Splicing conservation signals in plant long non-coding RNAs

José Antonio Corona-Gómez, Irving Jair García-Lopez, Peter F. Stadler, Selene L. Fernandez-Valverde

Abstract

Long non-coding RNAs (lncRNAs), With a length of at least 200 nt and little to no protein-coding potential, have recently emerged as prominent regulators of gene expression in eukaryotes. LncRNAs often drive the modification and maintenance of gene activation or gene silencing states via chromatin conformation rearrangements. In plants, lncRNAs have been shown to participate in gene regulation, and are essential to processes such as vernalization and photomorphogenesis. Despite their prominent functions, however, only over a dozen lncRNAs have been experimentally and functionally characterized.

Little is known about the evolutionary patterns of lncRNAs plants. The rates of divergence are much higher in lncRNAs than in protein coding mRNAs, making it difficult to identify lncRNA conservation using traditional sequence comparison methods. One of the few studies that has tried to address this found only 4 lncRNAs with positional conservation and 15 conserved at the sequence level in Brassicaceaee. Here, we characterised the splicing conservation of lncRNAs in Brassicaceaee. We generated a whole-genome alignment of 16 Brassica species and used it to identify synthetic lncRNA orthologues. Using a scoring system trained on transcriptomes from A. thaliana and B. oleracea, we identified splice sites across the whole alignment and measured their conservation. Our analysis revealed that 17.9% (112/627) of all intergenic lncRNAs display splicing conservation in at least one exon, an estimate that is substantially higher to previous estimates of lncRNA conservation in this group. Our findings agree with similar studies in vertebrates, suggesting that splicing conservation can be evidence of stabilizing selection and thus used to identify functional lncRNAs in plants.

Key words: long non-coding RNAs, lncRNA, splice sites, multiple sequence alignments, evolution, conservation, evolutionary plasticity.

1. INTRODUCTION

Long non-coding RNAs (lncRNAs), by definition, do not code for proteins. Over the last decade, a wide variety of mechanisms has been discovered by which lncRNAs contribute to the regulation of the expression of protein-coding genes as well as small RNAs, recently reviewed by Liu et al. (2015); Chekanova (2015); Ulit-ský (2016); Wang Chekanova (2017); Yamada (2017). The majority of the lncRNAs are found in the nucleus associated with the chromatin, regulating gene expression by recruiting components of the epigenetic machinery to specific genomic locations. Some lncRNAs also influence genome stability and nuclear domain organization. Serving as molecular sponges and decoys they act both a the transcription level by affecting RNA-directed DNA methylation, and in post-transcriptional regulation by inhibiting the interaction
between microRNAs (miRNAs) and their target messenger RNAs (mRNAs). Sequestering splicing factors, they are also involved in the control alternative splicing (Bardou et al., 2014). Hence they differ not only in size but also in their molecular mechanisms from small RNAs such as miRNA and siRNAs (Banfai et al., 2012). Instead, they are regulated and processed similar to mRNAs (Mercer Mattick, 2013). The expression patterns of IncRNAs are very specific for particular tissues or developmental stages. Recent data suggest that there appears to be a distinction between highly conserved, constitutively transcribed IncRNAs and tissue-specific IncRNAs with low expression levels (Deng et al., 2018b).

Systematic studies into the evolution of plant IncRNAs have been rare until very recently. An analysis of IncRNAs in five monocot and five dicot species (Deng et al., 2018b) found that the majority of IncRNAs well conserved at sequence level, while a minority is highly divergent but syntenically conserved. These positionally conserved IncRNAs were previously found to be locate near telomeres in A. thaliana (Mohammadin et al., 2015). Plant IncRNAs also display canonical splicing signals (Deng et al., 2018b).

Despite their often very poor sequence conservation, the majority of IncRNAs is well-conserved across animal families, as evidenced by the conservation of many of their splice sites (Nitsche et al., 2015). While well-conserved as entities, they show much more plasticity in their gene structure and sequence than protein-coding genes. The many lineage-specific differences have implicated IncRNAs as major players in lineage-specific adaptation (Lozada-Chávez et al., 2011): changes in transcript structure are likely associated with the inclusion or exclusion of sets of protein or mRNA binding sites and hence may have large effects on function and specificity of a particular IncRNA.

The systematic annotation of orthologous IncRNAs is important not only to provide complete maps of the transcriptome, but also as means of establishing that a particular IncRNA has a biological function. After all, conservation over long evolutionary timescales is used as the most important argument for the biological function of an open reading frame in the absence of direct experimental evidence for translation and experimental data characterizing the peptide product. While a large amount of work is available showing that vertebrate genome contain a large number of secondary elements that are under negative selection (Seemann et al., 2017; Smith et al., 2013; Hezroni et al., 2015; Nitsche Stadler, 2016) and the majority of human IncRNAs are evolutionary old (Nitsche et al., 2015), a much less systematic and complete picture is available for plants.

Nevertheless, there are some plant IncRNAs whose regulatory functions have been studied extensively and are understood at a level of detail comparable to most proteins (Kai et al., 2018): COOLAIR in Brassicaceae has a crucial role in the vernalization process (Hawkes et al., 2016) and its transcription accelerates epigenetic silencing of the flowering locus C (FLC) (Rosa et al., 2016). The IncRNA HID1, a key component in promoting photomorphogenesis in response to different levels of red light (Wang et al., 2014). HID1 is highly conserved and plays a role in controlling the nuclear translocation of the flowering repressor FLC (Hawkes et al., 2016). The lncRNA COOLAIR in Brassicaceae is one of the best-characterized plant IncRNAs (Franco-Zorrilla et al., 2007).

Although the functional characterization of IncRNAs is confined to a small number of cases, plant IncRNAs are being reported at a rapidly increasing pace (Nelson et al., 2016). As in the case of animals, it is important to find evidence for the functionality of individual transcripts. Differential expression alone, or correlations with important regulatory proteins or pathways alone does provide evidence to decide whether a transcript has a causal effect or whether its expression pattern is a coincidental downstream effect. As a first step towards prioritizing candidates, we advocate to use unexpected deep conservation of the gene structure as an indicator of biological function. While logically this still does not inform about function in an specific context, it is much less likely that changes expression patterns of a functional molecule are without biological consequence.

The much higher level of plasticity in plant genomes, compared to animal genomes, potentially makes it more difficult to trace the evolution of IncRNAs. We therefore concentrate here on a phylogenetically relatively narrow group, the Brassicaceae, with genomes that are largely alignable with each other. As a consequence we trace the conservation of functional elements, in particular splice junctions, through the entire data set. This provides direct evidence also in cases where transcriptome data are not available in sufficient coverage and or sufficient diversity of tissues and/or developmental stages. As a final result, this study produced a list of homologous IncRNAs in Brassicaceae as well as a detailed
map of the conservation of splice sites in this clade.

2. MATERIALS AND METHODS

2.1. Whole genome alignment

We selected sixteen genomes from genomes of plant from the Brassicaceae family available in NCBI, Phytoke and Ensembl-Plants (Supplemental Table S1) based on the quality of assembly, as measured by the number of contigs/scaffolds. All genomes were downloaded in fasta format. Mitochondrial and chloroplast sequences were excluded based on annotation.

The genomes were aligned using Cactus (Paten et al., 2011). Like other whole genome alignments (WGA) methods, Cactus uses small regions with very high sequence similarity as anchors. To resolve conflicts at this level, Cactus uses a specialized graph data structure that produces better overall alignments than other WGA approaches (Earl et al., 2014). The final WGA result were stored in HAL format (Hickey et al., 2013) for further processing.

2.2. Transcriptome data and assembly

We used four previously published base-line transcriptomes for *A. thaliana* (Liu et al., 2012) (GEO accession number GSE38612), as well as transcriptomes of shade response experiments from (Kohnen et al., 2016) (GEO accession number GSE81202). For *Brassica oleracea* we used transcriptomes from (Yu et al., 2014) (GEO accession number E-GEOD-42891). All transcriptomes were downloaded as raw reads in fastq format.

To generate our own lncRNA annotation, the 57 single end stranded sequencing libraries from (Kohnen et al., 2016) were quality-filtered using Trimomatic (Bolger et al., 2014), and mapped to the TAIR10 genome (Berardini et al., 2015) using tophat v2.1.1 (Trapnell et al., 2009) with parameters: -I 20 -I 1000 -read-edit-dist 3 -read-realign-edit-dist 0 -library-type fr-firststrand -g 1. Transcripts were assembled by Cufflinks v2.2.1 (Trapnell et al., 2010) with parameters: --overlap-radius 1 -p 8 -I 1000 -min-intron-length 20 -g TAIR10.GFF3.gff -library-type fr-firststrand and subsequently merged into a single reference transcriptome using Cuffmerge.

2.3. lncRNA Annotation

LncRNAs in the (Kohnen et al., 2016) dataset were annotated using two independent methods. First, coding and non-coding transcripts were identified with CPC (Coding Potential Calculator) (Kong et al., 2007), a support vector machine classifier. Additionally, we used Cabili’s strict stepwise annotation workflow (Cabili et al., 2011) on all transcripts. Specifically, we removed transcripts less than 200 nt in length, and identified ORFs 75 aminoacids (AA) or longer. Identified ORFs were compared against the NCBI non redundant (nr) database using blastx and blastp (Altschul et al., 1990) with E-value and cutoff of $E < 10$ for hits to be considered significant. In addition, we used HMMER (Finn et al., 2011) to search for Pfam protein domains, signalP (Nielsen Krogh, 1998) to identify signal peptides, and tmhmm (Krogh et al., 2001) for transmembrane helices. Only sequences that had no similarity with proteins in nr and no identifiable protein domains, signal peptides or transmembrane domains were annotated as bona fide lncRNAs.

To characterize the genomic context of identified lncRNAs, we used bedtools (Quinlan Hall, 2010) and compared the lncRNA annotation with the protein coding gene annotation in Araport11 (Cheng et al., 2017). All lncRNA candidates that overlapped a coding sequence or some other ncRNA (miRNA, snoRNA, snRNA) by at least 1 nt were discarded. We classified lncRNAs as adjacent if they were located within 500 nt upstream or downstream of a coding gene, and as intergenic, i.e., lincRNAs, otherwise. lncRNAs that were fully contained within intronic regions were annotated as intronic.

2.4. Splicing map

The construction of splicing maps requires a seed set of experimentally determined splice sites in at least one species as well as a statistical model to assess the conservation of splice donors and splice acceptors whenever no direct experimental evidence is available.

To obtain these data for Brassicaceae, we mapped the reference transcriptomes to the corresponding reference genome using STAR (Dobin et al., 2013) using default parameters. The table of splice junctions produced by STAR for each data set were concatenated. Only splice junctions that (a) had at least 10 uniquely mapped reads crossing the junction, and (b) showed the canonical GT/AG dinucleotides delimiting the intron (c) within an intron of size between 59 bp and 999 bp were retained for subsequent analyses. Since some of the
transcriptome datasets were not strand-specific we included GT/AC delimiters, interpreting these as reverse-complements.

For each identified splice site in *Arabidopsis thaliana*, we used the HalTools liftover tool (Hickey et al., 2013) to determine the corresponding orthologous positions in all other genome sequences in the Cactaceae generated WGA. For each of the retained splice-site we extracted the genomic sequences surrounding the donor and acceptor sites. If more than one homolog per species is contained in the WGA, we retained the candidate with the highest sequence similarity to *A. thaliana*. For each known splice site and their orthologous position, the MaxEntScan splice-site score (MES) (Yeo and Burge, 2004) was computed with either the donor or acceptor model provided the corresponding region contained neither gaps nor ambiguous nucleotides. Otherwise, the regions was treated as non-conserved. A MaxEntScan splice-site score cutoff of 0 was used. All positively predicted splice-site, i.e., those with $MES > 0$, were added to the splicing map. The pipeline implementing this analysis strategy is available at: https://bitbucket.org/JoseAntonioCorona/splicing_map

2.5. Data Availability

TrackHubs for all datasets and IncRNAs used in this study as well as WGA are available here: http://de.cyverse.org/d/d/I-1F42267D1-B51F-4E9F-80F4-1536834BB8EC2/hub.txt

Additional information as well as machine readable intermediate results are provided at http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/19-001

3. RESULTS

3.1. Identification of splice sites and IncRNAs

We predicted about 125,000 introns using the transcriptomes of Liu et al. 2012 (Liu et al., 2012) compared with 175,000 introns annotated in TAIR10 (Release 38) (Berardini et al., 2015). The smaller number was expected as only introns with convincing coverage by uniquely mapping reads were considered (see Methods). Additionally, not all *A. thaliana* genes are expressed in the four transcriptomes used for building the splice map. Consistent with previous reports (Hebsgaard, 1996; Brown et al., 1996), the vast majority of the detected splice junctions have the canonical GT/AG motif required for inclusion into our splice site map. In total, we identified 222,772 individual sites in *A. thaliana* (117,644 donor and 121,022 acceptor sites).

To characterize splicing conservation in IncRNAs, we focused solely on intergenic non-coding RNAs (lincRNAs), as conservation in splice sites of lincRNAs overlapping with coding genes may be confounded by the coding gene conservation signal, resulting in false positives. The lincRNAs described by Liu et al. (2012) comprise only 595 lincRNAs with annotated introns (Liu et al., 2012), while in Araport11 (Cheng et al., 2017) only 288 annotated lincRNAs out of 2,444 have introns. We therefore used an alternate set of lincRNAs expressed in *A. thaliana* cotyledons and hypocotyls in Col-0 plants in normal light or shade conditions (Kohnen et al., 2016). These libraries were stranded, and had three replicates as well as sufficient depth to produce a high confidence lincRNA annotation. We identified 2,375 lincRNAs transcripts, 1,465 of which overlapped with protein coding RNAs, while 808 were found in intergenic regions and were thus considered bona fide lincRNAs. In our analysis, we found 159 lincRNAs that were included in neither Araport11 nor TAIR10 (Cheng et al., 2017; Berardini et al., 2015). Given that the datasets come from only two experimental conditions (shadow and light) (Kohnen et al., 2016), they encompass only a fraction of the lincRNAs expressed throughout the *A. thaliana* life cycle. All 808 lincRNAs transcripts aggregated in 627 lincRNA genes, of which 58 have multiple isoforms. In contrast to the situation in animals, lincRNAs are therefore mostly mono-exonic in *A. thaliana*. Of the 627 lincRNA genes, only 173 had at least one intron and thus were used to test splice site conservation in lincRNAs.

3.2. Conservation of lincRNAs

In the WGS, the other Brassicaceae species cover between 69.6% to 44.2% of the *A. thaliana* genome. For the protein-coding genes annotated in Araport11 (Cheng et al., 2017) the coverage ranges from 95.3% (26,153/27,445) (A. lyrata) to 86.9% (23,856/27,445) (Aethionema arabicum). As expected, the values are substantially lower for the Araport11 lincRNAs, where we recover between 77.1% (1,885/2,444) in *A. lyrata* and 50.8% (1243/2444) in (A. arabicum). Using our annotation, we find between 62.0% (389/627) in *A. lyrata* and 38.1% (239/627) in (A. arabicum), i.e., values comparable to the overall coverage of the genome. This reflects the fact that lincRNA sequences experience very little constraint on their sequence. Conservation (as measured by alignability) is summarized in Fig. 1 for different types of RNA elements. These values are comparable to a previous estimate of about 22% of the lincRNA loci are at least partially conserved at the se-
sequence level in the last common ancestor of Brassicaceae (Nelson et al., 2016).

Conservation of splice sites is a strong indication for the functionality of the transcript. In order to evaluate splice site conservation quantitatively, we constructed a splicing map that identifies for every experimentally determined splice site the homologous position in the other genomes and evaluates them using the MES (see Methods for details). Fig. 2 shows the splicing map for the lincRNA TCONS00053212-00053217 as an illustrative example. Despite the unusually complex transcript structure and the conservation throughout the Brassicaceae, so far nothing is known of the function of this lincRNA. While not all splice sites are represented in all species in the WGA, almost all MES values are well above the threshold of $MES > 0$. Most of the isoforms therefore can be expected to present throughout the Brassicaceae, even though the locus is not annotated in Ensembl Plants (release 42) for B. oleracea, B. rapa, and A. lyrata. Only the the short first exon and the 5’ most acceptor of the last exon are poorly conserved even in close relatives of A. thaliana.

On a genome-wide scale, the conservation of splice sites in lincRNAs provides a lower bound on the fraction of lincRNAs that are under selective constraint as a transcript. We find that 112 of the 173 spliced A. thaliana lincRNAs have at least one conserved splice site in another species (Fig. 3).

As expected, we find that splice sites in lincRNAs are much less well conserved than splice sites in protein coding genes (Fig. 4). In total, we identified 39 lincRNAs conserved between the most distant species and A. thaliana and 26 lincRNAs with conservation in at least one splice site in the 16 species included in the WGA. These numbers are much lower than for coding genes. Albeit this is expected, given the high conservation of protein coding genes, one has to keep in mind that coding genes on average have at least 6 introns (Deng et al., 2018b), hence it is much more likely to observe conservation of at least one splice site and in lincRNAs with only one or two introns, see Fig. 2.

In comparison to vertebrates, we observe a much lower level of conservation as measured by gene structure. For instance, 35.2% of the transcripts are conserved between human and mouse (Nitsche et al., 2015), while we find only 6.2% (39/627) of total of own lincRNAs conserved between A. thaliana and A. arabicum and Araport11 lincRNAs 1.3% (32/2444). This difference is even more striking given the fact that the evolutionary distance between human and mouse (~75 Mya) (Waterston et al., 2002) is larger than between A. thaliana and A. arabicum (~54 Mya) (Beilstein et al., 2010).

Transposable elements (TEs) are important factors in lncRNA origin (Kapusta et al., 2013). In order to see if conserved lincRNAs have a relation with TEs, we compared our 627 lincRNAs with the genomic positions of TEs described in Araport11 database. We find only 149 of 627 lincRNAs overlap with TEs and these lincRNAs display significantly lower positional conservation than other lincRNAs in the WGA. Indeed, only 11 were found to be positionally conserved between A. thaliana and B. rapa. The number of TEs coincident lincRNAs with splicing sites is even smaller; of the 173 lincRNAs with introns only 11 overlapped with TEs. From the total of TEs in Araport11 database (3,897) only 450 are conserved for position in the WGA between A. thaliana and A. arabicum. This represents only 11.5% of the TEs, i.e., less than the conservation level of the lincRNAs by genomic position (Fig. 7).

4. Discussion

In this work we explore the conservation of lncRNAs in the Brassicaceae plant family and we find conservation at different levels: from 627 lincRNAs identified we have 38.1% (239/627) conserved by genomic position as determined by the presence of alignable sequence. Only a small fraction (27.6) of the lincRNAs contains introns. Of these, only 19.1% are conserved between A. thaliana and B. oleracea, the species with the lowest level of conservation in our data set. While sequence conservation may be a consequence of selective constraints on DNA elements, conservation of splice sites directly indicates selective constraints at the transcript level, and thus can be interpreted as evidence for an (unknown) functional role of the lincRNA. The 112 lincRNAs with conserved splice sites are therefore attractive candidates for studies into lincRNA function.

Comparing the 38.1% (239/627) of conservation of lincRNAs in Brassicaceae with others family of plants, for example Poaceae, we find that in maize and rice have around 20% of lincRNAs conserved by position in WGA (Wang et al., 2015). These numbers are roughly comparable given that the divergence times of the two families are similar: Brassicaceae 52.6 Mya (Kagale et al., 2014), Poaceae 60 Mya (Charles et al., 2009). This difference may be explained by the much large genome size, and thus higher content in repetitive elements and unconstrained sequence, leaving conserved sequence regions more “concentrated” – and thus easier to align – in the small genomes of Brassicaceae. Consistent with previous findings we find that only a small fraction of our lincRNAs associated with TEs compared
to a much stronger association in e.g. in Poacea (Wang et al., 2017). We interpret this to be consequence of the strong reduction of genome size in Brassicas. More detailed comparisons of lncRNA conservation among different families will have to await better assembled and annotated genomes as basis for WGAs.

There is clear evidence that the conservation of splicing sites is an important factor in vertebrates, where about 70% of the lncRNAs are conserved in placental mammals (Nitsche et al., 2015). In Brassicaceae we find a much lower level of conservation. At least in part this difference is the consequence of prevalence of single-exon lncRNAs in this clade and the small number of splice sites in those lncRNAs that contains introns. This reduced the power of the method and hints a reduced importance of introns in the small genomes of Brassicaceae. However, this may also be a result of using A. thaliana as a reference which, in addition to drastic genome reduction, may have been subjected to clade-specific intron-loss. Transcriptomes of other Brassicas and other plant families that have not undergone drastic genome reduction will clarify the actual prevalence on monoexonic and intron-gain-loss in plant lncRNAs.

A limitation of our work is the restriction to intergenic lncRNAs, caused by the need to avoid potential overlaps of the splice sites with other constrained elements. High quality transcriptomes for most the species could alleviate this shortcoming since it would allow us to construct the splicing map based on experimental evidence only. Spurious sequence conservation would then no longer influence the results. This is of particular relevance in Brassicaceae, since about 70% of transcript have antisense lncRNAs (Wang et al., 2014). These had to be excluded from our the analysis even though at least some of them, e.g. COOLAIR (Hawkes et al., 2016), are known to have important biological functions. Complementarily to the analysis of splice site conservation, conserved RNA secondary structure can serve as evidence of section constraints on the RNA level, see e.g. (Washietl et al., 2005). This would also be applicable to unspliced transcripts. So far, no genome-wide assessment of RNA secondary structure conservation has been reported for plants, however. recent structure-sequence data indicates RNA structure is under selection at genome-wide levels also in plants (Deng et al., 2018a).

Acknowledgments

This work was funded in part by CONACYT (JAG: CONACYT Ph.D. Scholarship - 338379, SLF-V: CONACYT Research Fellowship 2015 - 72223). We are grateful to Thomas Gatter for advice on transcriptome assembly.

References


<table>
<thead>
<tr>
<th>Own</th>
<th>LinRNA</th>
<th>LinRNA</th>
<th>NAT</th>
<th>mRNAs</th>
<th>Pseudocoding</th>
<th>miRNA</th>
<th>TE</th>
<th>snoRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>627</td>
<td>244</td>
<td>1037</td>
<td>27445</td>
<td>941</td>
<td>125</td>
<td>3897</td>
<td>287</td>
</tr>
<tr>
<td>Arabidopsis lyrata</td>
<td>389</td>
<td>1885</td>
<td>1010</td>
<td>26153</td>
<td>672</td>
<td>244</td>
<td>1402</td>
<td>278</td>
</tr>
<tr>
<td>Arabidopsis halleri</td>
<td>346</td>
<td>1662</td>
<td>957</td>
<td>25333</td>
<td>611</td>
<td>208</td>
<td>974</td>
<td>232</td>
</tr>
<tr>
<td>Camelina sativa</td>
<td>335</td>
<td>700</td>
<td>1009</td>
<td>25592</td>
<td>593</td>
<td>205</td>
<td>915</td>
<td>270</td>
</tr>
<tr>
<td>Capsella rubella</td>
<td>300</td>
<td>1538</td>
<td>996</td>
<td>25008</td>
<td>516</td>
<td>191</td>
<td>703</td>
<td>266</td>
</tr>
<tr>
<td>Boechera stricta</td>
<td>325</td>
<td>1650</td>
<td>1004</td>
<td>25382</td>
<td>548</td>
<td>203</td>
<td>730</td>
<td>258</td>
</tr>
<tr>
<td>Leavenworthia alabamica</td>
<td>253</td>
<td>350</td>
<td>973</td>
<td>24300</td>
<td>481</td>
<td>158</td>
<td>411</td>
<td>221</td>
</tr>
<tr>
<td>Arabis alpina</td>
<td>249</td>
<td>340</td>
<td>972</td>
<td>24301</td>
<td>489</td>
<td>173</td>
<td>506</td>
<td>211</td>
</tr>
<tr>
<td>Sisymbrium irio</td>
<td>244</td>
<td>333</td>
<td>987</td>
<td>24352</td>
<td>490</td>
<td>161</td>
<td>478</td>
<td>209</td>
</tr>
<tr>
<td>Eutrema salsugineum</td>
<td>252</td>
<td>333</td>
<td>979</td>
<td>24330</td>
<td>485</td>
<td>155</td>
<td>408</td>
<td>217</td>
</tr>
<tr>
<td>Eutrema parvulum</td>
<td>249</td>
<td>329</td>
<td>987</td>
<td>24333</td>
<td>488</td>
<td>158</td>
<td>404</td>
<td>216</td>
</tr>
<tr>
<td>Raphanus sativus</td>
<td>230</td>
<td>342</td>
<td>977</td>
<td>24177</td>
<td>469</td>
<td>148</td>
<td>462</td>
<td>208</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>229</td>
<td>123</td>
<td>965</td>
<td>23972</td>
<td>459</td>
<td>145</td>
<td>391</td>
<td>193</td>
</tr>
<tr>
<td>Brassica napus</td>
<td>234</td>
<td>132</td>
<td>970</td>
<td>24137</td>
<td>469</td>
<td>145</td>
<td>393</td>
<td>202</td>
</tr>
<tr>
<td>Brassica oleracea</td>
<td>232</td>
<td>131</td>
<td>964</td>
<td>23997</td>
<td>457</td>
<td>144</td>
<td>375</td>
<td>201</td>
</tr>
<tr>
<td>Aethionema arabiicum</td>
<td>239</td>
<td>1243</td>
<td>964</td>
<td>23856</td>
<td>457</td>
<td>147</td>
<td>450</td>
<td>237</td>
</tr>
</tbody>
</table>

Figure 1: Conservation of genes by position in WGA. Own: lincRNAs genes expressed in shade experiments (Kohnen et al., 2016). Araport11 database annotations (Cheng et al., 2017): lincRNAs (long intergenic non-coding RNAs), NAT (Natural antisense transcripts), mRNAs (messenger RNAs of coding genes), miRNA (microRNAs), Pseudocoding (Pseudocoding genes), TE (Transposable elements), snoRNA (Small nucleolar RNAs)
Figure 2: Histograms showing number of splicing sites, isoforms and conservation in WGA of the 173 lincRNAs genes with introns in Own dataset.
Figure 3: Conservation genes in the Brassicaceae family measured by the conservation of splice sites. *Blue:* own lincRNA set (627); *green:* lincRNAs in Araport11 (2,444); and *red:* coding RNA genes (27,445).
Figure 4: Splicing conservation map of lincRNA locus TCONS00053212-TCONS00053217.  
A) UCSC Genome browser screenshot of the TCONS00053212-TCONS00053217 locus, blocks denote exons and line with arrows introns. The arrow direction indicates direction of transcription. Splicing sites are shown in purple. Light blue blocks represent aligned regions as identified by Cactus.  
B) Heatmap of TCONS00053212-TCONS00053217 MES each splice sites (columns) in each species (rows), linked to its position in panel A with a purple line. MES are shown from more negative (light yellow) to more positive (dark blue). MES values > 0 were used to identify conserved splice sites.
### SUPPLEMENTAL MATERIAL

Table S1 - Genome versions used in this study.

<table>
<thead>
<tr>
<th>Species</th>
<th>Ensembl ID</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aethionema arabicum</td>
<td>GCA_000411095.1</td>
<td>NCBI</td>
</tr>
<tr>
<td>Arabidopsis halleri</td>
<td>GCA_900078215.1</td>
<td>Phytosome v12</td>
</tr>
<tr>
<td>Arabidopsis lyrata</td>
<td>GCA_000004255.1</td>
<td>Ensembl-Plants</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>GCA_000001735.1</td>
<td>Ensembl-Plants</td>
</tr>
<tr>
<td>Arabis alpina</td>
<td>GCA_000733195.1</td>
<td>NCBI</td>
</tr>
<tr>
<td>Boechera stricta</td>
<td>GCA_002079875.1</td>
<td>NCBI</td>
</tr>
<tr>
<td>Brassica napus</td>
<td>GCA_000686985.1</td>
<td>Ensembl-Plants</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>GCA_000309985.1</td>
<td>Ensembl-Plants</td>
</tr>
<tr>
<td>Camelina sativa</td>
<td>GCA_000633955.1</td>
<td>NCBI</td>
</tr>
<tr>
<td>Capsella rubella</td>
<td>GCA_000375325.1</td>
<td>Phytosome v12</td>
</tr>
<tr>
<td>Leavenworthia alabamica</td>
<td>GCA_000411055.1</td>
<td>NCBI</td>
</tr>
<tr>
<td>Raphanus sativus</td>
<td>GCA_000801105.2</td>
<td>NCBI</td>
</tr>
<tr>
<td>Schrenkiella parvula</td>
<td>GCA_000218505.1</td>
<td>NCBI</td>
</tr>
<tr>
<td>Sisymbrium irio</td>
<td>GCA_000411075.1</td>
<td>NCBI</td>
</tr>
<tr>
<td>Brassica oleracea</td>
<td>GCA_000695525.1</td>
<td>Ensembl-Plants</td>
</tr>
<tr>
<td>Eutrema salsugineum</td>
<td>GCA_000478725.1</td>
<td>NCBI</td>
</tr>
</tbody>
</table>