasing. Adjusted *A. lyrata* gain and loss counts were 4 and 2, respectively. Gains and losses in ath-*TAS2* 3'D6(-) / *aly-TAS2a 3'D*9(-)-initiated phasing were invariably due to gains or losses of targeting, rather than through conservation of targeting but loss of phasing. ath-*TAS2* 3'D6(-) / *aly-TAS2a 3'D*9(-) targeting almost always led to phased small RNA production. The only exception is in *Al-RFL1_4995*, a co-ortholog of *AT1G12300* (Appendix 5). This is interesting given that none of the *PPR* transcripts from the small genomic cluster produce any phased small RNA, even in the case where miR161 or miR400 targets are present.

miR161-initiated phasing is predominantly lost in *A. lyrata* when orthologs were compared, but often due to a loss in phasing rather than a loss of targeting. Among non-paralogous ortholog pairs, miR161.1 phasing was lost in eight *A. lyrata* transcripts, while miR161.2 phasing was lost in eleven. In 13 of these 19 losses, miR161.1 or miR161.2 is still predicted to target the transcript but phasing is not detected. There were only three conserved pairwise comparisons where miR161.1- or miR161.2-associated phased small RNA was found in both transcripts and only one comparison where phasing was *A. lyrata*-specific. Overall, within the non-pseudogenic pairs, 20 out of 23 (87%) pairwise comparisons uncovered a loss of phasing in *A. lyrata*. When the adjustment is made to account for paralog over-counting, the absolute numbers decrease somewhat but the proportion of losses to gains and conserved phasing is still high with non-conserved phasing in 16 of 18 transcript groups (~89%).

Many transcripts had phasing signals that were not coincident with the three known phase initiators. Among non-pseudogenic orthologs, cryptic phasing in all eleven *A. thaliana* transcripts was conserved in orthologous *A. lyrata* transcripts. However, nine pairwise comparisons revealed a gain in cryptic phasing in *A. lyrata*. These numbers were reduced somewhat when adjusted for paralogous groups, with eight groups showing conservation and five showing a gain. Among pseudogenic pairs cryptic phasing was conserved in 11 pairwise comparisons and unique to *A. lyrata* in 13 comparisons. These numbers drop to 8 and 9, respectively, when adjusted for paralogs. The increase in *A. lyrata* cryptic phasing may be due to the greater number of genes in *A. lyrata* and the greater number and variety of small RNA that will be generated from their associated transcripts.

Figure 19. Conservation of small RNA phasing.
A. Pairwise comparison of phasing. Gene models are shown as in Figure 13A and Figure 15A. Small RNA phasing initiators are labeled across at top. Cell values correspond to phasing present only in *A. thaliana* (Ath), *A. lyrata* (Aly), or found in both species (Cons). Cells containing Ath or Aly highlighted in grey indicate that the target is absent in the orthologous gene. Unhighlighted cells containing Ath or Aly have a target score consistent with targeting in both species but lack phasing in one. **B.** Number of pairwise comparisons that contained conserved and non-conserved phasing. Counts are presented as in Figure 13B and Figure 15B.

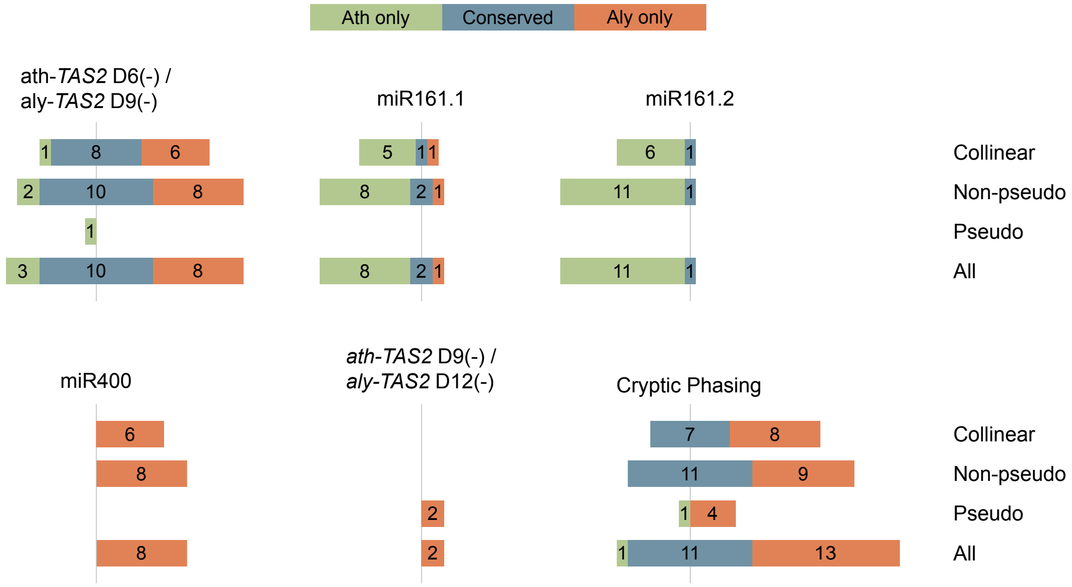
Figure 19.

A.

***A. thaliana* gene**	***A. lyrata* gene**	**ath-*TAS2* D6(-)**	**miR161.1**	**miR161.2**	**miR400**	**ath-*TAS2* D9(-)**	**Cryptic Phasing**
AT1G62590	*Al-RFL2_1611*	Cons					Aly
AT1G62670	*Al-RFL2_1557*	Aly			Aly		Aly
AT1G62670	*Al-RFL2_1561*	Aly			Aly		Aly
AT1G62670	*Al-RFL2_1566*						Aly
AT1G62670	*Al-RFL2_1571*	Aly					Aly
AT1G62670	*Al-RFL2_1575*	Aly					Aly
AT1G62910	*Al-RFL2_1417*	Cons	Ath	Cons	Aly		Aly
AT1G62914	*Al-RFL2_1415*	Cons	Aly	Ath	Aly		Cons
AT1G62930	*Al-RFL2_1411*	Aly	Cons	Ath			Cons
AT1G63070	*Al-RFL2_1215*	Cons			Aly		Cons
AT1G63080	*Al-RFL2_1201*	Cons	Ath	Ath			Cons
AT1G63080	*Al-RFL2_1205*	Cons	Ath	Ath	Aly		Cons
AT1G63130	*Al-RFL2_1159*	Ath	Ath	Ath			Cons
AT1G63330	*Al-RFL2_894*	Cons					Aly
AT1G63400	*Al-RFL2_894*	Cons	Ath	Ath			Cons
AT5G16640	*Al-RFL6_6831*	Aly					
AT1G62670	*Al-RFL2_5036*	Aly			Aly		Aly
AT1G62910	*Al-RFL2_629*	Ath	Ath	Ath	Aly		
AT1G62930	*Al-RFL2_1393*			Ath			Cons
AT1G62930	*Al-RFL2_1395*	Aly	Cons	Ath			Cons
AT1G63130	*Al-RFL2_729*	Cons	Ath	Ath	Aly		Cons
AT1G63150	*Al-RFL2_894*	Cons	Ath	Ath			Cons
AT1G62670	*Al-RFL2_1559*						Aly
AT1G63070	*Al-RFL2_1211*	Ath					Ath
AT1G63230	*Al-RFL2_1082*						Aly
AT1G63230	*Al-RFL2_700*					Aly	Aly
AT1G63630	*Al-RFL2_700*					Aly	Aly

Figure 19 (Continued).

B.



DISCUSSION

Expanded *RFL* Gene Complement in *A. lyrata*

In this study, we identified 51 potential *RFL* genes in *A. lyrata* that are orthologous to 30 *A. thaliana RFL* genes. Seven out of 51 *A. lyrata* genes were classified as probable pseudogenes. The high proportion of *RFL* genes in *A. lyrata* that contain introns is of potential concern. Twenty-four out of 44 *A. lyrata* genes categorized as functional contain introns, as compared with 4 out of 26 in *A. thaliana*. All four of the *RFL* pseudogenes in *A. thaliana* contain introns. In other gene families, the presence of an intron would not, on its own, raise concerns that a particular gene model is a potential pseudogene. However, *A. thaliana* and *Oryza sativa* generally have a high proportion of intron-less *PPR* genes, especially when compared to *Physcomitrella patens,* and it is thought that the angiosperm expansion of the *PPR* gene family is, at least in part, the result of retrotransposition of spliced *PPR* transcripts (Lurin et al., 2004; Geddy and Brown, 2007; O'Toole et al., 2008). Geddy and Brown (2007) found that *RFL* genes in particular are prone to transposition and propose that rapid birth and death of retrotransposed genes would explain both the predominance of intron-less genes, the frequent identification of *RFL* genes with no collinear orthologs in related species, and the frequent inversion of *RFL* genes. If this were the dominant mechanism of *RFL* gene duplication, the presence of an intron in a gene could be the attempt by gene-calling programs to bypass an internal stop codon. Further support for the proposal that *RFL* genes with introns are likely to be pseudogenes is provided by the fact that the three *A. thaliana RFL* genes for which a function has been identified are intron-less (Jonietz et al., 2010; Hölzle et al., 2011; Jonietz et al., 2011). These patterns of duplication raise the possibility that many of the genes we categorized as functional in *A. lyrata* actually contain internal stop codons and are pseudogenes. Counterbalancing this concern is the possibility that *A. thaliana* is under selective pressure for a reduced genome size when compared to *A. lyrata* (Hu et al., 2011) and that direct comparison of intron gains and losses between *A. thaliana* and *A. lyrata* is biased towards intron losses in *A. thaliana* (Fawcett et al., 2011). Although the *A. lyrata* gene set used by Fawcett et al. (2011) presumably did not include many of the *PPR* genes found in this study (because the JGI annotation masked many PPR motif-rich regions of the genome), it is possible that an accelerated rate of intron loss in *A. thaliana* is responsible for some of the intron losses between *RFL* orthologs in *A. thaliana* and *A. lyrata*. It is also possible that the duplication process prevalent in the non-*RFL PPR* genes is somewhat different than for *RFL* genes. We identified at least one inversion (*AT1G63330* and *AT1G62590*) that is unlikely to be the result of retrotransposition, since the duplication spanned at least two genes. Non-*RFL PPR* genes tend to be conserved across lineages and appear as single-gene orthologs rather than paralogous groups of genes (O'Toole et al., 2008; Yuan et al., 2009), which indicates that a different mechanism of duplication may predominate in *RFL* genes. It is possible that the retrotransposition process proposed as the means of duplication for the original expansion of *PPR* genes in flowering plants is not as prevalent or is modified in the *RFL* clade. For these reasons, we were hesitant to categorize genes as non-functional simply based on the presence of introns in the gene model, though we recognize that they may indeed be pseudogenes. At this juncture, it is also unclear how stringent the tandem arrangement of PPR motifs must be for a PPR protein to be functional. Further work on the protein folding constraints on PPR proteins could help differentiate functional genes from pseudogenes.

The similarity of some genes to paralogs in distant collinear groups is also curious. There are several possible explanations for these phylogenetic groupings. One is that the phylogeny is confounded by the repeat structure of the *PPR* genes or the recent duplication of these genes. Indeed, the relatively poor bootstrap values associated with the key branch points within PPR trees do not provide great confidence about the specific configuration. However, there are several collinear groups that clade distinctly from other collinear groups, indicating that phylogenetic inference is working for at least some *RFL* genes. Further investigation into the age of the paralogous groups might help elucidate the difference between these phylogenetic patterns.

Another explanation for the similarity observed between paralogs is that there is gene conversion between paralogs with similar sequences within *A. lyrata.* It is interesting to note that the paralogous groupings take place only for *PPR* genes within the large and small gene clusters, not between clusters or between clusters and non-clustered *RFL* genes, perhaps because the higher level of similarity and physical proximity of these genes promotes gene conversion. Indeed, one proposed model for *RFL* gene evolution is that individual domains or groups of domains are exchanged through recombination (Brown et al., 2003; Li et al., 1998) and there is evidence that the *PLS* subfamily of *PPR* genes may change the number of PPR-related motifs in a gene by expansion of whole motifs or blocks of motifs (Rivals et al., 2006). Brown et al. (2003) also suggested that the presence of an intron in one of the paralogs of the Radish *Rfo* gene might be a signature of recombination, which is interesting given the number of *RFL* genes models in *A. lyrata* that contain at least one intron. Recombination as a mechanism of creating diversity within a gene family has been observed in resistance genes, where variation in the number and types of Leucine-rich repeats can be important for pathogen defense (reviewed by Ellis et al., 2000). It is possible that a similar mechanism could be functioning to provide motif diversity in *RFL* genes. A multiplicity of paralogs could provide a pool of diversity from which an appropriate restorer allele could emerge in response to new CMS variants or other variation in the mitochondrial genome. The presence of multiple CMS variants in single plant lineages is consistent with the need to respond to new CMS phenotypes on a somewhat frequent basis.

Interchange of sequence between closely related *RFL* genes could explain the phylogenetic proximity of paralogs. Recent studies have provided evidence that PPR domains and their mRNA targets sequences form a one-to-one correspondence such that each PPR domain provides specificity for a single nucleotide. One study found that the number of whole PPR domains in the maize protein PPR10 matched the number of nucleotide residues that were minimally required for *PPR* transcript binding (Prikryl et al., 2010). Another study noted that amino acid residues at specific positions within the PPR motif were more likely to be under diversifying selection (Fujii et al., 2011), which fits the hypothesis that particular amino acid residues will be important in determining the target specificity of the motif. This implies that partial gene conversion could actually have a diversifying effect on the targeting capability of

PPR proteins. If multiple paralogous genes were divergent at the site associated with nucleotide affinity, the swapping of these specific domains could modify the target site specificity of the PPR protein. Whole or partial domain swapping between paralogs could create homogeneity within lineages but still produce diversity at the amino acid residues that are critical RNA binding sites. Lineage-specific exchanges would also cause relative divergence between lineages and cause the paralogs on one lineage to group separately from *RFL* genes on other lineages. This explanation assumes that the expansion of *RFL* clades within each species is a result of a requirement for a CMS restorer function, though, as was shown in *A. thaliana*, individual *RFL* genes can have non-restorer functions.

Stability of Small RNA Targeting

We found several differences in miRNA and tasiRNA targeting and phasing patterns, but these changes were often specific to particular miRNA and tasiRNA and not part of an overall pattern between species. Targeting of *PPR* transcripts by miR400 and miR161.1 was largely conserved. The ratio of non-conserved to conserved targets in non-pseudogenic target pairs was 10:23 and 10:17 in miR400 and miR161.1, respectively. This is consistent with our prior study that found a ratio of non-conserved to conserved targets of roughly 2:3 across all miRNA targets (Fahlgren et al., 2010). miR161.2 has a slightly higher proportion of lineage-specific targets with a ratio of 13:18, but the miR161.2 ratio is not significantly different from that of miR400 or miR161.1 (p=0.4358 and p=0.7909, respectively, by Fisher's Exact test). More dramatic are the changes in tasiRNA targeting between species. ath-*TAS2* 3'D6(-) / aly-*TAS2* 3'D9(-) are conserved in proportions similar to miRNA conservation. However, ath-*TAS2* 3'D9(-) / aly-*TAS2* 3'D12(-) targets have a non-conserved to conserved ratio of 16:2, which is significantly different than the 10:17 ratio of miR161.1 (p= 0.0189 by Fisher's Exact test). ath-*TAS2* 3'D11(-) / aly-*TAS2* 3'D14(-) has a ratio of 14:8 (non-conserved are all specific to *A. thaliana*) and ath-*TAS2* 3'D12(-) / aly-*TAS2* 3'D15(-) has a ratio of 11:11 (all but one non-conserved targets are specific to *A. lyrata*). Neither of these ratios was significantly different from the miR161.1 ratio. These distinctions in conservation ratios are much less pronounced if orthologous tasiRNA target clusters are classified as conserved targets (all p-values > 0.05 by Fisher's Exact test with a Bonferroni correction). While it is unlikely that every tasiRNA target cluster contains functional targets, tasiRNA targeting of *RFL* transcripts may be more highly conserved than a straightforward analysis of validated tasiRNA targets might suggest. The most dramatic changes take place in the tasiRNA targets that are not associated with any known validated tasiRNA, with a non-conserved to conserved ratio of 91:4. One possible explanation for this change is that *TAS* genes are derived from *PPR* genes and that the highly lineage-specific target predictions represent independent, evolutionarily neutral degradation of sequence similarity in regions of the *TAS* genes that are not producing functional tasiRNA. This explanation of these target sites is consistent with the conservation observed in tasiRNA target clusters that include validated tasiRNA target sites as well as the relatively high target scores and small cluster sizes of the target clusters without validated tasiRNA target sites. It is also possible that the tasiRNA target clusters which include validated tasiRNA target sites represent a pool of lineage-specific tasiRNA targets that are no longer functional or even new target sites that are recent adaptations to some selective pressure. Both of these explanations would require a more rapid divergence in *RFL* and *TAS* gene sequences than a completely neutral explanation.

It was surprising to find a detectable signal of miR400-dependent phasing in *A. lyrata*. It is possible that these phasing signals are not the result of miRNA-guided cleavage at all but rather other siRNA-guided cleavage sites that happen to be in the same phase as the miRNA-guided cleavage sites. With 23 phased transcripts, *A. lyrata* has the potential to produce a greater variety of small RNA than *A. thaliana*, which has only 10 phased transcripts. These small RNA may trigger a cascade of *RFL*-generated phased small RNA, some of which might target in the same phase as the miRNA-guided cleavage sites. The phasing signal from ath-*TAS2* 3'D6(-) / *aly-TAS2 3'D*9(-) is usually of greater amplitude and physically distinct from other phasing signals from the same transcript. In *A. lyrata*, the signals coincident with *MIRNA* cleavage are similar in amplitude to cryptic phasing from the same region of the transcript. The fact that our *A. lyrata* small RNA sequence libraries contained no miR400 reads adds further uncertainty to the origin of miR400-directed phasing in particular. Sequencing libraries from pollen and sperm cells have yielded roughly twice the number of miR400 reads as from inflorescence libraries (reviewed by Borges et al., 2011) so it is possible that conducting phasing analysis using tissue-specific libraries would shed additional light on miR400-initiated phasing. In spite of the difficulty in sequencing mature miR400, conservation of the mR400 sequence and many of its associated targets, often with a target scores in *A. lyrata* that are lower than those in *A. thaliana*, seem to point towards a functional system.

The observation that aly-*TAS1b*- and aly-*TAS1c*-derived tasiRNA target in clusters orthologous to ath-*TAS1a* targets is also interesting. This suggests that duplication of an ancestral *TAS1* gene may have led to *TAS1a* subfunctionalization in *A. thaliana*. As we have shown, *A. lyrata* target clusters with *TAS1b*-derived tasiRNA targets also tend to be targeted by *TAS1c*-derived tasiRNA. By contrast, in *A. thaliana* tasiRNA derived from ath-*TAS1b* and ath-*TAS1c* do not tend to target in the same clusters as ath-*TAS1a*. ath-*TAS1b* in particular produces few tasiRNA that target *RFL* transcripts. One scenario for this targeting pattern is that a single *TAS1* gene was duplicated before the two lineages split and then duplicated once again in the *A. thaliana* lineage soon after the split. Alternatively, both duplication events may have occurred prior to the species split and the ortholog to ath-*TAS1a* subsequently deleted in *A. lyrata*. However, under both scenarios, tasiRNA derived from all *TAS1* copies present in each respective genome would have been similar to each other and targeted the same *RFL* sites. Subsequent deletions and mutations in the ath-*TAS1b* and ath-*TAS1c* genes led to specificity of ath-*TAS1a* targeting. In *A. lyrata*, no fewer deletions and mutations were retained and thus similar tasiRNA from both *TAS1* transcripts are able to target the same *RFL* target clusters. It is also interesting to speculate on whether the hypothesized pressure for reduced genome size in *A. thaliana* (Hu et al., 2011) could drive this kind of subfunctionalization. If the deletions in the ath-*TAS1b* and ath-*TAS1c* genes were a slight selective advantage due to the pressure on genome size, ath-*TAS1a* may have been left as the only *TAS1* copy that remained capable of targeting certain *RFL* transcripts. On the other hand, there are often multiple small RNA targets for each *RFL* transcript, which would presumably ensure the ability to regulate a transcript in the event that targeting by any single source of small RNA is lost.

Uncertain Cellular Role for Small RNA Regulation of *RFL* Transcripts

The precise function of small RNA regulation in fertility restoration is still unknown. Mitochondrion-related CMS and the restoration of fertility by nuclear genes has often been cited as a case of cytonuclear conflict (Charlesworth, 2002; Frank, 1989; Chase, 2007; Budar et al., 2003; Fujii et al., 2011) and small RNA regulation could conceivably benefit either side. CMS confers a reproductive advantage to female plants because energy that would have been spent on producing pollen can be used to accelerate other growth or reproductive processes (del Castillo and Trujillo, 2009; Dufaÿ et al., 2008; Thompson and Tarayre, 2000). However, as the male sterile portion of the population increases in frequency, a nuclear *Rf* allele also becomes advantageous and is indeed necessary for the survival of the population. Small RNA down-regulation of *PPR* genes could benefit the cytoplasmic genotype, allowing the defective pollen to persist. However, the presence multiple potential small RNA targets in most *RFL* transcripts and the variety of mechanisms by which these small RNA are produced would require a successful restorer to escape a multi-layered set of potential small RNA regulators. It seems more likely that small RNA targeting provides for regulatory control over the production of PPR proteins, perhaps by allowing the plant to produce restorer proteins at a specific time or location (e.g. male reproductive organs). It is also possible that small RNA regulation is primarily useful in reducing the production of “experimental” RFL proteins. If restorer genes are produced through exchanges of sequence between *RFL* genes, non-restorer proteins could arise that target a functional mitochondrial transcript rather than the CMS-causing transcript. The presence of a flexible and multi-layered regulatory mechanism might help prevent aberrant targeting of functional mitochondrial transcripts but still allow restorer gene to arise in the nuclear genome. None of these potential roles for small RNA are mutually exclusive. The conflicting pressures of cytoplasmic and nuclear genomes could both affect small RNA, as could the need for experimentation and CMS restoration. Experiments that target the timing, location, and conditions under which *RFL* genes fall under small RNA regulation could shed light on these questions.

METHODS

PPR Motif Identification

Whole chromosome six-frame translations of *A. thaliana* (TAIR10) and *A. lyrata* (Hu et al., 2011; repeat masked as in Fahlgren et al., 2010) genome assemblies and were performed with the Transeq program from EMBOSS software suite version 6.3.1 (Rice et al., 2000). A search for PPR motifs was performed for each translation using PPR_ls and PPR_fs models from PFAM (accession PF01535.12, Pfam release 23.0, available at ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/; Finn et al., 2009) and hmmsearch from the HMMER package (version 2.3.2, http://hmmer.janelia.org/). The PPR_ls model is better able to identify partial matches whereas the PPR_fs model is more specific to whole motif alignments. Hmmer alignments to the 6-frame translations were mapped back to genomic positions and compared with exonic sequences to identify gene candidates (see below for details of the gene annotation pipeline). Genes with minimum of 2 exon-overlapping PPR motifs less than 200 nt apart were considered *PPR* genes.

Gene Annotation

In *A. lyrata*, The JGI FilteredModels6 collection of gene models (Hu et al., 2011) was initially used in the effort to identify *PPR* genes. Surprisingly, many clusters of PPR motifs found on the 6-frame translation of the genome did not overlap JGI gene models. A review of several genomic regions with clusters of predicted PPR motifs using the JGI *A. lyrata* genome browser (http://genome.jgi.doe.gov/cgi-bin/browserLoad?db=Araly1) revealed extensive masking of the PPR-rich genomic sequences, with many marked as repeat elements. To compensate for this masking, we re-annotated the entire genome using GENSCAN (Burge and Karlin, 1997) on the same masked genome as was used in the motif search. A customized script was used to break each scaffold of the genome into non-overlapping sections of sizes that were suitable for analysis by GENSCAN. We used the GeneMark (GeneMark.hmm-E version 3.3, Lukashin and Borodovsky, 1998) and fgenesh (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=g

find, using the “Dicot Plants (Arabidopsis)” model) programs to generate additional gene models in *RFL* orthologous regions when JGI and GENSCAN models were deemed poor, usually because they were unreasonably long. In some cases, longer models were manually shortened to include the shorter coding sequence that contained all of the predicted PPR motifs for that locus. Preference was given to JGI models if the number of PPR motifs covered by each model was the same. Although many *A. lyrata* RFL proteins were identified in a previous study (Fujii et al., 2011), their genomic coordinates could not be ascertained with certainty and we thus decided to use our own naming convention that includes the chromosome and the start position of the gene.

With one exception, *A. thaliana* annotations and genomic sequences were from Version 10 of The Arabidopsis Information Resource (TAIR10), (ftp://ftp.arabidopsis.org/home/tair/), on www.arabidopsis.org, (February 8-28, 2011). *A. thaliana* genes are referred to by their accession name in TAIR10. The gene model for *AT1G62860* was based on the TAIR6 annotation.

Separation of *P*-type and *PLS*-type *PPR* Genes

HMMER (version 2) models for each PPR sub-type (P, S, L, L+, DYW, E, and E+) were provided by Dr. Ian Small. Each model was run separately on the *PPR* genes using hmmsearch (without the –cut_ga option). As noted by Lurin et al. (2004), because the PPR motif subtypes are closely related, a single locus was often identified by multiple subtype models. In order to determine the best scoring subtype, hits were clustered based on position along the transcript, with hits within 10 amino acid residues of the previous hit considered part of the same cluster. The hit with the lowest e-value from each cluster was used to determine the sequence of PPR subtype motifs present on each gene. Genes with 50% P hits were considered *P*-type genes, while genes with less than 50% P hits were considered to be *PLS*-type genes.

Identification and Comparison of Gene Orthologs

A. lyrata RFL genes that were collinear with *A. thaliana RFL* genes were identified by using a whole genome alignment and the MERCATOR and MAVID programs (Fahlgren et al., 2010; Dewey, 2007). *PPR* genes found within 25kb of either the orthologous locus or another collinear ortholog were considered collinear orthologs. In all but two cases, the maximum distance from the orthologous locus or another collinear ortholog was less than 5kb. Non-collinear *RFL* orthologs were identified using phylogenetic analysis and BLAST alignment. Clustalw version 1.83 (Thompson et al., 1994) was used to align all P-type PPR amino acid sequences and to create a neighbor-joining tree. We created trees using default parameters and with customized settings which included the BLOSUM substitution matrix, and with gap extension penalty of 0.4 across both BLOSUM and GONNet matrices. BLAST alignments were performed using default settings. Alignments with the highest bitscore were generally considered the most likely orthologs. In some cases, a sequence with a worse bitscore but a better sequence similarity over a slightly shorter sequence space was used as the ortholog.

For the phylogenetic analysis in Figure 8, *RFL* collinear non-pseudogenic nucleotide sequences from both species were combined in the same file and aligned using the linsi algorithm of the MAFFT program (Katoh et al., 2002; Katoh and Toh, 2008). RAxML v7.3.0 (Stamatakis, 2006; Ott et al., 2007) was run with the parameters: "raxmlHPC-PTHREADS -f a -x 16055 -p 16055 -# 1000 -m GTRGAMMA" to generate a maximum likelihood tree with 1,000 bootstraps. Global pairwise alignments for alignment plots were performed on orthologous transcript sequences using the linsi algorithm in the MAFFT (Katoh et al., 2002; Katoh and Toh, 2008) and custom Perl scripts. Diagrams for the RAxML and neighbor joining trees were drawn using Dendroscope (Huson et al., 2007).

Alignment diagrams in Figure 16 were created in Geneious (Drummond et al., 2011).

Small RNA Sequencing and Target Comparison

A. lyrata small RNA libraries were previously described (Fahlgren et al., 2010). Briefly, seven small RNA sequence libraries were used, four from flower tissue, two from seedling tissue, and one from leaf tissue. Two flower libraries and both seedling libraries were sequenced by pyrosequencing (454 Life Sciences) while the other three libraries were sequenced using an Illumina GAI sequencer. *A. thaliana* small RNA sequences from flower tissue were previously described (Montgomery, Howell, et al., 2008; Fahlgren et al., 2009). *A. thaliana* leaf material was collected at 8, 12, and 19 days post-germination from leaves 1 and 2. Small RNAs were isolated from leaf tissue and small RNA sequence libraries were constructed as previously described (Fahlgren et al., 2009) and sequenced using an Illumina GAI sequencer.

miRNA sequences were adopted from Fahlgren et. al. (2010). *A. lyrata TAS* genes were identified by sequence similarity and collinearity with *A. thaliana TAS* genes. We found that no ortholog for *ath-TAS1a* exists in *A. lyrata* but that *TAS1b*, *TAS1c*, *TAS2*, *TAS3a-b*, and *TAS4* were present. TasiRNA sequences for both genomes were identified by aligning all 18-30nt small RNA to *TAS1* and *TAS2* transcripts using CASHX pipeline v2.3 (Fahlgren et al., 2009). Potential *PPR* targets of tasiRNA were identified using TargetFinder v1.4 (Fahlgren et al., 2007; available at http://carringtonlab.org/resources/targetfinder). TargetFinder provides a penalty-based score for the base pairing between a small RNA sequence and its potential target. Mismatches between a small RNA sequence and target are weighted by position on the alignment and the type of mismatch. Briefly, a G-U mismatch adds 0.5 point to the overall score and other mismatches add one point to the score. Scores for mismatches that occur between positions 2-13 on the 5’ end of the small RNA are doubled. Thus, a score of zero indicates a perfect match. A score of four or less was considered a potential target in all analyses.

Presence or absence of orthologous target sites was identified using pairwise alignments and mapping targets onto aligned sequences. Transcript sequences from ortholog pairs were aligned using the linsi algorithm from MAFFT (Katoh et al., 2002; Katoh and Toh, 2008). miRNA targets were considered conserved if the cleavage positions were in the same relative position based on pairwise alignments. TasiRNA targets were evaluated in a similar fashion, but an additional step was taken to evaluate target clusters. Cleavage sites within 10 nucleotides of each other on the same transcript were considered part of the same target cluster. Start and end positions for each target cluster were compared to target clusters on orthologous transcripts to identify orthologous clusters. In only one ortholog pair, *Al-RFL2_1553* and *AT1G62680*, did two clusters from one transcript map to a single cluster on the ortholog, and in this case the two clusters on *AT1G62680* were treated as a single cluster.

Statistical analysis for 5' terminal nucleotide ratios and targeting conservation ratios was conducted using the fisher.test package in RStudio (R Development Core Team, 2011; RStudio available from http://rstudio.org/). In both sets of analyses, a Bonferroni correction was used account for multiple comparisons. In the 5' nucleotide analysis, a Bonferroni correction of 21 was used for pairwise comparisons across 7 *TAS* transcript sources, which yields an alpha value of 0.0024. In the targeting conservation ratio analysis, a Bonferroni correction of 28 was used for pairwise comparisons across 8 conserved small RNA sequences, which yields an alpha value of 0.0018.

Phasing analysis was conducted using the method developed by Howell et al. (2007). Phasing plots in Figures 80 and 115 were created using custom Perl scripts. Conservation of phasing was determined manually.

CONCLUSIONS

We have examined the *Arabidopsis lyrata* genome and identified 539 *PPR* genes, of which 51 are probable *RFL* genes. Forty-one *A. lyrata RFL* genes are collinear with *A. thaliana RFL* genes while 13 have transposed. Aly-TAS1 and aly-*TAS2* transcripts produced phased tasiRNA in a manner very similar to *A. thaliana*, though *A. lyrata* has one fewer copies of the TAS1 gene and only three out of eight RFL-targeting tasiRNA. In spite of these differences, miRNA and tasiRNA initiated phasing is observed in many *RFL* transcripts. miRNA initiated phasing in particular was different in *A. lyrata* in that it appears to occur bi-directionally from the initiation site, although this could be simply non miRNA initiated phasing that happens to be in the same phasing register as miRNA initiated phasing. The absence of *TAS1a* derived tasiRNA in *A. lyrata* is counterbalanced by tasiRNA derived from aly-*TAS1b* and aly-*TAS1c*, which target loci orthologous to ath-*TAS1a* derived tasiRNA targets. *RFL* transcripts in both species are often the targets of both miRNA and validated tasiRNA but the specific composition of small RNA targets was somewhat fluid. miRNA targets were conserved at rates similar to those observed for miRNA targets as a whole. Conservation of validated tasiRNA targets varied depending on the specific tasiRNA in question. Many target loci with non-conserved validated tasiRNA target sites contained predicted target sites for other tasiRNA, which raises the possibility that targeting at these loci is actually conserved. While there are many difference between small RNA targeting of *RFL* transcripts, the overall picture is one of conserved target loci but in many cases diverged target and, for tasiRNA, small RNA sequences.

BIBLIOGRAPHY

Addo-Quaye, C. et al. (2008). Endogenous siRNA and miRNA Targets Identified by Sequencing of the *Arabidopsis* Degradome. *Current Biology* **18**: 758–762.

Addo-Quaye, C. et al. (2009). Sliced microRNA Targets and Precise Loop-First Processing of *MIR319* Hairpins Revealed by Analysis of the *Physcomitrella patens* Degradome. *RNA* **15**: 2112–2121.

Adenot, X. et al. (2006). DRB4-Dependent *TAS3 trans*-Acting siRNAs Control Leaf Morphology through AGO7. *Current Biology* **16**: 927–932.

Allen, E. et al. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nat Genet* **36**: 1282–1290.

Allen, E. et al. (2005). microRNA-Directed Phasing during *Trans*-Acting siRNA Biogenesis in Plants. *Cell* **121**: 207–221.

Altschul, S.F. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389 –3402.

Aubourg, S. et al. (2000). In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. *Plant Molecular Biology* **42**: 603–613.

Axtell, M.J. et al. (2006). A Two-Hit Trigger for siRNA Biogenesis in Plants. *Cell* **127**: 565–577.

Barr, C.M. and Fishman, L. (2010). The Nuclear Component of a Cytonuclear Hybrid Incompatibility in Mimulus Maps to a Cluster of Pentatricopeptide Repeat Genes. *Genetics* **184**: 455–465.

Baumberger, N. (2005). *Arabidopsis* ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proceedings of the National Academy of Sciences* **102**: 11928–11933.

Bentolila, S. et al. (2002). A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 10887 –10892.

Borges, F. et al. (2011). MicroRNA activity in the *Arabidopsis* male germline. *Journal of Experimental Botany* **62**: 1611 –1620.

Brodersen, P. et al. (2008). Widespread Translational Inhibition by Plant miRNAs and siRNAs. *Science* **320**: 1185–1190.

Brown, G.G. et al. (2003). The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. *The Plant Journal* **35**: 262–272.

Budar, F. et al. (2003). The Nucleo-Mitochondrial Conflict in Cytoplasmic Male Sterilities Revisited. *Genetica* **117**: 3–16.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**: 78–94.

Cai, X. et al. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10**: 1957 –1966.

del Castillo, R.F. and Trujillo, S. (2009). Evidence of restoration cost in the annual gynodioecious *Phacelia dubia*. *Journal of Evolutionary Biology* **22**: 306–313.

Charlesworth, D. (2002). Plant Population Genetics: What maintains male-sterility factors in plant populations? *Heredity* **89**: 408–409.

Chase, C.D. (2007). Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. *Trends in Genetics* **23**: 81–90.

Chateigner-Boutin, A.-L. and Small, I. (2007). A rapid high-throughput method for the detection and quantification of RNA editing based on high-resolution melting of amplicons. *Nucleic Acids Research* **35**: e114.

Chen, H.-M. et al. (2010). 22-nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proceedings of the National Academy of Sciences* **107**: 15269 –15274.

Chi, W. et al. (2010). Interaction of the pentatricopeptide-repeat protein DELAYED GREENING 1 with sigma factor SIG6 in the regulation of chloroplast gene expression in *Arabidopsi*s cotyledons. *The Plant Journal* **64**: 14–25.

Chitwood, D.H. et al. (2009). Pattern Formation Via Small RNA Mobility. *Genes Dev.* **23**: 549–554.

Chitwood, D.H. and Timmermans, M.C.P. (2010). Small RNAs are on the move. *Nature* **467**: 415–419.

Cuperus, J.T. et al. (2011). Evolution and Functional Diversification of *MIRNA* Genes. *The Plant Cell Online* **23**: 431 –442.

Cuperus, J.T. et al. (2010). Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in *Arabidopsis*. *Nat Struct Mol Biol* **17**: 997–1003.

Delannoy, E. et al. (2007). Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. *Biochem. Soc. Trans* **35**: 1643.

Dewey, C.N. (2007). Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol. Biol* **395**: 221–236.

Drummond, A. et al. (2011). Geneious v5.4.

Dufaÿ, M. et al. (2008). Variation in pollen production and pollen viability in natural populations of gynodioecious *Beta vulgaris* ssp. *maritima*: evidence for a cost of restoration of male function? *Journal of Evolutionary Biology* **21**: 202–212.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics* **14**: 755 – 763.

Elbashir, S.M. et al. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Development* **15**: 188 –200.

Ellis, J. et al. (2000). Structure, function and evolution of plant disease resistance genes. *Current Opinion in Plant Biology* **3**: 278–284.

Fahlgren, N. et al. (2009). Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA* **15**: 992–1002.

Fahlgren, N. et al. (2007). High-Throughput Sequencing of *Arabidopsis* microRNAs: Evidence for Frequent Birth and Death of *MIRNA* Genes. *PLoS ONE* **2**: e219.

Fahlgren, N. et al. (2010). MicroRNA Gene Evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell* **22**: 1074–1089.

Fahlgren, N. et al. (2006). Regulation of *AUXIN RESPONSE FACTOR3* by *TAS3* ta-siRNA Affects Developmental Timing and Patterning in *Arabidopsis*. *Current Biology* **16**: 939–944.

Fawcett, J.A. et al. (2011). Higher intron loss rate in *Arabidopsis thaliana* than *A. lyrata* is consistent with stronger selection for a smaller genome. *Mol Biol Evol* **29**: 849–859.

Finn, R.D. et al. (2009). The Pfam protein families database. *Nucleic Acids Research* **38**: D211–D222.

Fire, A. et al. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**: 806–811.

Frank, S.A. (1989). The evolutionary dynamics of cytoplasmic male sterility. *American Naturalist* **133**: 345–376.

Fujii, S. et al. (2006). Retrograde regulation of nuclear gene expression in CW-CMS of rice. *Plant Mol Biol* **63**: 405–417.

Fujii, S. et al. (2011). Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. *Proceedings of the National Academy of Sciences* **108**: 1723 –1728.

Garcia, D. et al. (2006). Specification of Leaf Polarity in *Arabidopsis* via the *trans*-Acting siRNA Pathway. *Current Biology* **16**: 933–938.

Geddy, R. and Brown, G. (2007). Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. *BMC Genomics* **8**: 130.

German, M.A. et al. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotech* **26**: 941–946.

Hammani, K. et al. (2011). An *Arabidopsis* Dual-Localized Pentatricopeptide Repeat Protein Interacts with Nuclear Proteins Involved in Gene Expression Regulation. *Plant Cell*: tpc.110.081638.

Hammond, S.M. et al. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**: 293–296.

Hanson, M.R. and Bentolila, S. (2004). Interactions of mitochondrial and nuclear genes that affect male gametophyte development. *Plant Cell* **16**: S154–S169.

Hashimoto, M. et al. (2003). A nucleus-encoded factor, CRR2, is essential for the expression of chloroplast *ndhB* in *Arabidopsis*. *The Plant Journal* **36**: 541–549.

Hölzle, A. et al. (2011). A RESTORER OF FERTILITY-like PPR gene is required for 5′-end processing of the *nad4* mRNA in mitochondria of *Arabidopsis thaliana*. *The Plant Journal* **65**: 737–744.

Howell, M.D. et al. (2007). Genome-Wide Analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 Pathway in *Arabidopsis* Reveals Dependency on miRNA- and tasiRNA-Directed Targeting. *Plant Cell* **19**: 926–942.

Hu, T.T. et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* **43**: 476–481.

Huson, D.H. et al. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**: 460.

Ikegaya, Y. (1986). Frequent appearance of cytoplasmic male sterile plants in a radish cultivar Kosena. *Jpn J Breed* **36 (Suppl 2)**: 106–107.

Jones-Rhoades, M.W. and Bartel, D.P. (2004). Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA. *Molecular Cell* **14**: 787–799.

Jonietz, C. et al. (2010). RNA PROCESSING FACTOR2 Is Required for 5′ End Processing of *nad9* and *cox3* mRNAs in Mitochondria of *Arabidopsis thaliana*. *The Plant Cell Online* **22**: 443 –453.

Jonietz, C. et al. (2011). RNA PROCESSING FACTOR3 Is Crucial for the Accumulation of Mature *ccmC* Transcripts in Mitochondria of Arabidopsis Accession Columbia. *Plant Physiology* **157**: 1430 –1439.

Kasschau, K.D. et al. (2007). Genome-Wide Profiling and Analysis of *Arabidopsis* siRNAs. *PLoS Biol* **5**: e57.

Katoh, K. et al. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**: 3059 –3066.

Katoh, K. and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9**: 286 –298.

Kim, V.N. et al. (2009). Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10**: 126–139.

Koch, M.A. et al. (2000). Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in *Arabidopsis*, *Arabis*, and Related Genera (Brassicaceae). *Mol Biol Evol* **17**: 1483–1498.

Koizuka, N. et al. (2003). Genetic characterization of a pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosena radish. *Plant J* **34**: 407–415.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet* **39**: 309–338.

Koprivova, A. et al. (2010). Identification of a Pentatricopeptide Repeat Protein Implicated in Splicing of Intron 1 of Mitochondrial *nad7* Transcripts. *Journal of Biological Chemistry* **285**: 32192 –32199.

Kotera, E. et al. (2005). A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature* **433**: 326–330.

Kurihara, Y. and Watanabe, Y. (2004). *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 12753 – 12758.

Lanet, E. et al. (2009). Biochemical Evidence for Translational Repression by *Arabidopsis* MicroRNAs. *Plant Cell*: tpc.108.063412.

Laser, K.D. and Lersten, N.R. (1972). Anatomy and cytology of microsporogenesis in cytoplasmic male sterile angiosperms. *Bot. Rev* **38**: 425–454.

Laughnan, J.R. and Gabay-Laughnan, S. (1983). Cytoplasmic Male Sterility in Maize. *Annual Review of Genetics* **17**: 27–48.

Leppälä, J. and Savolainen, O. (2011). Nuclear-cytoplasmic Interactions Reduce Male Fertility in Hybrids of *Arabidopsis lyrata* Subspecies. *Evolution* **65**: 2959–2972.

Li, X.-Q. et al. (1998). Restorer genes for different forms of *Brassica* cytoplasmic male sterility map to a single nuclear locus that modifies transcripts of several mitochondrial genes. *Proceedings of the National Academy of Sciences* **95**: 10032 –10037.

Liu, X., Rodermel, S., et al. (2010). A *var2* leaf variegation suppressor locus, *SUPPRESSOR OF VARIEGATION3*, encodes a putative chloroplast translation elongation factor that is important for chloroplast development in the cold. *BMC Plant Biology* **10**: 287.

Liu, X., Yu, F., et al. (2010). An Arabidopsis Pentatricopeptide Repeat Protein, SUPPRESSOR OF VARIEGATION7, Is Required for FtsH-Mediated Chloroplast Biogenesis. *Plant Physiol.* **154**: 1588–1601.

Llave, C. et al. (2002). Endogenous and Silencing-Associated Small RNAs in Plants. *The Plant Cell Online* **14**: 1605 –1619.

Lukashin, A.V. and Borodovsky, M. (1998). GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research* **26**: 1107 –1115.

Luo, Q.-J. et al. (2011). An autoregulatory feedback loop involving *PAP1* and *TAS4* in response to sugars in Arabidopsis. *Plant Molecular Biology*.

Lurin, C. et al. (2004). Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis. *Plant Cell* **16**: 2089–2103.

Ma, Z. et al. (2010). *Arabidopsis lyrata* Small RNAs: Transient *MIRNA* and Small Interfering RNA Loci within the *Arabidopsis* Genus. *Plant Cell* **22**: 1090–1103.

Mallory, A. and Vaucheret, H. (2010). Form, Function, and Regulation of ARGONAUTE Proteins. *Plant Cell* **22**: 3879–3889.

Manavella, P.A. et al. (2012). Plant Secondary siRNA Production Determined by microRNA-Duplex Structure. *PNAS* **109**: 2461–2466.

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

Montgomery, T.A., Yoo, S.J., et al. (2008). AGO1-miR173 complex initiates phased siRNA formation in plants. *Proceedings of the National Academy of Sciences* **105**: 20055–20062.

Montgomery, T.A., Howell, M.D., et al. (2008). Specificity of ARGONAUTE7-miR390 Interaction and Dual Functionality in *TAS3 Trans*-Acting siRNA Formation. *Cell* **133**: 128–141.

Nakamura, T. et al. (2003). RNA-binding properties of HCF152, an *Arabidopsis* PPR protein involved in the processing of chloroplast RNA. *European Journal of Biochemistry* **270**: 4070.

O’Toole, N. et al. (2008). On the Expansion of the Pentatricopeptide Repeat Gene Family in Plants. *Mol Biol Evol* **25**: 1120–1128.

Ogura, H. (1968). Studies on the new male sterility in Japanese radish with special reference to the utilization of sterility towards the practical raising of hybrid seeds. *Memoirs of the Faculty of Agriculture, Kagoshima University* **6**: 39–78.

Okuda, K. et al. (2007). Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. *Proceedings of the National Academy of Sciences* **104**: 8178–8183.

Ossowski, S. et al. (2010). The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. *Science* **327**: 92 –94.

Ott, M. et al. (2007). Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. *Proceedings of the 2007 ACMIEEE conference on Supercomputing SC 07*: 1.

Prikryl, J. et al. (2010). Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proceedings of the National Academy of Sciences* **108**: 415–420.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing: Vienna, Austria).

Rajagopalan, R. et al. (2006). A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes & Development* **20**: 3407–3425.

Reinhart, B.J. et al. (2002). MicroRNAs in plants. *Genes & Development* **16**: 1616 –1626.

Rhoades, M.W. et al. (2002). Prediction of Plant MicroRNA Targets. *Cell* **110**: 513–520.

Rice, P. et al. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**: 276–277.

Rivals, E. et al. (2006). Formation of the Arabidopsis Pentatricopeptide Repeat Family. *Plant Physiol.* **141**: 825–839.

Rivas, F.V. et al. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat Struct Mol Biol* **12**: 340–349.

Schmitz-Linneweber, C. et al. (2006). A Pentatricopeptide Repeat Protein Facilitates the *trans*-Splicing of the Maize Chloroplast *rps12* Pre-mRNA. *Plant Cell* **18**: 2650–2663.

Schmitz-Linneweber, C. and Small, I. (2008). Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in Plant Science* **13**: 663–670.

Schnable, P.S. and Wise, R.P. (1998). The molecular basis of cytoplasmic male sterility and fertility restoration. *Trends in Plant Science* **3**: 175–180.

Schwab, R. et al. (2005). Specific Effects of MicroRNAs on the Plant Transcriptome. *Developmental Cell* **8**: 517–527.

Schwarz, D.S. et al. (2003). Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell* **115**: 199–208.

Small, I.D. and Peeters, N. (2000). The PPR motif – a TPR-related motif prevalent in plant organellar proteins. *Trends in Biochemical Sciences* **25**: 45–47.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688 –2690.

Sunkar, R. and Zhu, J.-K. (2004). Novel and Stress-Regulated MicroRNAs and Other Small RNAs from Arabidopsis. *The Plant Cell Online* **16**: 2001 –2019.

Takeda, A. et al. (2008). The Mechanism Selecting the Guide Strand from Small RNA Duplexes Is Different Among Argonaute Proteins. *Plant Cell Physiol* **49**: 493–500.

Thompson, J.D. et al. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673 –4680.

Thompson, J.D. and Tarayre, M. (2000). Exploring The Genetic Basis And Proximate Causes Of Female Fertility Advantage In Gynodioecious *Thymus Vulgaris*. *Evolution* **54**: 1510–1520.

Uyttewaal, M. et al. (2008). Characterization of *Raphanus sativus* Pentatricopeptide Repeat Proteins Encoded by the Fertility Restorer Locus for Ogura Cytoplasmic Male Sterility. *Plant Cell* **20**: 3331–3345.

Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes & Development* **20**: 759 –771.

Vazquez, F. et al. (2004). Endogenous *trans*-Acting siRNAs Regulate the Accumulation of *Arabidopsis* mRNAs. *Molecular Cell* **16**: 69–79.

Wright, S.I. et al. (2002). Rates and Patterns of Molecular Evolution in Inbred and Outbred *Arabidopsis*. *Molecular Biology and Evolution* **19**: 1407 –1420.

Wu, L. et al. (2009). Rice MicroRNA Effector Complexes and Targets. *Plant Cell* **21**: 3421–3435.

Xie, Z. et al. (2005). Expression of Arabidopsis *MIRNA* Genes. *Plant Physiology* **138**: 2145 –2154.

Yoshikawa, M. et al. (2005). A pathway for the biogenesis of *trans*-acting siRNAs in *Arabidopsis*. *Genes & Development* **19**: 2164 –2175.

Yuan, Y.-W. et al. (2009). The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. *New Phytologist* **182**: 272–283.

Zamore, P.D. et al. (2000). RNAi: Double-Stranded RNA Directs the ATP-Dependent Cleavage of mRNA at 21 to 23 Nucleotide Intervals. *Cell* **101**: 25–33.

Zehrmann, A. et al. (2011). PPR proteins network as site-specific RNA editing factors in plant organelles. *RNA Biol* **8**: 67–70.

APPENDICIES

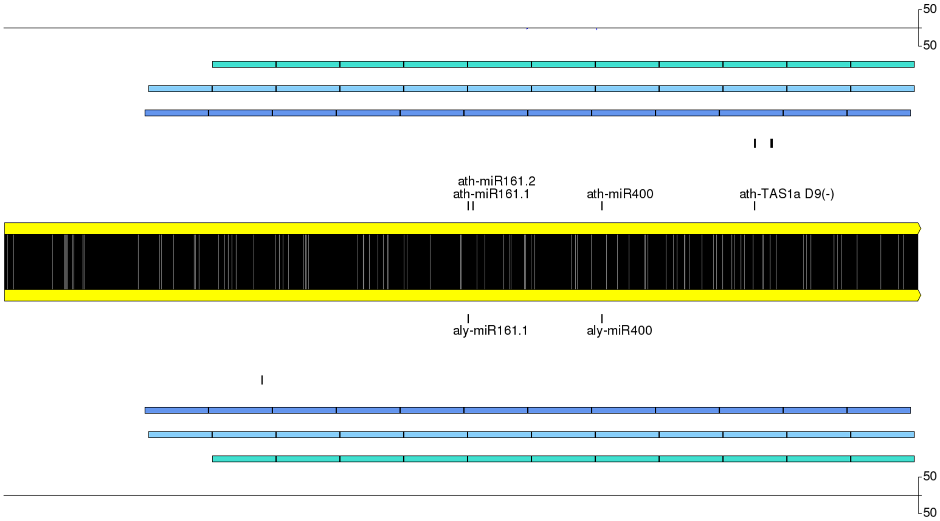
Query	Target	% Identity	Align Len	Mismatches	Gaps	Query Start	Query End	Target Start	Target End	E-value	Bitscore
Al-RFL2_1395	Al-RFL2_1395	100	1815	0	0	1	1815	1	1815	0	3538
Al-RFL2_1395	Al-RFL2_1571	93.11	1436	97	2	62	1496	14	1448	0	2046
Al-RFL2_1395	Al-RFL2_1415	91.17	1303	113	2	195	1496	174	1475	0	1655
Al-RFL2_1395	Al-RFL2_1201	91.32	1198	104	0	187	1384	154	1351	0	1550
Al-RFL2_1395	Al-RFL2_1561	89.64	1303	133	2	195	1496	174	1475	0	1497
Al-RFL2_1395	Al-RFL2_894	89.23	1319	142	0	154	1472	136	1454	0	1489
Al-RFL2_1395	Al-RFL2_1557	88.99	1353	147	2	145	1496	103	1454	0	1485
Al-RFL2_1395	Al-RFL2_1215	89.01	1310	144	0	187	1496	439	1748	0	1455
Al-RFL2_1395	Al-RFL2_1205	89.27	1239	133	0	215	1453	209	1447	0	1402
Al-RFL2_1395	AT1G62930	88.38	1317	153	0	73	1389	58	1374	0	1398
Al-RFL2_1395	Al-RFL2_1411	88.41	1286	149	0	195	1480	174	1459	0	1368
Al-RFL2_1395	Al-RFL2_629	87.82	1330	162	0	154	1483	157	1486	0	1352
Al-RFL2_1395	Al-RFL2_1575	87.15	1300	158	2	195	1494	180	1470	0	1251
Al-RFL2_1395	AT1G62670	87.14	1291	166	0	193	1483	181	1471	0	1243
Al-RFL2_1395	AT1G62910	87.02	1287	167	0	197	1483	191	1477	0	1227

Appendix 1. BLAST alignment of *Al-RFL2_1395* to other *RFL* transcripts. *Al-RFL2_1395* transcript sequence was aligned to other *RFL* transcript sequences from *A. thaliana* and *A. lyrata* using BLASTN. Results sorted by bitscore. Candidate orthologs are shaded.

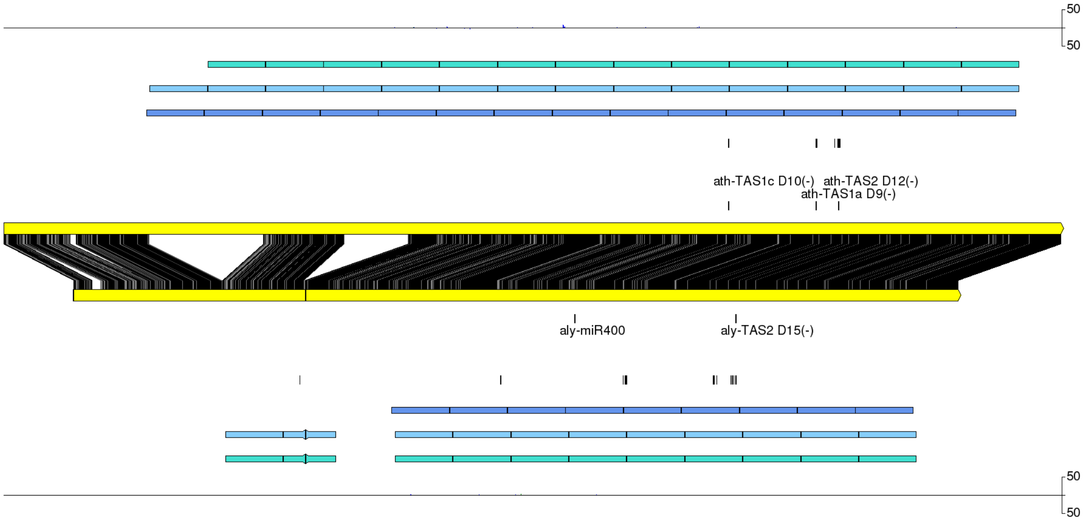
Appendix 2. Pairwise *RFL* transcript alignments.

Pairwise alignments plots of *RFL* orthologs. Plots include PPR motifs, small RNA density, and small RNA targeting, as described in Figures 4 and 12. Only plots absent in the main document are shown here. Plots are ordered by position in the *A. thaliana* genome.

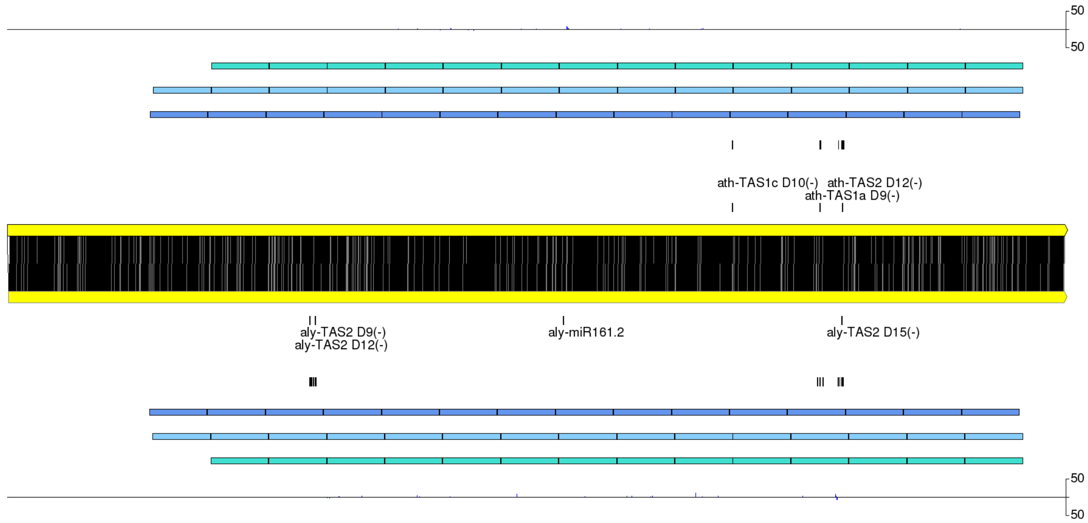
AT1G06580 / Al-RFL1_2385



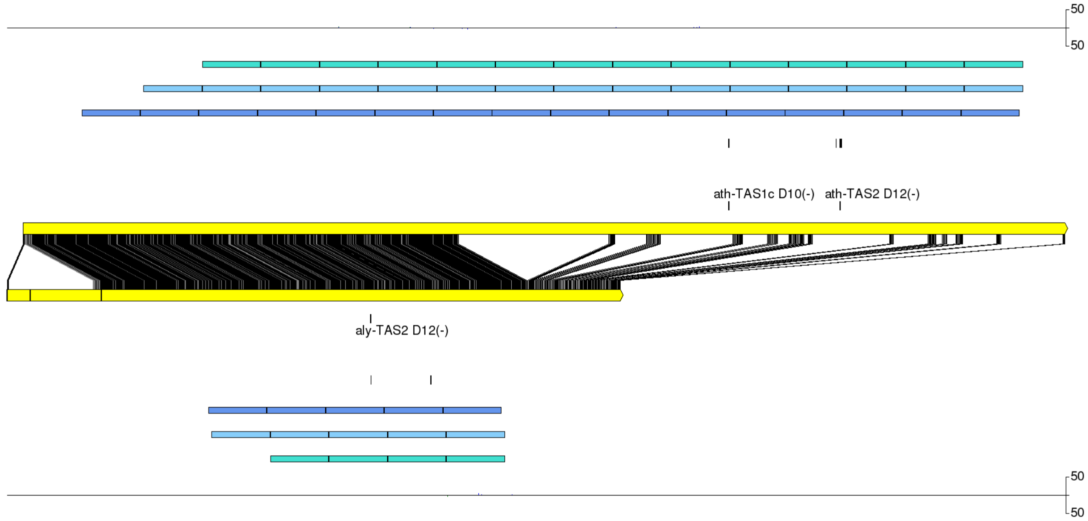
AT1G12300 / Al-RFL1_5181



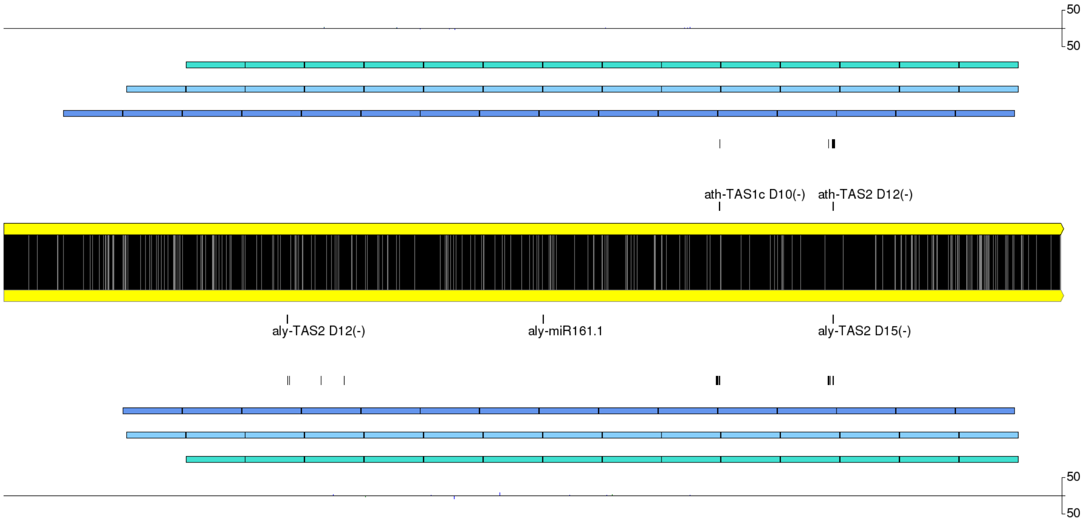
AT1G12300 / Al-RFL1_4995



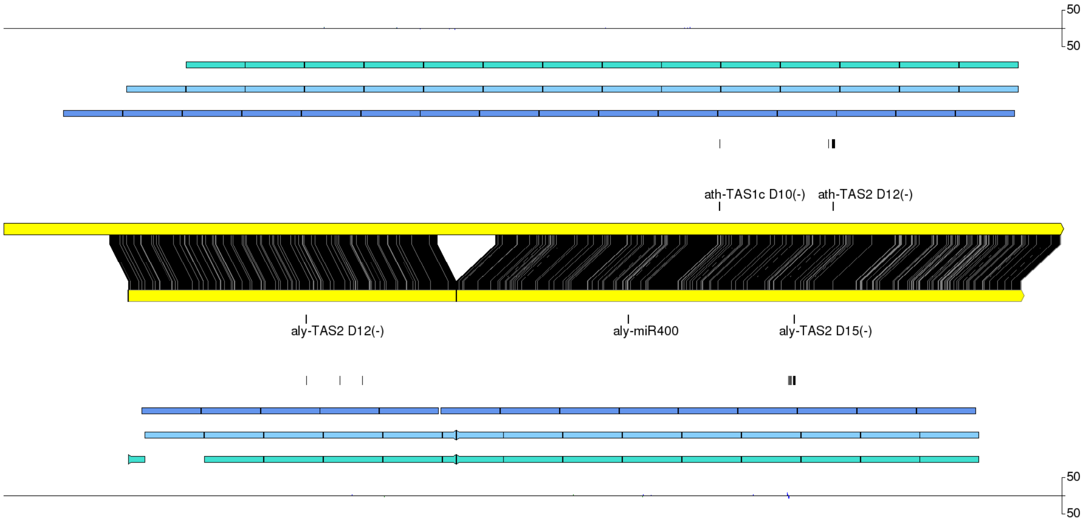
AT1G12620 / Al-RFL1_5179



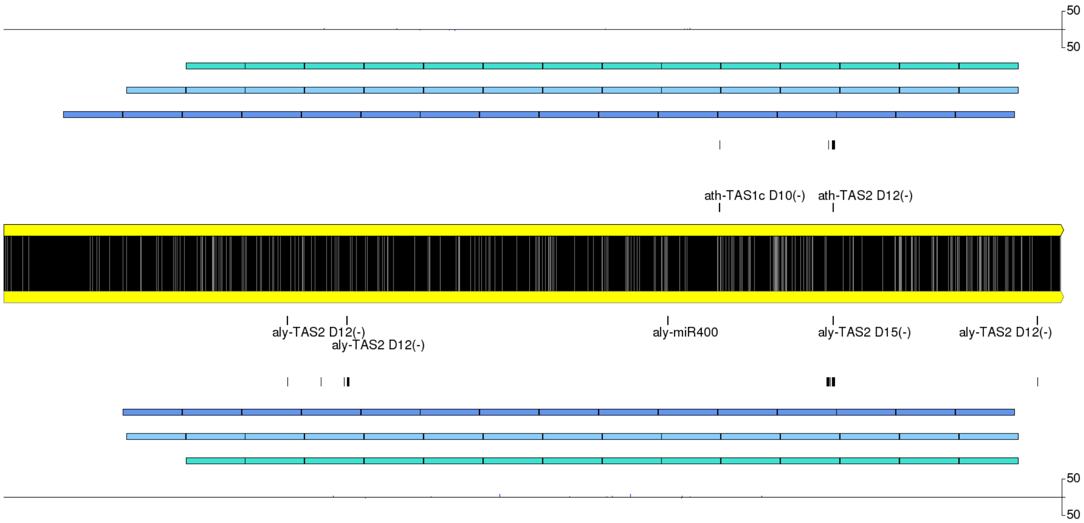
AT1G12620 / Al-RFL1_5122



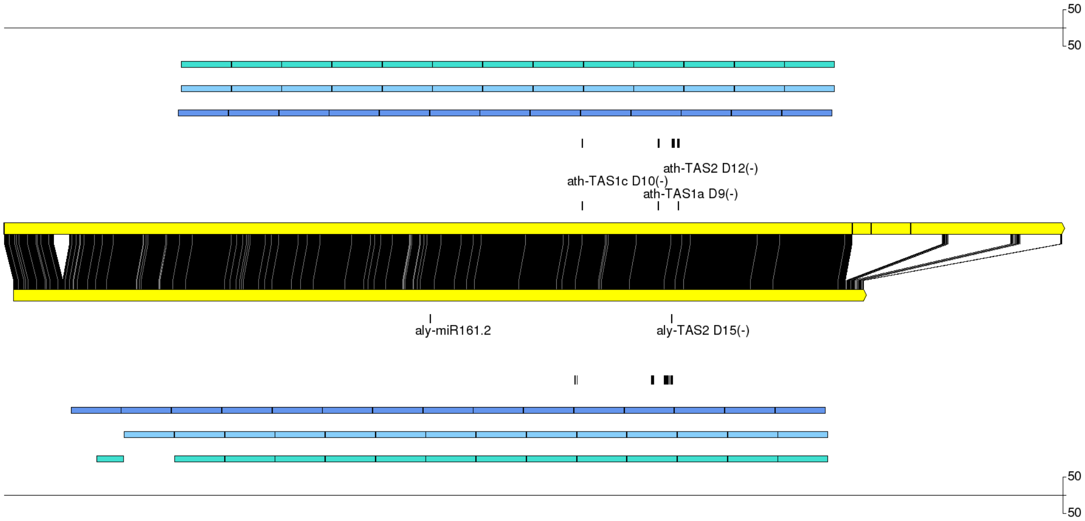
AT1G12620 / Al-RFL1_5125



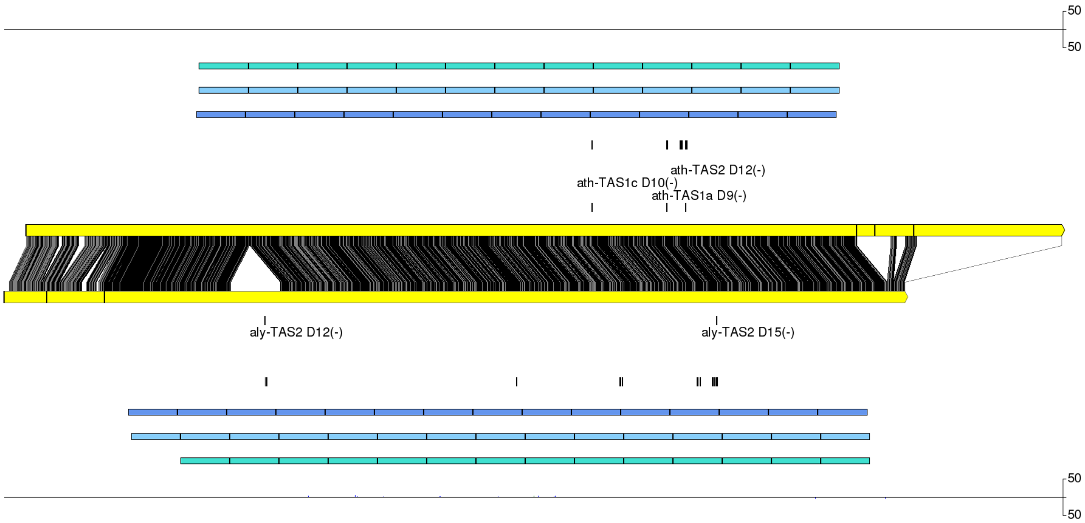
AT1G12620 / Al-RFL1_5127



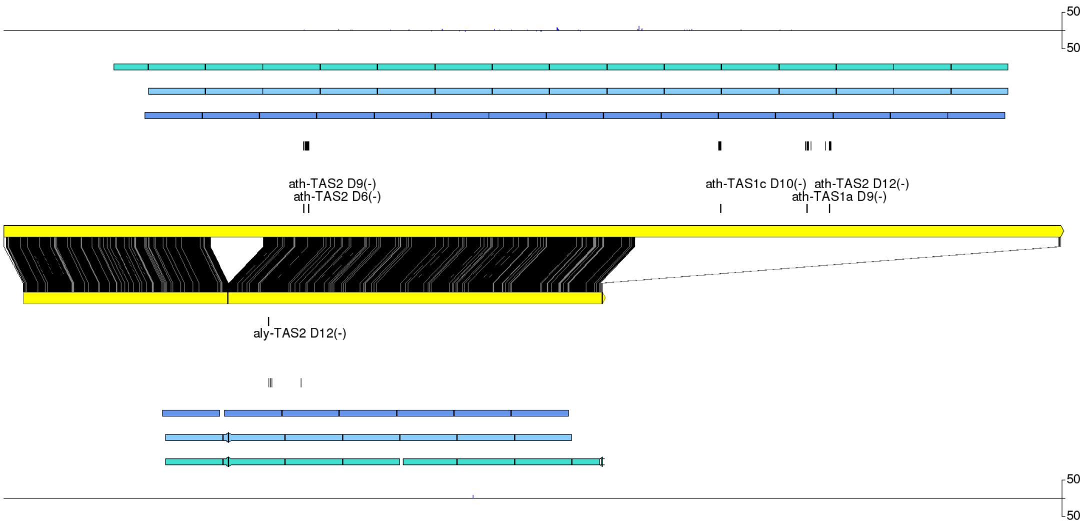
AT1G12700 / Al-RFL1_5216



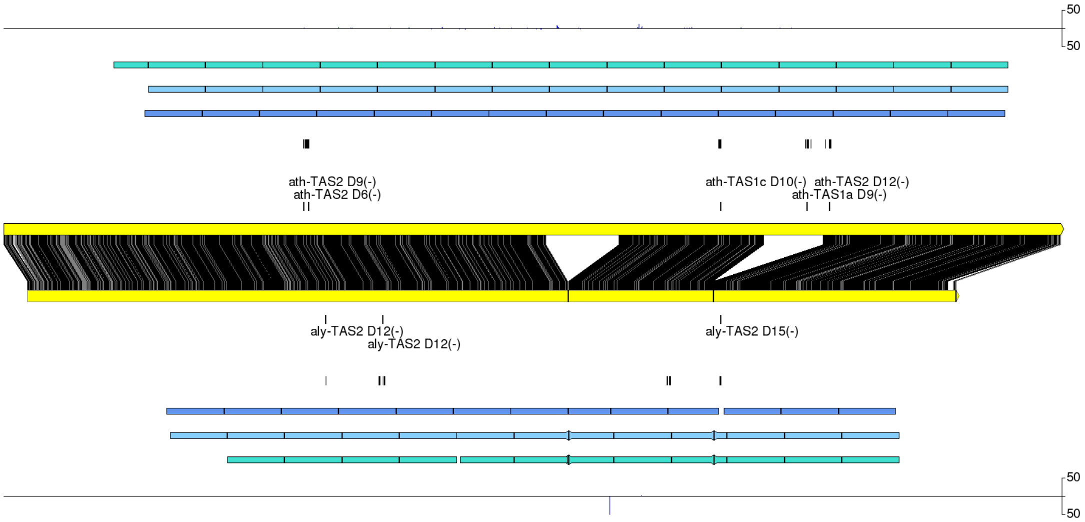
AT1G12700 / Al-RFL1_5213



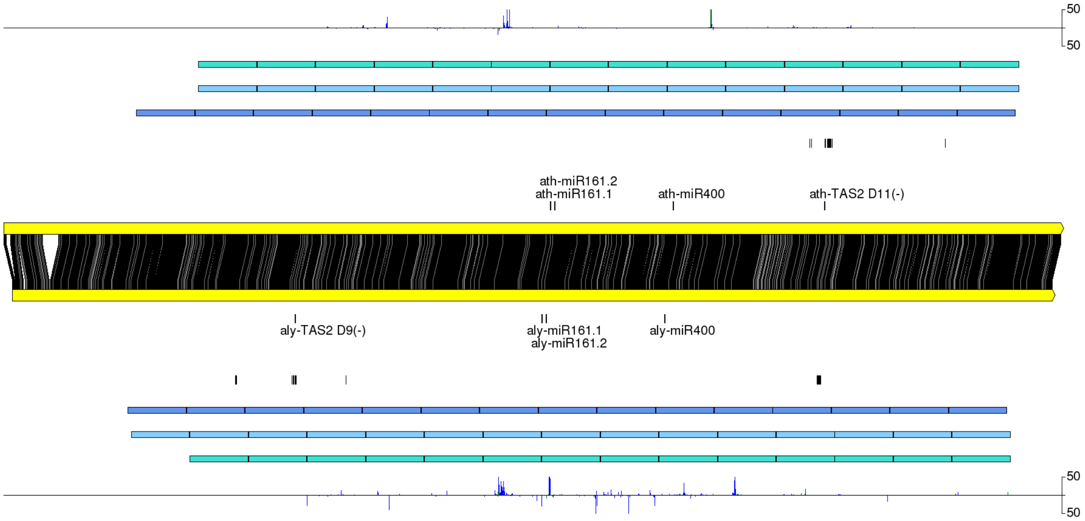
AT1G12775 / Al-RFL1_5291



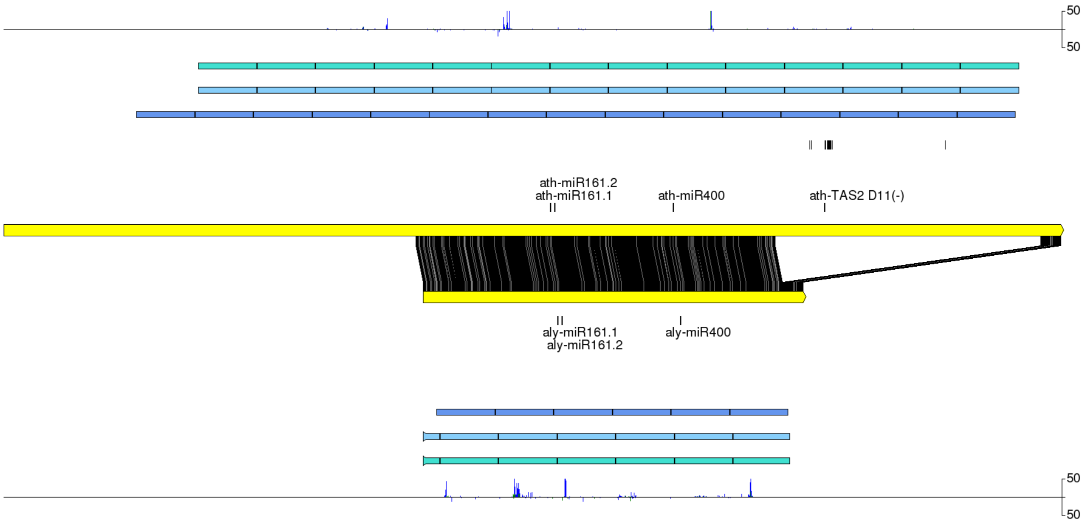
AT1G12775 / Al-RFL1_5296



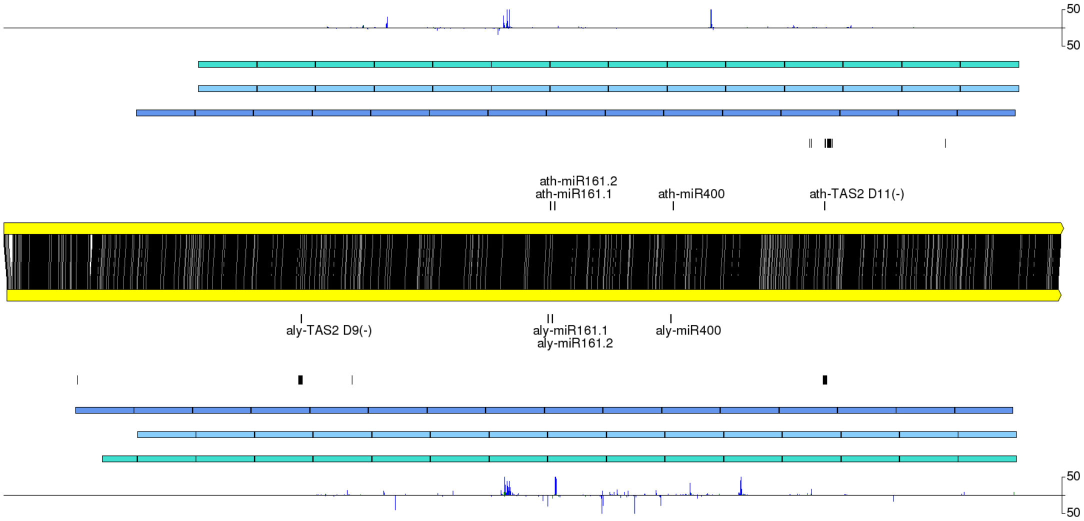
AT1G62670 / Al-RFL2_1557



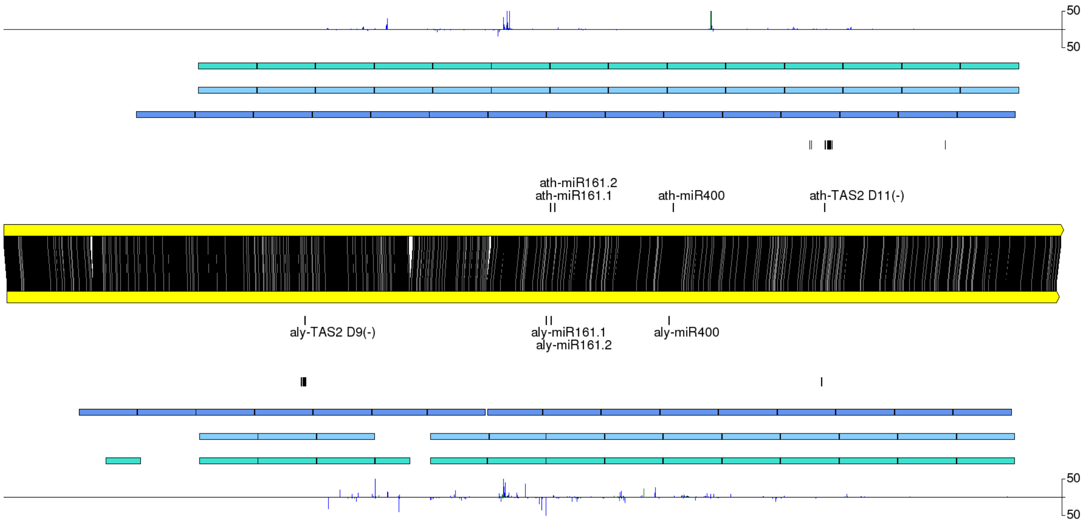
AT1G62670 / Al-RFL2_1559



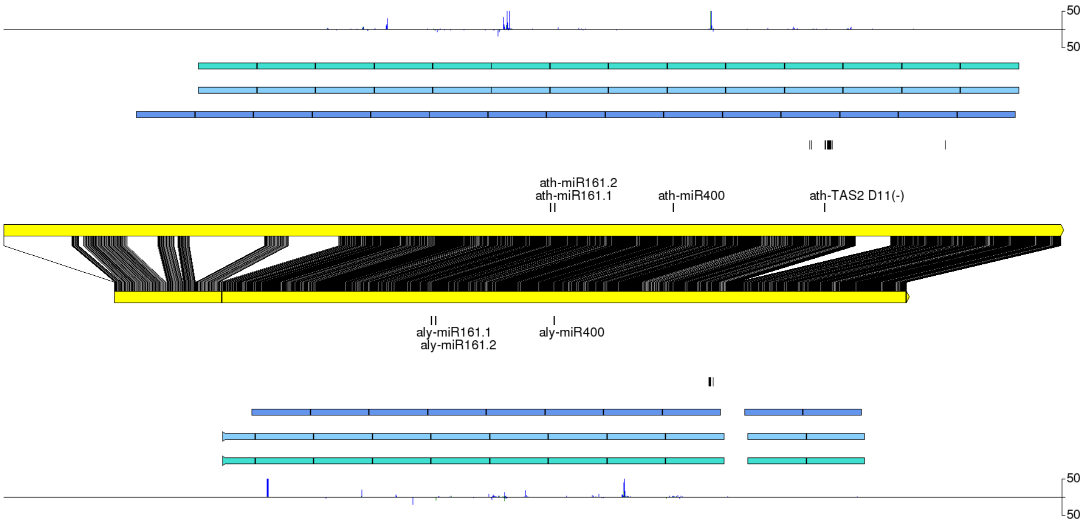
AT1G62670 / Al-RFL2_1561



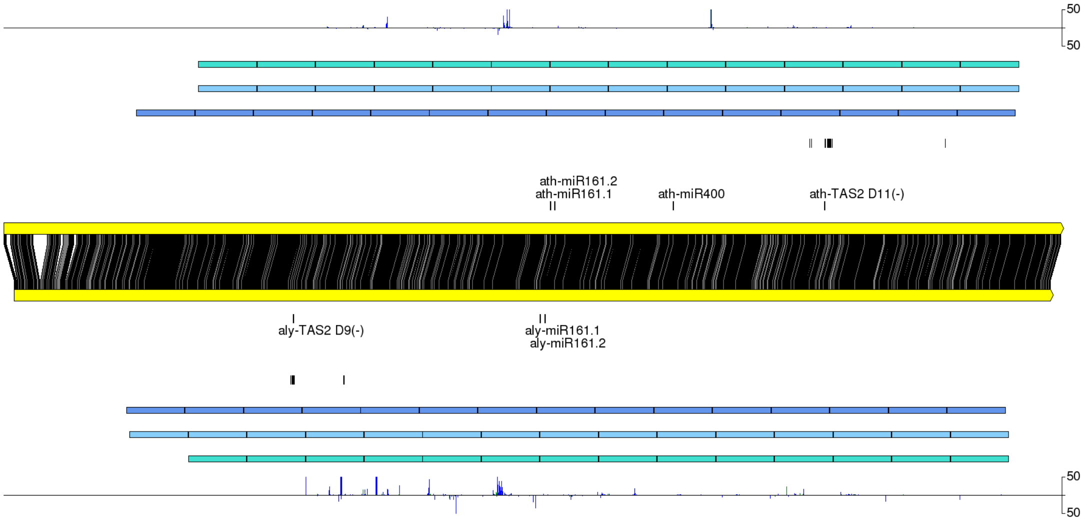
AT1G62670 / Al-RFL2_1575



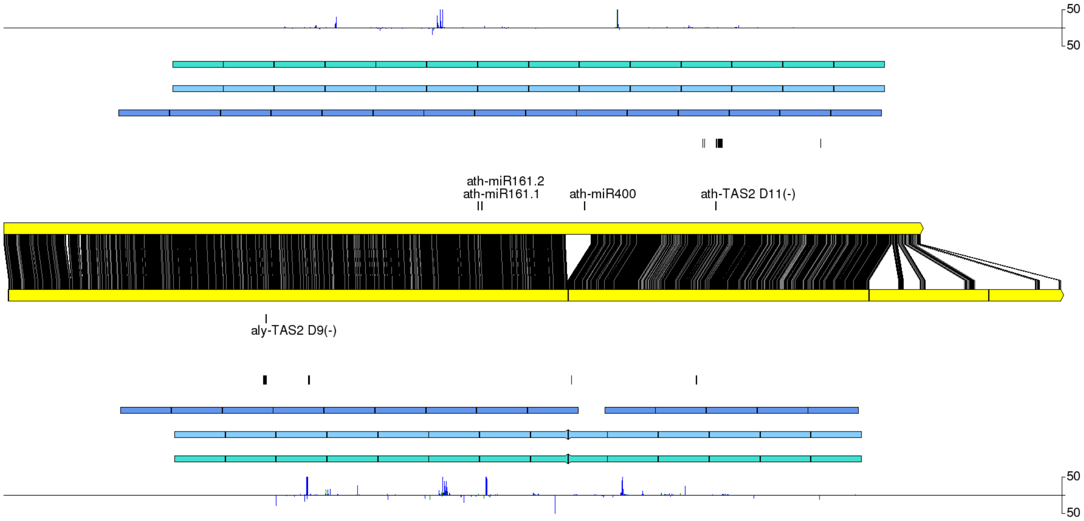
AT1G62670 / Al-RFL2_1566



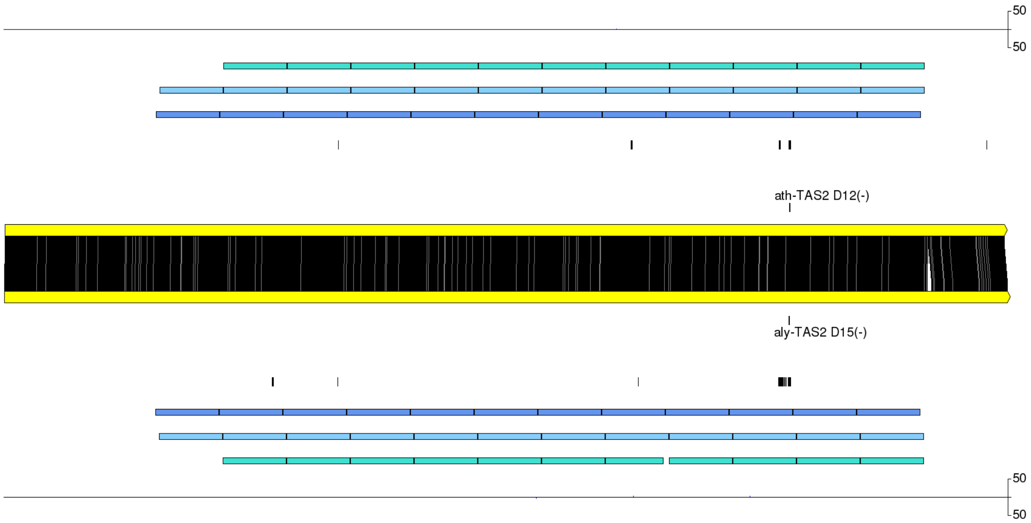
AT1G62670 / Al-RFL2_1571



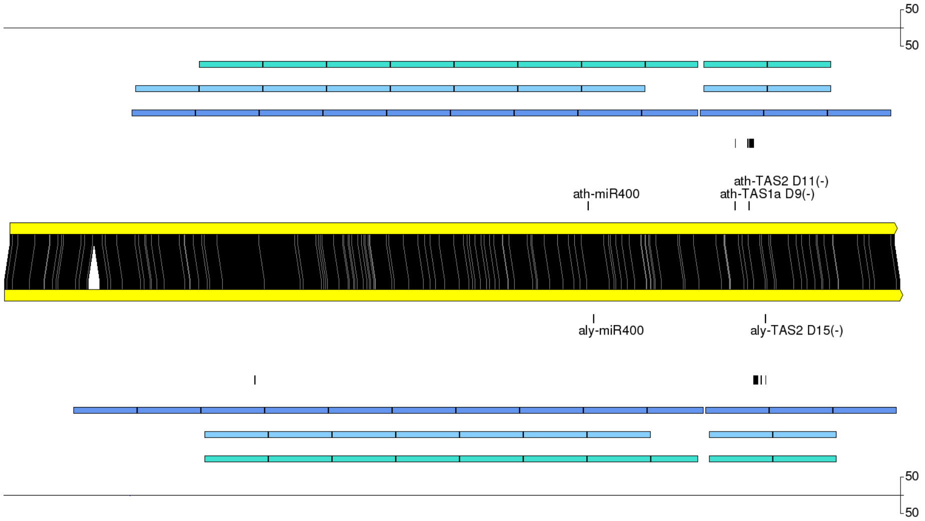
AT1G62670 / Al-RFL2_5036



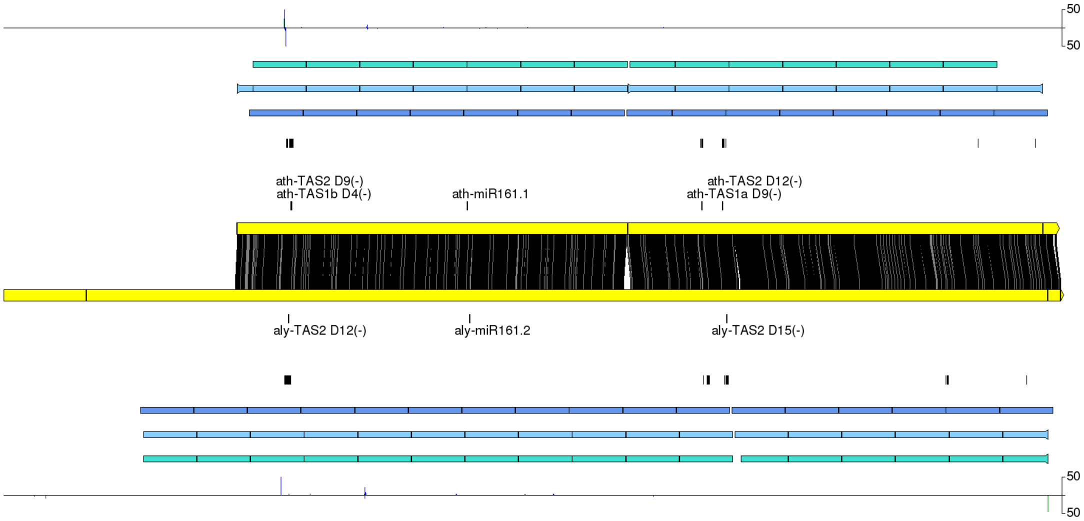
AT1G62680 / Al-RFL2_1553



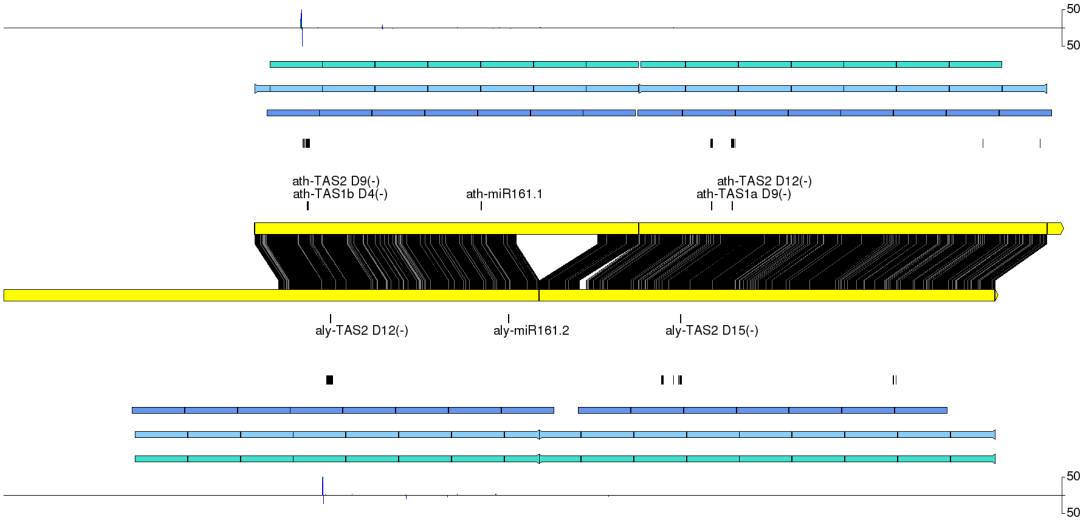
AT1G62720 / Al-RFL2_1521



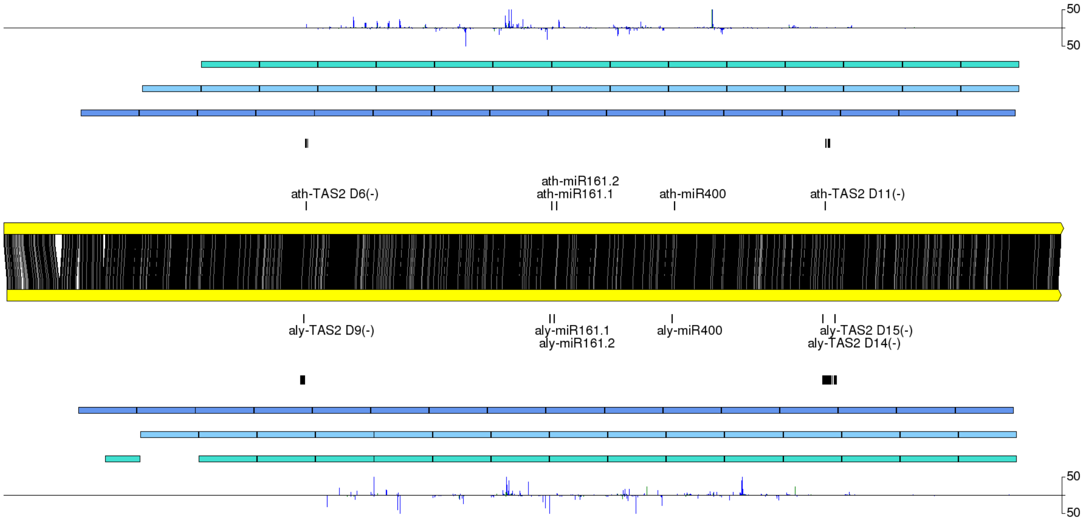
AT1G62860 / Al-RFL2_1458



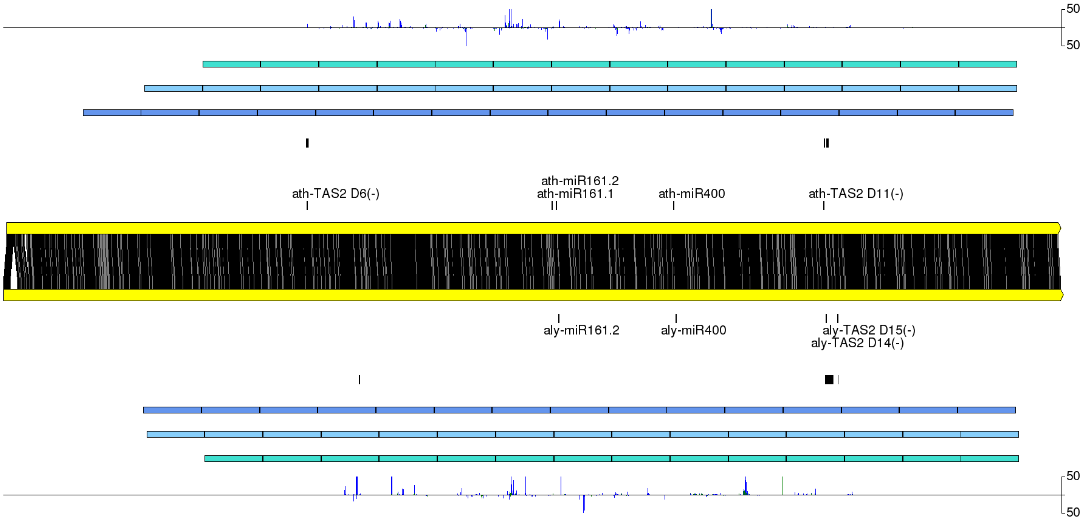
AT1G62860 / Al-RFL2_1453



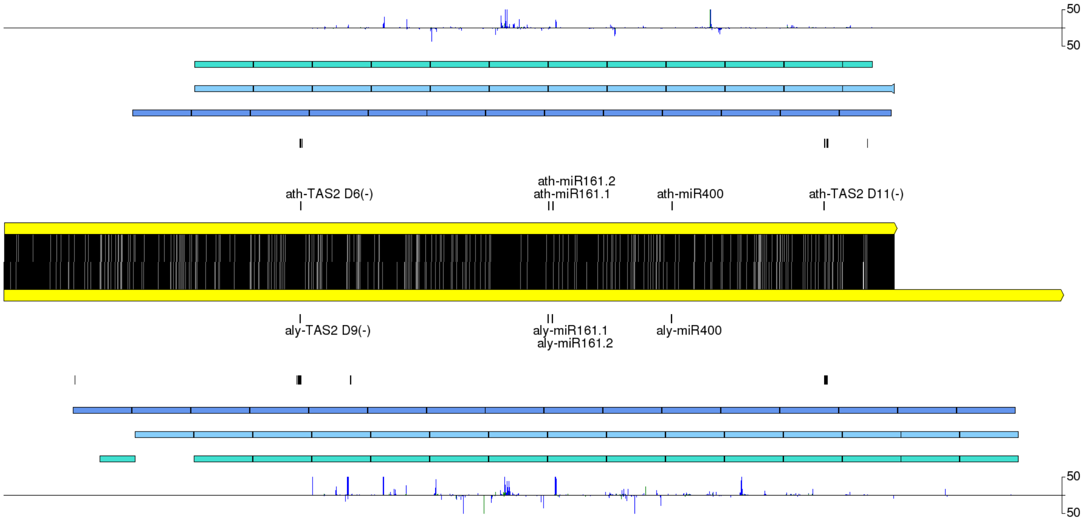
AT1G62910 / Al-RFL2_1417



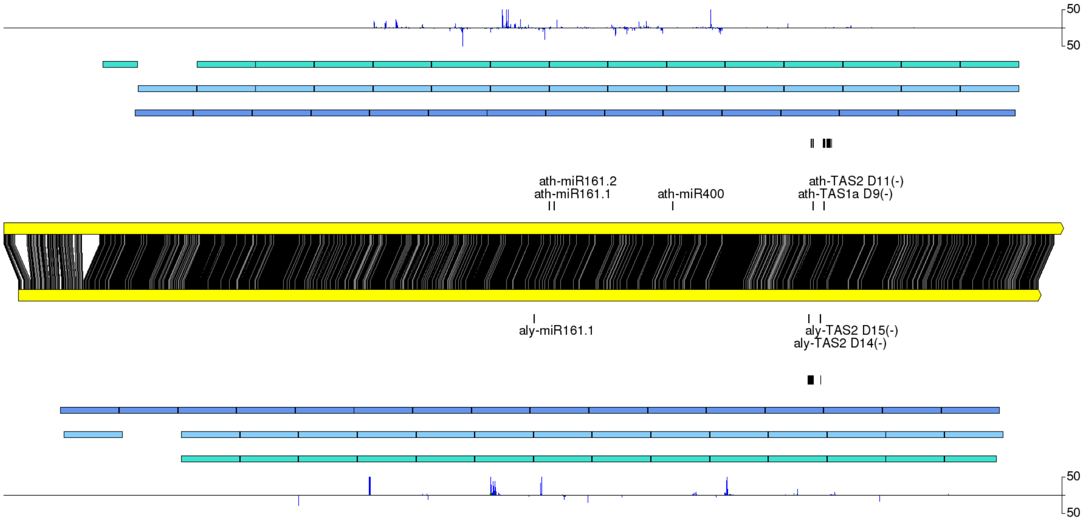
AT1G62910 / Al-RFL2_629



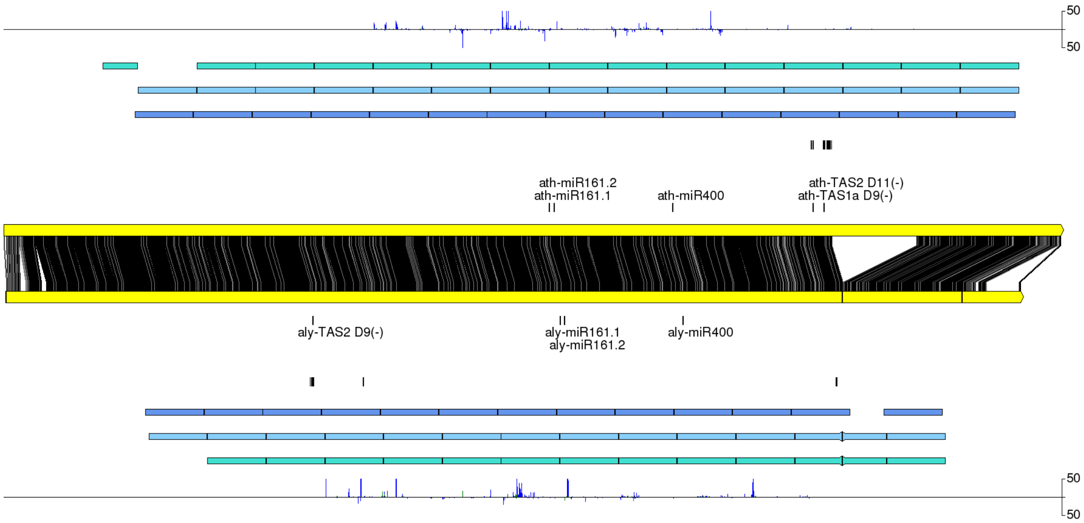
AT1G62914 / Al-RFL2_1415



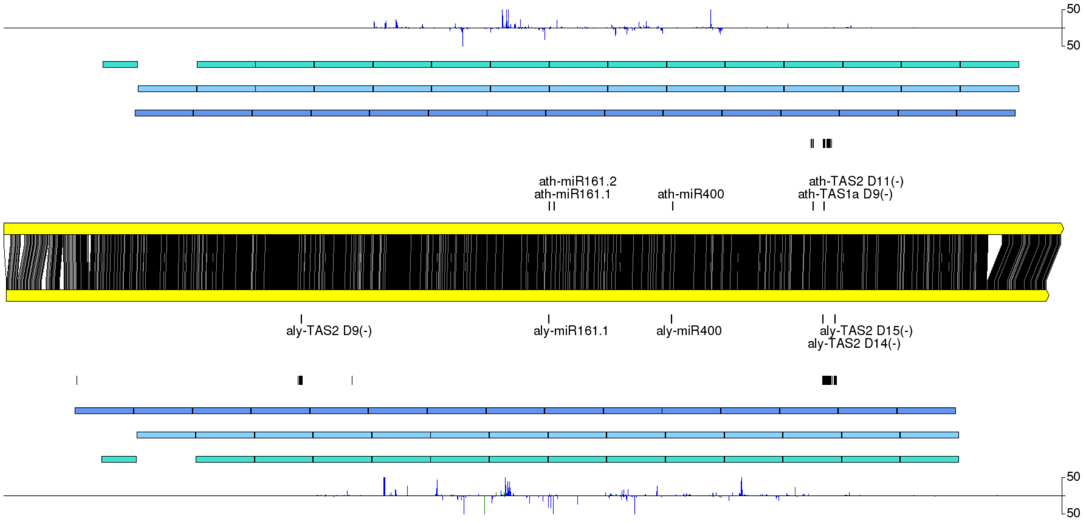
AT1G62930 / Al-RFL2_1393



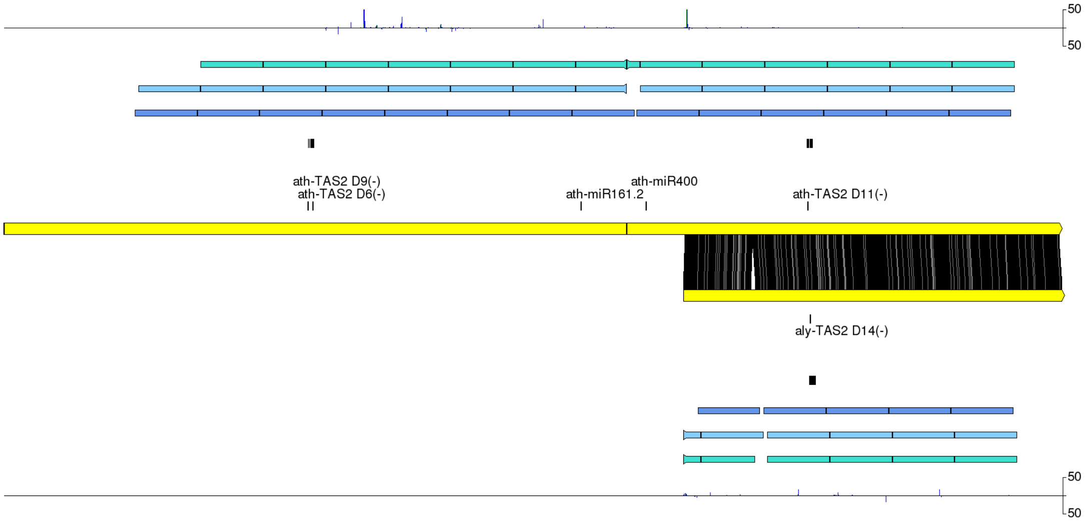
AT1G62930 / Al-RFL2_1395



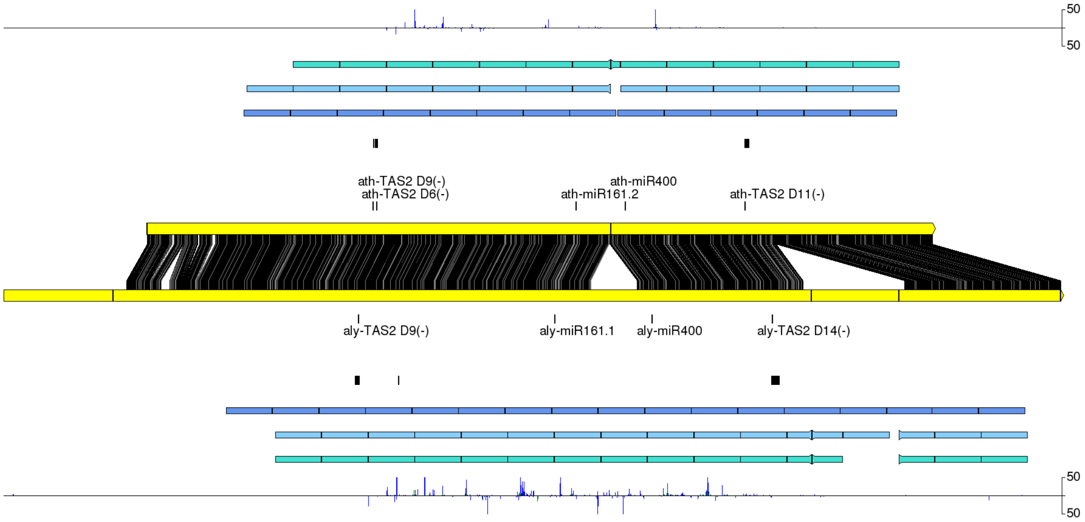
AT1G62930 / Al-RFL2_1411



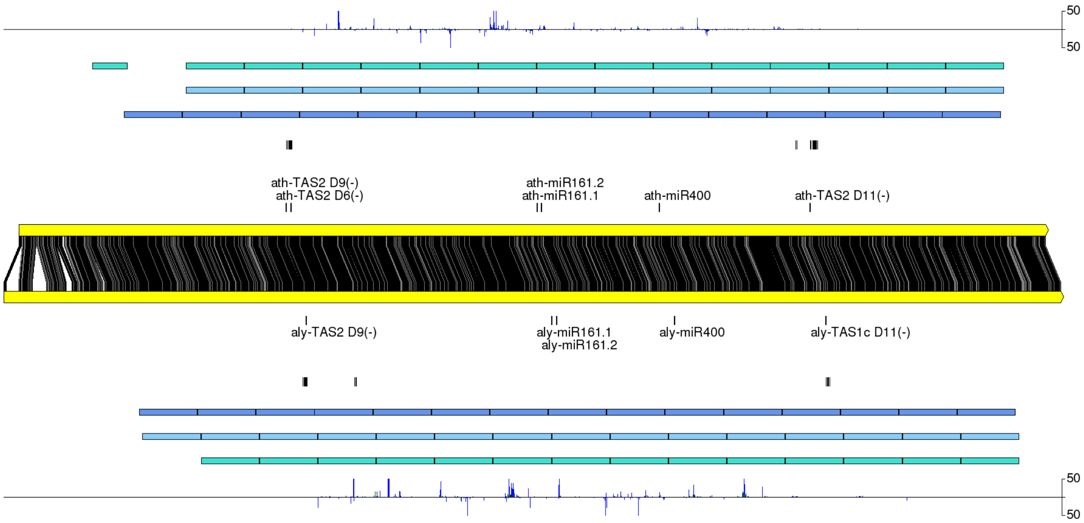
AT1G63070 / Al-RFL2_1211



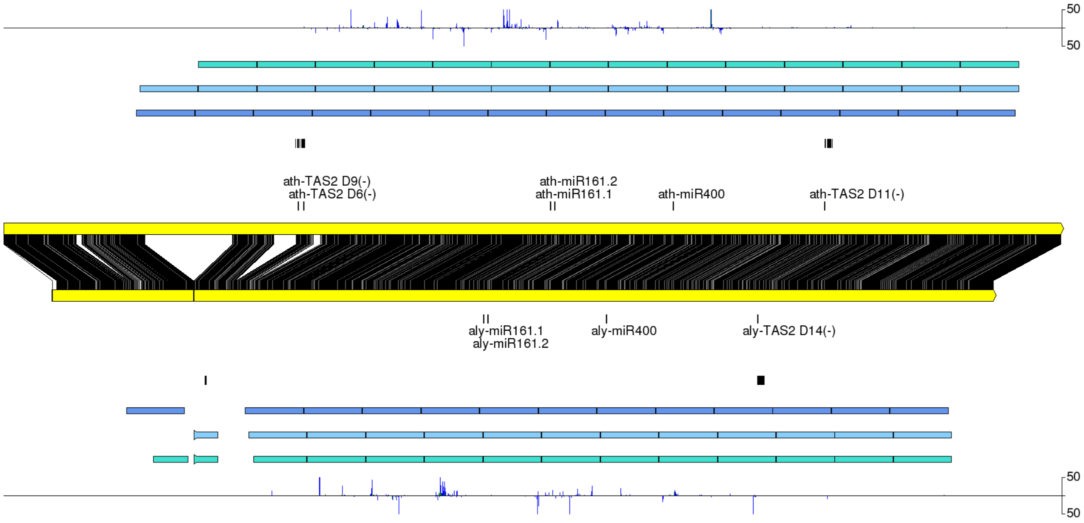
AT1G63070 / Al-RFL2_1215



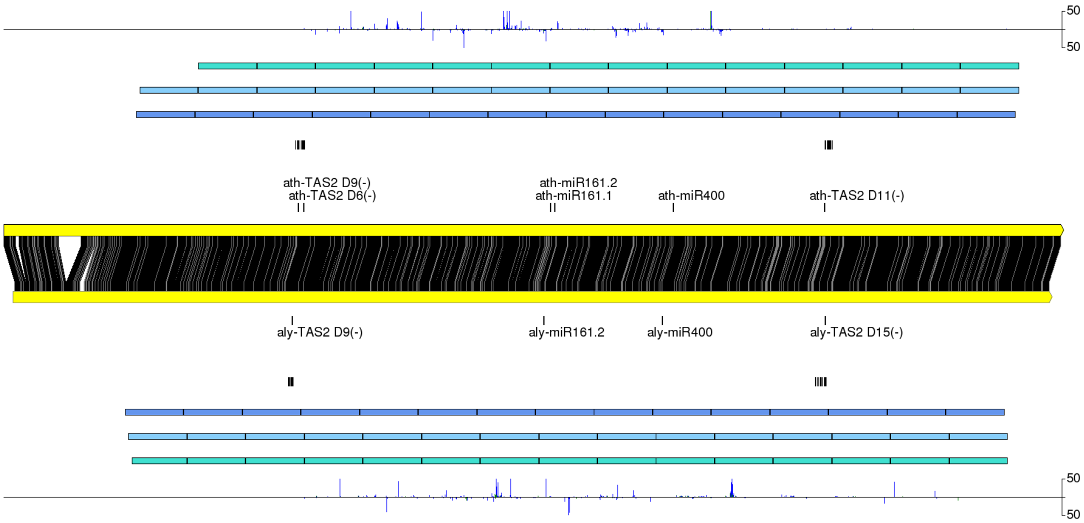
AT1G63080 / Al-RFL2_1205



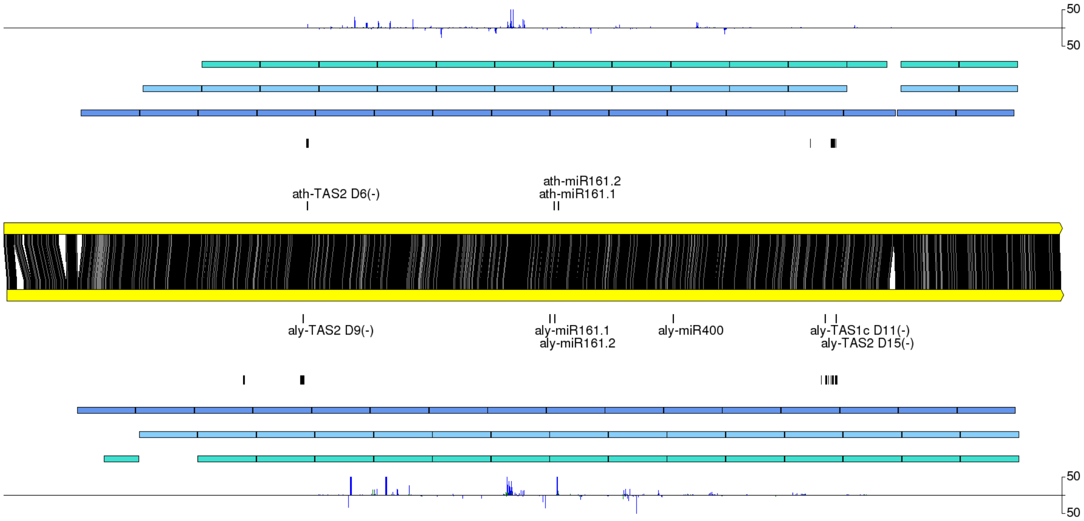
AT1G63130 / Al-RFL2_1159



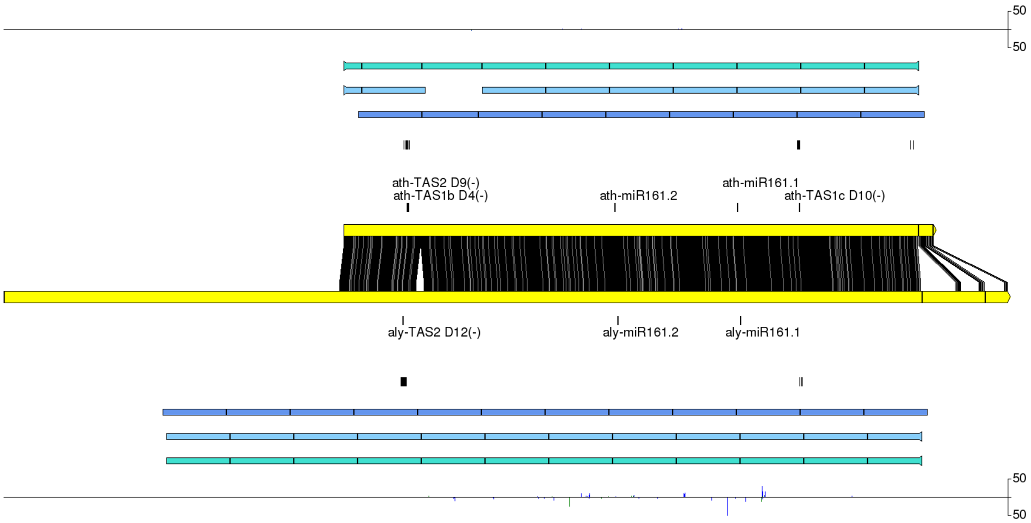
AT1G63130 / Al-RFL2_729



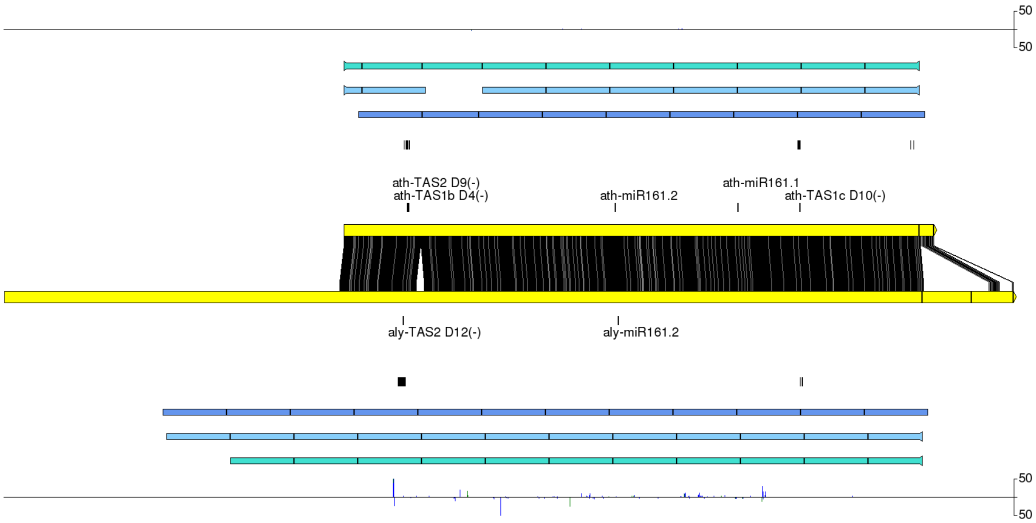
AT1G63150 / Al-RFL2_894



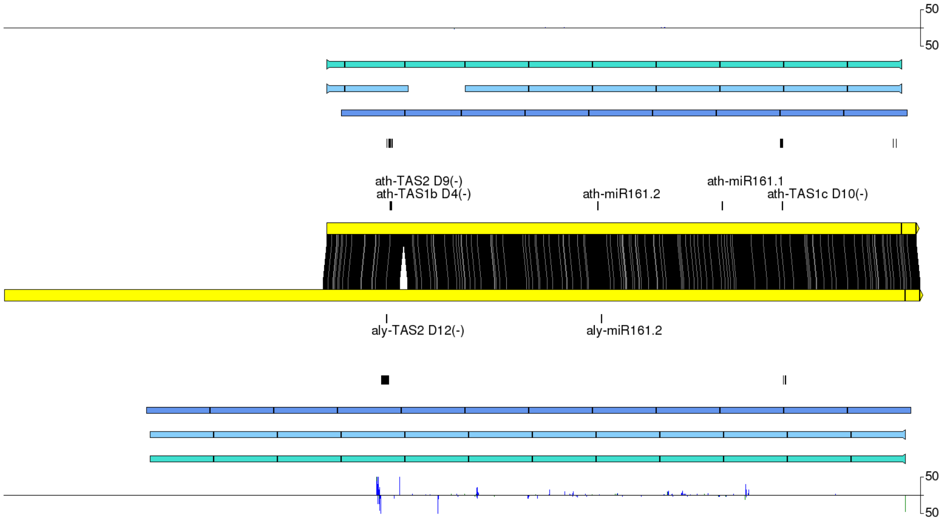
AT1G63230 / Al-RFL2_1054



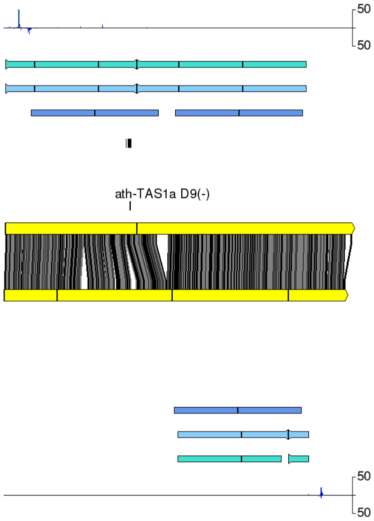
AT1G63230 / Al-RFL2_1082



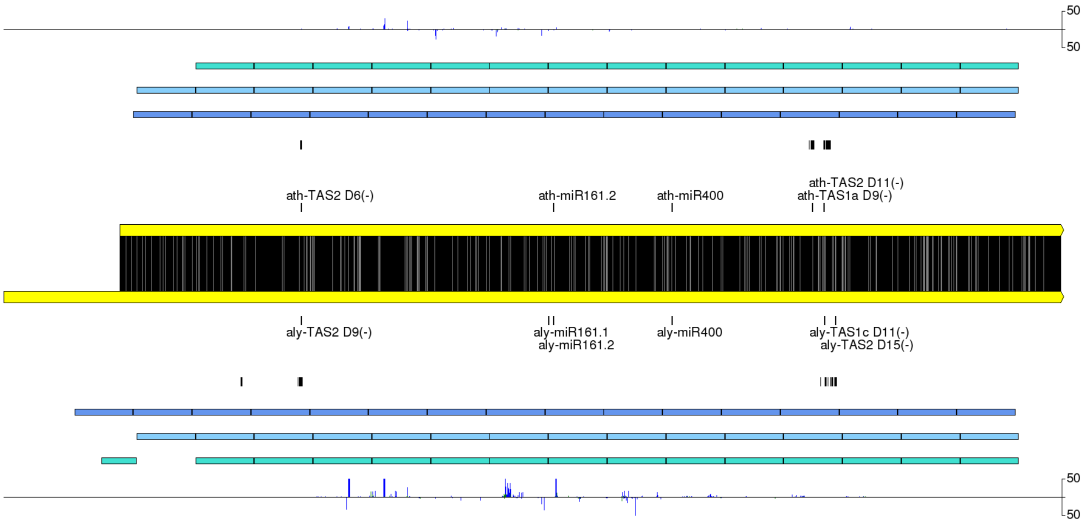
AT1G63230 / Al-RFL2_700



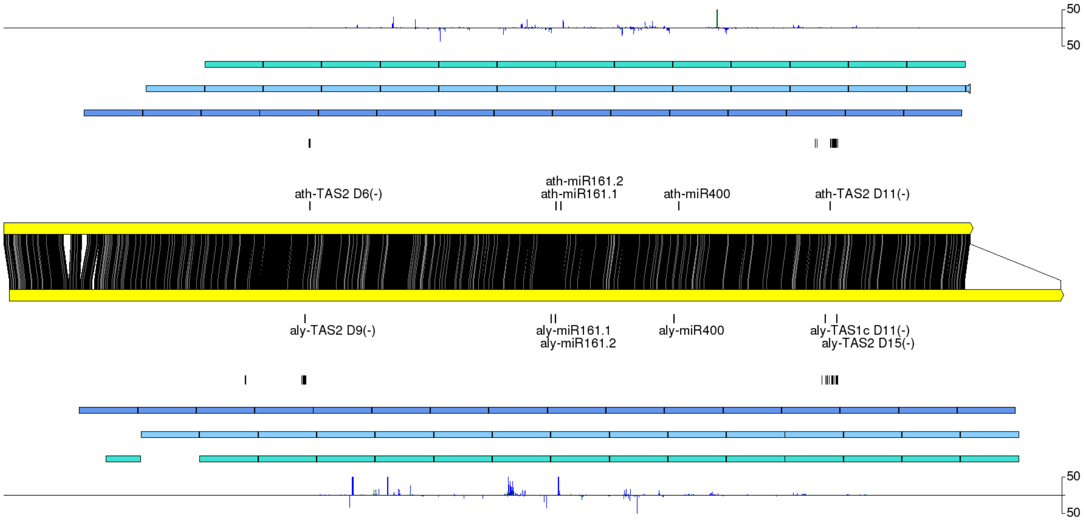
AT1G63320 / Al-RFL2_911



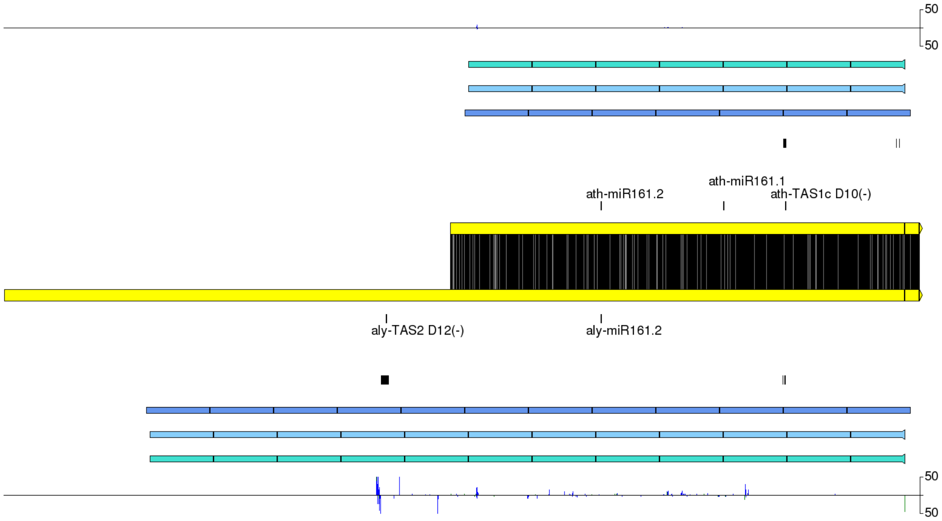
AT1G63330 / Al-RFL2_894



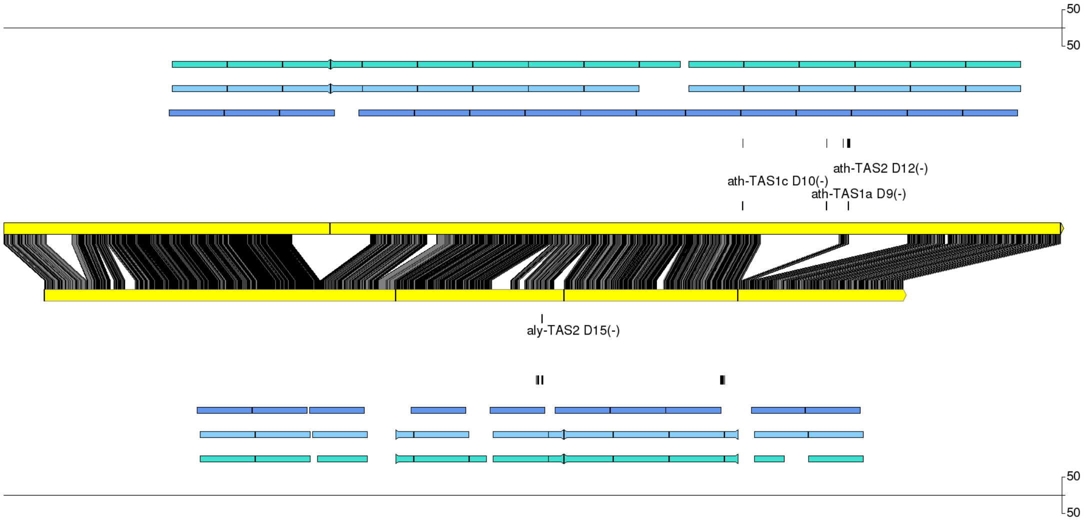
AT1G63400 / Al-RFL2_894



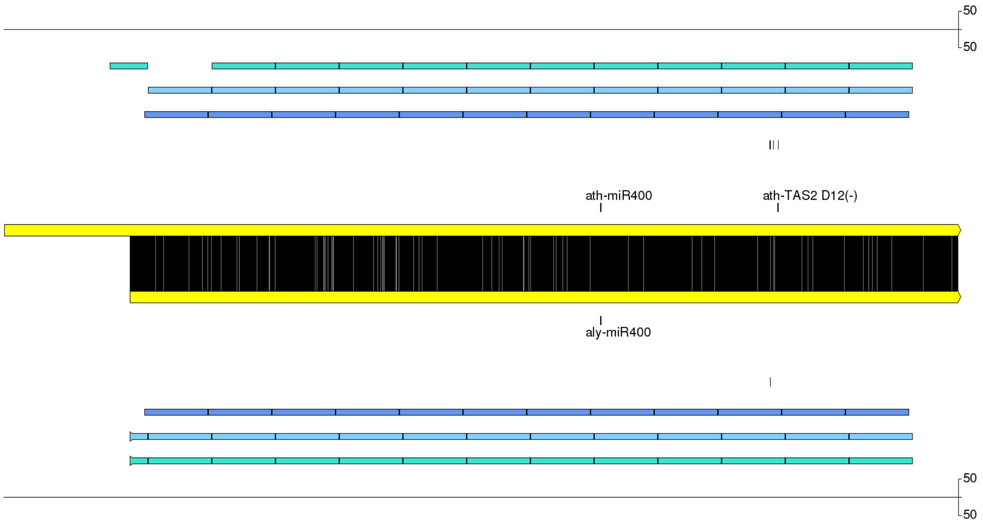
AT1G63630 / Al-RFL2_700



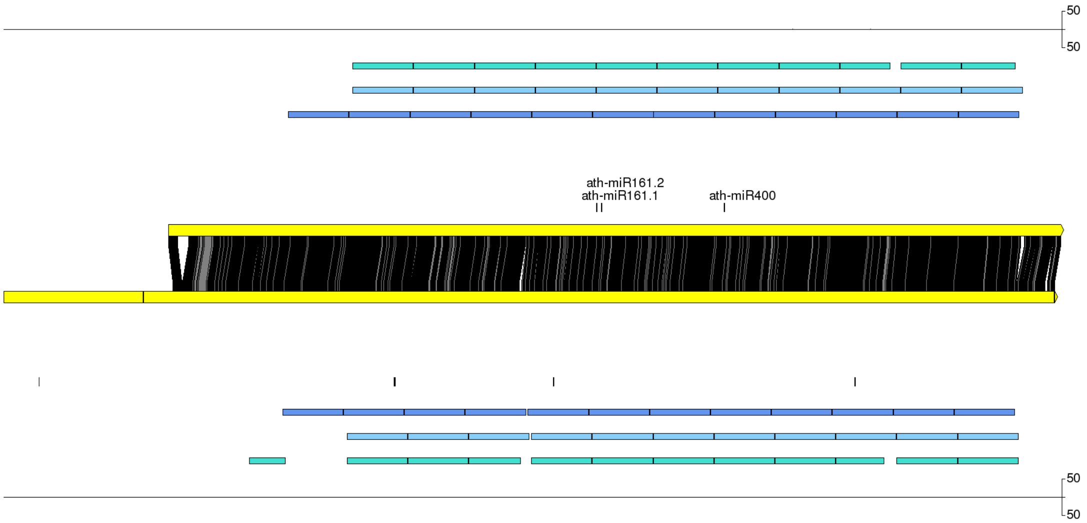
AT1G64100 / Al-RFL2_898



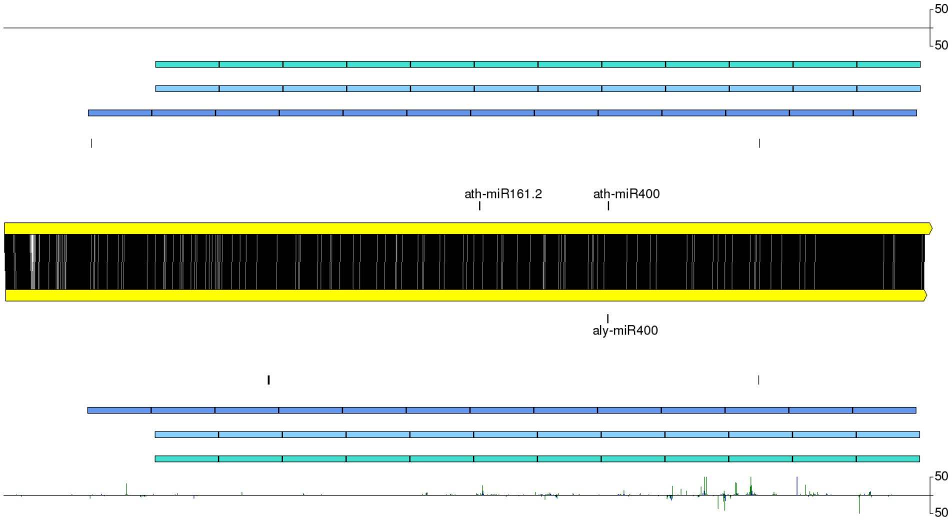
AT1G64580 / Al-RFL2_112



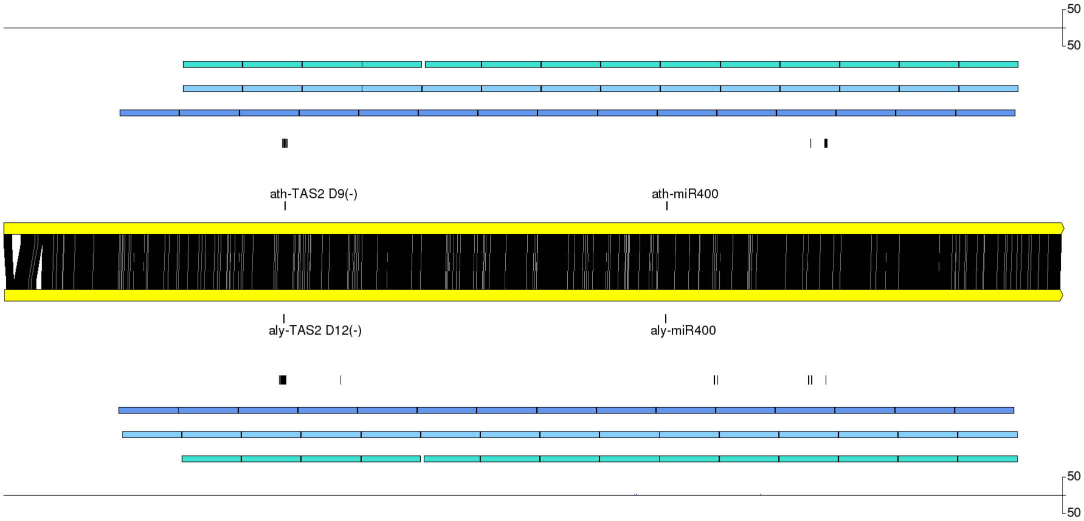
AT1G64583 / Al-RFL2_109



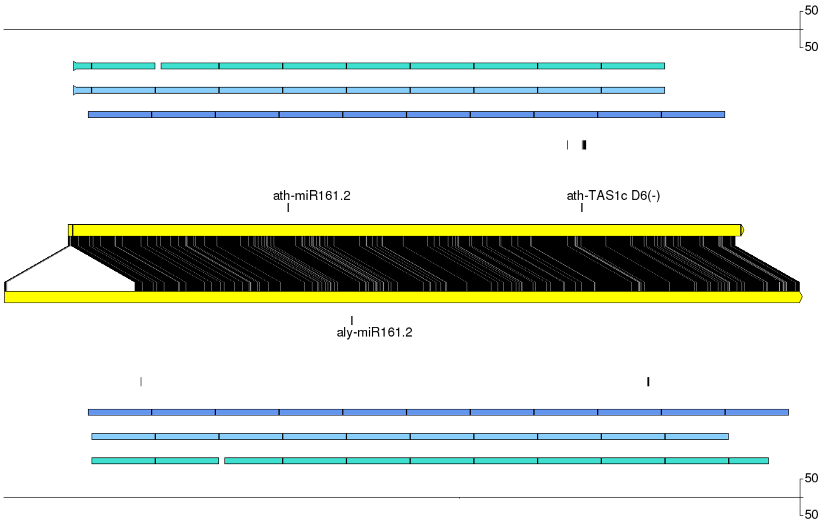
AT3G16710 / Al-RFL3_7093



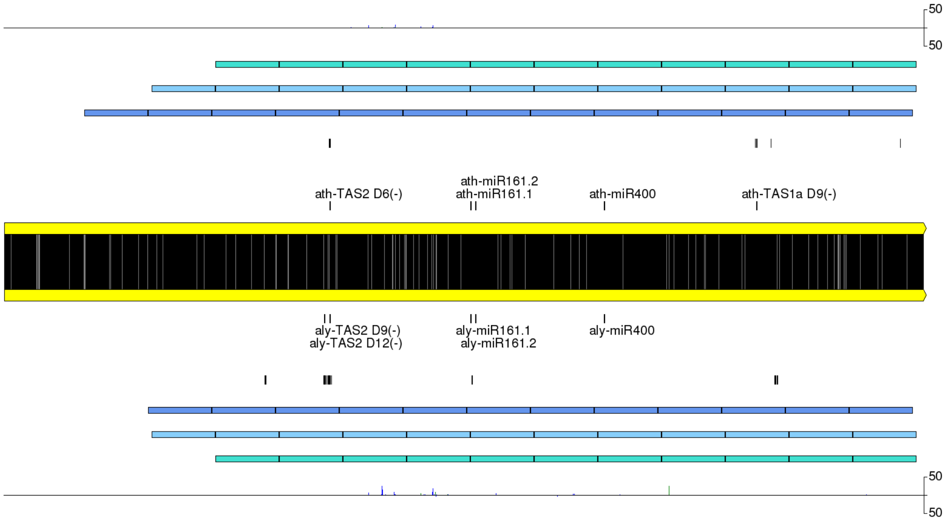
AT3G22470 / Al-RFL3_10057



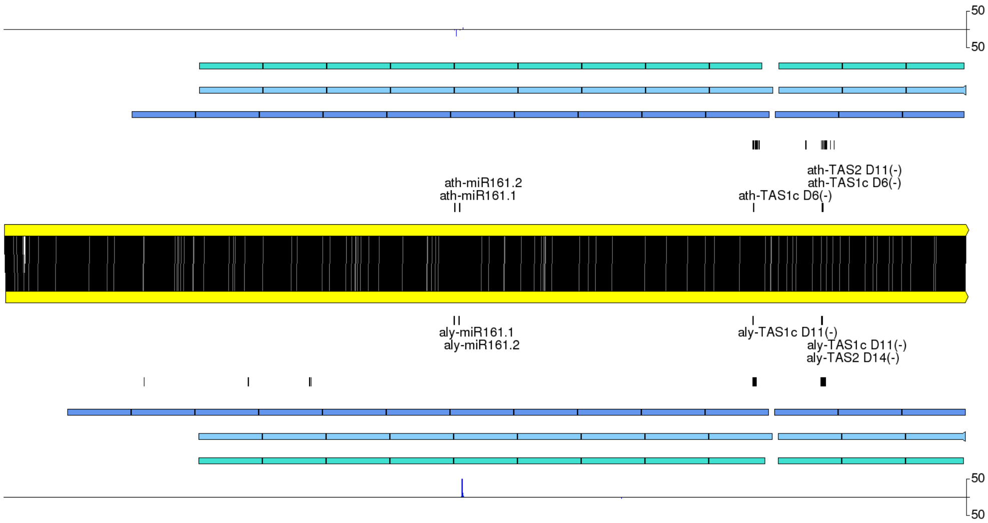
AT4G26800 / Al-RFL7_6500



AT5G16640 / Al-RFL6_6831



AT5G41170 / Al-RFL7_21693



Query	Target	% Identity	Align Len	Mismatches	Gaps	Query Start	Query End	Target Start	Target End	E-value	Bitscore
Al-RFL2_894	Al-RFL2_894	100	628	0	0	1	628	1	628	0	1252
Al-RFL2_894	Al-RFL2_629	83.94	635	95	2	1	628	1	635	0	1069
Al-RFL2_894	Al-RFL2_1205	83.76	628	96	3	7	628	5	632	0	1044
Al-RFL2_894	Al-RFL2_1201	83.23	632	93	5	1	628	1	623	0	1023
Al-RFL2_894	Al-RFL2_1415	82.99	623	101	2	10	628	6	627	0	1020
Al-RFL2_894	AT1G62910	81.16	637	106	6	1	628	1	632	0	1013
Al-RFL2_894	AT1G62670	80.25	628	120	1	5	628	3	630	0	1011
Al-RFL2_894	AT1G63130	80.54	627	116	3	7	628	5	630	0	1004
Al-RFL2_894	Al-RFL2_1417	79.78	628	122	2	5	628	3	629	0	993
Al-RFL2_894	AT1G62930	81.07	618	111	2	11	628	18	629	0	992
Al-RFL2_894	Al-RFL2_1411	80.9	623	106	3	10	628	6	619	0	990
Al-RFL2_894	Al-RFL2_1571	79.55	621	121	2	8	628	4	618	0	984
Al-RFL2_894	AT1G62590	77.95	635	132	3	1	628	1	634	0	983
Al-RFL2_894	Al-RFL2_1611	78.45	631	132	2	1	628	4	633	0	983
Al-RFL2_894	Al-RFL2_1561	79.45	623	123	2	10	628	6	627	0	980
Al-RFL2_894	Al-RFL2_729	79.77	618	116	2	12	628	9	618	0	978
Al-RFL2_894	AT1G63080	78.16	618	130	2	11	628	2	614	0	971
Al-RFL2_894	Al-RFL2_1557	81.34	595	107	1	38	628	26	620	0	968
Al-RFL2_894	Al-RFL2_1575	77.07	628	136	4	5	628	3	626	0	959
Al-RFL2_894	Al-RFL2_1215	70.34	708	124	7	1	628	93	794	0	942
Al-RFL2_894	AT1G63150	76.14	637	135	7	1	628	1	629	0	936
Al-RFL2_894	AT1G63400	79.41	578	111	3	1	571	1	577	0	921
Al-RFL2_894	AT1G63330	80.86	559	107	0	70	628	1	559	0	917

Appendix 3. BLAST alignment of Al-RFL2_894 to other RFL sequences. Al-RFL2_894 peptide sequence was aligned to other RFL peptide sequences from *A. thaliana* and *A. lyrata* using BLASTP*.* Results sorted by bitscore. Orthologs of *Al-RFL2_894* are shaded.

Sorted by Bitscore

Query	Target	% Identity	Align Len	Mismatches	Gaps	Query Start	Query End	Target Start	Target End	E-value	Bitscore
RsRfo	Al-RFL2_1458	61.31	716	210	6	1	682	1	683	0	826
RsRfo	AT1G62860	65.6	532	154	3	175	682	1	527	0	679
RsRfo	Al-RFL2_464	59.65	575	200	5	1	568	1	550	0	668
RsRfo	AT1G64100	54.81	644	252	7	20	644	43	666	0	655
RsRfo	Al-RFL2_700	65.16	488	161	2	15	494	8	494	0	635
RsRfo	Al-RFL2_1054	63.78	508	165	4	1	494	1	503	0	629
RsRfo	Al-RFL2_1453	47.66	705	298	7	1	681	1	658	3.00E-179	617
RsRfo	Al-RFL1_5184	49.47	570	275	3	49	618	43	599	2.00E-159	551
RsRfo	Al-RFL1_5127	48.6	570	280	3	49	618	43	599	7.00E-158	546
RsRfo	Al-RFL2_894	47.69	585	290	4	60	644	60	628	3.00E-157	544
RsRfo	AT1G12300	47.45	609	297	5	20	618	20	615	4.00E-157	543
RsRfo	AT1G62930	49.29	564	274	2	60	623	61	612	4.00E-157	543
RsRfo	Al-RFL2_729	45.99	574	297	3	46	619	37	597	1.00E-150	522

Sorted by Percent Identity

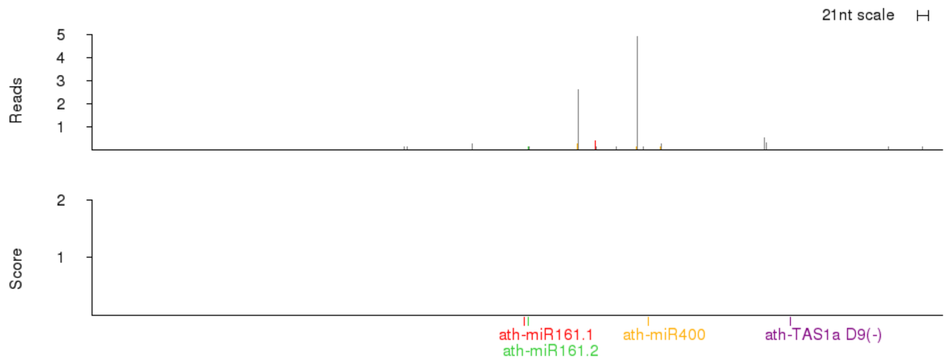
Query	Target	% Identity	Align Len	Mismatches	Gaps	Query Start	Query End	Target Start	Target End	E-value	Bitscore
RsRfo	AT1G62860.1	65.6	532	154	3	175	682	1	527	0	679
RsRfo	AT1G63230.1	63.24	321	113	2	175	495	1	316	1.00E-116	409
RsRfo	AT1G63630.1	61.54	247	95	0	249	495	4	250	2.00E-87	312
RsRfo	AT1G64100.1	54.81	644	252	7	20	644	43	666	0	655
RsRfo	AT1G62914.1	50.63	478	223	3	50	527	50	514	7.00E-139	483
RsRfo	AT1G62930.1	49.29	564	274	2	60	623	61	612	4.00E-157	543
RsRfo	AT1G63400.1	49.16	537	260	3	36	572	43	566	1.00E-151	526
RsRfo	AT1G12620.1	49.12	570	277	3	49	618	43	599	2.00E-156	541
RsRfo	AT1G63130.1	47.58	559	281	2	60	618	62	608	2.00E-152	528
RsRfo	AT1G12300.1	47.45	609	297	5	20	618	20	615	4.00E-157	543

Appendix 4. BLAST alignment of Radish Rfo to other RFL sequences. Al-RFL2_729 peptide sequence was aligned to other RFL peptide sequences from *A. thaliana* and *A. lyrata* using BLASTP. The *A. lyrata* gene *Al-RFL2_729* is collinear with the Radish *Rfo* gene but aligns poorly by BLAST. Alignments in top section sorted by bitscore, lower section sorted by percent identity. The best alignment to Al-RFL2_729 is shown at the end of the top section after a blank line. Bottom section includes the product of *AT1G63630*, which is within 40kb of the Radish *Rfo* orthologous region in *A. thaliana*.

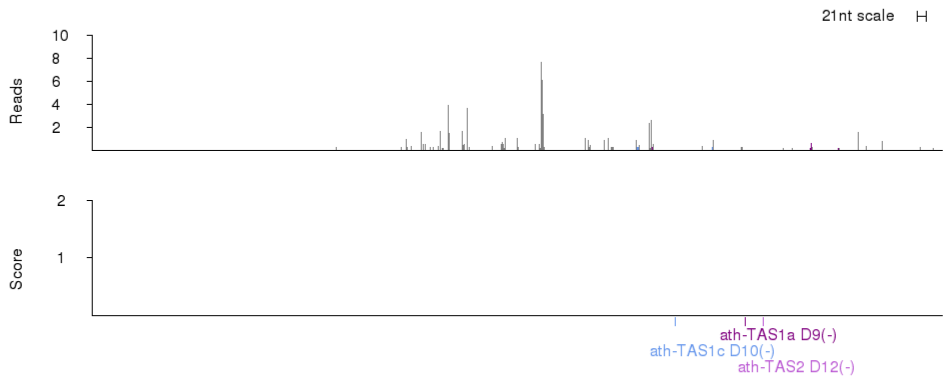
Appendix 5. *RFL* transcript phasing plots.

Small RNA density and phasing plots for those *RFL* transcripts not shown in the main document, as described in Figures 11 and 18. *A. thaliana* transcripts are shown prior to *A. lyrata* transcripts. Plots are ordered by position in the genome.

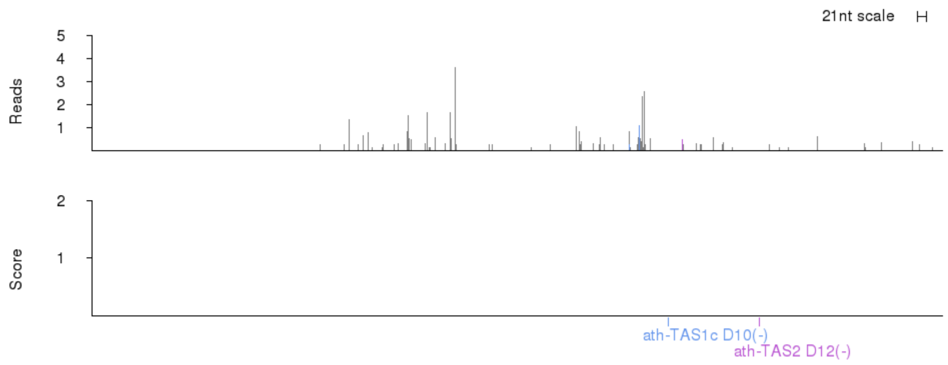
AT1G06580



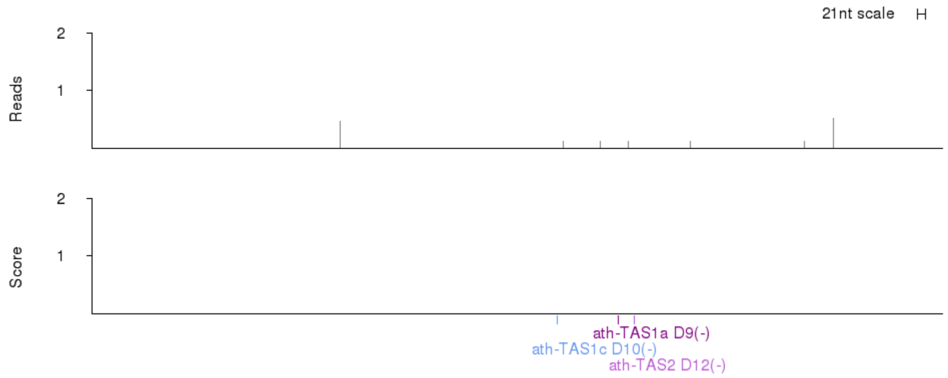
AT1G12300



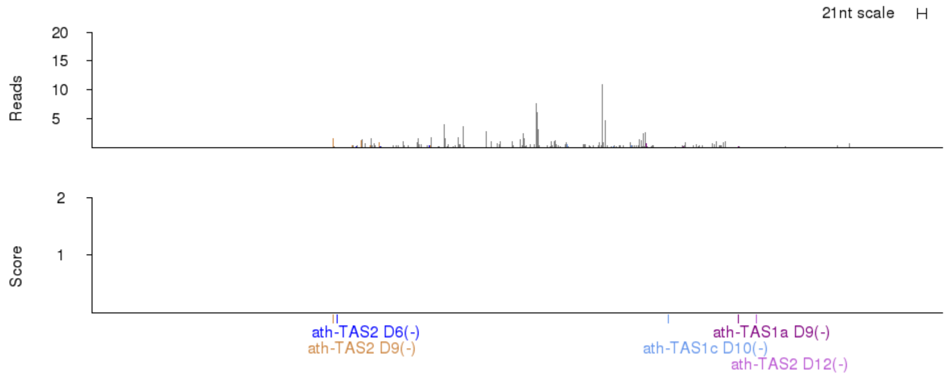
AT1G12620



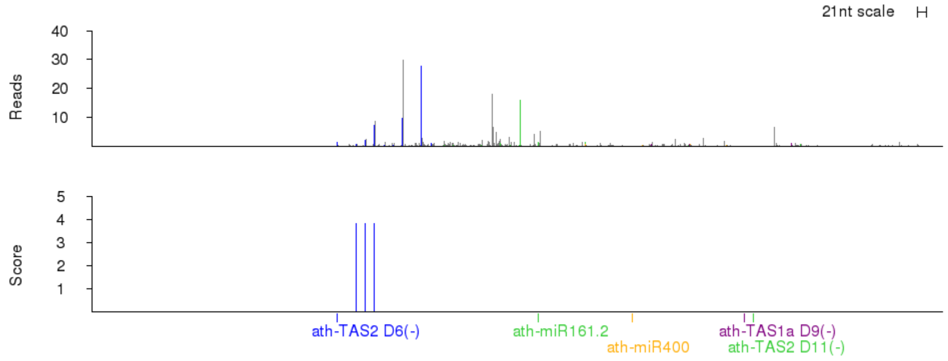
AT1G12700



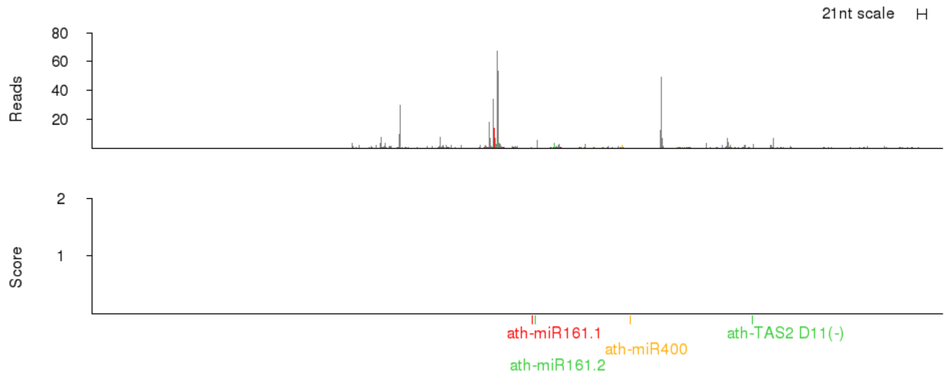
AT1G12775



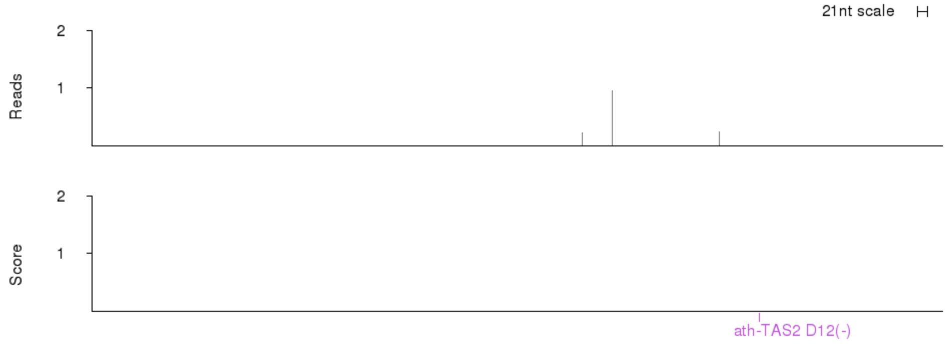
AT1G62590



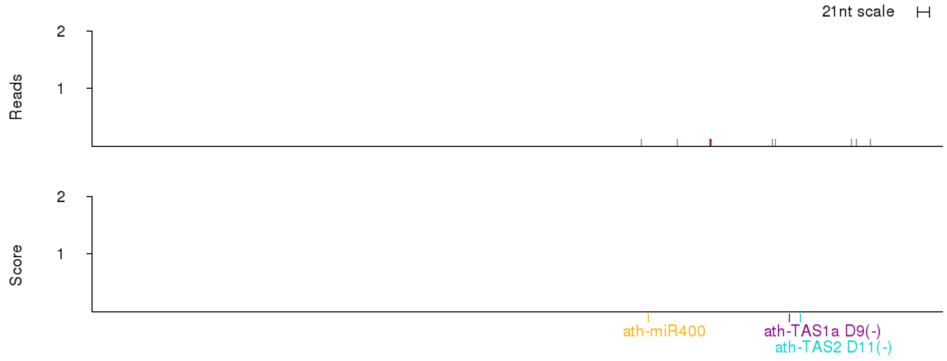
AT1G62670



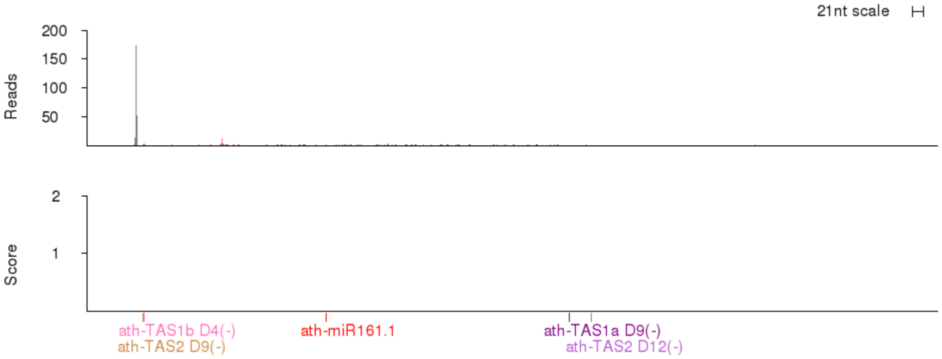
AT1G62680



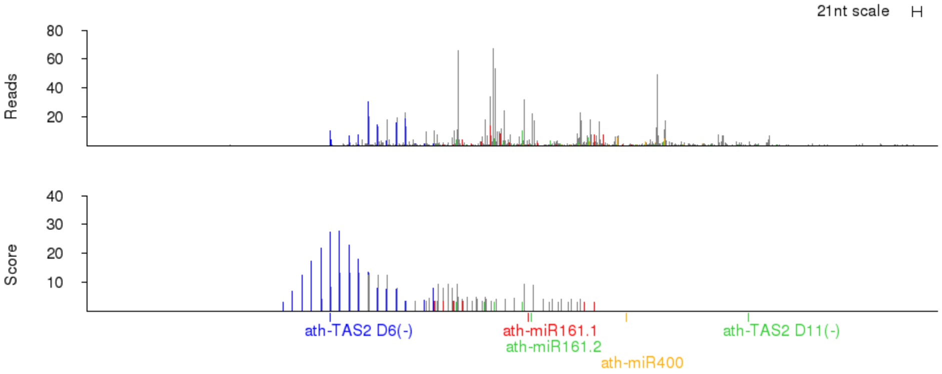
AT1G62720



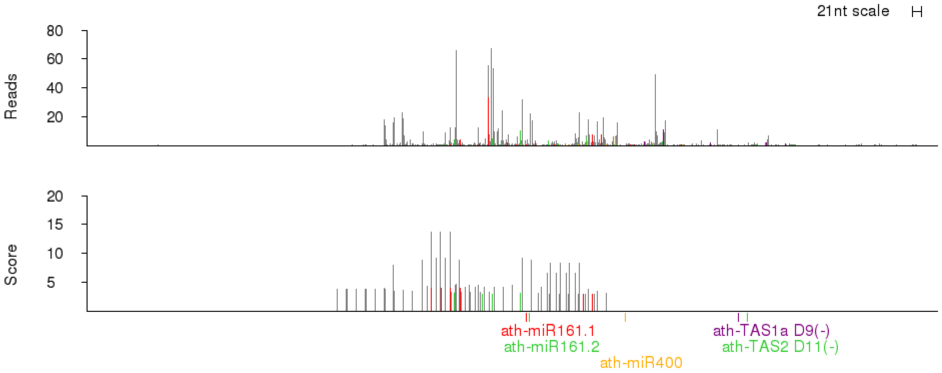
AT1G62860



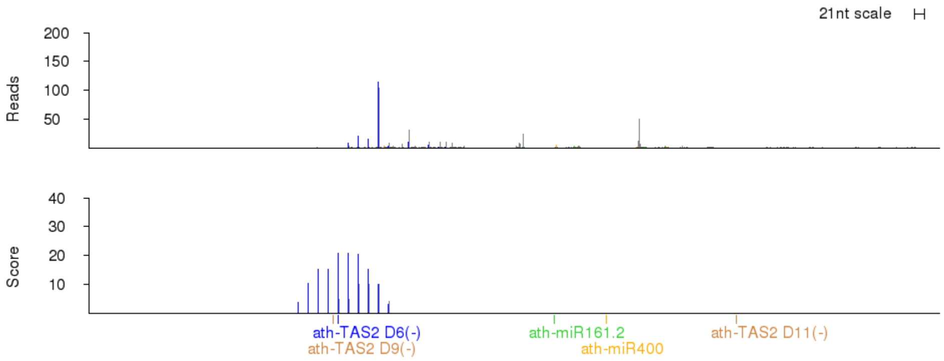
AT1G62910



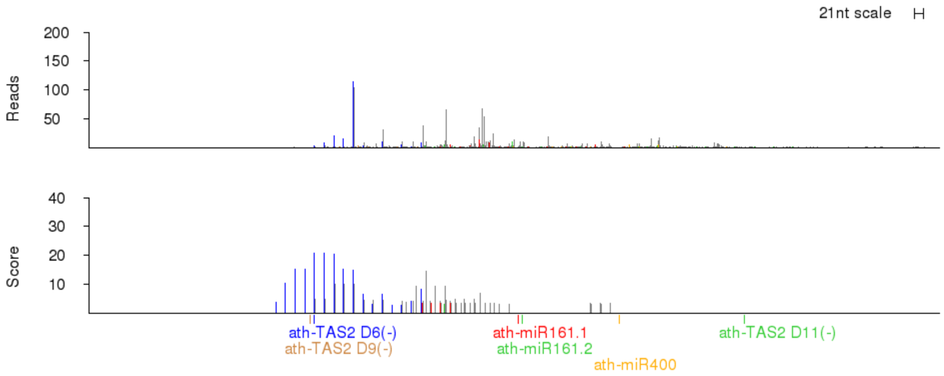
AT1G62930



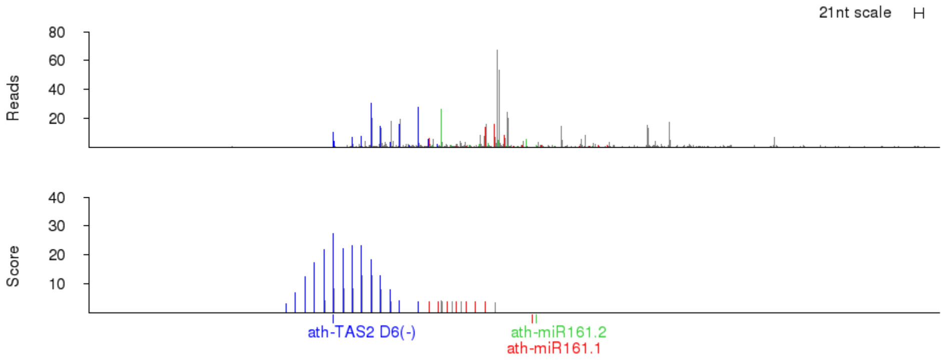
AT1G63070



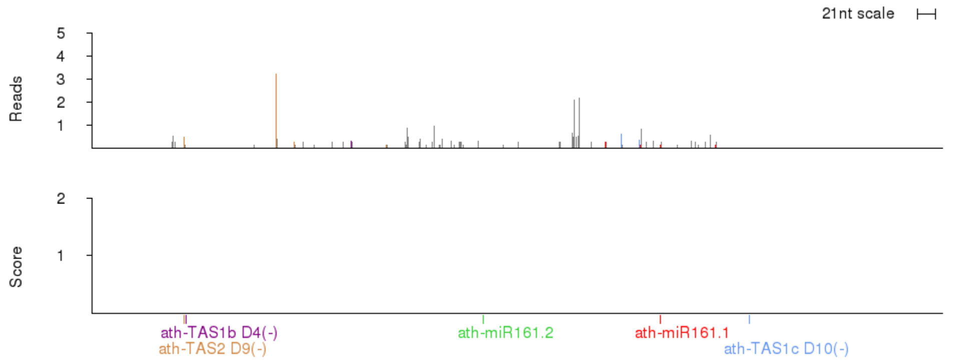
AT1G63080



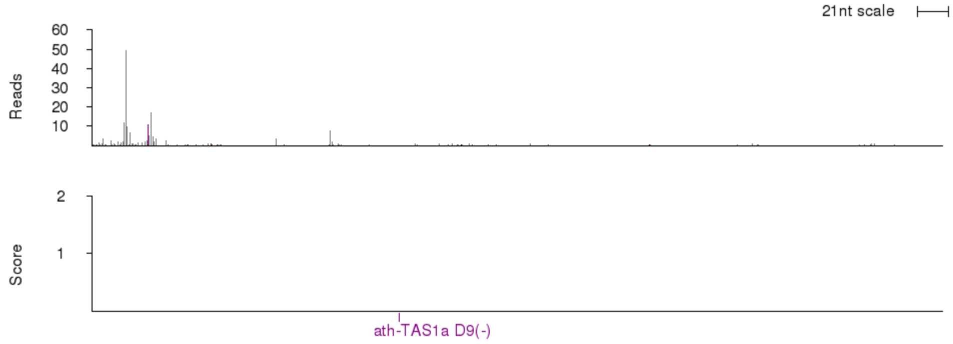
AT1G63150



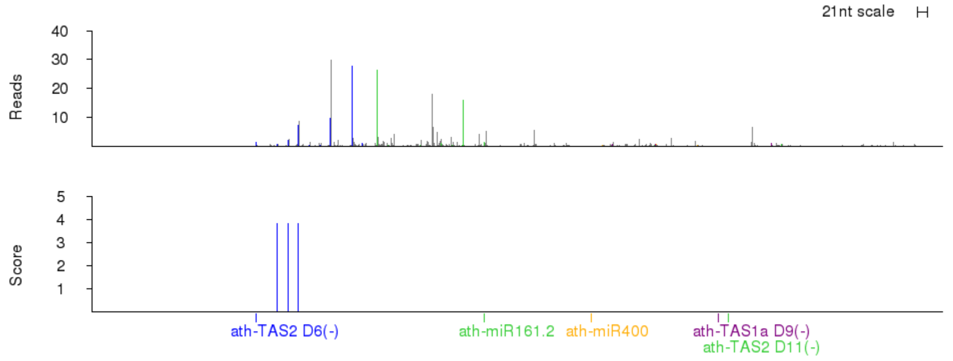
AT1G63230



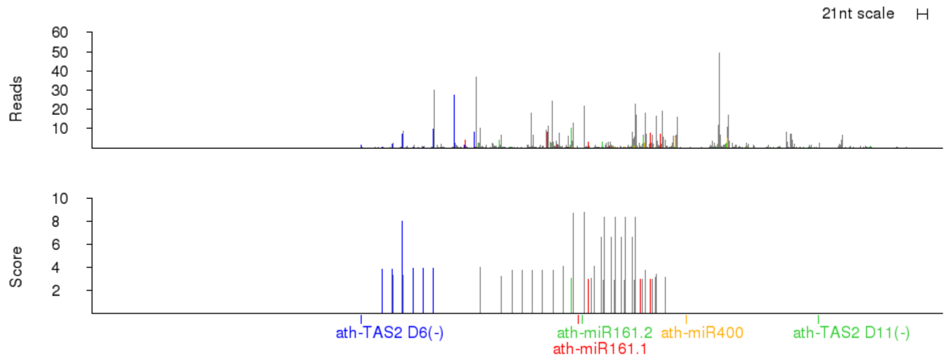
AT1G63320



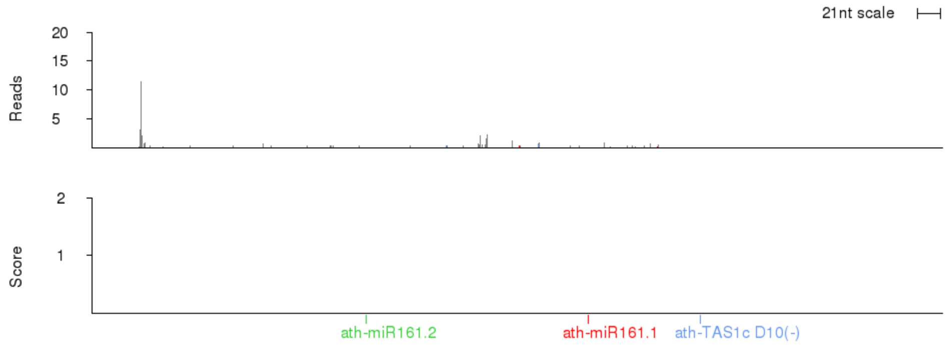
AT1G63330



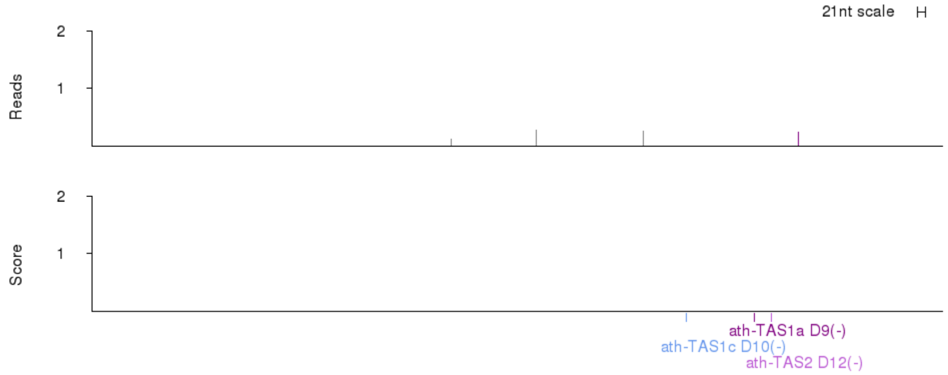
AT1G63400



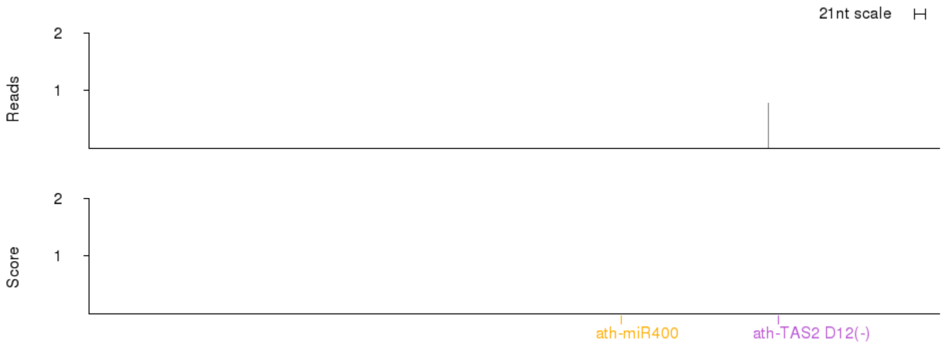
AT1G63630



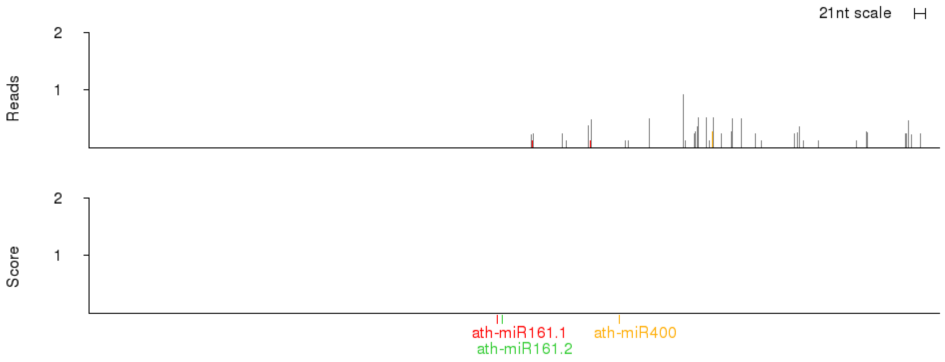
AT1G64100



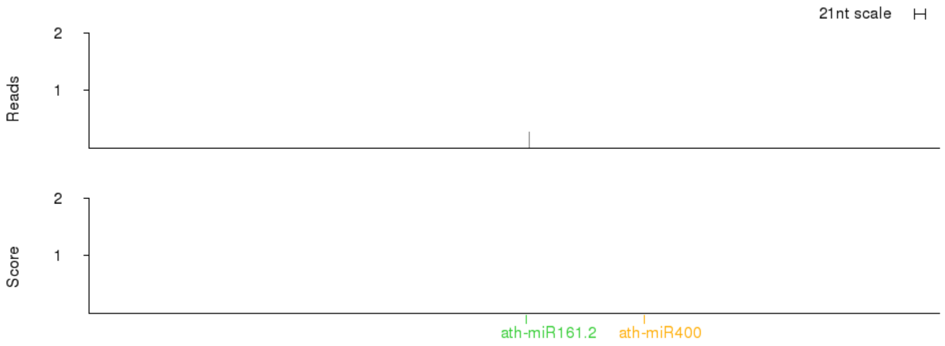
AT1G64580



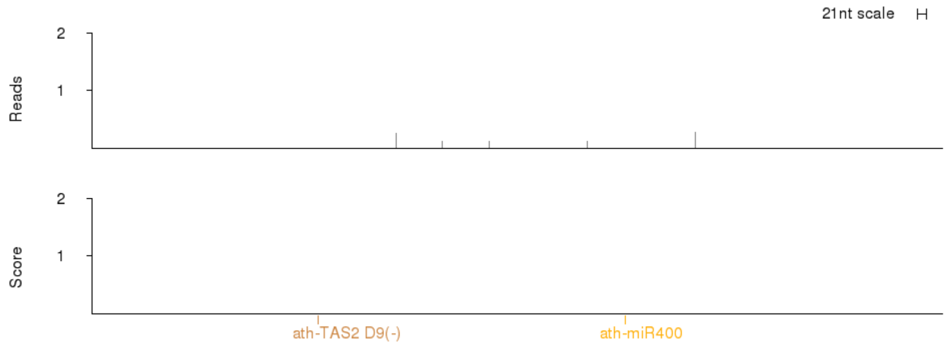
AT1G64583



AT3G16710



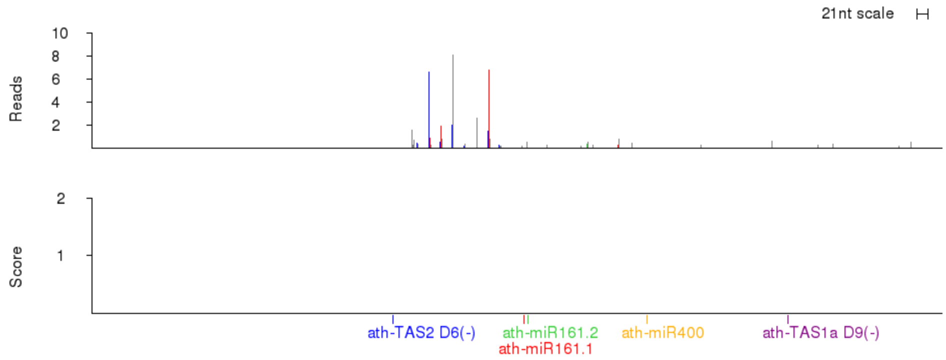
AT3G22470



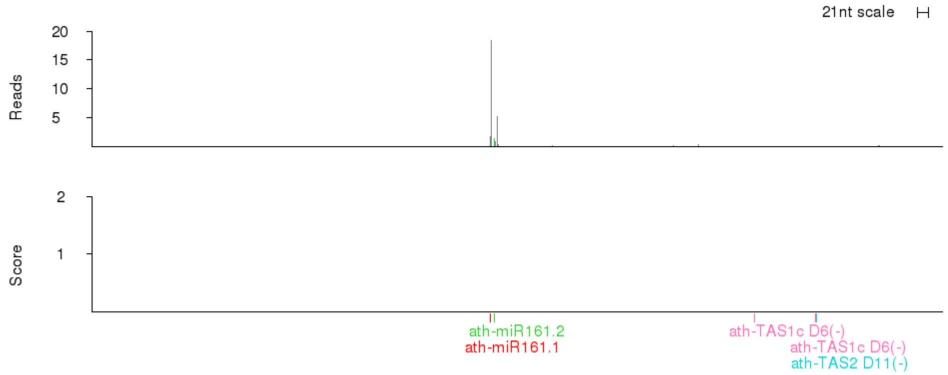
AT4G26800

There were no small RNA reads that aligned to this transcript

AT5G16640



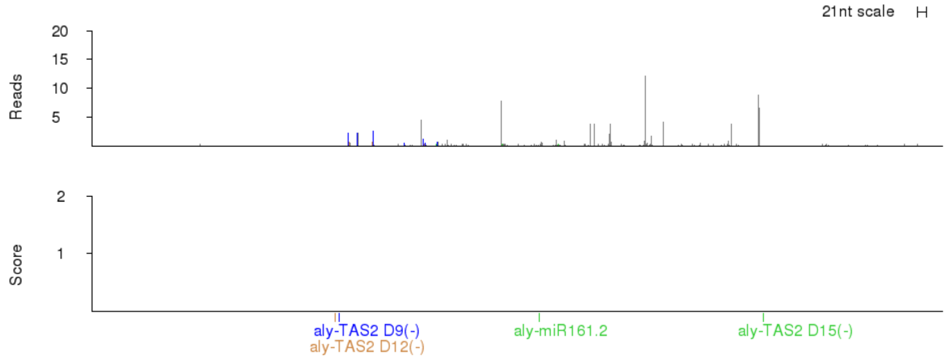
AT5G41170



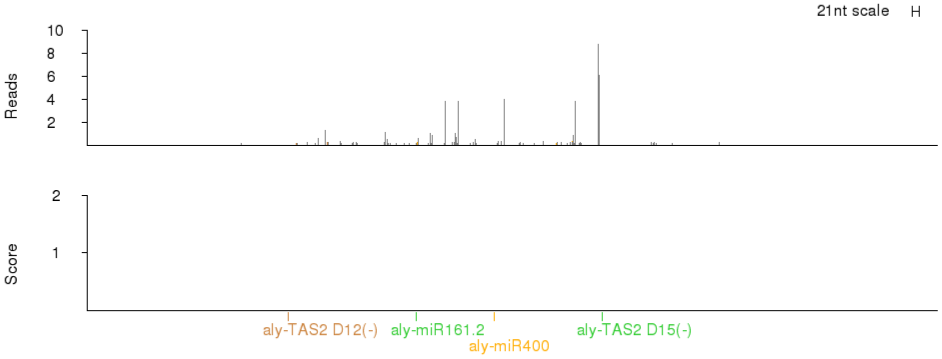
Al-RFL1_2385

There were no small RNA reads that aligned to this transcript

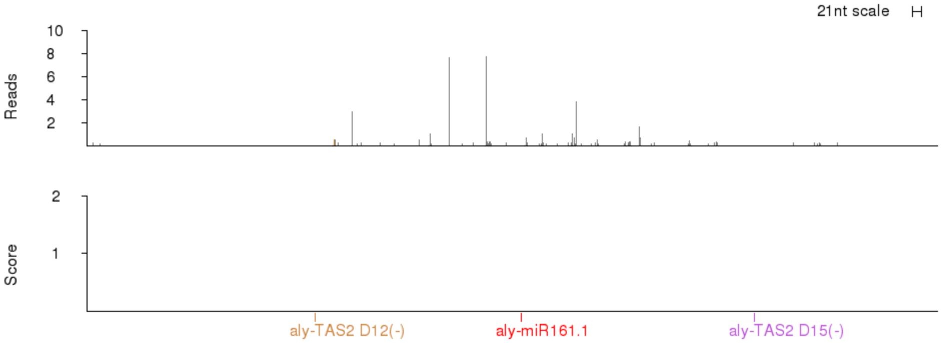
Al-RFL1_4995



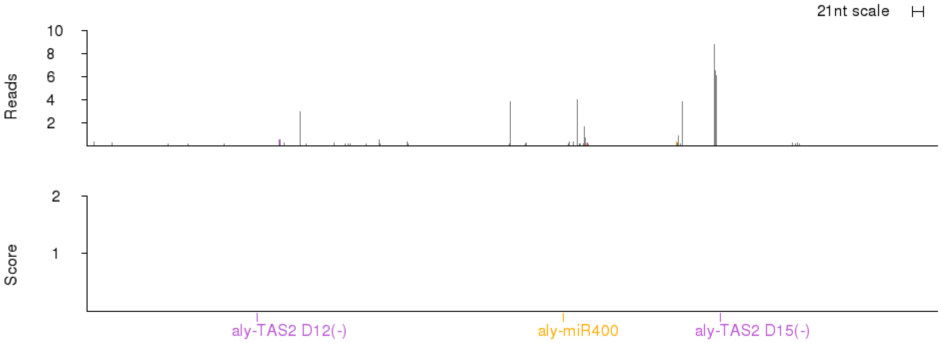
Al-RFL1_4999



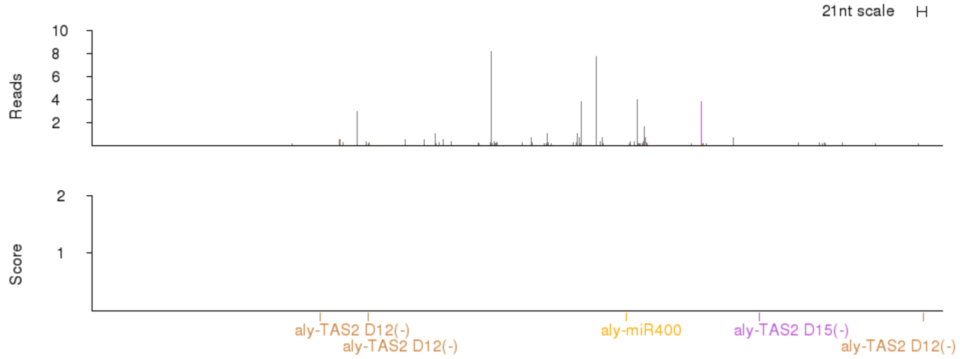
Al-RFL1_5122



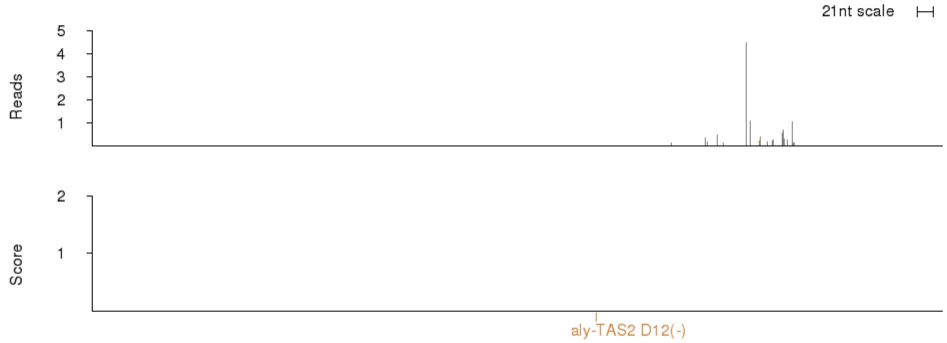
Al-RFL1_5125



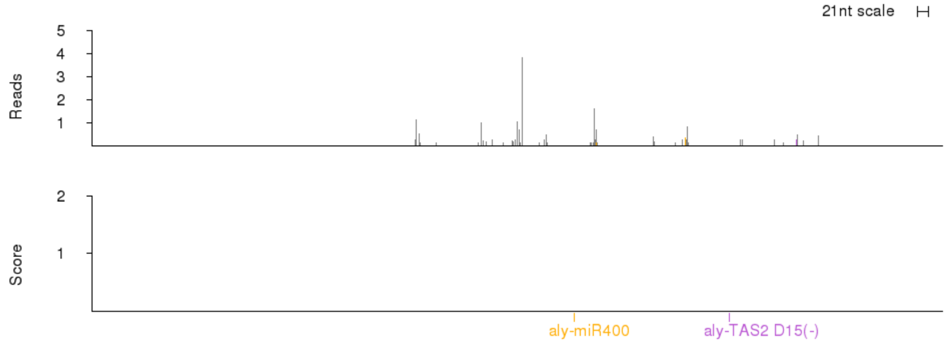
Al-RFL1_5127



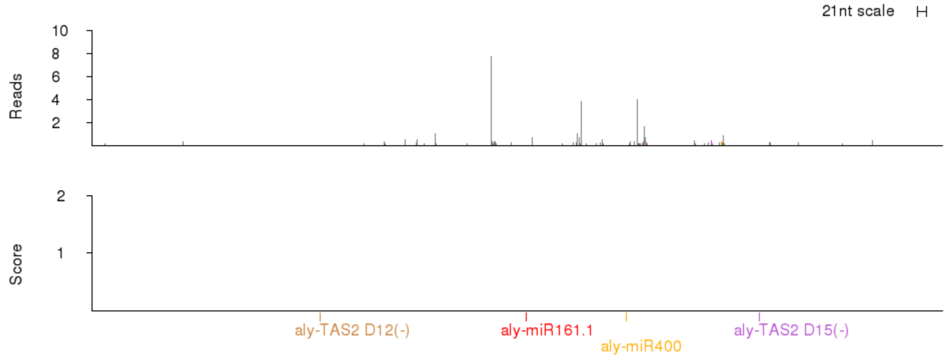
Al-RFL1_5179



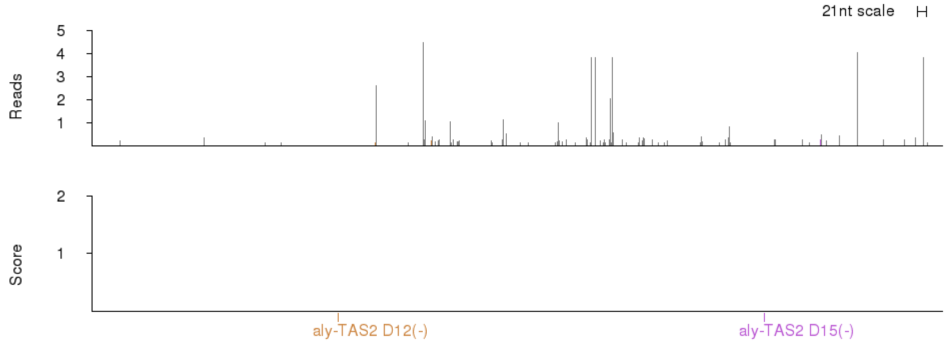
Al-RFL1_5181



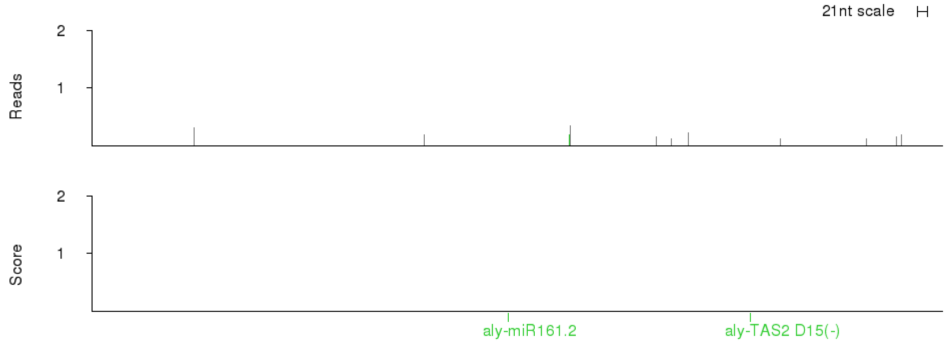
Al-RFL1_5184



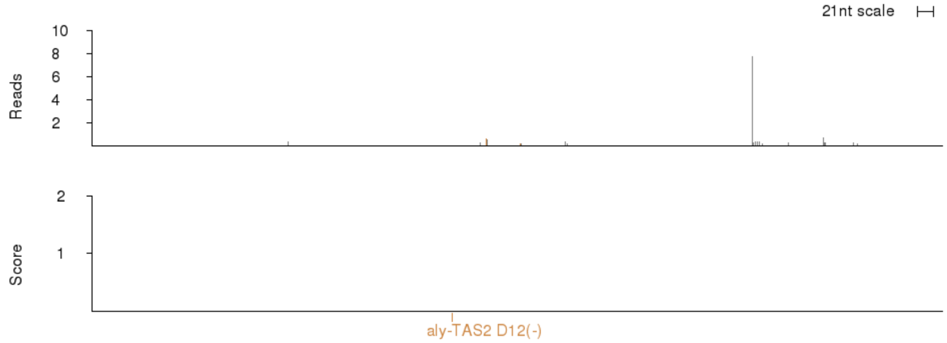
Al-RFL1_5213



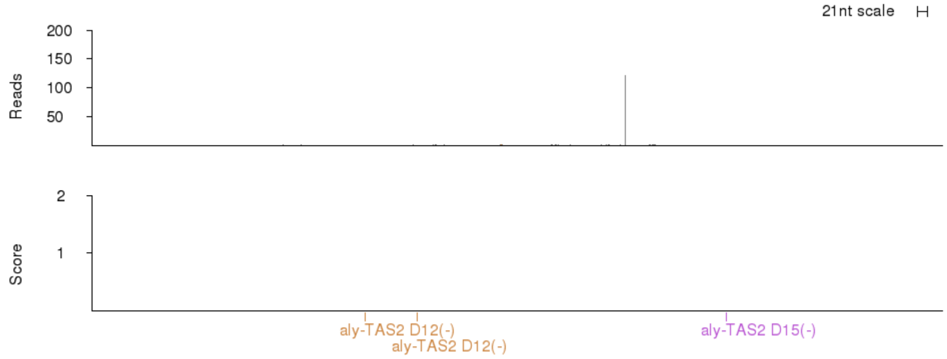
Al-RFL1_5216



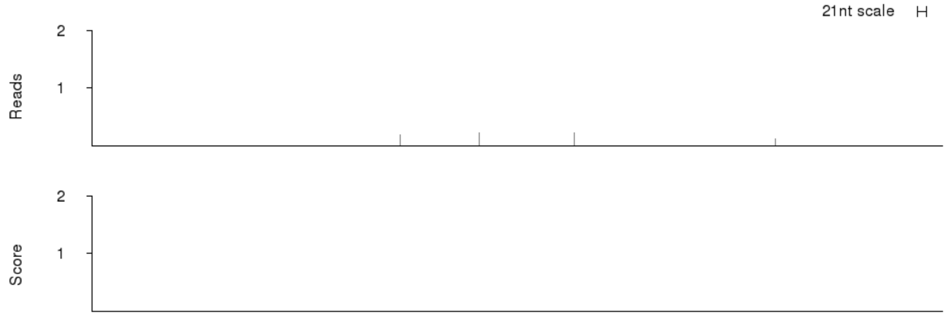
Al-RFL1_5291



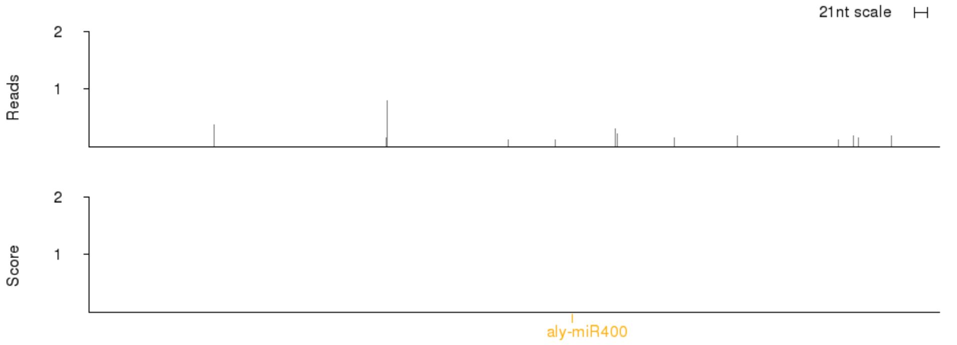
Al-RFL1_5296



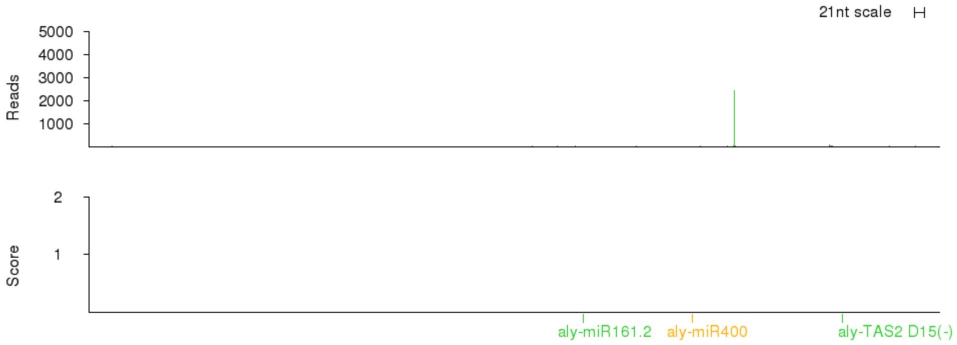
Al-RFL2_109



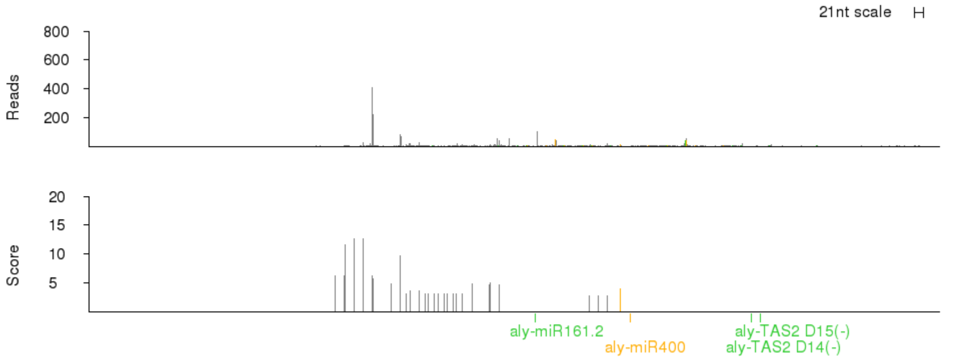
Al-RFL2_112



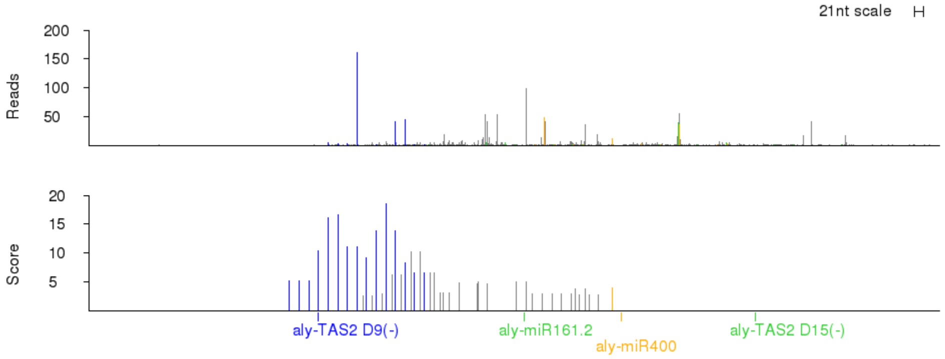
Al-RFL2_464



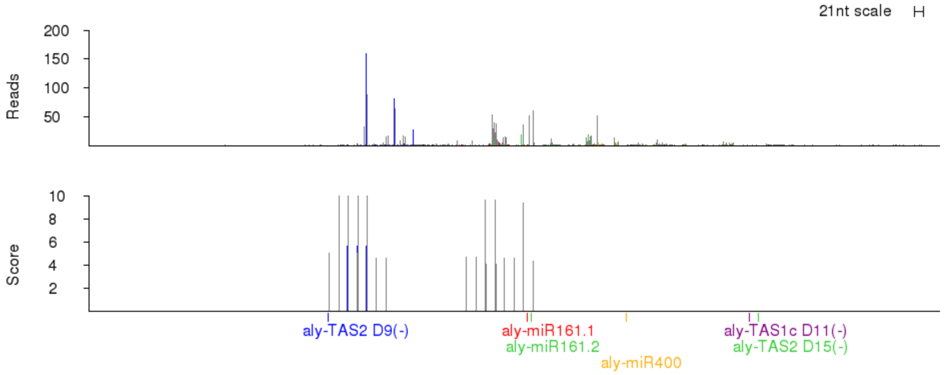
Al-RFL2_629



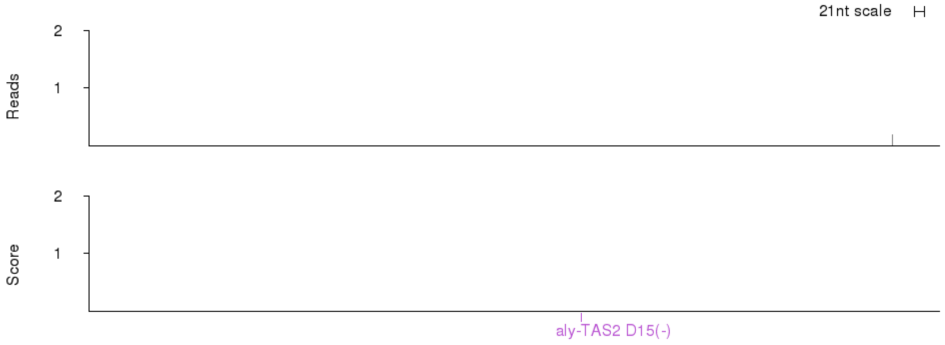
Al-RFL2_729



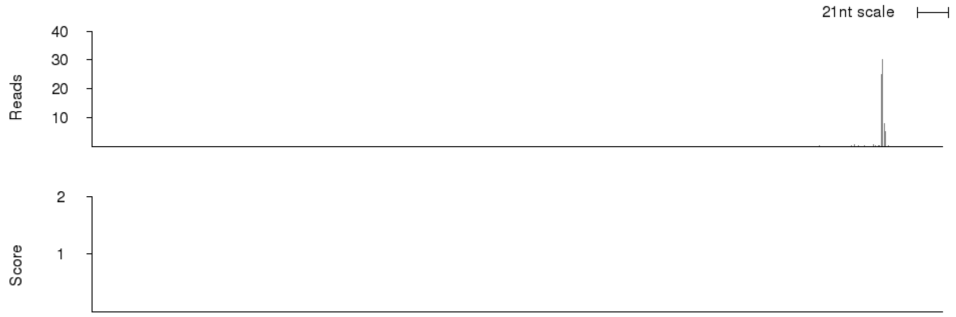
Al-RFL2_894



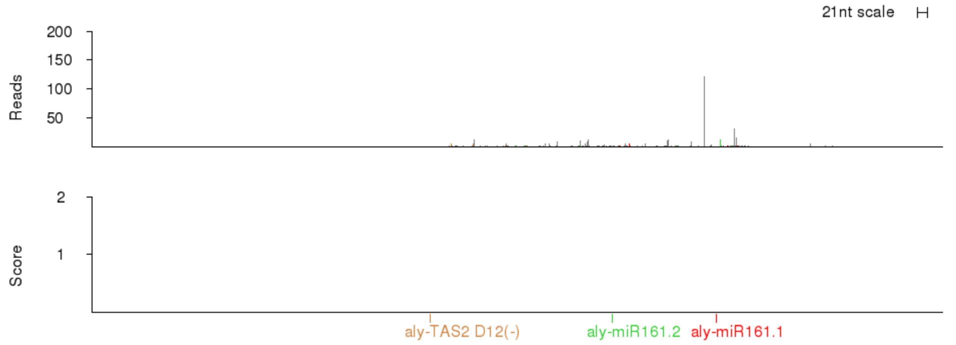
Al-RFL2_898



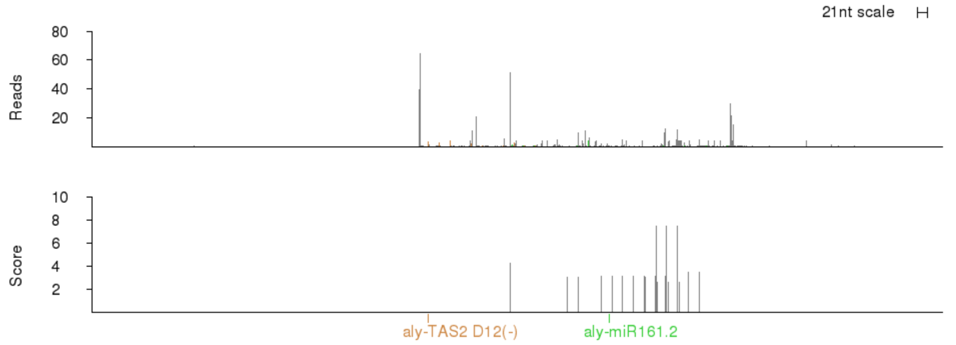
Al-RFL2_911



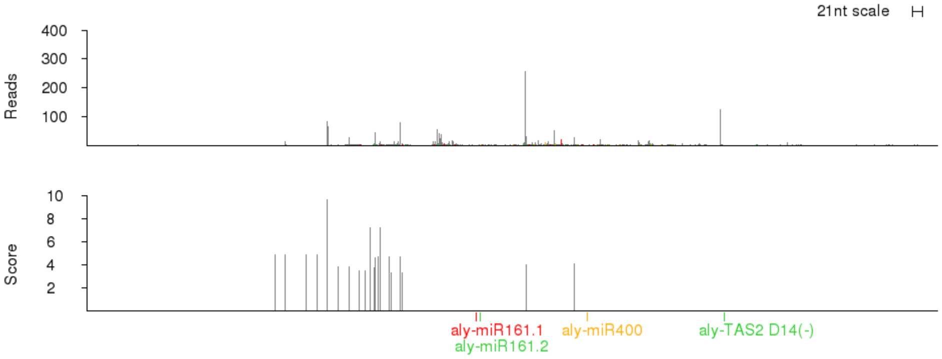
Al-RFL2_1054



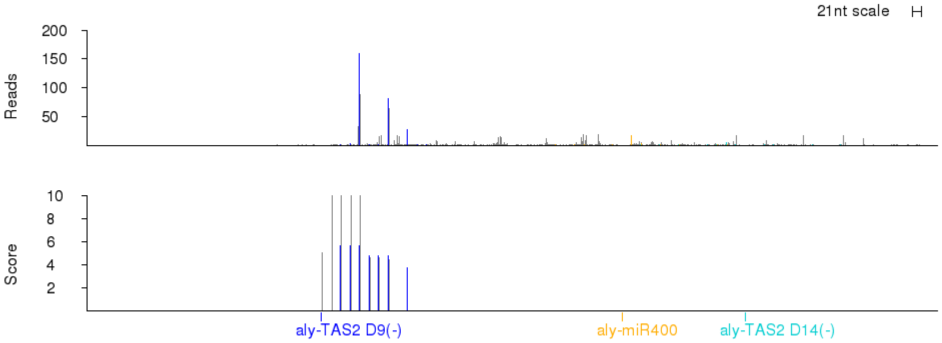
Al-RFL2_1082



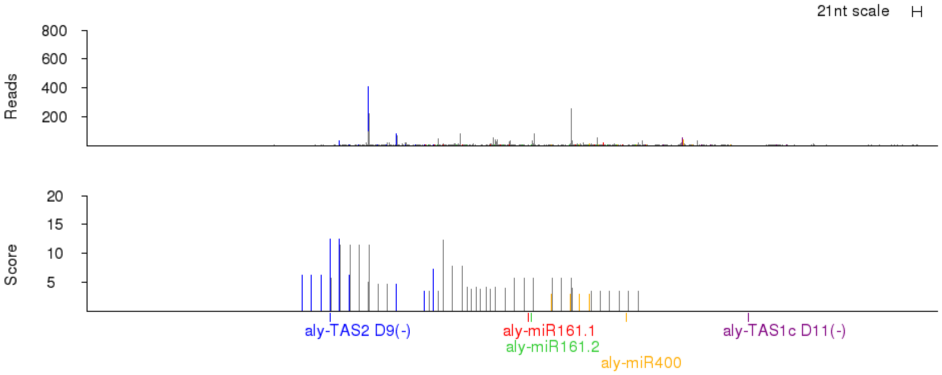
Al-RFL2_1159



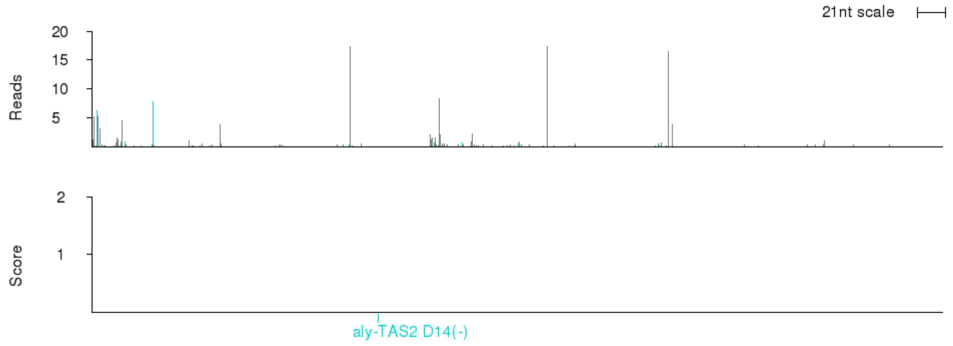
Al-RFL2_1201



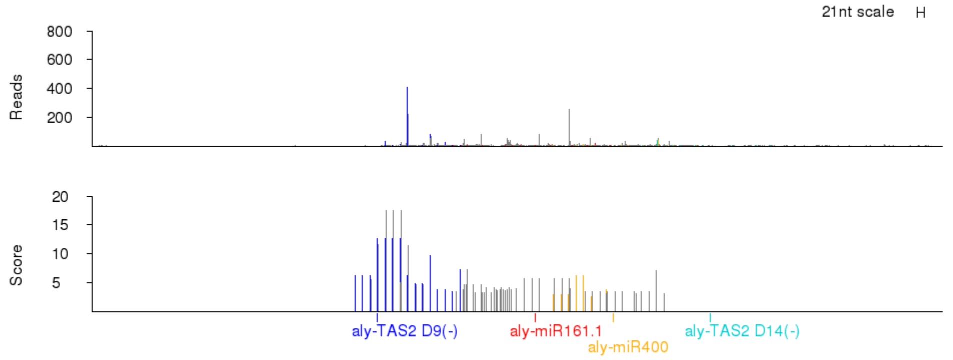
Al-RFL2_1205



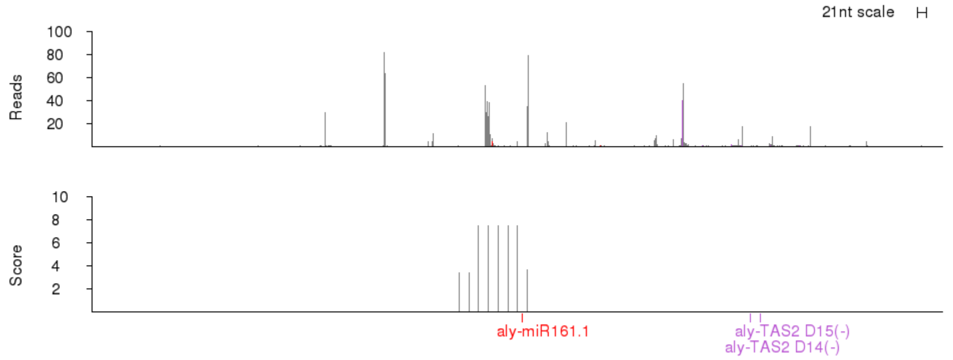
Al-RFL2_1211



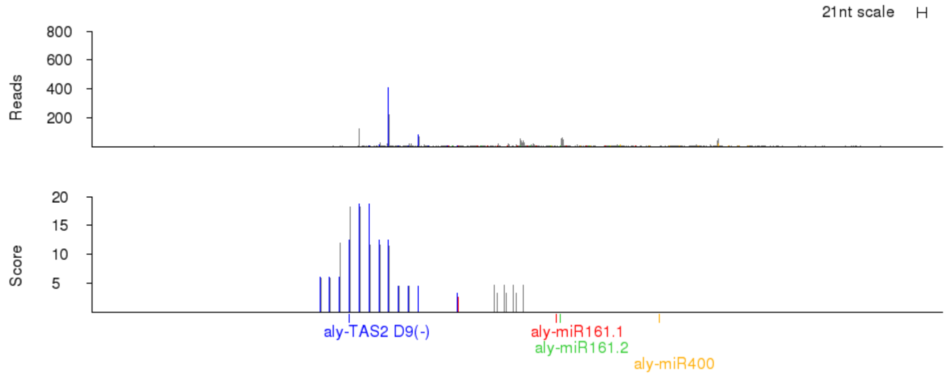
Al-RFL2_1215



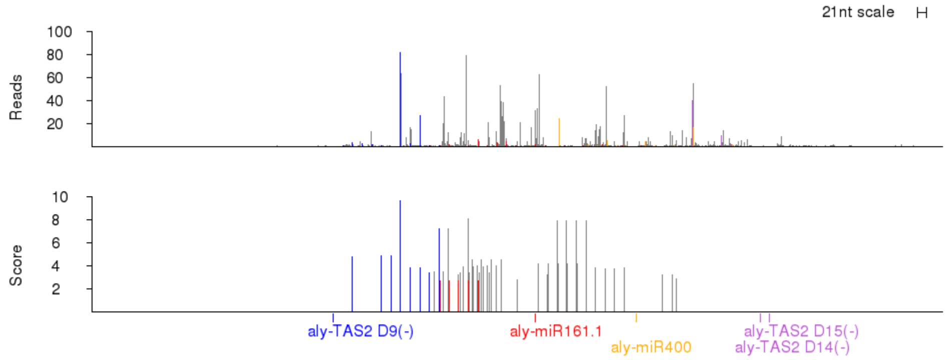
Al-RFL2_1393



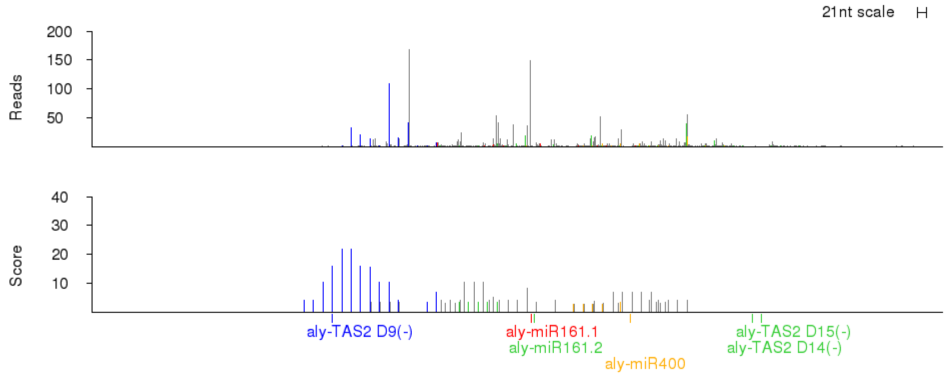
Al-RFL2_1395



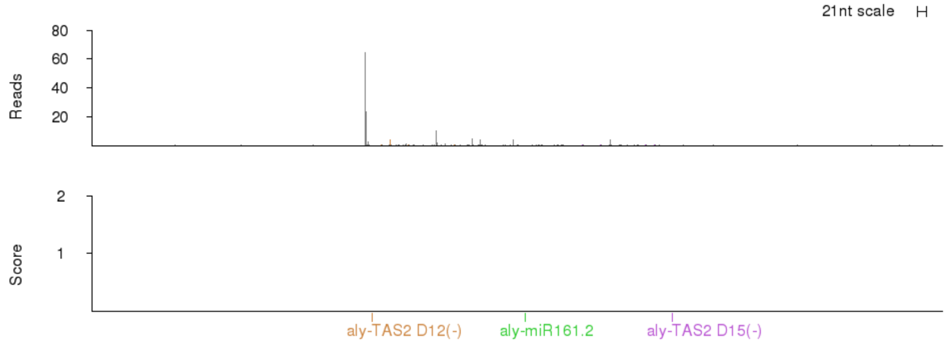
Al-RFL2_1411



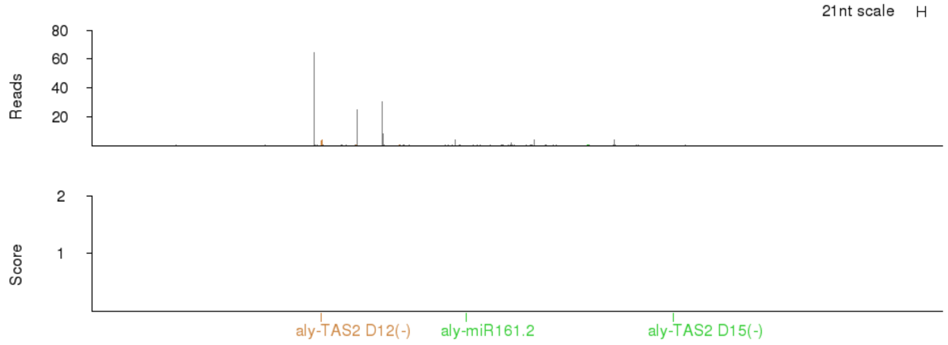
Al-RFL2_1417



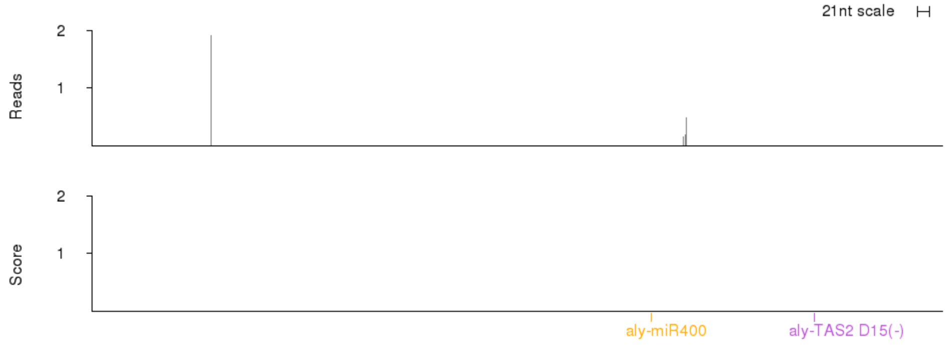
Al-RFL2_1453



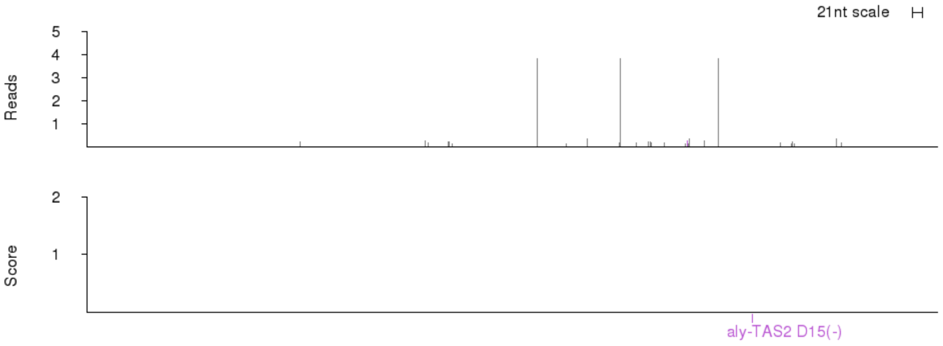
Al-RFL2_1458



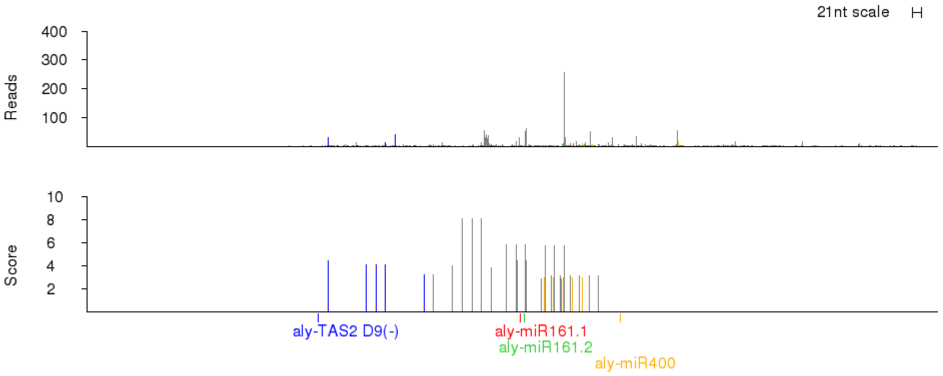
Al-RFL2_1521



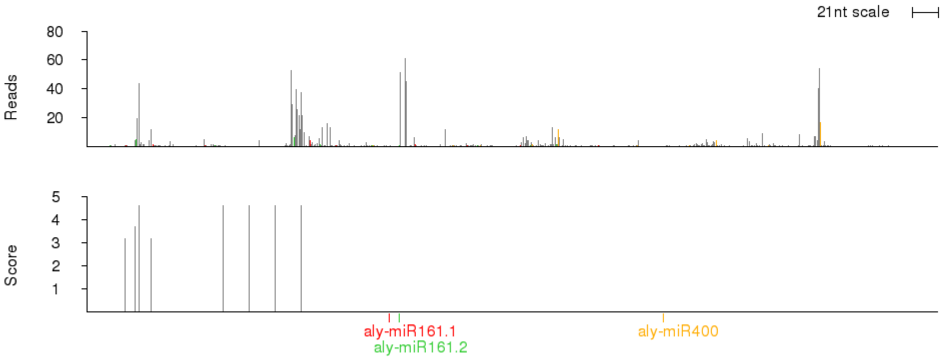
Al-RFL2_1553



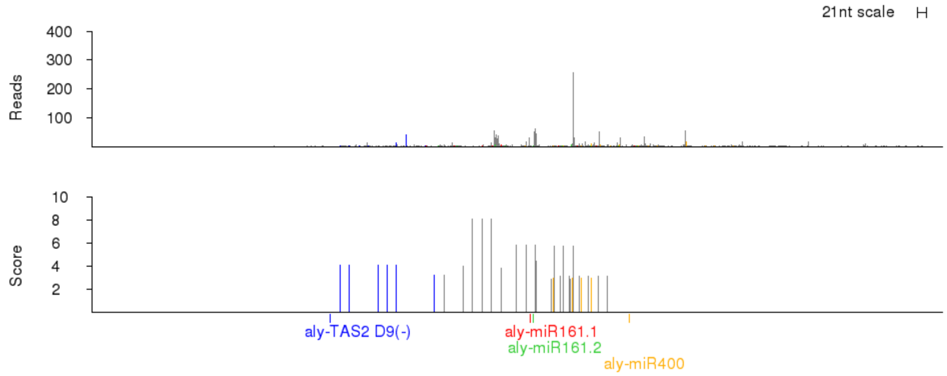
Al-RFL2_1557



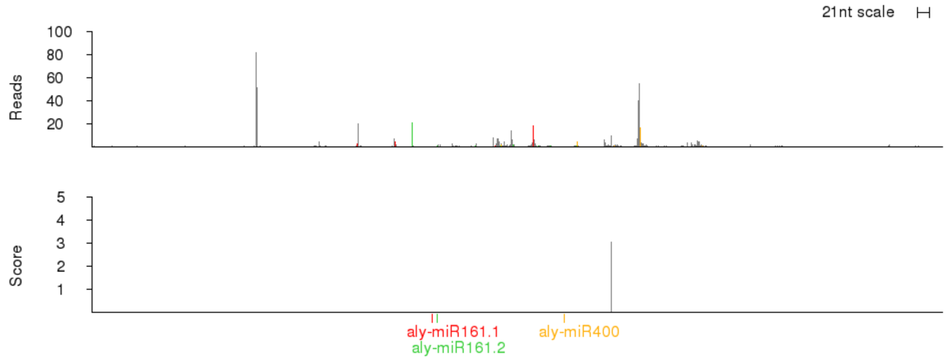
Al-RFL2_1559



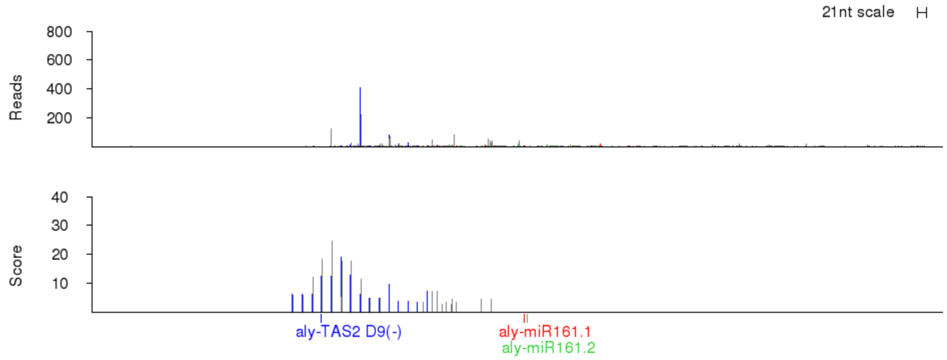
Al-RFL2_1561



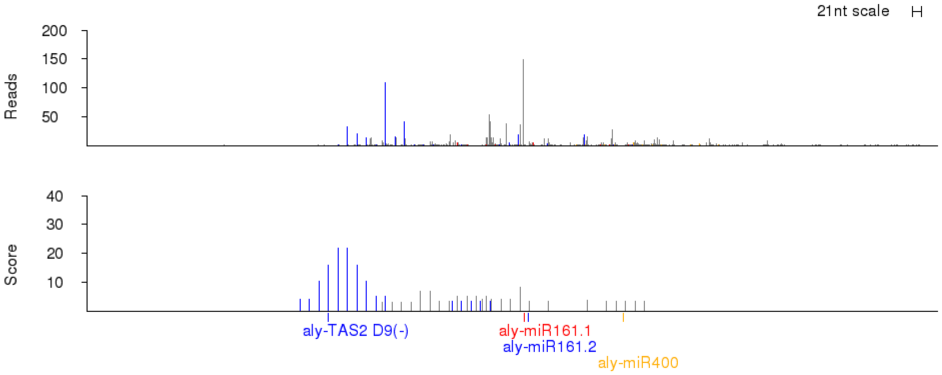
Al-RFL2_1566



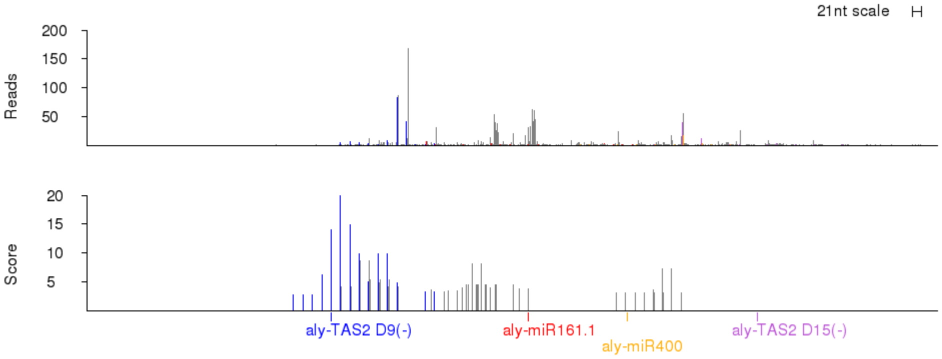
Al-RFL2_1571



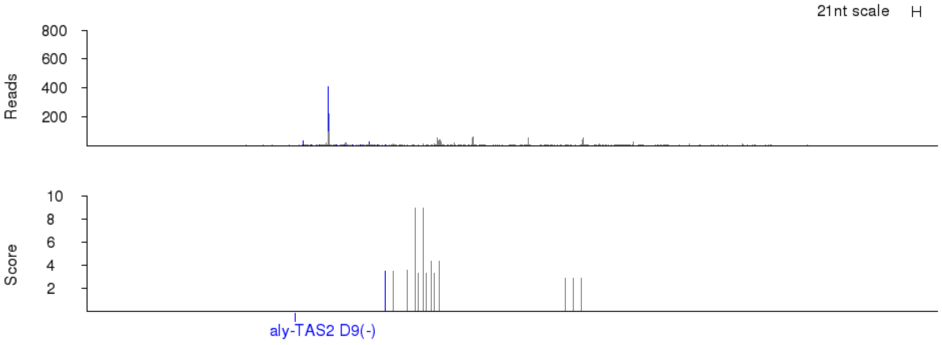
Al-RFL2_1575



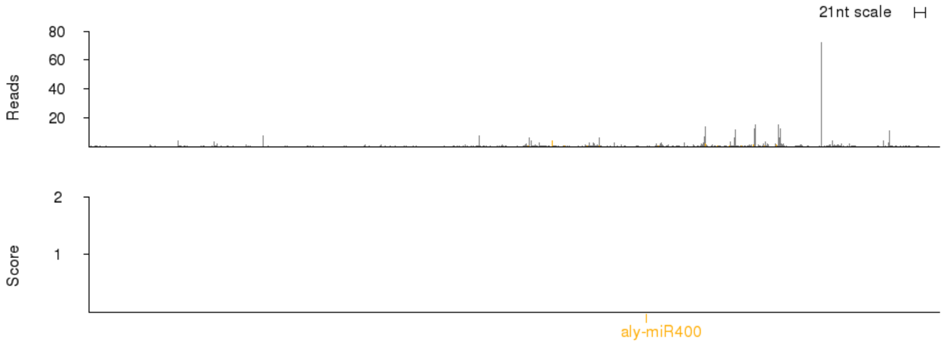
Al-RFL2_1611



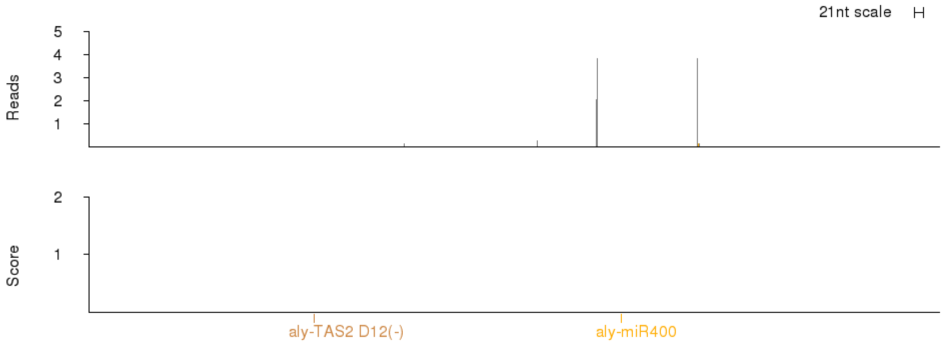
Al-RFL2_5036



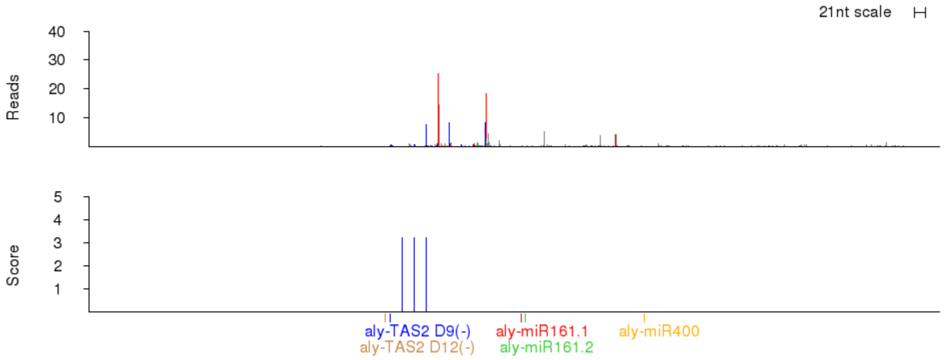
Al-RFL3_7093



Al-RFL3_10057



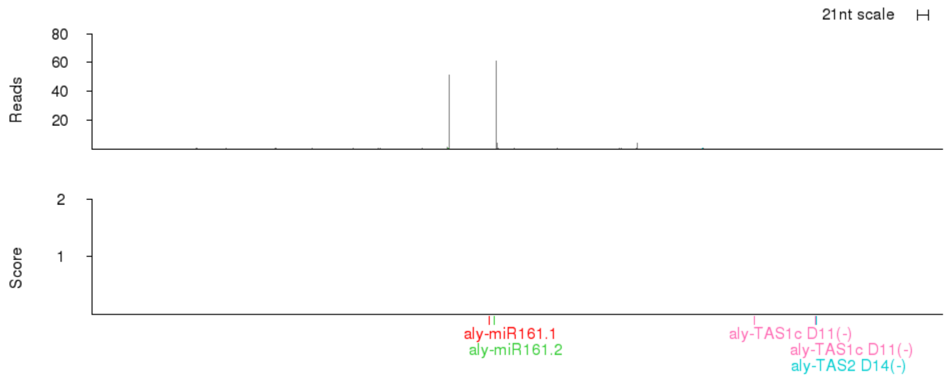
Al-RFL6_6831



Al-RFL7_6500

There were no small RNA reads that aligned to this transcript

Al-RFL7_21693


21nt scale
Reads
Score
aly-miR161.1
aly-miR161.2
aly-TAS1c D11(-)
aly-TAS1c D11(-)
aly-TAS2 D14(-)