Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Atypical DNA methylation, sRNA size distribution and female gametogenesis correlate with genome compaction in *Utricularia gibba*

Sergio Alan Cervantes-Pérez
Centro de Investigación y de Estudios Avanzados del IPN
Lenin Yong-Villalobos
Texas Tech University
Nathalia M.V. Florez-Zapata
Centro de Investigación y de Estudios Avanzados del IPN
Araceli Oropeza-Aburto
Centro de Investigación y de Estudios Avanzados del IPN
Felix Rico-Reséndiz
Centro de Investigación y de Estudios Avanzados del IPN
Itzel Amasende-Morales
Centro de Investigación y de Estudios Avanzados del IPN
Tianying Lan
University at Buffalo, State University of New York
Octavio Martínez
Centro de Investigación y de Estudios Avanzados del IPN
Jean-Philippe Vielle-Calzada
Centro de Investigación y de Estudios Avanzados del IPN
Victor A. Albert
University at Buffalo, State University of New York
Luis Herrera-Estrella (lherrerae@cinvestav.mx)
Texas Tech University

Research Article

Keywords: DCL3, DNA methylation, female gametogenesis, lncRNAs, RdDM, siRNAs.

Posted Date: February 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-235397/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at Scientific Reports on August 3rd, 2021. See the published version at https://doi.org/10.1038/s41598-021-95054-y.

Abstract

The most studied DNA methylation pathway in plants is the RNA Directed DNA Methylation (RdDM), a conserved mechanism that involves the role of noncoding RNAs to control the expansion of the noncoding genome. Genome-wide methylation levels have been reported to correlate with genome size. However, little is known about the catalog of noncoding RNAs and the impact on DNA methylation in compact plant genomes. Because the small genome size of the carnivorous plant *Utricularia gibba* we investigate the noncoding RNA landscape and global DNA methylation in a compact genome. Here, we report that, compared to other angiosperms, *U. gibba* has an unusual distribution of noncoding RNAs and reduced global DNA methylation levels, as determined by a novel strategy based on long-read DNA sequencing with the Pacific Bioscience platform and confirmed by whole-genome bisulfite sequencing. Moreover, reduced DNA methylation correlates with lack of a functional RdDM pathway, as *U. gibba* DICER-LIKE 3 (DCL3), encoding a DICER endonuclease that produces 24-nt small-interfering RNAs lost key domains required for complete function. Our findings unveil that lack of a functional *DCL3* in *U. gibba* correlates with a decreased proportion of 24-nt small-interfering RNAs, low genome methylation levels, and developmental abnormalities during female gametogenesis that are reminiscent of RdDM mutant phenotypes in *Arabidopsis thaliana*. It would be interesting to further study the biological implications of the DCL3 truncation in *U. gibba*, as it could represent an initial step in the evolution of apomixis in compact genomes.

Background

Epigenetic modifications are chemical additions to DNA and/or histones that are associated with changes in gene expression and can be heritable [1]. One of these epigenetic modifications is DNA methylation at the 5′ position of the cytosine base (m5C), an ancient evolutionary trait associated with gene and transposable element (TE) silencing in eukaryotes [2]. In plant genomes, m5C is a widely conserved epigenetic mark that modulates gene expression and plays a key role in many developmental processes and environmental responses [3,4].

Non-coding RNAs (ncRNAs) are fundamental in regulating DNA methylation and the accessibility to genetic information [5]. For instance, mobile short RNAs (sRNAs) underlie shoot to root communication of methylation status [6] and long-noncoding RNAs (lncRNAs) serve as scaffolds in *de novo* DNA methylation [7]. RNA-directed DNA methylation (RdDM) is the major small RNA-mediated epigenetic pathway in plants. The canonical pathway can be subdivided into 3 different phases: 1) RNA polymerase IV (Pol-IV) dependent biogenesis of small interfering RNAs (siRNAs), 2) RNA polymerase V (Pol-V) mediated *de novo* methylation, and 3) chromatin modifications [8]. In addition to Pol-IV and Pol-V, other key components in RdDM and other sRNA biogenesis pathways key protein families are the ARGONAUTE (AGO), DICER-LIKE (DCL), and RNA DEPENDENT RNA POLYMERASE (RDR). Specifically, in the production of siRNAs, RDRs synthesize the second RNA strand, DCLs process RNA precursors, and AGOs select one DNA strand and load sRNAs to a specific target [8,9].

A considerable number of plant methylomes have been analyzed by whole-genome bisulfite sequencing (BS-Seq) or by high-performance liquid chromatography [10,11], where the vast majority of these methylomes are from model flowering plants. Methylome analysis has revealed variation of methylation patterns in intergenic and gene body regions among different plant species [11,12] and these patterns could be related to genome architectural features such as genome size, rearrangements, duplications, and content/expansion of transposable elements (TEs), among others. Little is known about the relationship between genome methylation, number, type of ncRNA *loci*, and genome size, particularly for plant species with small genomes with reduced TEs and other repetitive sequences.

Utricularia gibba is an aquatic, carnivorous plant belonging to the asterid lineage that despite having undergone two recent whole genome duplication (WGD) events, has a remarkably small genome size (ca. 100 Mb) [13]. *U. gibba* is a rootless plant that harbors slender green stems that grow like stolons with alternate thread-like leaves and numerous complex trapping bladders that catch small invertebrate prey. *U. gibba* has a gene repertoire similar to other plants species with larger genomes, albeit with a reduced non-coding genome harboring short intergenic regions (IGR) and a low TE content [13]. Here, we took advantage of the small *U. gibba* genome to study the impact of genome size on the conservation of “non-coding DNA”, especially on nature of genes coding for ncRNAs and the impact of the ncRNA landscape on genome methylation. Here, we report that *U. gibba* has, compared to Arabidopsis, a functionally incomplete canonical pathway for siRNAs biogenesis and *de novo* methylation that correlates with a rather unusual content and proportion of sRNA. In addition, our data suggests the interesting hypothesis that the unusual siRNA content and altered female gametogenesis in *U. gibba* could be due to the lack of a functional *DCL3*. Finally, we suggest that Pacific Bioscience (PacBio) Single Molecule Real-Time (SMRT) long read sequencing data to determine m5C methylomes is a useful alternative to BS-seq.

Results

U. gibba lncRNAs has a reduced number of intergenic lncRNAs

lncRNAs were obtained from sequencing data of 12 RNA-Seq libraries, 10 from vegetative and 2 from trap tissue from plants subjected to the contrasting abiotic stress or hormone treatments (Additional file 1: Dataset S1). We produced over 19 million mapped reads per library (95.37% mapping to the genome) for a total of 330,258,977 million mapped reads (Additional file 1: Dataset S1). To identify lncRNAs, assembled transcripts were translated into the three potential reading frames and filtered by homology to proteins encoded in the *U. gibba* genome and other

plants in a non-redundant protein database [14], yielding 10,386 putative lncRNAs of at least 200-nt in length. Putative lncRNAs were then filtered to eliminate precursors of tRNAs, rRNAs, miRNAs among others. Putative lncRNAs were further filtered for coding potential in sense-direction, resulting in 4,295 putative lncRNAs (Additional file 2: Fig. S1). The vast majority of these lncRNA *loci* were relatively short, 89.05% were smaller than 500-nt, and only 1.81% longer than 800-nt (Fig. 1a). The lncRNA mean length was 336.9-nt and the largest one was 2134-nt (Additional file 2: Table S1; Additional file 3: Dataset S2). We also found that in *U. gibba* 86.51% of lncRNAs had a single exon structure and 13.49% had two or more exons, percentages like other plant species for which lncRNAs have been reported (Fig. 1b; Additional file 1: Dataset S1). When putative lncRNAs were mapped onto the *U. gibba* genome, we determined that 37.03% mapped to regions corresponding to exons of protein coding genes in antisense orientation, 10.89% to intronic regions, 25.26% overlapped gene bodies and IGRs, and 26.82% were located only in IGRs (Fig. 1c). As it is more complex to determine whether a non-coding transcript that completely overlaps with the transcript of a coding gene indeed corresponds to a lncRNA, we made only a comparison of intergenic lncRNAs (lincRNAs) with other plant species. By mapping directly RNA-Seq reads and assembled contigs, we found that this carnivorous plant, under all conditions tested, expresses a total of 1152 lincRNAs, which is significantly lower than that reported for other plant species that ranges between 1580 and 3100 lincRNAs (Additional file 2: Fig. S2). Additionally, were evaluated the expression profiles for coding and noncoding transcripts and we found a higher level of expression for mRNA than for the entire catalog of lncRNAs or lincRNAs (Fig. 1d). We found no structural differences between lncRNAs annotated as intergenic or intragenic; in both cases the predominant structure was 1 exon (Additional file 2: Fig. S3, S4). To visualize the distribution lincRNAs *loci*, reads were mapped onto the 18 largest contigs of the *U. gibba* genome (>1Mb), which included 4 putatively entire chromosomes. *loci* encoding lncRNAs were distributed across the entire genome with prevalence in high gene density regions, and low frequency in pericentromeric regions (Fig. 2), which can be more clearly seen in the four complete chromosomes (Fig. 2; dark grey color). lncRNAs density was similar in 17 of the largest contigs of the *U. gibba* genome [15], except in Unitig_8 (6.8 Mb) that has considerably lower density of lncRNA *loci* (Fig. 2).

U. gibba has an atypical abundance of 24-nt sRNAs compared to other angiosperms

To further characterize noncoding RNA diversity in *U. gibba*, we carried out small RNA-Seq analysis of RNAs extracted from green tissue and traps from plants grown under the same conditions described for lncRNA identification. We obtained a total of 23.6 million mapped reads, of which 19.9 million had lengths between 20 to 25-nt, which were selected for further analyses (Additional file 1: Dataset S1). Upon mapping the reads onto the *U. gibba* genome, we found that *loci* for 20 to 25-nt sRNAs are mainly located at pericentromeric regions, where TE density is higher, but with some peaks in regions of high gene content (Fig. 2). Similar results have been published previously for other plant species [16].

An interesting finding was that most sRNA sequencing reads corresponded to 21-nt sRNAs (52.8%) and only 14.4% to 24-nt sRNAs (Additional file 2: Fig. S5; Additional file 1: Dataset S1). This contrasts with sRNA size distribution for other angiosperms, for which the most abundant sRNA class is 24-nt [17,18]. To confirm that the sRNA size distribution in *U. gibba* differs from that of other angiosperms, we performed an analysis of sRNA size abundance in representative plants from different clades for which sRNAs have been characterized (Additional file 1: Dataset S1). In total, we analyzed sRNA datasets for 30 plant species, of which 21 were angiosperms and 9 were representative plants outside the angiosperms. Our results confirm that in both monocot and eudicot species, except for *U. gibba*, the most abundant sRNA class is 24-nt (Fig. 3a). For the case of green algae, the most prevalent sRNAs classes are of 21 to 23-nt, whereas in *Volvox carteri* the most abundant sRNAs are of 21-nt and 22-nt, in *Chara coraline* the most abundant are 22-nt and 23-nt (about 30% of each size), and for *Chlamydomonas reinhardtii* the 21-nt class predominates (Fig. 3a). In early-branching land plants (*Marchantia polymorpha, Physcomitrella patens*, and *Marsilea quadrifolia*) and gymnosperms (*Picea abies*, *Gynkgo biloba*) the most abundant size class of sRNAs is 21-nt, except for *Cycas rumphii*, which has similar amounts of 21 and 24-nt sRNAs (Fig. 3a).

A large number of 24-nucleotide small RNA *loci* produce a small proportion of sRNA reads and are associated preferentially with intergenic regions in *U. gibba*

In general, sRNA sequence distribution in angiosperms is characterized by a major 24-nt peak containing primarily unique reads, and a 21-nt peak comprising many redundant reads [19]. As expected this is also true in model angiosperm species such as Arabidopsis [20], tomato [21] and rice [22]. For the three samples we sequenced (Additional file 2: Fig. S6), 21-nt sRNAs had higher redundancy (up to 97% of the reads are redundant) in comparison with 24-nt sRNAs, of which 30% were unique and 70% redundant (Additional file 1: Dataset S1). To better classify sRNA *loci*, we performed an analysis with ShortStack V2.0 [23] to identify *in silico* which *DICER*-like (*DCL*) genes are involved in the biogenesis of miRNAs or siRNAs. ShortStack analysis identified 7478 siRNA *loci* and only 80 miRNA *loci* (Fig. 3b; Additional file 4: Dataset S3), which produce nearly 1.5 million and 8 million mapped reads respectively, suggesting that a large number of siRNA loci accumulate a low number of reads (Fig. 3c). Of the 80 miRNA *loci* identified, 78 already were annotated in the miRBase V22.0 catalog of eukaryotic miRNAs, and 2 represent putative Utricularia specific miRNAs (Additional file 4: Dataset S3; Additional file 2: Fig. S6). miRNA *loci* grouped into 17 families, of which the miR166, miR156, miR159, miR319 and miR858 families made up 94% of the miRNA reads (Additional file 2: Fig. S7). These miRNA families are conserved in most plant species and produce high levels of mature miRNAs [18].

sRNA *loci* annotation reveals that the 24-nt sRNA *loci* are preferentially located at IGRs with 80% of them, while the sRNA *loci* of 20-nt to 22-nt were located in similar proportions at genic and intergenic regions and for the 23-nt sRNA class, 57% of *loci* are located at IGRs and 31% in genic

regions (Additional file 2: Fig. S8).The annotation is consistent with the distribution of sRNAs at the genome scale and can be clearly seen in unitig_0, in which 21-nt sRNA *loci* are distributed across the chromosome with significant peaks in regions with high gene density and 24-nt sRNA *loci* were found to be predominantly located in regions with high TE density, presumably pericentromeric regions (Fig. 3d).

sRNA biogenesis and the RdDM pathway in *U. gibba*

The unusual proportions of 24-nt and 21-nt sRNAs observed in *U. gibba* suggest that its sRNA production machinery could differ somehow from those of other angiosperms. To explore this possibility, we focused on the presence of genes involved canonical miRNA biogenesis, genes that are part of the subunits of RNA Pol-IV and RNA Pol-V, homologs of AGO, DCL, RDR, and other key genes involved in siRNA biogenesis and *de novo* DNA methylation in *U. gibba*. We searched based on sequence homology (transcript and protein), protein domain conservation, synteny, and through phylogenetic analysis. Furthermore, we performed a comparison with representative plant species (both angiosperm and non-angiosperm) for which key RdDM pathway genes [24] and RNA polymerase compositions were previously reported [25].

We found that key genes involved in canonical miRNA biogenesis such as *DICER LIKE 1 (DCL1)*, *SERRATE* (*SE*), *HYPONASTIC LEAVES 1* (*HYL1*), *HASTY 1* (*HST1*) and *ARGONAUTE 1* (*AGO1*) are conserved in *U. gibba* (Additional file 2: Table S2). Pol-IV and Pol-V are crucial in the RdDM pathway and are constituted by diverse DNA-DIRECTED RNA POLYMERASES IV AND V (NRPD/NRPE) proteins. Pol-IV and Pol-V differ from RNA polymerase II in their second, fourth, fifth, and seventh subunits (NRPD2, NRPD4, NRPE5, NRPD7, respectively). The *U. gibba* genome encodes *NRPD1, NRPE1 and NRPD2* genes (Additional file 2: Fig. S9), which is consistent with the presence of siRNAs and their strong conservation in the land plant lineage. We also identified NRPD7 in the *U. gibba* genome (Additional file 2: Fig. S9), a subunit previously identified in green eukaryotes except in the algae Chlamydomonas, and NRPD5, found in gymnosperms and angiosperms but not in algae and ferns [25,26]. Only after an extensive search did we find evidence for an NRPD4 ortholog in *U. gibba*. Although this gene is classified as an orphan in Plaza 4.0 (HOM04D168668) for *Arabidopsis thaliana*, our analysis showed that this categorization is likely related to the *Arabidopsis thaliana* homolog being extremely divergent; even the *A. lyrata* ortholog was readily placed into a clear NRPD4 gene family (Additional file 2: Fig. S10). Interestingly, this subunit has been reported only in angiosperms and not in gymnosperms, early land plants or algae [25,26].

The AGO protein family in Arabidopsis has been subdivided into 4 clades: AGO2/3/7, AGO4/6/8/9, AGO1/10 and AGO5. The number of family members of this protein family ranges from 2 AGO proteins in *C. reinhardtii* to 10 in Arabidopsis and 20 in *Z. mays* [27]. We searched in the Phylome database v4 [28] for possible homologous genes in *U. gibba* and found evidence for at least one AGO for each clade, with the exception of AGO5 for which no homologs were found (Additional file 2: Fig. S11). Two genes were found for AGO clade 2/3/7 (Additional file 2: Fig. S12), two genes represented the AGO 4/6/8/9 clade (Additional file 2: Fig. S13), and 4 homologs grouped in the AGO 1/10 clade (Additional file 2: Fig. S14). Additionally, we performed our own exhaustive phylogenetic analysis with many plant genomes to assign each *U. gibba AGO* gene with more certainty into specific clades. We were able to identify the same 8 AGOs described above, wherein the 2 homologs of clade 2/3/7 apparently are AGO7 copies, the two 4/6/8/9 copies appear closer to AGO4 than AGO6. AGO8/9 are sister genes only in Brassicaceae. There is one *U. gibba* AGO10 and three homologs for AGO1 in the AGO1/10 clade (Additional file 2: Fig. S15).

In seed plants there are three *RDR* ortholog genes; two are conserved in all land plants, *RDR1* and *RDR6,* and *RDR2* is required for production of Pol-IV-siRNAs and is specific to seed plants. Aside from these RDRs, three additional members of this protein family, RDR3, RDR4, and RDR5, are present in Arabidopsis and other plants [29–31]. We found RDR6 (two copies), RDR1 and RDR2, but no evidence for the presence of RDR3/4/5 in the genome of *U. gibba* (Additional file 2: Fig. S16a). In angiosperms the DCL family has 4 members: DCL1, DCL2, DCL3, and DCL4 [32]; however, in lycophytes and ferns there is only evidence for the presence of DCL1, DCL3 and DCL4 [26]. Phylogenetic analysis of the DCL family permitted the identification in *U. gibba* of 4 DCL proteins (DCL1, DCL2, DCL3, DCL4), suggesting that has a DCL repertoire similar to other angiosperms (Additional file 2: Fig. S16b). Globally, were able to assign *U. gibba* homologs for the remaining key genes in the canonical RdDM pathway (Additional file 1: Dataset S1). However, although UgDCL3 phylogenetically groups very closely to tomato and Mimulus DCL3, UgDCL3 is missing its N-terminal region, where the conserved DEAD/DEAH, Helicase and Dicer dimerization domains are located (Fig. 4a; Additional file 2: Fig. S17). The absence of about 600 amino acids of the N-terminal region in UgDCL3 is evident in a multiple sequence alignment of DCL3 with other angiosperms (Fig. 4b; Additional file 2: Fig. S18). To confirm that the incomplete DCL3 does not represents mistake in the assemble/annotation of the *U. gibba* genome, we searched for the presence of DCL3 transcripts in the different RNA-Seq libraries and we confirm with a 5´ Rapid Amplification of cDNA Ends (5´RACE) that in both cases the sequence of transcripts and 5´ RACE sequence that the *U. gibba* DCL3 gene is missing the DEAD/DEAH domain (Additional file 2: Fig. S19).

Female gametogenesis in *U. gibba* is reminiscent of Arabidopsis mutants affected in the RdDM pathway

Mutations affecting most of the genes involved in the RdDM pathway have no obvious phenotype during the vegetative development of plants, but show defects in female gametogenesis, including the differentiation of supernumerary gametic precursors that often give rise to ectopic female gametophytes within the ovule [33,34]. To determine if the truncation of DCL3 and the unusual distribution of sRNAs could be related with female gametogenesis in *U. gibba* as has been reported for Arabidopsis, we first analyzed female gametogenesis in whole-mounted developing ovules, as no descriptions of ovule development have been previously reported for this species. Our results are described and illustrated in Fig. 4c-i and in Additional file 2: Table S3.

As for other species of *Utricularia* [35,36], the ovule of *U. gibba* is unitegmic, with a funiculus forming a raphe and merging into a voluminous placenta. The formation of differentiated gametes occurs after the formation of meiotically derived megaspores (megasporogenesis). Subsequent rounds of mitotic divisions give rise to the female gametophyte (megagametogenesis). Megasporogenesis occurs in ovules within ovaries having a diameter of 0.3 to 0.5 mm. Whereas 51.4% (n=142) of pre-meiotic ovules showed a single megaspore mother cell (MMC; Fig.4c), 42.9% showed from two to six differentiated cells resembling the MMC (Fig. 4d and e). In 5.7% of the ovules examined we could not identify a pre-meiotic precursor. Ovules contained in 0.5-1 mm ovaries had already undergone meiosis and often show a chalazal functional megaspore (FM; Fig. 4f) within which mitotic divisions will give rise to an 8-nucleated syncytium (Fig. 4g) that will cellularize before differentiating into a mature female gametophyte (FG). Although the degeneration of the FG prior to the end of megagametogenesis is not uncommon (13.1%; n=76), in most cases (40.7%; n=76) the ovule contains a FG in which the micropylar region containing the egg apparatus expands outside the integument and grows within the placenta (Fig. 4g and 4h). Interestingly, 22.4% of ovules examined showed supernumerary gametic cells in the chalazal region, independent of the developing FG (Fig. 4h), and containing two independently developing FGs (Fig. 4i), suggesting that supernumerary gametic precursors can give rise to female gametophytes that may or may not originate from a meiotically derived cell. The presence of supernumerary gametic precursor cells and ectopic female gametophytes in *U. gibba* is reminiscent of phenotypes found in Arabidopsis mutants *dicer-like3* (*dcl3*), *argonaute4* (*ago4*), *argonaute9* (*ago9*), *rna-dependent rna polymerase6* (*rdr6*), and *nrpd1a*, all affected in key components of the RdDM pathway.

U. gibba has a reduced levels of DNA methylation

In plants, 24-nt siRNAs from repetitive DNA and TEs that are loaded by AGO4/6 trigger DNA methylation, which results in histone modifications such as the H3K9me2 [37,38]. Because of the unusual distribution of 24-nt sRNAs and the low proportion of TEs and other repetitive DNA in *U. gibba*, we decided to explore preliminarily the global DNA methylation patterns in this carnivorous plant using long-read DNA sequencing data with the technology of SMRT-Seq, recently reported [15]. This technology measures each base addition as an interpulse duration (IPD) or retention time ratio. The IPD will depend on whether the new base is incorporated by pairing to a modified or non-modified base in the template and the nature of the modification [39]. Therefore, analysis of the IPDs during SMRT-Seq can allow the identification of m5C in the DNA template without the need for commonly used bisulfite DNA chemical conversion methods. Since there are no previous reports of using SMRT-Seq data to determine m5C global methylation in plants, we tested as preliminarily the m5C identifications in *U. gibba* taking advantage of PacBio data from its genome (http://merlion.scelse.ntu.edu.sg/shares/pbio_HGYDGSKAA23/). Raw SMRT-Seq data was aligned against the reference genome and base kinetics information analyzed using the program SMRT-link V4.0 to identify base modifications and the theoretical IPD value for non-modified bases which is 1 was used.

The genome-wide depth with SMRT sequencing was ~ 70X and the mean coverage for all bases was 34X. We identified 1,590,729 putative methylated cytosines and 1,088,032 high-confidence m5Cs (IPD >= 1.7), which represents 3.88% and 2.69% of the Cs in the *U. gibba* genome, respectively. DNA methylation corresponding to all methylated cytosines in high confidence m5Cs for each context was scored: 37.30% CG methylation, 22.22% in CHG context, and 40.45% in CHH (Additional file 2: Fig. S20a). These preliminarily results suggests that PacBio data could be useful to obtain a methylation landscape and we decided to explore the methylation levels and gene body methylation (GbM). The methylation levels for each context were 10% for CG, 6% for CHG and 2.4% for CHH (Additional file 2: Fig. S20b), which are reduced levels in comparison with other plant genomes [40]. On other hand, in the GbM we found lower DNA methylation density levels in upstream/downstream regions than in gene bodies, with major peaks located near start/stop codon sites (Additional file 2: Fig. S20c).

Moreover, to confirm these preliminary m5C identifications we performed BS-Seq for two replicates of the whole-plant of *U. gibba*. The bisulfite conversion rates in both replicates were greater than 99.85% and the mean coverage for base was 26X. After sequencing, the clean reads were mapped against the reference genome obtaining around 85% of mapping rate. During the base calling we found a total of 1,789,535 and 1,777,410, m5Cs for each replicate. Of these, we assigned as high-confidence 1,281,545 and 1,302,422 m5Cs, respectively. In percentages of total cytosines in the genome, the global methylation is ranged from 4.09%-4.057% to 2.92%-3.17%, from total and high-confidence m5C in both replicates, respectively. The distribution of m5Cs at the chromosome level shows peaks near centromeric regions for both replicates of BS-Seq and for SMRT-Seq (Fig. 5a) with very similar distributions, similar to that reported in other plant genomes but with one of the lowest global methylation rates reported for a land plant [11,12]. Of the total of m5Cs identified, for the methylation context CG correspond 39.1%, 38.67%, for CHG 27.4% and 27.1% and for CHH context 33.5% and 34.23% (Fig. 5b). When comparing the methylation levels of each context related with the total of these contexts sequence at the genome, the methylation levels were 12% for CG, 8% for CHG and 2% for CHH (Fig. 5c), reduced levels in comparison with other angiosperms including *A. thaliana* [40]. The GbM in *U. gibba* was calculated 800bp upstream, downstream and in genic region and the results indicate lowest methylation in upstream and downstream region compared with gene body region where the CG methylation is the highest (Fig. 5d). These results correlates with GbM densities previously reported for other plants [40,41]. The analysis of TE body methylation showed lower methylation levels near upstream and downstream regions such is in GbM and high methylation density is in CHH context (Fig. 5e), which can be explain the TE silencing in this genome.

To experimentally validate the predicted methylation densities, we performed Chop-qPCR analysis on two regions from one *U. gibba* chromosome (Unitig 0), one predicted to be highly methylated (GPoor, near the presumed centromeric region) and the other with lower methylation density at a

proposed euchromatin region (GRich) (Additional file 2: Fig. S21 above). PCR amplification of undigested DNA was used as positive controls to contrast with the amplification obtained from digested samples, where a decrease in amplification levels reveals a higher degree of methylation at CG and CHG sites in the tested regions. We observed for both enzymes that DNA methylation at the analyzed sites was over 2-fold more frequent in the GPoor region than in the GRich region (Additional file 2: Fig. S21). Our findings suggest that the SMRT-sequencing technology could be used effectively to estimate global methylation patterns in plant genomes and BS-Seq confirm several results obtained with SMRT-Seq.

Discussion

Genome rearrangements after WGD in *U. gibba* and their role in ncRNA content

The causes and mechanisms of WGD events and genome fractionation processes that lead to genome expansion and contraction in *U. gibba*, are still poorly understood. Moreover, the consequences of these processes on the repertoire of non-coding RNAs and epigenetic processes remain obscure. *U. gibba*, a carnivorous plant with an unusually small but dynamic genome that has experienced two relatively recent WGD events, represents a highly illustrative model to study the processes of genome contraction and its consequences on the diversity of ncRNAs, as well as its consequences on epigenetic processes. Our analysis of ncRNAs provides insights into possible reasons why this carnivorous plant has an unusual siRNAs distribution. One of our interesting observations is that in contrast to other angiosperms, such as Arabidopsis, maize, and tomato, for which lncRNA *loci* are preferentially located in centromeric regions, in *U. gibba* lncRNA *loci* are well distributed across the genome but with a lower density in centromeric regions (Fig. 2). Centromeres are genomic sites of spindle attachment essential for ensuring proper chromosome segregation during cell division. Despite their recognized functional importance, centromeres are not well defined at the sequence level in eukaryotic genomes except for some small fungal genomes [42]. In general, centromeres are accepted to be composed of high-copy tandem satellite repeats and/or the presence of centromeric chromoviruses, a lineage of Ty3/gypsy retrotransposons. Also, these sequences may be merely parasitic and tend to accumulate in recombination-poor centromeric regions to escape negative selection [43]. The low density of lncRNAs observed in the putative centromeric regions of the *U. gibba* genome could be related to the absence of both satellite repeats and paucity of centromeric chromoviruses in these regions [15]. Centromeres without long arrays of satellite DNA have been referred to as evolutionarily new centromeres [44]. It is possible that after the last WGD event in *U. gibba*, genome fractionation resulted in the formation of neo-centromeres targeted by few lncRNAs. These findings are consistent with the notion that epigenetic marks, in the form of stable, self-propagating chromatin states, rather than sequence specific structural features, define a functional centromere.

A truncated *DCL3* gene correlates with the distribution of sRNAs and impact the developmental biology of *U. gibba*

TE proliferation shapes genome sizes among eukaryotes. In plants, one of the mechanisms to control this expansion is through TE silencing via 24-nt siRNAs [45,46]. In the siRNA biosynthesis pathways, at least three RNA-dependent RNA polymerases (RDR1, RDR2, and RDR6) are functional in plants [29,30]. RDR2 works with DCL3 to form chromatin associated siRNAs (mainly 24-nt) that are involved in sRNA mediated DNA methylation [8]. As expected from its known role in the biogenesis of 24-nt small RNAs, Arabidopsis and maize mutants are defective in RDR2 or DCL3 show a severe reduction in the production of the 24-nt size siRNA class, whereas the 21-nt class (miRNA) is not affected and indeed becomes overrepresented [47,48]. In rice, DCL3 mutations also affect the production of 24-nt siRNAs associated with TEs [49]. Here, we report that *U. gibba* has an unusually a sRNA size distribution more similar to early-branching land plants and gymnosperms than to other angiosperms and low content of 24-nt sRNAs (Fig. 3a). However, the accumulation of 24-nt siRNAs in *U. gibba* is similar to *dcl3* mutants in Arabidopsis, rice and tomato [47,49,50], which correlates well with the presence in *U. gibba* of a truncated form of DCL3 (Fig. 4a,b). The loss of a fully functional DCL3 in *U. gibba* may be responsible for the reduction in the 24-nt class of siRNAs, as well as the reduced level of DNA methylation in its genome. It is possible that the loss of a fully functional DCL3 in *U. gibba* might have originally reflected lower selective pressure to silence fewer TEs. Furthermore, we report the first quantitative analysis of female gametogenesis in a member of the *Utricularia* genus and these developmental ovule defects have not been reported previously for *Utricularia* species [36,51] and this study reveals severe abnormalities in the ovule development of *U. gibba*.

Since the truncated DCL3 is structurally preserved by purifying selection, some function, perhaps regulatory, likely remains. Animal Dicer and plant DCL proteins dimerize at their RNase III domains [52,53]. It is possible that the truncated DCL3 acts as an interfering subunit in the DCL protein complex to generate a dominant negative phenotype [54]. As hypothesis, this dominant negative effect could suppress the production of the 24-nt class of siRNAs accompanied by defects that include the absence of a female gametophyte that lead to partial female sterility, or the differentiation of female ectopic precursors that give rise to supernumerary gametophytes. It would be interesting to further study the biological implications of the DCL truncation in *U. gibba*, as it could represent an initial step in the evolution of apomixis and further developmental and genetic studies are required to determine if the latter could eventually result in variable frequencies of unreduced gamete formation, polyploidization or apomixis.

SMRT-Seq as an alternative to decipher plant methylomes

All plant methylomes sequenced to date have been generated by BS-Seq [39]. In recent years, however, new techniques for sequencing has generated the possibility of direct DNA base modification detection without any chemical conversion [39]. Recently, genome-wide DNA methylation

on N-6 adenine (m6A) using SMRT-sequencing technology was reported for some animals and plants [55–57]. Our results of *U. gibba* BS-Seq support the use of SMRT sequencing technology to determine m5Cs on whole genomes. At least in qualitative terms, SMRT technology produces whole-genome m5C data similar in the percentages of methylated cytosines, the proportions of methylation context (CG, CHG and CHH), genome-wide m5C distribution and GbM to those obtained by bisulfite sequencing. Similar findings we obtain with SMRT-Seq preliminarily data and BS-Seq in the model plant *A. thaliana,* which strongly support our results (Additional file 2: Fig. S22). Some of the methylation variability between identification methods can be explained by de dynamic DNA methylation among cells, individuals or natural variation [58–60], bias in libraries preparation strategies [61] or the differences between methods. BS-Seq has so far been the standard for methylome determination and has been quite useful to derive insights into the roles of DNA methylation on gene expression and chromatin remodeling. However, because BS-Seq is based on short-length sequencing reads that provide a statistical probability that a cytosine is methylated, it is difficult, if not impossible to determine which m5C modifications need to be present in the same molecule to influence gene expression and/or chromatin remodeling. Although further analyses will be required to ascertain the extent to which SMRT technology can generate robust quantitative data on m5C DNA modifications, the fact that methylation patterns can be established linearly for long-reads could serve to establish which m5C modifications are truly necessary to influence gene expression or chromatin modifications.

Lower DNA methylation levels in *U. gibba* in comparison with other plant genomes

Focusing on canonical genes in RdDM pathway, *de novo* DNA methylation involves Pol-IV, Pol-V and several key genes [8,25]. However, we could not uncover a direct AGO6 ortholog, and the DCL3 ortholog is truncated in its catalytic domains. In Arabidopsis *dcl3* and *ago6* mutants have reduced DNA methylation levels at RdDM targets [62,63]. Moreover, AGO4 and AGO6 in Arabidopsis are not as redundant as was originally thought. Indeed, physical interactions demonstrate that they have specific functions: AGO4 is colocalized with Pol-II, and AGO6 interacts with Pol-V [63]. Global DNA methylation of *U. gibba* ranges from 2.65 to 3.9%, lower than that reported for other plant species, which range between 5% in *Theobroma cacao* to 43% in *Beta vulgaris* [11]. Our results suggest that the reduced level of 24-nt siRNAs in *U. gibba* could be due to the presence of a truncated DCL3 and/or the absence of AGO6, in turn impacting DNA methylation, mainly in methylation level of CHH context which in this work reported as reduced compared with *A. thaliana* and other angiosperms.

Conclusions

First, this study on *Utricularia gibba*, a plant with a remarkably small genome size that has undergone drastic genome contraction after two recent events of WGD, sheds light on noncoding RNA landscape and its impact on DNA methylation in plants with compact genome size, which provides valuable information for future studies of plant genome engineering and functional genomics. Second, they provide a genome-wide picture of the DNA methylation in *U. gibba*, estimated by long-read sequencing with SMRT technology of PacBio and tested in the model plant *A. thaliana*, which also confirm that PacBio data potentially provide an overall picture of methylation status. Third, they report that *U. gibba* genome lacks a functional RdDM pathway - where a truncated DCL3 correlates with an unusual repertoire of sRNAs, low global DNA methylation and may impact the reproductive developmental biology of this carnivorous plant-. Finally, we propose that alterations in the RdDM pathway could be the result a relaxed DNA methylation control of the low content of transposable elements in *U. gibba*.

Methods

Plant material

U. gibba was collected at the Umécuaro village in the municipality of Morelia, Mexico, in 2009 (Ibarra-Laclette et al, 2013) and since clonally propagated in our laboratory on liquid media MS (0.25X) at 22 C with 16 hours light and 8 hours of darkness. Subcultures, at day 14 of growth, were used in our experiments. We employed the following growth conditions and treatments: nutrient deficiency, exposure to plant hormones, high and low temperatures, continuous illumination and darkness, treatment with acidic and alkaline pH, osmotic stress, and saline stress. Traps and plant tissue were collected separately after short (48 and 72 hours) and long (7 and 14 days) time treatments for total RNA extraction and subjected to RNA-Seq (Additional file 1: Dataset S1).

Total RNA extraction and sequencing

Total RNA from three independent biological replicates for all the analyzed treatments was isolated using the TRIzol reagent (Life technologies), except for trap tissue, which was isolated using the Direct-zol RNA kit (Zymo Research) because this protocol optimizes RNA extraction for low tissue quantities and improved RNA quality for these samples. Twelve RNA-Seq libraries were prepared with the TrueSeq (Illumina technologies) kit and sequenced by non-directional single-end mode.

Mapping reads and lncRNAs identification

Trimmed reads were mapped against the latest available version of the *U. gibba* genome (https://genomevolution.org/coge/GenomeInfo.pl?gid=28800). To potentiate the identification of lncRNAs in *U. gibba* we decided to merge mapped reads for each library in a single file to obtain a

core transcriptome assembly with Trinity v2.0 [64]. The identification works with three major filters: I) One of the main features for lncRNAs is size, then during the assembly the transcripts with less than 200 nucleotides in size were cut-off. The first step was to search protein orthologs against a database of non-redundant eukaryotic proteins [14], this step eliminates those transcripts that encode for proteins. II) Putative lncRNAs transcripts were filtered finding homologs with housekeeping RNAs such are transfer-RNAs (tRNAs), ribosomal-RNAs (rRNAs), micro-RNAs (miRNAs) and other types of structural RNA using the database Rfam v12.0 [65] and sequence homologs with precursors of small RNAs from *U. gibba* dataset. III) Third step relays in the evaluation of the coding potential of the remaining transcripts using the tools CPC2 [66] and CPAT v1.2.4 [67] using the corrected training for *U. gibba* model. The lncRNAs identification is described and summarized in Additional file 2: Fig. S1.

Small RNA sequencing and annotation

RNA was extracted independently from each sample and then equimolarly pooled to produce three small RNA libraries (Additional file 1: Dataset S1). Small RNA libraries were constructed using the TrueSeq Small RNA kit (Illumina Technologies) and sequenced using NextSeq (Illumina Technologies) using a 50-nucleotide single end runs. Cleaned reads were mapped against the *U. gibba* reference genome and the alignments were then run in ShortStack V2.0 [23]. ShortStack-count mode was used to find relative small RNA abundances of *de novo*-identified sRNAs *loci*. The sRNAs reads were mapped against mature sequences in the miRNA database miRBase V22.0 (blastn –word_size 7 –dust no –identity 90 –evalue 1000).

Phylogenetic analysis

Were identified the homologs among plant genomes in Phytozome version 12.1 (http://phytozome.jgi.doe.gov/; [68]). For all sequences, homology analysis was performed (Blastx E-value < 0.001, Bit-Score > 70). Additionally, we performed a bidirectional blast analysis against Arabidopsis protein database from TAIR V10 (www.arabidopsis.org) using previous parameters. Protein multiple alignments were made using MAFFT V7.0 [69] and trimmed by trimAL [70]. Finally, the best model fit was selected to execute the phylogenetic inference with ProtTest V3.0 [71]. The Bayesian phylogenetic reconstruction was executed using the software MrBayes V3.2 [72] with 300000 generations. Also, likelihood phylogenetic reconstruction was estimated using RAxML back box on website CIPRES (https://www.phylo.org/) with GTR + G model and 1000 cycles for bootstrapping.

Cytological analysis of ovule development

For cytological examination of ovules, whole flowers were harvested and fixed in formalin acetic acid-alcohol solution (40% formaldehyde, glacial acetic acid, 50% ethanol; in a 5:5:90 volume ratio) for 24 hours at room temperature, and subsequently stored in 70% ethanol at 4° C. Fixed ovaries were dissected with hypodermic needles (1 mm insulin syringes), cleared, and observed by differential interference contrast microscopy using a Leica DMR microscope.

Identification of m5C with SMRT-sequencing and validation

Raw data was obtained from two sources: 1) The PacBio *U. gibba* SMRT sequencing data were provided by the authors of genome sequencing paper (http://merlion.scelse.ntu.edu.sg/shares/pbio_HGYDGSKAA23/) (4). 2) The Arabidopsis PacBio data was downloaded from PacBio public database for ler-0 ecotype (http://datasets. pacb.com.s3.amazonaws.com/2014/Arabidopsis/list.html). Each PacBio SMRT-cell provides raw data in bax.h5 format, these raw data were aligned against plant reference genome (*A. thaliana* or *U. gibba*) using pbaling program in base modification mode from SMRT-Link program designed by PacBio. For each case, to obtain more coverage and depth were merged the SMRT-cells (10 SMRT-cells for *U. gibba* and 40 SMRT-cell for Arabidopsis) and with loadPulses scripts (SMRT-link) the polymerase kinetics information for each base was obtained as average of IPD-ratio. Finally, the modification of cytosine was identified using ipdSummary.py script (SMRT-link).

The theoretical IPD value for non-modified bases is 1. We found that for non-modified cytosines the median of IPD ratio was 1.02 (S=0.19), whereas for modified cytosine (m5C) the median was 1.6699 and the mean 1.84, whereas for high confidence modified cytosine (m5C) the median was 1.89 (S=0.33) and average IPD was 2.02, which corresponds to the expected IPD for m5C. Automatically, the m5C high confidence position is identified and the m5C putative positions (modified bases) were identified performing a filter related with IPD-ratio >=1.79 (between median and average) of m5C high-confidence for non-centromeric regions. Raw data in bax.h5 format was aligned using pbaling in base modification mode. Polymerase kinetics information was further loaded after alignment using loadPulses scripts. Finally, m5Cs were identified using the ipdSummary.py script. We used a filter of 15X coverage to select for m5Cs. Parts of the analysis were done using customized scripts in R, MySQL, and Perl.

Whole-Genome Bisulfite Sequencing

Genomic DNA from *U. gibba* fresh tissue 14-days old was grinded with liquid nitrogen and purified with DNeasy Plant Maxi Kit (Qiagen). For the library preparation, DNA samples were fragmented into 200-400 bp using a sonicator S220 (Covaris). Then DNA fragments were blunt ended and, a dA 3'-end addition was performed prior to sequencing adapter ligation. Illumina methylated adapters were used according to the manufacturer's instructions (Illumina). The DNA fragments were bisulfite treated with EZ DNA methylation Gold Kit (Zymo Research). The final DNA library was

obtained by size selection and PCR amplification. High-throughput pair-end sequencing was carried out using the Illumina HiSeq 2500 system according to manufacturer instructions. Clean reads were aligned to reference genome using Bismark software [73]. To identify the true methylated sites, methylated and unmethylated counts at each site from Bismark output was tested by binomial distribution.

Comparison between sequencing platforms in Arabidopsis and *U. gibba* methylomes

Once obtaining the m5C identification with SMRT-Seq in Arabidopsis, we used previous bisulfite sequencing dataset from our group [74] to compare both methods. We calculated the cytosine coverage per chromosome (the number of cytosines identified by sequencing method) with bisulfite sequencing and SMRT-Sequencing to compare the depth for each identification method. Also, the percentages of m5C identification per chromosome (number of cytosines methylated/total of cytosines) were calculated. The methylation contexts were calculated from motif (3-bases) generated. Finally, the comparisons were reported as metaplot graphs of m5C density per chromosome corrected by Z-score analysis and barplot for each library of Bisulfite sequencing and SMRT-sequencing.

Correlation analysis

Breafly, data consist of methylation density estimation by BS-Seq and SMRT-Seq in 604 syntenic genomic windows, each one covering 100,000 bp for a total genome coverage of 60.4 Mb. Pearson correlation coefficient, r, was calculated for the whole set of syntenic regions to estimate correlation at genomic level as well as dividing the syntenic set by chromosome of origin. The normalized methylation density data were plotted in R. In each case the null hypothesis of zero correlation was tested.

Declarations

Acknowledgments

The authors thank G. Alejo-Jacuinde, B. Pérez-Sánchez and Q. Ortiz-Vasquez for help collecting plants and flowering specimens. Special thanks to R. Purbojati and S. Schuster for providing us the PacBio raw sequencing data. We thank J. Cervantes-Luevano for informatic support, Luis Delaye for evolutionary analysis advice and Koribinian Schneeberger for provided us syntenic dataset generating by SyRI. Authors acknowledge Dr. T. Markow for critical review of this manuscript. S.A.C.-P., I.A.-M. and F.R.-R are the recipients of a graduate scholarship from Conacyt. This work was supported in part by a Senior Scholar grant from the Howard Hughes Medical Institute and a Basic Science grant from CONACytT Mexico to L.H.-E.

Data availability

All files containing reads and quality scores were deposited in the National Center for Biotechnology Information (NCBI) archive [BioProject: PRJNA526734; SRA: SRR8717102, SRR8717105, SRR8717107, SRR8717103, SRR8717106, SRR8717097, SRR8717101, SRR8717104, SRR8717093, SRR8717098, SRR8717099, SRR8717100 and BioProject: PRJNA633566; SRA: SRR8717102, SRR8717105].

Author contributions: S.A.C.-P. and L.H.-E. designed research; S.A.C-P., L.Y.-V, N.M.V.F.-Z, A.O.-A., I.A.-M., and F.R.-R. performed research; L.H.-E., VAA and J.P.V.-C. contributed new reagents/analytic tools; S.A.C.-P., L.Y.-V., N.M.V.F.-Z., T.L, O.M., J.P.V.C., V.A.A., and L.H.-E. analyzed data; and S.A.C.-P., VAA and L.H.-E. wrote the paper with inputs from all authors.

Ethics approval and consent to participate

Original sampling site of *Utricularia gibba* is not classified as protected natural area and *U. gibba* is not included as an endangered species in the Mexican Official Norms (NOM-059-SEMARNAT-2010), therefore permits were no required for collecting the plants. Sample collection and preservation were done according to institutional regulations and following Mexican and international regulations and guidelines. *U. gibba* was collected with prior consent of landowner (Mr. Gerardo Negrete Negrete)

Competing interests

The authors declare that they have no competing interests.

References

1. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11.
2. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet [Internet]. 2007;8:272–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17363976
3. Henderson IR, Jacobsen SE. Epigenetic inheritance in plants. Nature [Internet]. Nature Publishing Group; 2007;447:418. Available from: http://dx.doi.org/10.1038/nature05917

4. Zhang H, Lang Z, Zhu J-K. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol [Internet]. Springer US; 2018;19:489–506. Available from: http://www.nature.com/articles/s41580-018-0016-z
5. Wierzbicki AT, Haag JR, Pikaard CS. Noncoding Transcription by RNA Polymerase Pol IVb/Pol V Mediates Transcriptional Silencing of Overlapping and Adjacent Genes. Cell [Internet]. Elsevier; 2008;135:635–48. Available from: https://doi.org/10.1016/j.cell.2008.09.035
6. Lewsey MG, Hardcastle TJ, Melnyk CW, Molnar A, Valli A, Urich MA, et al. Mobile small RNAs regulate genome-wide DNA methylation. Proc Natl Acad Sci [Internet]. 2016;113:E801 LP-E810. Available from: http://www.pnas.org/content/113/6/E801.abstract
7. Ariel F, Romero-Barrios N, Jégu T, Benhamed M, Crespi M. Battles and hijacks: noncoding transcription in plants. Trends Plant Sci [Internet]. 2015;20. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1360138515000552
8. Matzke MA, Mosher RA. RNA-directed DNA methylation: An epigenetic pathway of increasing complexity. Nat Rev Genet [Internet]. Nature Publishing Group; 2014;15:394–408. Available from: http://dx.doi.org/10.1038/nrg3794
9. Bologna NG, Voinnet O. The Diversity, Biogenesis, and Activities of Endogenous Silencing Small RNAs in *Arabidopsis*. Annu Rev Plant Biol [Internet]. 2014;65:473–503. Available from: http://www.annualreviews.org/doi/10.1146/annurev-arplant-050213-035728
10. Alonso C, Pérez R, Bazaga P, Herrera CM. Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. Front Genet [Internet]. Frontiers Media S.A.; 2015;6:4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310347/
11. Vidalis A, Živković D, Wardenaar R, Roquis D, Tellier A, Johannes F. Methylome evolution in plants. Genome Biol. 2016;
12. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim K Do, Li Q, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biol [Internet]. Genome Biology; 2016;17:1–19. Available from: http://dx.doi.org/10.1186/s13059-016-1059-0
13. Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang T-H, et al. Architecture and evolution of a minute plant genome. Nature. 2013;498:94–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23665961
14. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium the U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics [Internet]. Oxford University Press; 2015;31:926–32. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4375400/
15. Lan T, Renner T, Ibarra-Laclette E, Farr KM, Chang T-H, Cervantes-Pérez SA, et al. Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. Proc Natl Acad Sci [Internet]. 2017;114:E4435–41. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1702072114
16. Morin RD, Aksay G, Dolgosheina E, Ebhardt HA, Magrini V, Mardis ER, et al. Comparative analysis of the small RNA transcriptomes of Pinus contorta and Oryza sativa. Genome Res [Internet]. Cold Spring Harbor Laboratory Press; 2008;18:571–84. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2279245/
17. Chen D, Meng Y, Ma X, Mao C, Bai Y, Cao J, et al. Small RNAs in angiosperms: sequence characteristics, distribution and generation. Bioinformatics [Internet]. 2010;26:1391–4. Available from: http://dx.doi.org/10.1093/bioinformatics/btq150
18. Montes R a C, de Fátima Rosas-Cárdenas F, De Paoli E, Accerbi M, Rymarquis L a, Mahalingam G, et al. Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. Nat Commun [Internet]. 2014;5:3722. Available from: http://www.nature.com/ncomms/2014/140423/ncomms4722/full/ncomms4722.html%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/24759728
19. Lelandais-Brière C, Sorin C, Declerck M, Benslimane A, Crespi M, Hartmann C. Small RNA diversity in plants and its impact in development. Curr Genomics. 2010;11:14–23.
20. Wang H, Zhang X, Liu J, Kiba T, Woo J, Ojo T, et al. Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions. Plant J. 2011;67:292–304.
21. Gao C, Ju Z, Cao D, Zhai B, Qin G, Zhu H, et al. MicroRNA profiling analysis throughout tomato fruit development and ripening reveals potential regulatory role of RIN on microRNAs accumulation. Plant Biotechnol J. 2015;13:370–82.
22. Jeong D-H, Park S, Zhai J, Gurazada SGR, De Paoli E, Meyers BC, et al. Massive Analysis of Rice Small RNAs: Mechanistic Implications of Regulated MicroRNAs and Variants for Differential Target RNA Cleavage. Plant Cell. 2011;23:4185–207.
23. Axtell MJ. ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA [Internet]. Cold Spring Harbor Laboratory Press; 2013;19:740–51. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3683909/
24. Ma L, Hatlen A, Kelly LJ, Becher H, Wang W, Kovarik A, et al. Angiosperms are unique among land plant lineages in the occurrence of key genes in the RNA-Directed DNA methylation (RdDM) Pathway. Genome Biol Evol. 2015;7:243–67.
25. Huang Y, Kendall T, Forsythe ES, Dorantes-Acosta A, Li S, Caballero-P??rez J, et al. Ancient origin and recent innovations of RNA polymerase IV and V. Mol Biol Evol. 2015;
26. You C, Cui J, Wang H, Qi X, Kuo LY, Ma H, et al. Conservation and divergence of small RNA pathways and microRNAs in land plants. Genome Biol. Genome Biology; 2017;18:1–19.
27. Rodríguez-Leal D, Castillo-Cobián A, Rodríguez-Arévalo I, Vielle-Calzada J-P. A Primary Sequence Analysis of the ARGONAUTE Protein Family in Plants. Front Plant Sci [Internet]. Frontiers Media S.A.; 2016;7:1347. Available from:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5007885/

28. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res [Internet]. 2013;42:D897–902. Available from: https://doi.org/10.1093/nar/gkt1177
29. Willmann MR, Endres MW, Cook RT, Gregory BD. The Functions of RNA-Dependent RNA Polymerases in Arabidopsis. Arab B [Internet]. American Society of Plant Biologists; 2011;9:e0146–e0146. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22303271
30. Hunter LJR, Brockington SF, Murphy AM, Pate AE, Gruden K, MacFarlane SA, et al. RNA-dependent RNA polymerase 1 in potato (Solanum tuberosum) and its relationship to other plant RNA-dependent RNA polymerases. Sci Rep [Internet]. The Author(s); 2016;6:23082. Available from: https://doi.org/10.1038/srep23082
31. Qin L, Mo N, Muhammad T, Liang Y. Genome-Wide Analysis of DCL, AGO, and RDR Gene Families in Pepper (Capsicum Annuum L.). Int J Mol Sci [Internet]. MDPI; 2018;19:1038. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29601523
32. Mukherjee K, Campos H, Kolaczkowski B. Evolution of Animal and Plant Dicers: Early Parallel Duplications and Recurrent Adaptation of Antiviral RNA Binding in Plants. Mol Biol Evol [Internet]. Oxford University Press; 2013;30:627–41. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3563972/
33. Olmedo-Monfil V, Durán-Figueroa N, Arteaga-Vázquez M, Demesa-Arévalo E, Autran D, Grimanelli D, et al. Control of female gamete formation by a small RNA pathway in Arabidopsis. Nature. 2010;464:628–32.
34. Hernández-Lagana E, Rodríguez-Leal D, Lúa J, Vielle-Calzada J-P. A Multigenic Network of ARGONAUTE4 Clade Members Controls Early Megaspore Formation in Arabidopsis. Genetics [Internet]. 2016/09/01. Genetics Society of America; 2016;204:1045–56. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27591749
35. Jhori BM, Ambegaokar KB SP. Comparative Embryology of the Angiosperms. New York: Springer Verlag; 1992.
36. Płachno BJ, Swiątek P. Actin cytoskeleton in the extra-ovular embryo sac of Utricularia nelumbifolia (Lentibulariaceae). Protoplasma [Internet]. 2011/07/24. Springer Vienna; 2012;249:663–70. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21786167
37. Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, Richard McCombie W, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature [Internet]. Macmillan Magazines Ltd.; 2004;430:471. Available from: http://dx.doi.org/10.1038/nature02651
38. Xu C, Tian J, Mo B. siRNA-mediated DNA methylation and H3K9 dimethylation in plants. Protein Cell. 2013;4:656–63.
39. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods [Internet]. Nature Publishing Group; 2010;7:461. Available from: http://dx.doi.org/10.1038/nmeth.1459
40. Takuno S, Ran J-H, Gaut BS. Evolutionary patterns of genic DNA methylation vary across land plants. Nat Plants [Internet]. Macmillan Publishers Limited; 2016;2:15222. Available from: http://dx.doi.org/10.1038/nplants.2015.222
41. Bewick AJ, Schmitz RJ. Gene body DNA methylation in plants. Curr Opin Plant Biol [Internet]. 2017;36:103–10. Available from: http://dx.doi.org/10.1016/j.pbi.2016.12.007
42. Buscaino A, Allshire R, Pidoux A. Building centromeres: home sweet home or a nomadic existence? Curr Opin Genet Dev [Internet]. Elsevier Current Trends; 2010 [cited 2019 Feb 27];20:118–26. Available from: https://www.sciencedirect.com/science/article/pii/S0959437X10000195?via%3Dihub
43. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res [Internet]. Cold Spring Harbor Laboratory Press; 2008;18:359–69. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18256242
44. Piras FM, Nergadze SG, Magnani E, Bertoni L, Attolini C, Khoriauli L, et al. Uncoupling of satellite DNA and centromeric function in the genus Equus. PLoS Genet [Internet]. Public Library of Science; 2010;6:e1000845–e1000845. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20169180
45. Lisch D. Epigenetic Regulation of Transposable Elements in Plants. Annu Rev Plant Biol. 2008;60:43–66.
46. Tenaillon MI, Hollister JD, Gaut BS. A triptych of the evolution of plant transposable elements. Trends Plant Sci [Internet]. Elsevier Ltd; 2010;15:471–8. Available from: http://dx.doi.org/10.1016/j.tplants.2010.05.003
47. Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, et al. Genome-Wide Profiling and Analysis of Arabidopsis siRNAs. PLoS Biol. 2007;5:0479–93.
48. Nobuta K, Lu C, Shrivastava R, Pillay M, De Paoli E, Accerbi M, et al. Distinct size distribution of endogenous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 2008;105:14958–63. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2567475/
49. Wei L, Gu L, Song X, Cui X, Lu Z, Zhou M, et al. 24-nt siRNAs that control agricultural traits in rice. Proc Natl Acad Sci. 2014;111:3877–82.
50. Kravchik M, Damodharan S, Stav R, Arazi T. Generation and characterization of a tomato DCL3-silencing mutant. Plant Sci [Internet]. Elsevier Ireland Ltd; 2014;221–222:81–9. Available from: http://dx.doi.org/10.1016/j.plantsci.2014.02.007
51. Płachno BJ. Female germ unit in Genlisea and Utricularia, with remarks about the evolution of the extra-ovular female gametophyte in members of Lentibulariaceae. Protoplasma. 2011;248:391–404.

52. Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W. Single Processing Center Models for Human Dicer and Bacterial RNase III. Cell [Internet]. Elsevier; 2004;118:57–68. Available from: https://doi.org/10.1016/j.cell.2004.06.017
53. Mickiewicz A, Sarzyńska J, Miłostan M, Kurzyńska-Kokorniak A, Rybarczyk A, Łukasiak P, et al. Modeling of the catalytic core of Arabidopsis thaliana Dicer-like 4 protein and its complex with double-stranded RNA. Comput Biol Chem [Internet]. 2017;66:44–56. Available from: http://www.sciencedirect.com/science/article/pii/S1476927116303371
54. Veitia RA, Caburet S, Birchler JA. Mechanisms of Mendelian dominance. Clin Genet [Internet]. John Wiley & Sons, Ltd (10.1111); 2018;93:419–28. Available from: https://doi.org/10.1111/cge.13107
55. Greer EL, Blanco MA, Gu L, Sendinc E, Liu J, Aristizábal-Corrales D, et al. DNA methylation on N6-adenine in C. elegans. Cell. 2015;161:868–78.
56. Wu TP, Wang T, Seetin MG, Lai Y, Zhu S, Lin K, et al. DNA methylation on N6-adenine in mammalian embryonic stem cells. Nature [Internet]. Nature Publishing Group; 2016;532:329–33. Available from: http://dx.doi.org/10.1038/nature17640
57. Liang Z, Shen L, Cui X, Bao S, Geng Y, Yu G, et al. DNA N6-Adenine Methylation in Arabidopsis thaliana. Dev Cell [Internet]. Elsevier Inc.; 2018;45:406-416.e3. Available from: https://doi.org/10.1016/j.devcel.2018.03.012
58. Busche S, Shao X, Caron M, Kwan T, Allum F, Cheung WA, et al. Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. Genome Biol [Internet]. Genome Biology; 2015;16:1–18. Available from: http://dx.doi.org/10.1186/s13059-015-0856-1
59. Yagound B, Smith NMA, Buchmann G, Oldroyd BP, Remnant EJ, Costantini M. Unique DNA Methylation Profiles Are Associated with cis-Variation in Honey Bees. Genome Biol Evol. 2019;11:2517–30.
60. Zhang Y, Wendte JM, Ji L, Schmitz RJ. Natural variation in DNA methylation homeostasis and the emergence of epialleles. Proc Natl Acad Sci U S A. 2020;117:4874–84.
61. Olova N, Krueger F, Andrews S, Oxley D, Berrens R V., Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. Genome Biology; 2018;19:1–19.
62. Stroud H, Greenberg MVC, Feng S, Bernatavichute Y V., Jacobsen SE. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell [Internet]. Elsevier Inc.; 2013;152:352–64. Available from: http://dx.doi.org/10.1016/j.cell.2012.10.054
63. Duan C-G, Zhang H, Tang K, Zhu X, Qian W, Hou Y-J, et al. Specific but interdependent functions for Arabidopsis AGO4 and AGO6 in RNA-directed DNA methylation. EMBO J. 2014;34:581–92.
64. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;8:1494. Available from: http://dx.doi.org/10.1038/nprot.2013.084
65. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res [Internet]. Oxford, UK: Oxford University Press; 2003;31:439–41. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC165453/
66. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res [Internet]. 2017;45:W12–6. Available from: http://dx.doi.org/10.1093/nar/gkx428
67. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res [Internet]. 2013;41:e74–e74. Available from: http://dx.doi.org/10.1093/nar/gkt006
68. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res [Internet]. 2011/11/22. Oxford University Press; 2012;40:D1178–86. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22110026
69. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol [Internet]. 2013/01/16. Oxford University Press; 2013;30:772–80. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23329690
70. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics [Internet]. 2009/06/08. Oxford University Press; 2009;25:1972–3. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19505945
71. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics [Internet]. 2011/02/17. 2011;27:1164–5. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21335321
72. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol [Internet]. 2012/02/22. Oxford University Press; 2012;61:539–42. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22357727
73. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics [Internet]. 2011/04/14. Oxford University Press; 2011;27:1571–2. Available from: https://pubmed.ncbi.nlm.nih.gov/21493656

74. Yong-Villalobos L, González-Morales SI, Wrobel K, Gutiérrez-Alanis D, Cervantes-Peréz SA, Hayano-Kanashiro C, et al. Methylome analysis reveals an important role for epigenetic changes in the regulation of the *Arabidopsis* response to phosphate starvation. Proc Natl Acad Sci [Internet]. 2015;201522301. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1522301112
75. Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, et al. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci U S A. 2016;113:E4052–60.
76. Jiao W, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis thaliana accessions reveal hotspots of genomic rearrangements. 2019;1–25.
77. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. bioRxiv. Genome Biology; 2019;1–13.

Figures

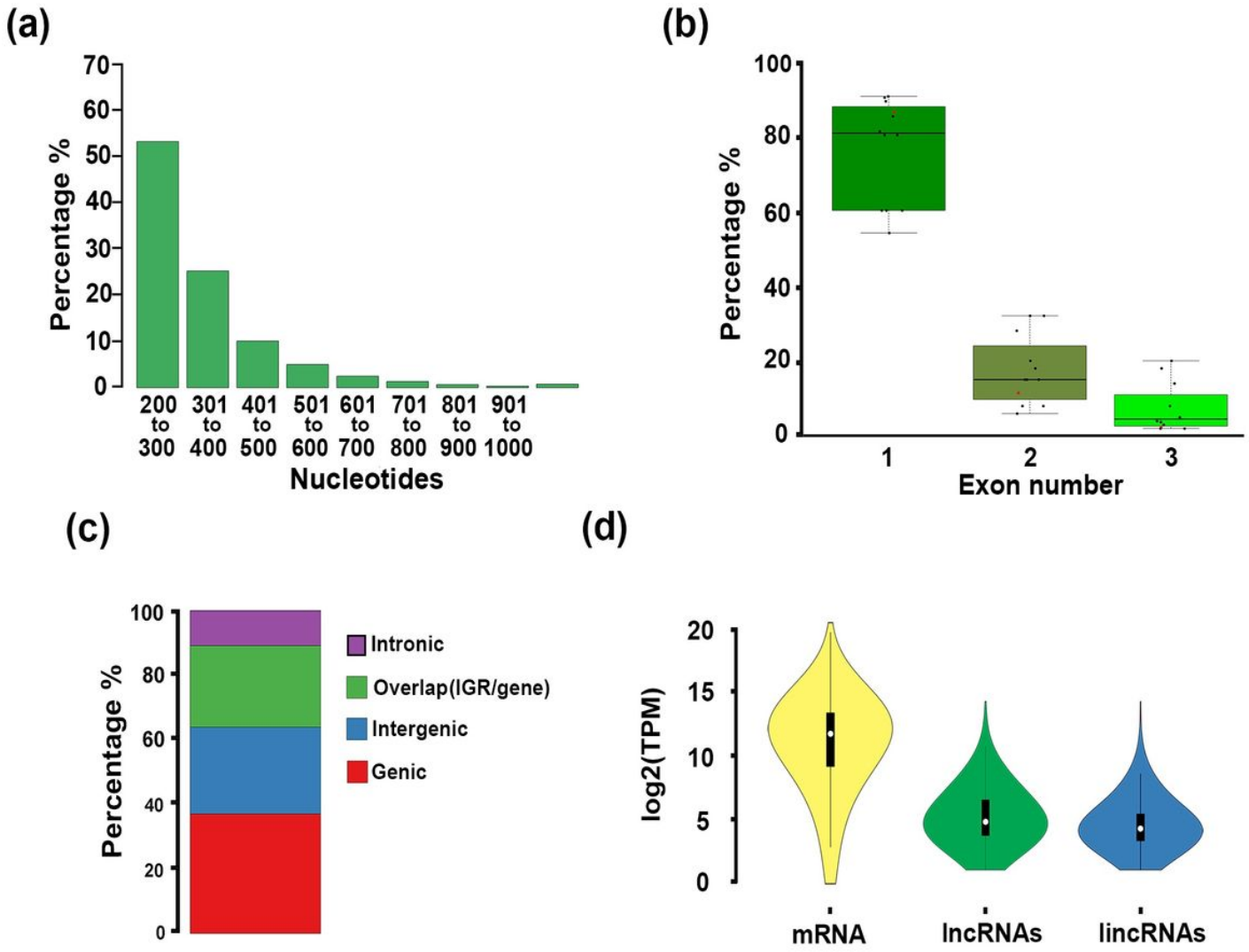

Figure 1

Architecture and annotation of lncRNAs in U. gibba and other plants. a lncRNAs size distribution in U. gibba is shown in a barplot grouping each 100-nt in size (X axis) and by percentage of the total (Y axis). b Structure of lncRNAs in selected plants (Additional file 1: Dataset S1). The box plot groups the proportion of exon number (X axis) by percentage (Y axis) among lncRNAs identified from various plant genomes. c Genomic annotation of U. gibba lncRNAs. The proportion is from 4,295 putative lncRNAs loci identified. Red represents the proportion of lncRNAs mapped to body gene regions; blue shows lncRNAs located in intergenic regions; green indicates those that overlap gene body and intergenic regions. d Coding and noncoding transcripts expression. The violin plot represents the total of mRNAs, lncRNAs and lincRNAs identified in this study with correspondent log2(tpm).

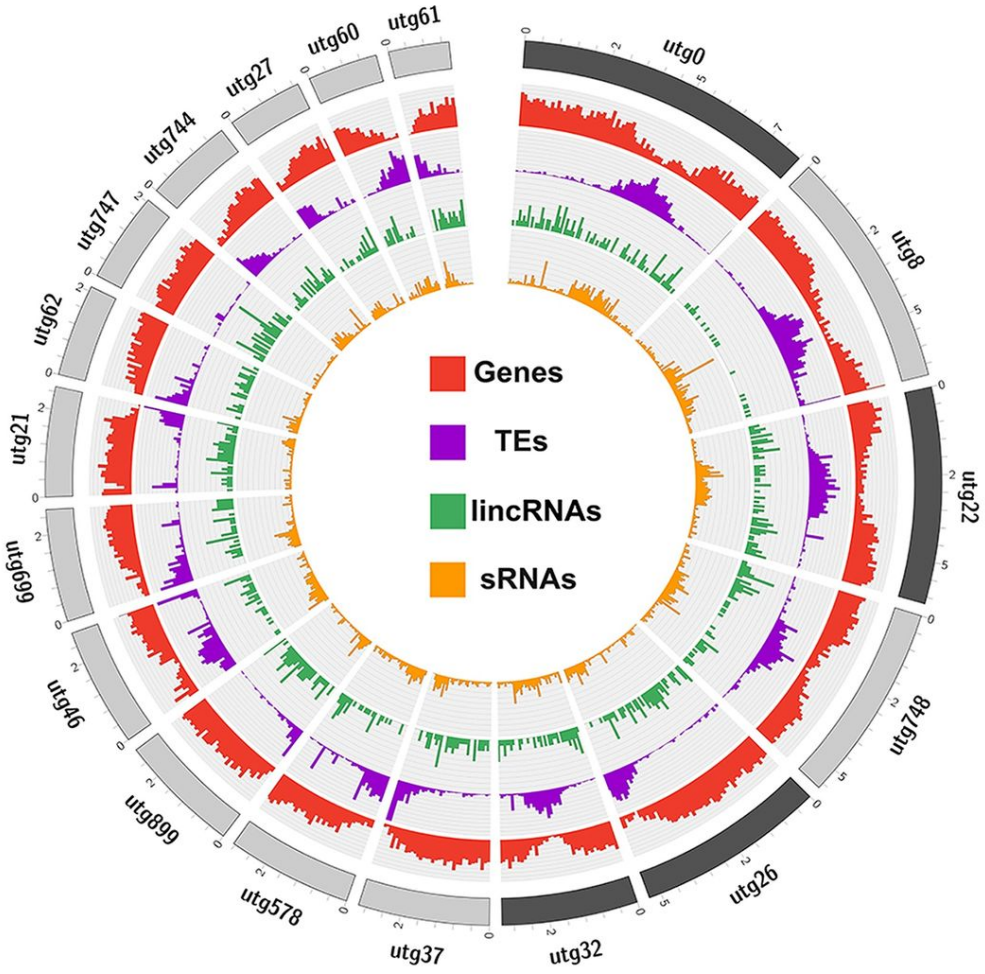

Figure 2

Noncoding RNA landscape in Utricularia gibba. Representation of the U. gibba genome in a circus plot where the outside blocks represent complete chromosomes (dark gray) and Unitigs or contigs (gray) larger than 1Mb in size. The density plots were calculated in 10Kb windows. Red histogram shows the gene density. The Purple histogram displays transposable element density. lincRNA density is represented in the green histogram, and small RNA density for sRNAs 20-nt to 24-nt in size are shown in orange.

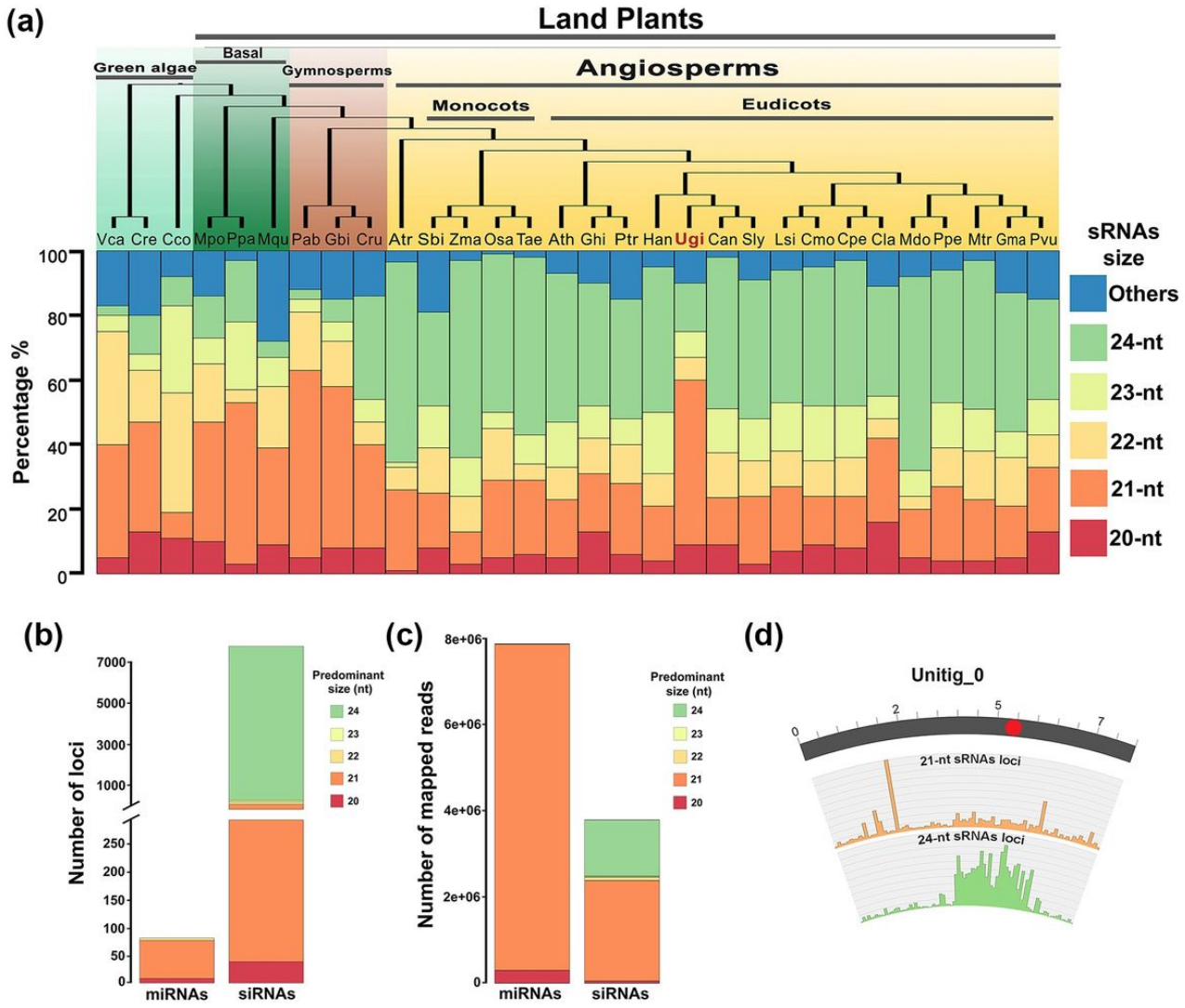

Figure 3

Annotation of sRNAs in U. gibba and distribution of small RNA size abundance in plants. a Top, a phylogenetic tree of plants with representative species for green algae, early branching land plants, gymnosperms, and angiosperms. Bottom, a stacked plot of small RNA abundances (20-nt to 24-nt in size) from various small RNA-seq studies (ataset S1). b Total locus predictions from ShortStack for miRNAs and siRNAs at different sizes. (c) Numbers of reads mapped per loci at different sizes from ShortStack. d The largest complete chromosome (Unitig_0) representation of gene density, TE density, 21-nt small RNAs loci density distribution and 24-nt small RNAs loci distribution. Vca: Volvox carteri; Cre: Chlamydomonas reinhardtii; Cco: Chara coraline; Mpo: Marchantia polymorpha; Ppa: Physcomitrella patens; Mqu: Marsilea quadrifolia; Pab: Picea abies; Ginkgo biloba; Cru: Cycas rumphii; Atr: Amborella trichopoda; Sbi: Sorghum bicolor; Zma: Zea mays; Osa: Oryza sativa; Tae: Triticum aestivum; Ath: Arabidopsis thaliana; Ghi: Gossypium hirsutum; Ptr: Populus trichocarpa; Han: Helianthus annuus; Ugi: Utricularia gibba; Can: Capsicum annuum; Sly: Solanum lycopersicum; Lsi: Lagnaria siceraria; Cmo: Cucurbita moschata; Cpe: Cucurbita pepo; Cla: Citrullus lanatus; Mdo: Malus domestica; Ppe: Prunus persica; Mtr: Medicago truncatula; Gma: Glycine max; Pvu: Phaseolus vulgaris.

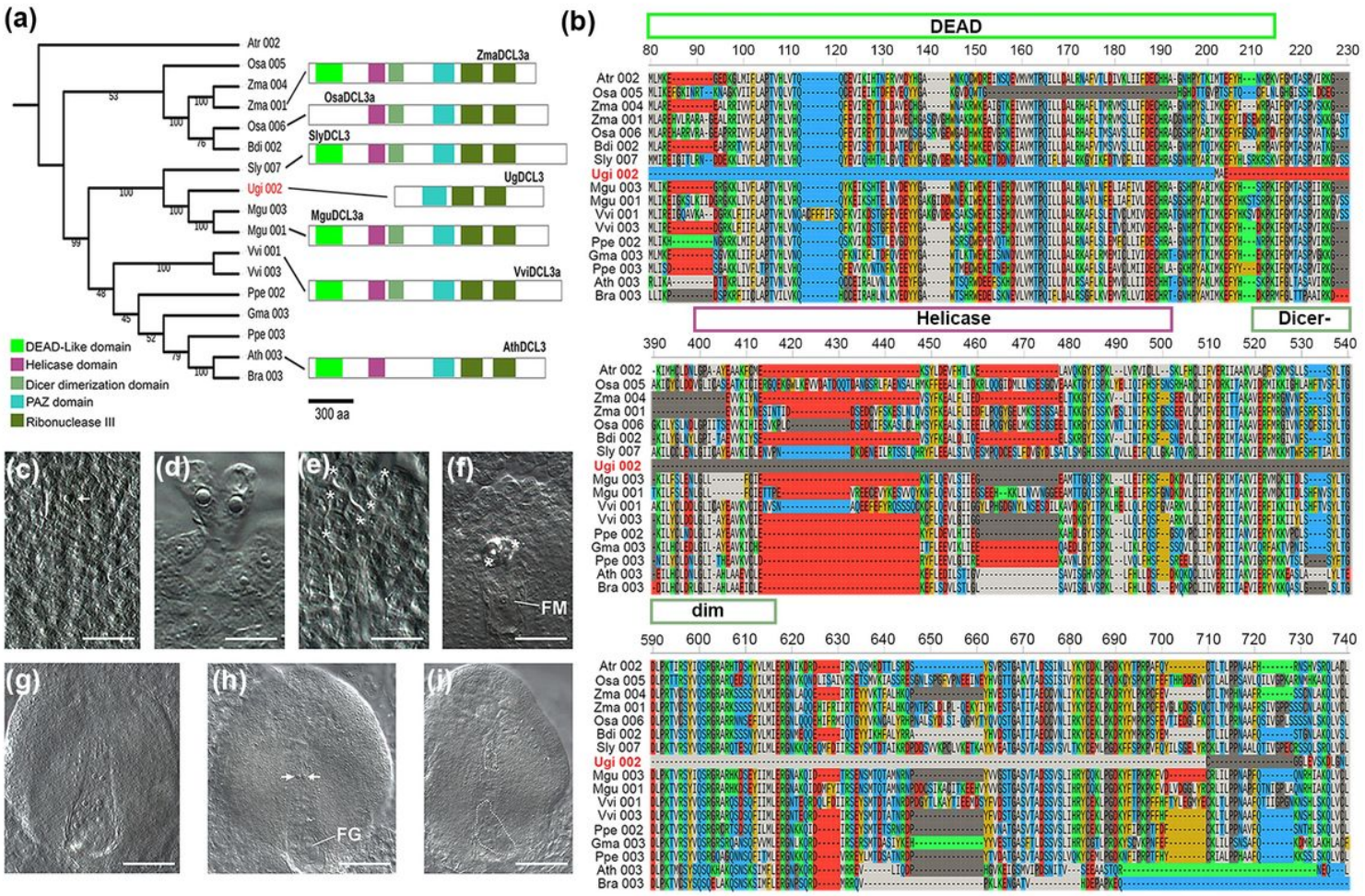

Figure 4

Truncated DCL3 and female gametophyte development in the ovule of U. gibba. a Phylogenetic analysis of DCL3 in some angiosperms including U. gibba. A typical DCL3 protein has the DEAD, Helicase, Dicer dimerization, PAZ, and two Ribonuclease III domains, while UgDCL3 only contains the PAZ and Ribonuclease III domains. b Multiple sequence alignment of DCL3 proteins in blocks of 150 aa showing the missing domains in UgDCL3. c Developing ovules showing a single pre-meiotic precursor corresponding to the MMC. d Twin pre-meiotic precursors prior to megasporogenesis. e Six differentiated cells (asterisk) corresponding to gametic precursors prior to megasporogenesis. f Functional megaspore (FM) and two degenerated megaspores (asterisk) following megasporogenesis. g Developing ovule showing a single 8-nuclear female gametophyte. h Developing ovule showing a female gametophyte (FG) and two ectopic gametic precursors (arrows) at the chalazal region. i Developing ovule showing two independent female gametophytes (dashed). Scale bars: 12.5 μm in c to f; and 20 μm in g and i.

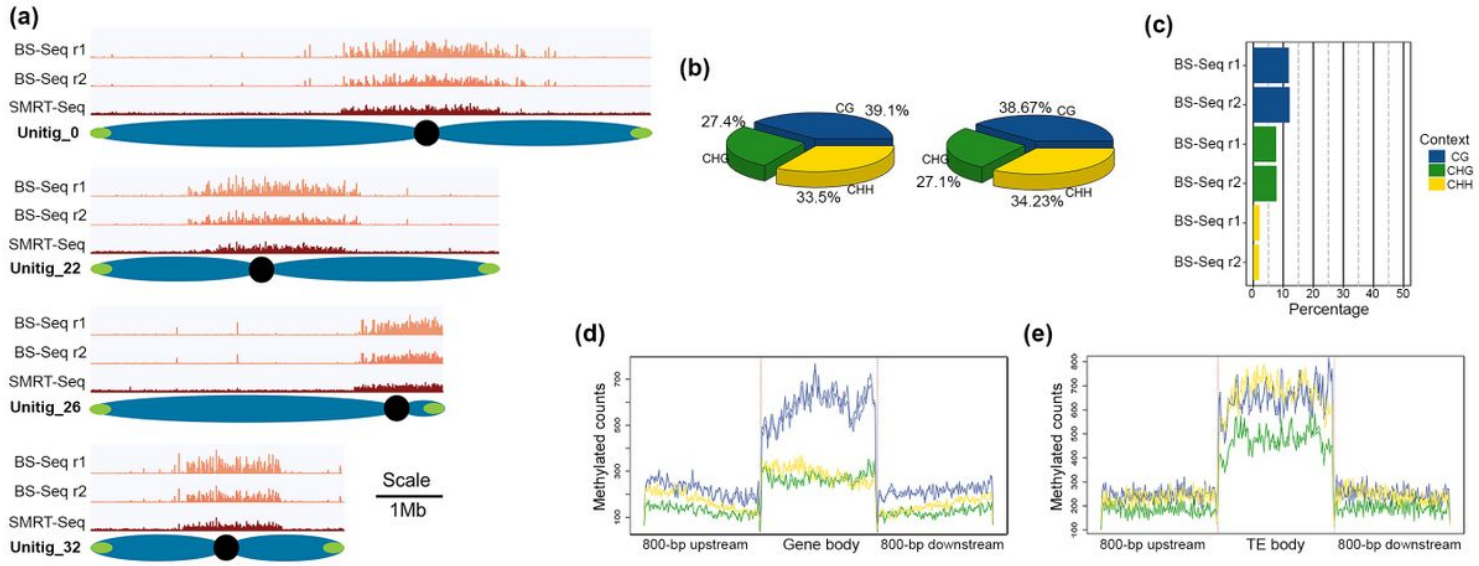

Figure 5

Genome-wide methylation in Utricularia gibba. a Distribution of methylation (m5C) representation only in the complete chromosomes in U. gibba for location of pericentromeric regions. Here, the blue light red shows the methylation density in BS-Seq replicates and the red histogram represent the methylation distribution in SMRT-Seq with normalized data (Scale 1Mb). b Methylation context at the genome level for two replicates. CG methylation in blue, CHG methylation in green, and CHH methylation in yellow. c Methylation levels for CG, CHG and CHH context. d Gene body methylation and TE body methylation representation in all contexts (CG, CHG and CHH) 800bp upstream and downstream. e TE body methylation and TE body methylation representation in all contexts (CG, CHG and CHH) 800bp upstream and downstream.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DATASETS1.xlsx
- DATASETS2.xlsx
- DATASETS3.xlsx
- SupplementaryInformationandfigurescorrected.docx