**Atypical DNA methylation, sRNA-size distribution, and female gametogenesis in the compact genome of *Utricularia gibba****.*

**Running title:** Methylation and noncoding RNA landscape in *Utricularia gibba*.

**Authors:**

Cervantes-Pérez Sergio Alana; Yong-Villalobos Leninb, Florez-Zapata Nathalia M. V.a,1

Oropeza-Aburto Aracelia; Rico-Reséndiz Félixa; Amasende-Morales Itzela; Lan Tianyingc;

Martínez Octavioa; Vielle-Calzada Jean-Philippea; Albert Victor A.c,d and Herrera-Estrella Luisa,b,2

aLaboratorio Nacional de Genómica para la Biodiversidad, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 36824, Irapuato, Guanajuato, México; bInstitute of Genomics for Crop Abiotic Stress Tolerance, Plant and Soil Department, Texas Tech University. cDepartment of Biological Sciences, University at Buffalo, 14260, Buffalo, New York, USA; dSchool of Biological Sciences, Nanyang Technological University, Singapore 637551;

1Current address: Universidad EIA, Km 2 + 200 vía al aeropuerto José María Córdova Alto de Las Palmas, Envigado, Antioquia (Colombia).

2Corresponding author: lherrerae@cinvestav .mx; [luis.herrera-estrella@ttu.edu](mailto:luis.herrera-estrella@ttu.edu)

**Supplementary Information and figures**

The following Additional file 2 is available for this article:

**Fig. S1** Workflow of lncRNAs identification.

**Fig. S2** Experimental validation of Pac-Bio methylome data.

**Fig. S3** Intergenic lncRNAs identification in some plants.

**Fig. S4** Small RNAs size distribution in the three libraries of sRNA-Seq.

**Fig. S5** Known and novel miRNAs loci in comparison with miRBase.

**Fig. S6** Proportion of sRNA reads found for each miRNA family in the three sequenced libraries.

**Fig. S7** Annotation of sRNAs loci.

**Fig. S8** Distribution of Polymerase (IV and V) holoenzymes in plants.

**Fig. S9** NRPD4 phylogenetic tree.

**Fig. S10** AGO5 clade homolog identification.

**Fig. S11** AGO2/3/7 clade in the *U. gibba* phylome.

**Fig. S12** AGO6/4/8/9 clade in the *U. gibba* phylome.

**Fig. S13** AGO1/10 clade in *U. gibba* phylome.

**Fig. S14** Phylogenetic tree of AGO family in *Viridiplantae*.

**Fig. S15** Phylogenetic analysis of DCL and RDR proteins.

**Fig. S16** Conserved protein domains among DCL3 for some angiosperms.

**Fig. S17** Multiple Sequence Alignment of some DCL3 proteins.

**Fig. S18** Ug*DCL3* in the transcriptomic datasets and RACE-PCR alignment.

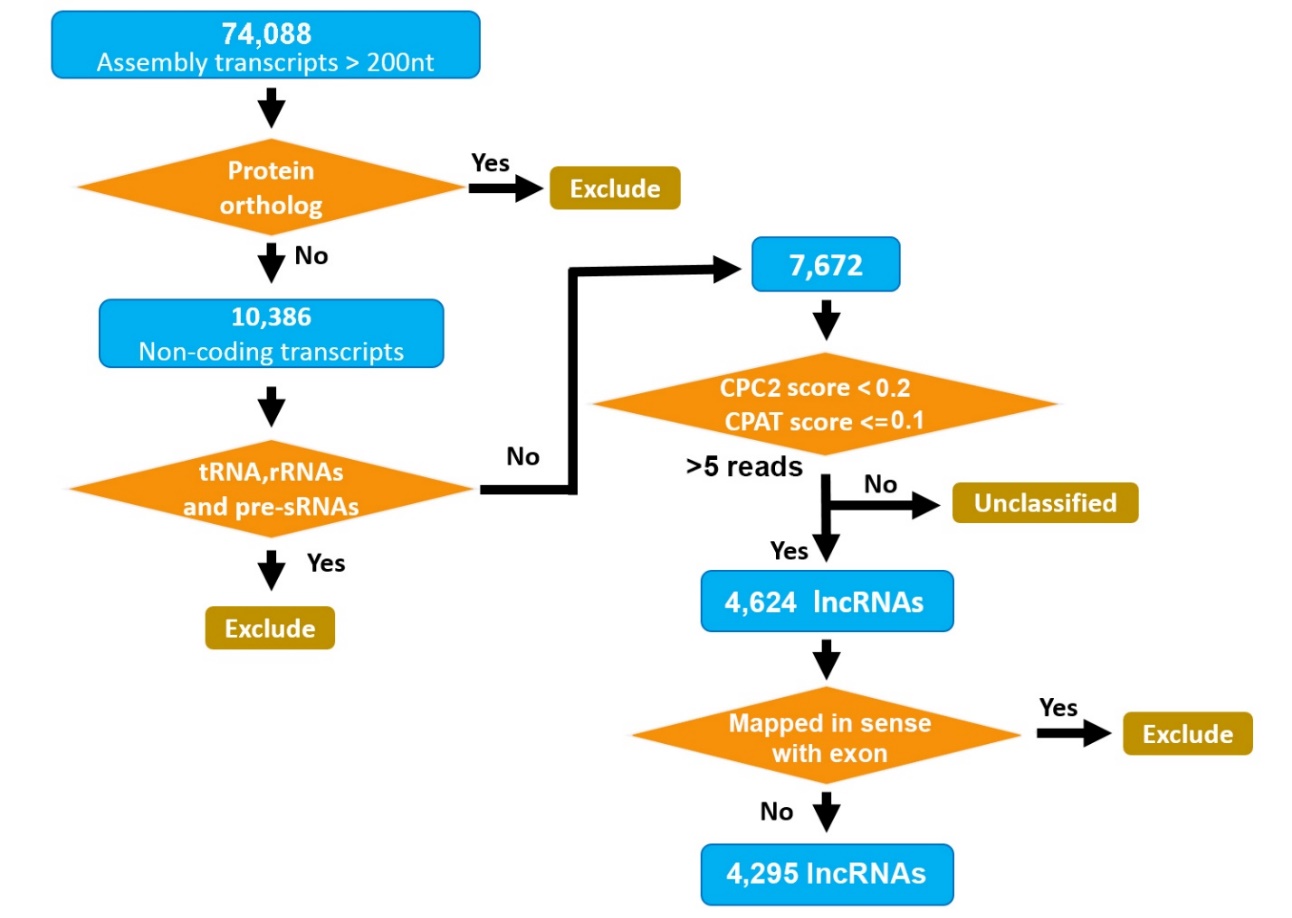
**Fig. S19** Comparison between m5C identification in *A. thaliana* with BS-Seq and SMRT-Seq.

**Fig. S20** Methylation identification with SMRT-Seq.

**Table S1** lncRNAs features. Summary of lncRNAs size and exon number.

**Table S2** Canonical genes involved in miRNAs biogenesis.

**Table S3** Ectopic gametic precursors and female gametophytes in the ovule of *Utricularia gibba*.



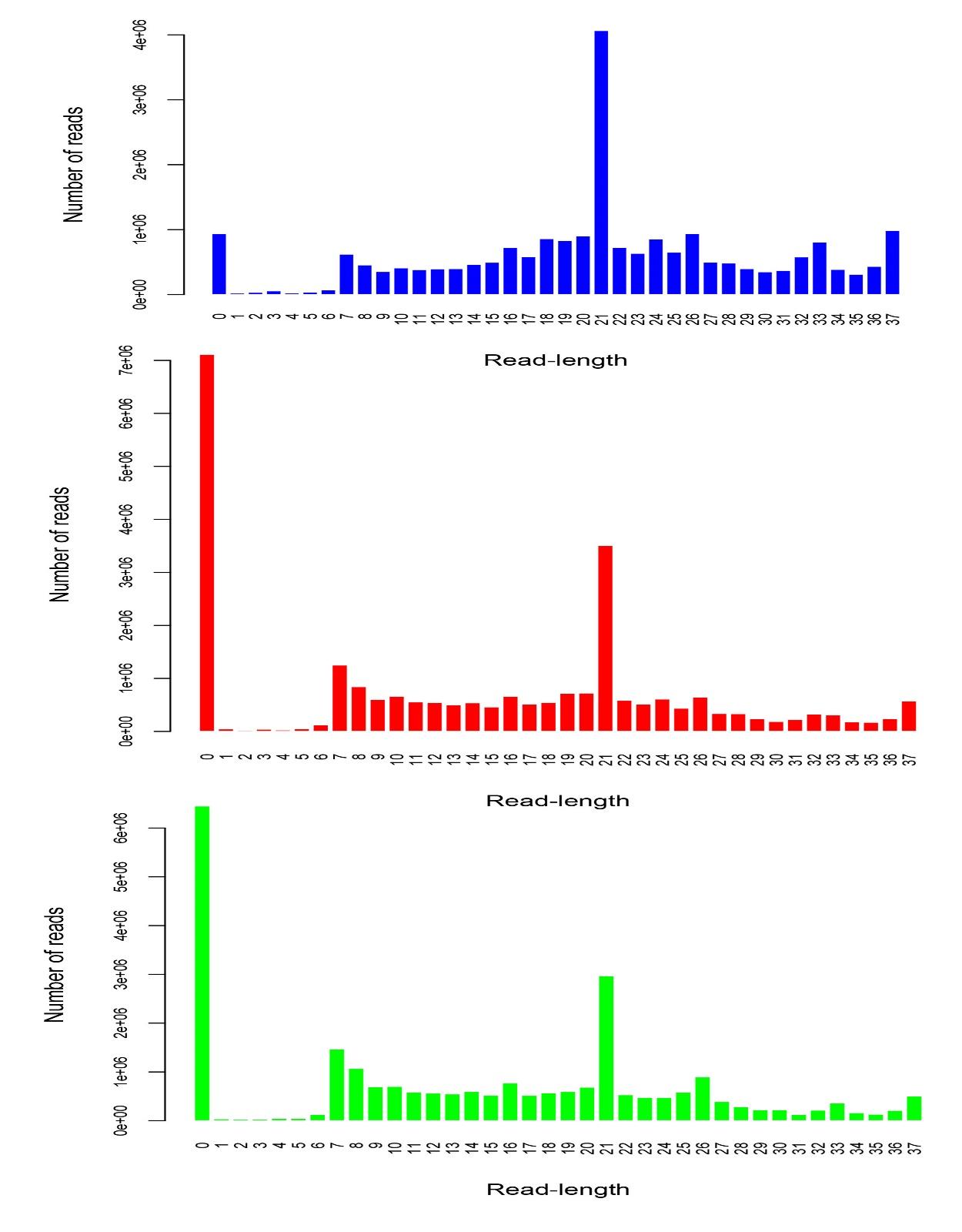
**Fig. S1. Workflow of lncRNAs identification**. Schematization for the identification of lncRNAs in this work. In the workflow, the input/output are representing by blue oval shapes, the arrows are the connector lines that show relationships between the representative shapes, the orange diamonds (which indicates a decision) are the key filters, and brown ovals that are connected to a decision represents all the exclude. A summary with the key steps are described following: the first input of assembled transcripts with size ≥ 200-nt; a decision for protein orthologs must be taken by sequence homology against plant protein database where exclude those transcripts potentially proteins, then the output/input are the transcripts that not found match with proteins, and a second decision implicates the elimination for the similar transcripts to housekeeping and structural transcripts and maintains these without similarity which is the input for the third decision. The final decision calculates the coding potential for a protein in a form of functional ORF eliminating the potentially coding proteins and retaining the final number of putative lncRNAs in *U. gibba* transcriptional landscape.



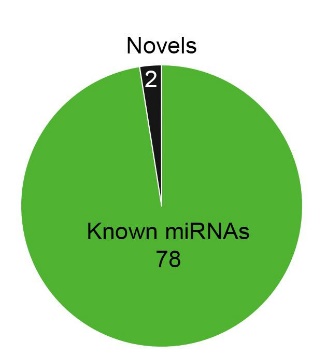
**Fig. S2. Experimental validation of Pac-Bio methylome data.** **(a)** Selection of genomic regions for Chop-qPCR Analysis. Two genomic regions (~ 9 Kb each) were selected from Utricularia gibba Unitig 0 for analysis, one gene rich region (GRich, purple line) and one gene poor region (GPoor, red line (pericentromeric)), within each of these particular regions two independent regions with similar DNA methylation density values presenting restriction sites for either Hpa II (green line) and Hae II (blue line) (GRich B, GPoor B) DNA methylation sensitive enzymes or absence of the same restriction sites (GRichA, GRichB) were selected. **(b)** Eschematic representation of the Chop-qPCR experiment workflow. Vegetative tissue explants fom *Utricularia gibba* plants grown in MS media 0.5 X (MS 0.5X) for X days (1), were transferred to new MS 0.5X recipients containing either 0 (non-Zeb treatment) or 50 M Zebularine (Zeb treatment) for 8 days (2). Tissue of both Zeb and non-Zeb treatment was collected for genomic DNA isolation and purification (3). Purified Utricularia gibba genomic DNA was either non-digested or digested with methylation sensitive restriction enzymes Hae III or Hpa II (restriction is decreased or blocked if there is 5mC methylation inside recognition sequence) (4). Chop-qPCR was performed using digested and nondigested DNA from Zeb and non-Zeb samples (5). **(c)** Chop-qPCR of *Utricularia gibba* Unitig 0 GRich and Gpoor regions. Average of the amplification values of Unitig 0 regions without Hae III and Hpa II sites were used as internal controls (GRich A, GPoor A); non-Zebularine treatment values for each time point were used as calibrator samples. Values are foldchange of either digested (Hae III, Hpa II) and non-digested (non-Hae III, non-Hpa II) Zebularine treatments compared with non-Zebularine treatments, eg. a fold-change value lower than 1 indicates a decrease in the amplification of a fragment in the Zebularine treatment. Data presented are means of two biological replicates ± SEM (of two technical replicates each).



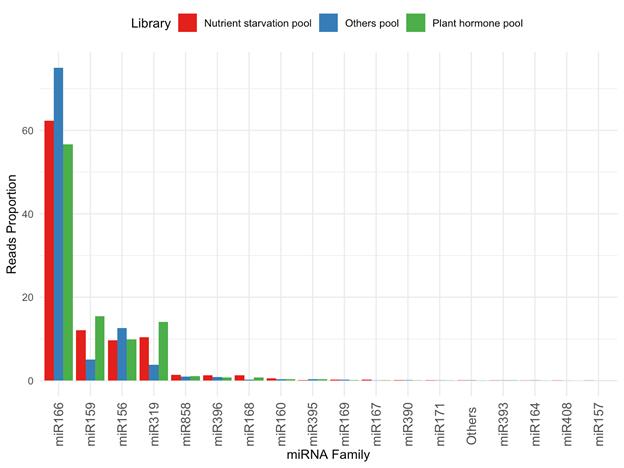
**Fig. S3. Intergenic lncRNAs identification in some plants**. A comparison of intergenic lncRNAs identification for plants only at genomic level searched in the reports for each one of them. For some cases only high confidence lncRNAs were used. The data were obtained from the follow reports: Maize (4), Cotton (5), tomato (6), Populus (7), Arabidopsis (8) and U. gibba (this work).

****

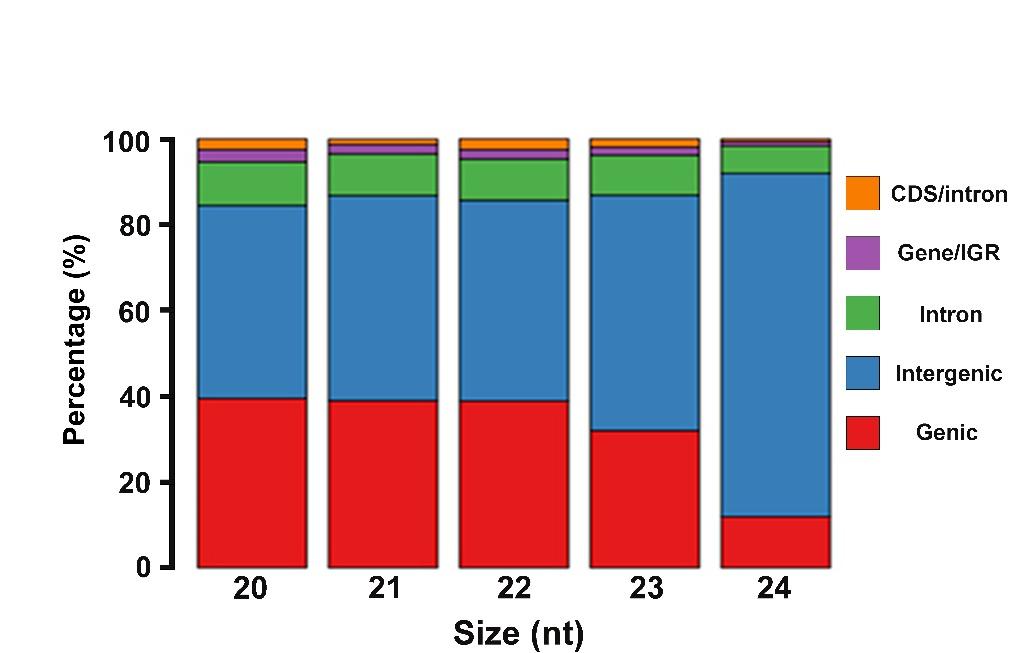
**Fig. S4. Small RNAs size distribution in the three libraries of sRNA-Seq.** The distribution of reads size in the small RNA-Seq libraries in *U. gibba* are show in different barplots. Blue barplot represents the nutrients library, the red barplot shows the plant hormones mix library and green barplot represents other conditions library. In all barplot, the X axis is for size in the range of 0 to 37-nt and Y axis show the number of reads per each size.



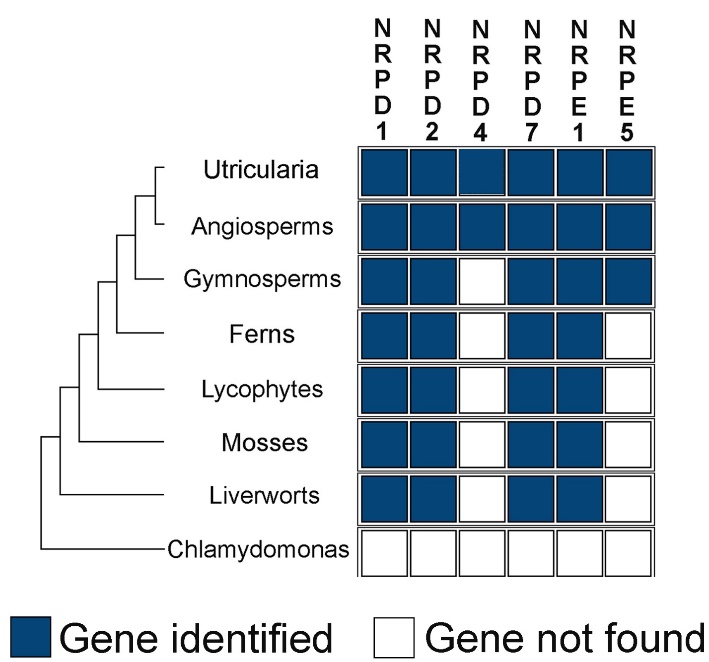
**Fig. S5. Known and novel miRNAs loci in comparison with miRBase.** From 80 miRNA loci predicted by ShortStack (17), 78 miRNAs loci were identified with sequence homologs in miRBase database for other organism and only for 2 miRNAs loci evidence not found in other plant.



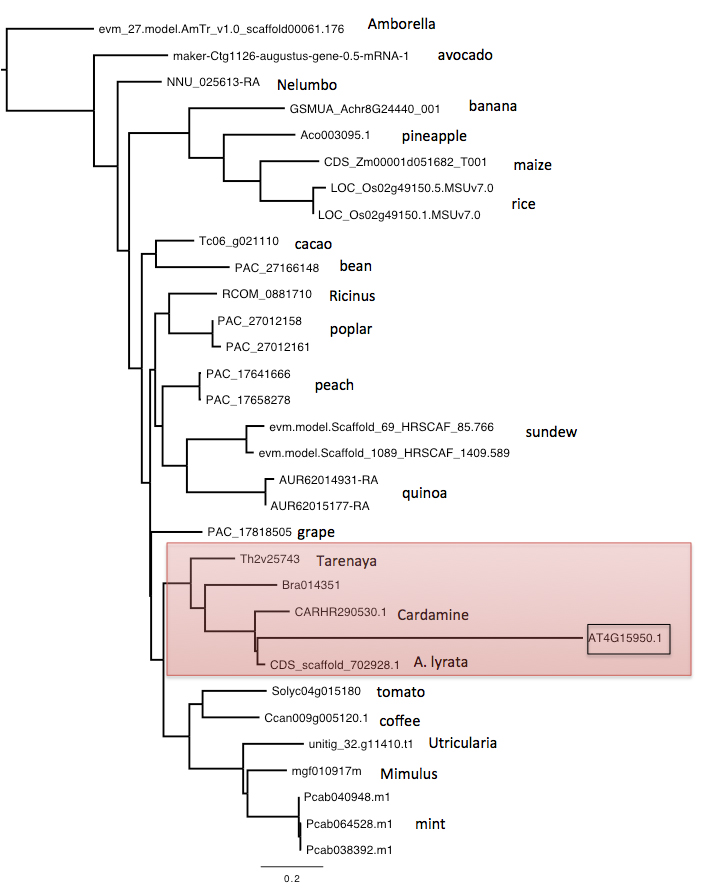
**Fig. S6. Proportion of sRNA reads found for each miRNA family in the three sequenced libraries.** The 17 miRNA families found in *U. gibba* are show in the X axis ordered by descendent proportion of reads (Y axis) the different colors represents the abundance of each miRNA family per library sequenced, nutrients library (red barplot), plant hormone treatments (green barplot) and other conditions library (blue barplot).



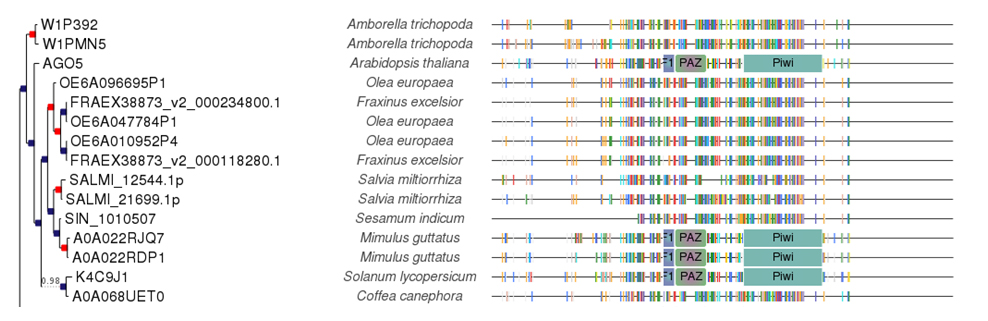
**Fig. S7. Annotation of sRNAs loci.** The distribution of a range between 20-nt to 24-nt in size for sRNA loci compared with their respective genome annotation. X axis represents the sRNA sizes in nucleotides, the different colors show the genome annotation, and the stacked plot shows the percentage corresponding for each group.



**Fig. S8. Distribution of Polymerase (IV and V) holoenzymes in plants.** An illustrative cladogramwith representative plants in each clade is shown and taken from a recent report (19). The modified figure including *U. gibba* (an angiosperm) show the presence (blue square)/absence (white square) of the different subunits of polymerases in plants.



**Fig. S9. NRPD4 phylogenetic tree.** Maximum likelihood tree for NRDP4 homolog genes in some angiosperms (Amborella, avocado, nelumbo, banana, pineapple, maize, rice, cacao, bean ricinus, poplar, peach, sundew, quinoa, grape, Tarenaya, Cardamine, *A. thaliana*, *A. lyrata*, tomato, coffee, Utricularia, Mimulus, and mint). The red rectangle shows NRPD4 for *A. thaliana* and other members of Brasicaceae where *A. thaliana* is divergent even *A. lyrata*. Numbers beside nodes are bootstrap support values showing confidence.

****

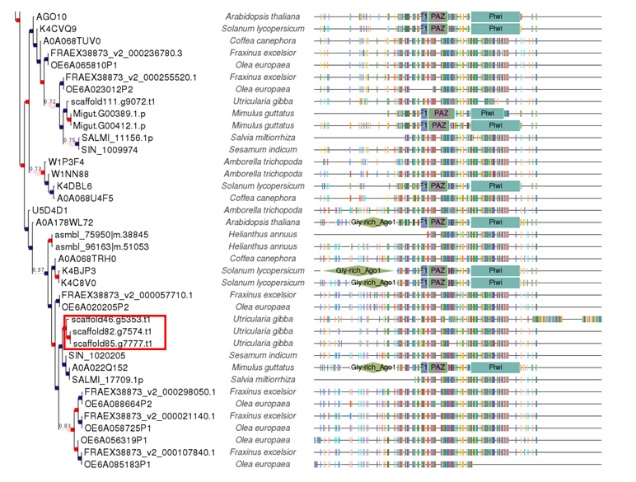
**Fig. S10. AGO5 clade homolog identification.** Phylogenetic tree representing the *U. gibba* phylome in the Phylome DB to find the homolog of AGO5 protein.



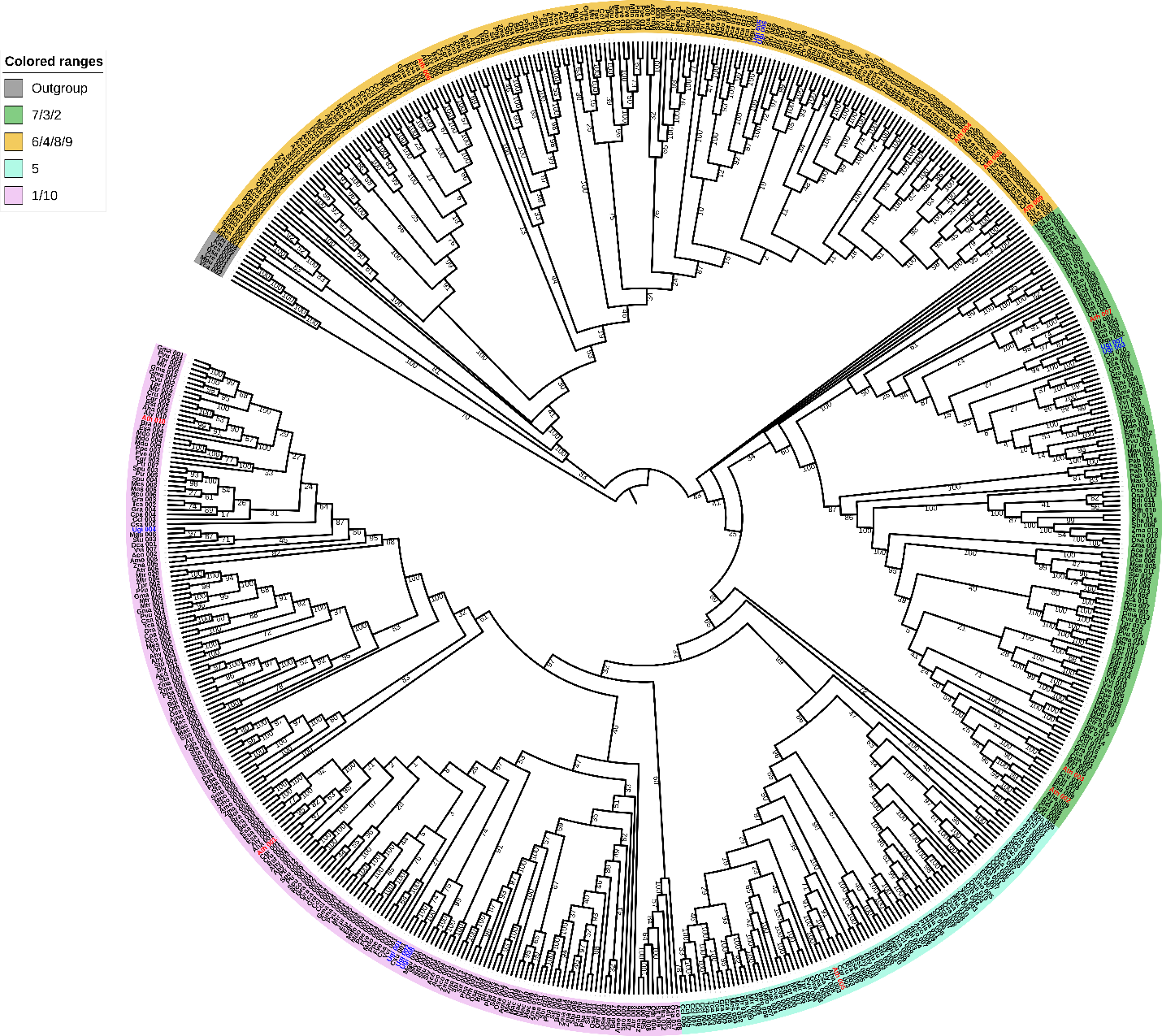
**Fig. S11. AGO2/3/7 clade in the *U. gibba* phylome. (a)** Phylogenetic tree representing the evolutionary history of AGO2, AGO3. **(b)** Phylogenetic tree representing the evolutionary history of AGO7.



**Fig. S12. AGO6/4/8/9 clade in the *U. gibba* phylome**. **A** Phylogenetic tree representing the evolutionary history in AGO4, AGO8, AGO9 in Phylome DB and **B** evolutionary history of AGO4.



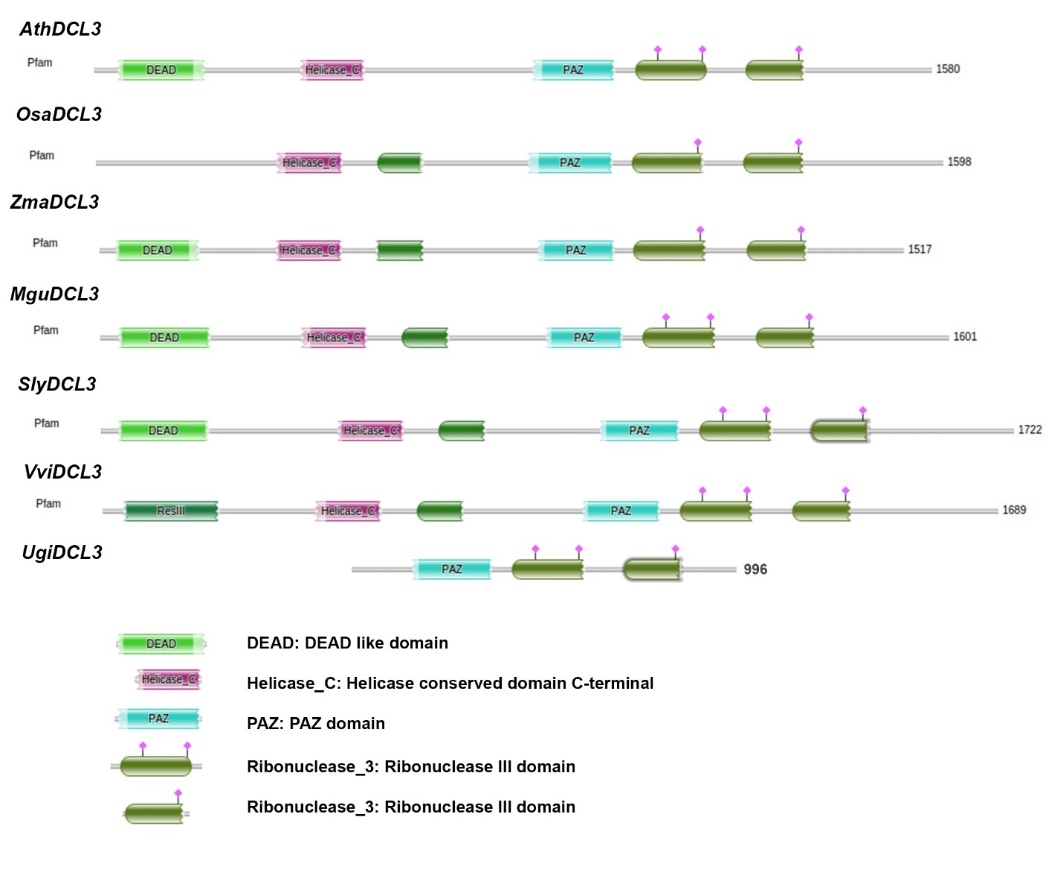
**Fig. S13. AGO1/10 clade in *U. gibba* phylome**. Phylogenetic tree representing the evolutionary history of AGO1 and AGO10 in Phylome DB.



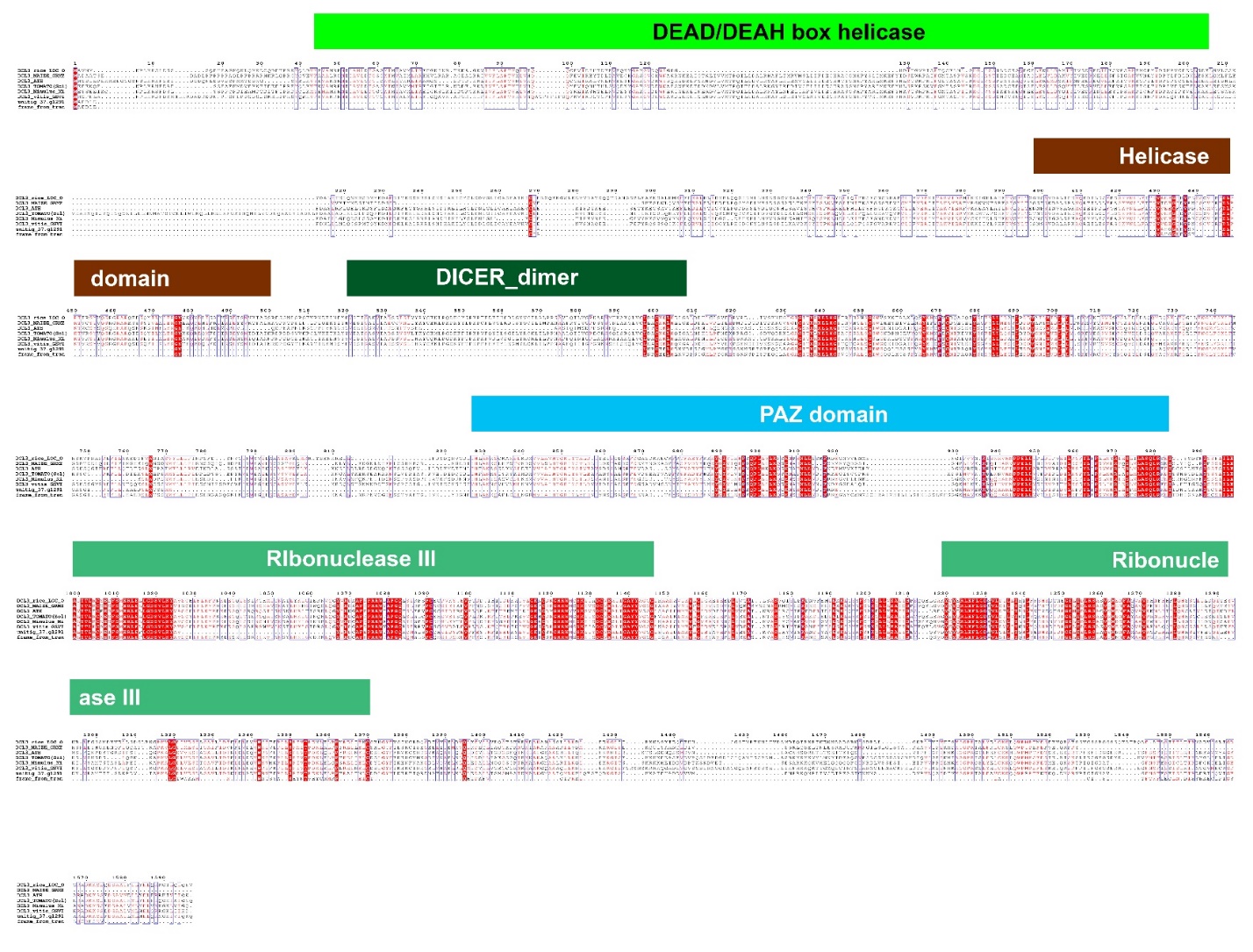
**Fig. S14. Phylogenetic tree of AGO family in Viridiplantae**. Phylogenetic reconstruction of AGO proteins for Viridiplantae to identify AGO in *U. gibba*. A maximum likelihood tree (ML + GTR + bootstrapp=1000) for *AGO1, AGO2, AGO3, AGO4, AGO5, AGO6, AGO7, AGO8, AGO9, AGO10* and *AGO-like* genes from representative genomes of land plants. The sequences from *Charophyte green algae (Coccomixa subellipsoidea, Volvox carteri, Dunaliella salina and Chlamydomonas reinhardtii)* are clustered as out group including only AGO-like genes. Moreover, all AGO sequences from land plants form four clades which are named after with th*e AGO* genesfrom *A. thaliana (*yellow clade *AGO 6/4/8/9,* green clade *AGO 7/3/2,* cyan clade AGO5 *and* purple *AGO1/10,* respectively*).* Blue flags represent the AGO identified in Arabidopsis and the red flags show the specific AGO for *U. gibba*.Numbers beside nodes are bootstrap support values showing confidence.



**Fig. S15. Phylogenetic analysis of DCL and RDR proteins. A** Maximum likelihood (ML + GTR + bootstrap = 1000) tree for RDR1, RDR2, RDR3, RDR4, RDR5 and RDR6 from representative genomes of land plants. The Phylogenetic reconstruction reveals an arrangement of three clades which are named after with the *RDR* genes from *A. thaliana (*light red clade *RDR6,* light blue *RDR 3/4/5* and light green *RDR 1/2,* respectively*).* Blue flag indicates the *RDR* sequences which belong to *Arabidopsis thaliana* whereas the flags in red point to *Utricularia gibba* sequences*.* Numbers beside nodes are bootstrap support values showing confidence*.* **B** A maximum likelihood (ML + GTR + boostrap = 1000) tree fo*r DCL1, DCL2, DCL3 and DCL4* from representative genomes of land plants*.* The Phylogenetic reconstruction reveals an arrangement of three clades which are named after with theDCL proteins from *A. thaliana* (Light red clade DCL4, light purple clade DCL2, green clade DCL3 and Blue clade DCL4, respectively).Blue flag indicates the *RDR* sequences which belong to *Arabidopsis thaliana* whereas the flags in red point to *Utricularia gibba* sequences*.* Numbers beside nodes are bootstrap support values showing confidence



**Fig. S16. Conserved protein domains among DCL3 for some angiosperms**. Identification of conserved protein domains in different DCL3 proteins in plants in HMMER database (<http://hmmer.org/>).



**Fig. S17. Multiple Sequence Alignment of some DCL3 proteins**. A multiple sequence alignment where red block represented the conserved protein domains with a representation in color blocks the typical protein domains for DCL3.



**Fig. S18. Ug*DCL3* in the transcriptomic datasets and RACE-PCR. (a)** Synteny analysis between *U. gibba* two gene models (two latest genome version) and A. thaliana DCL3, the colored blocks represent the syntenic regions. **(b)** The protein of ugDCL3 comprises 996 aminoacids (aa) and has three conserved protein domains (PAZ domain, and two contiguous Ribonuclease III domains). The gene model has a coding sequence of size 2991 bp and contains 13 exons which can be visualize in the protein conserved domains. The ug*DCL3* is located on the unitig\_37 one of the largest contig. In the bottom we show the search of largest transcripts in our transcriptomic data of contrasting conditions. The conditions with largest transcripts were Control (Normal conditions), BAP (treatment with the hormone Benzil-aminopurine), ABA (treatment with Abscisic acid plant hormone) and PEG (Polyethylene glycol, an osmotic treatment). The transcript size varies among these conditions, but the coding sequence is the same. **(c)** 5´Race-PCR sequence alignment against reference genome highlighting in red blocks the 5´RACE-PCR and in green blocks de CDS sequence of *DCL3* gene (**d**) left: IGV browser location of *DCL3* gene and compared with reads coverage of some RNA-Seq libraries. Right: Table of transcripts per million (TPM) in WT libraries in *DCL3* gene.



**Fig. S19. Comparison between m5C identification in *A. thaliana* with BS-Seq and SMRT-Seq. (a)** Distribution of m5C in the Arabidopsis genome in syntenic regions. Normalized methylation distribution estimated by BS-Seq and SMRT-Seq.In the histogram were represented 604 syntenic windows at genome-wide level divided by chromosome. The red line represents the density distribution of BS-Seq, and blue line represents the SMRT-Seq density. The estimated rˆ is given in purple for each chromosome. **(b)** Global methylation context for total m5C identification in 4 libraries of BS-Seq and SMRT-Seq. The figure shows the methylation context for individual BS-Seq libraries and SMRT-Seq of Arabidopsis. For BS-Seq were used 4 libraries named B1 (shoot 7days), B2 (Root 7 days), B3 (Shoot 16 days), B4 (Root 16 days) grown in optimal conditions and SMRT. In the X axis can see the methylation context (CG, CHG, CHH) grouped for different libraries. The Y axis represents the percentage for each methylation context. **(c)** Global m5C distribution at chromosome scale.

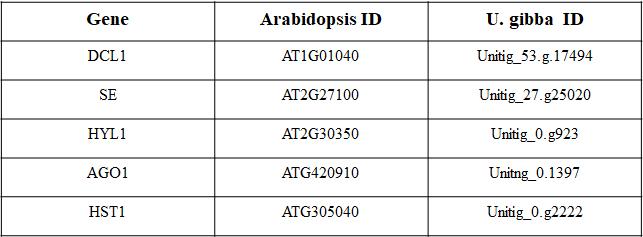


**Fig. S20. Methylation identification with SMRT-Seq. (a)** Pie plot representing global methylation context for m5Cs in the *U. gibba* genome. **(b)** Methylation levels for each methylation context. **(c)** Gene body methylation and 800 bp upstream and downstream the gene body.

**Table S1. lncRNAs features. Summary of lncRNAs size and exon number.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mean**  **Size (nt)** | **Median Size (nt)** | **Largest**  **Size (nt)** | **Median exon number** | **Exon number** | **Number**  **lncRNAs** | **Mean**  **Size (nt)** |
| 339.96 | 294 | 2134 | 1 | 1 | 3716 | 329.8652 |
|  |  |  |  | 2 | 482 | 380.4295 |
|  |  |  |  | 3 | 72 | 502.6944 |
|  |  |  |  | 4 or more | 25 | 592.4 |

**Table S2. Canonical genes involved in miRNAs biogenesis.**



**Table S3. Ectopic gametic precursors and female gametophytes in the ovule of *Utricularia gibba*. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

                             Ovary size (mm)

0.3 – 0.5a 0.5 – 1a

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

One gametic precursor 51.4% (73)

Several gametic precursors 42.9% (61)

One female gametophyte 40.7% (31)

One female gametophyte

+ ectopic gametic precursors 22.4% (17)

Two female gametophytes 23.7% (18)

Degenerated female gametophyte 13.1% (10)

NDb 5.7% (8)

Total 100% (142) 100% (76)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

a Brackets indicate the number of ovules observed with a given phenotype.

b Not determined