8 Repetitive Sequences: Impacts and Uses in the Spirodela Genome

Paul Fourounjian

Abstract

Repetitive DNA, consisting of small and large satellite repeats and transposable elements, comprises over 50% of most plant genomes. The Lemnaceae family demonstrates a ~12-fold difference in genome size and relatively similar number of genes, indicating a wide variability in repeat content. The best studied genome of the family *Spirodela polyrhiza* had a normal total satellite DNA content, yet a surprisingly high 50% of those were dinucleotide microsatellite repeats. The telomeres and 119 bp centromere repeats were typical, although ribosomal repeats appear scarce. Genomic studies showed a small number of 24nt heterochromatic siRNAs accompanied by the lowest rate of DNA methylation seen in any plant sequenced at 9% and low rates of heterochromatin formation. Despite this low level of regulation, the transposable elements are unexpectedly rare and old. In fact, they even show high rates of DNA methylation and high rates of inactivation through illegitimate recombination. This suggests that the scarce 24nt siRNAs are surprisingly effective and an intriguing topic of further research.

P. Fourounjian (✉)
Waksman Institute of Microbiology, Rutgers
University, Piscataway 08854, USA
e-mail: pjf99@scarletmail.rutgers.edu

In the early years of DNA and chromosome research, structural components of chromosomes were noticed as patterns in DNA and protein stains, often in the centromeric or telomeric regions. Once DNA sequencing began it was uncovered that virtually all eukaryotic genomes contain significant portions of repetitive DNA, previously thought of as "junk DNA" (Biscotti et al. 2015). In plants, repetitive elements comprise the majority of most genomes sequenced, ranging from a mere 14% in the grain teff to 85% in maize (Wendel et al. 2016). These repetitive elements can be categorized into tandem repeats which aid in chromosome structure, and longer interspersed repeats derived from transposable elements (TEs). As of 2018 there are two published sequences for *Spirodela polyrhiza* clones 7498 and 9509, and the *Lemna minor* 5500, along with draft genomes of two *Lemna* species *minor* and *gibba* and the *Wolffia* species *australiana* (Unpublished), (Wang et al. 2014; Van Hoeck et al. 2015; Ernst and Martienssen 2016; Michael et al. 2017). Similar to other angiosperms as a whole, these genomes vary considerably in size, but not significantly in gene number (Table 8.1). The Lemnaceae family displays a 12-fold difference in genome size from the smallest sequenced monocot *Spirodela polyrhiza* to the 1881 megabase *Wolffia arrhiza* (Wang et al. 2011). A recent review on plant genome architecture summarized that these size variations between genomes are due to common whole genome duplication, followed by

X. H. Cao et al. (eds.), *The Duckweed Genomes*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-030-11045-1_8

Table 8.1 Lemnaceae genome size and gene content

Species, clone	Genome size (Megabases)	Gene copy #
S. polyrhiza, 7498	158	19,623
S. polyrhiza, 9509	158	18,507
L. minor, 5500	481	22,382
L. minor, 8627	800	NA
L. gibba, 7742	450	21,830
W. australiana	~380	NA

reduction of coding genes, and proliferation of transposable elements (Wendel et al. 2016). Taken in summary, these repeats play a large role in genomic size and composition and chromosomal structure, in the duckweeds and eukaryotes as a whole.

When DNA was separated by density gradient centrifugation tandem repeats with differential AT/GC content created satellite bands above and below the majority of DNA eventually leading to the name satellite DNA. These tandem repeats range in size from the 180 bp corresponding to a nucleosome to tiny 2 nucleotide microsatellite repeats. They were found to have structural implications in centromeres and telomeres where they maintain heterochromatic structure, and disruptions of their expression have been shown to lead to genomic instability and cancer (Biscotti et al. 2015). The strain 7498 genome study showed that the small *Spirodela polyrhiza* genome had a normal number of satellite DNA repeats, at 1.3% of the genome. Yet while most plants have 10–100 bp minisatellites making up roughly half of the total satellite DNA, strain 7498 satellite DNA was 50% microsatellite repeats, largely comprised of GA repeats, which may have been mutated from methylated CG heterochromatin sequences (Wang et al. 2014; Michael et al. 2017). For the *Lemna minor* 5500 genome, we know that satellite and microsatellite repeats made up 0.6 and 3% of the genome, indicating a similar enrichment of microsatellite repeats (Van Hoeck et al. 2015). In a follow-up study assembling the 32 pseudo-molecules into 20 chromosomes relied on the telomeric repeats of TTTAGGG and the suspected centromeric repeats to help support the confidence of the pseudomolecule assembly (Cao et al. 2016). Another analysis of the 7498 and 9509 strains of *Spirodela* was run using longer reads for better resolution of repeat regions and found a high homology with few indels and less than 0.06% heterozygosity in SNPs. They found that a previously reported 138 bp centromeric repeat was found at 1 centromere and that 19 of 20 chromosomes contained large numbers of a 119 bp centromeric repeat (Melters et al. 2013; Michael et al. 2017). Additionally, they found an extremely low ribosomal DNA copy number of 81 compared to 570 in the similarly sized *Arabidopsis thaliana* genome. In summary, while the centromeres and telomeres of *Spirodela polyrhiza* are consistent with other plant genomes, the microsatellite repeats are very abundant and the ribosomal repeats are very rare.

Probably, the most interesting repeat elements are the transposable elements (TEs), which include DNA copying transposons, RNA copying retrotransposons with autonomous versions capable of replicating themselves and non-autonomous versions of each. Thanks to this replication potential, these selfish genes are always attempting to proliferate, while the plant host genome is perpetually suppressing them and removing them through illegitimate recombination. This push and pull occurring in countless plant species shows that of our crop plants TEs can comprise as little as 14% of the genome in teff and as much as 85% in maize (Wendel et al. 2016). In the annotation of the 7498 genome, LTR retrotransposons were annotated based on homology and found to be 15.5% of the genome, which agreed with its size, while the transposons were too distant from their homologs in their genomes an unable to be annotated (Wang et al. 2014). This lack of homology is due to the age of the transposons, which mutate over time. In *Spirodela,* the relatively few LTRs (264) had an average age of 4.3 million years, while the average in Brachypodium and rice was found to be 1.8 and 0.7 million years, respectively. In the later analysis of the 9509 genome, TEs were annotated by homology to other known TEs, and by mapping 22–24nt siRNAs known to regulate them through methylation. This showed that the genome is 25% TEs, with a Gypsy/Copia ratio of 1.5. In accordance with the age of the LTRs, the *Spirodela* genome was found to be purging them through illegitimate recombination resulting in the highest ratio of deactivated solo to intact LTRs seen in any plant genome.

After the *Spirodela* 7498 genome was published, the draft genome of *Lemna minor* 5500 was published due to its importance in ecotoxicological studies (Van Hoeck et al. 2015). While *Lemna minor* strains vary in genome size from 323 to 760 Mb strain 5500 is 481 Mb in size and only has 14% more annotated genes than *Spirodela polyrhiza* 7498 (Table 8.1). Compared to *Spirodela* 94.5% of the difference in genome size is due to repeats. These repeats make up 61% of the genome and 36% of the genome is TEs, mainly retrotransposons, which is slightly higher than *Spirodela*. The count of LTRs increased ~10-fold to 210,531. There was a final category of unclassified repeats that made up 21% of the genome. In strain, 7498 DNA-based transposons were difficult to annotate based on their old age and low homology, and in strain, 9509 the annotation relied on siRNAs. Therefore, the unclassified repeats may include many ancient unannotated transposons.

The relative lack of TEs in *Spirodela* brought attention to the RNA directed DNA methylation (RdDM) pathway. This is a mechanism of silencing transposons through siRNAs where Pol IV creates a ssRNA transcript and RDR2 makes it a dsRNA (Matzke et al. 2015). Then, DCL3 cleaves it into 24nt het-siRNAs (heterochromatic) that are loaded onto AGO4, which binds to DRM2 and RDM1 proteins that methylate the 5’ end of cytosine in GC, CHG, and CHH sequences. To finish the process a collection of proteins in a histone-modifying complex converts the methylated TE sequence to silenced heterochromatin. This pathway is highly conserved across all land plants, with the notable outlier of the Norway Spruce, which has relatively few 24nt het-siRNAs, mainly localized to reproductive organs (Matzke et al. 2015).

In *Spirodela polyrhiza,* it was noticed that 24nt sRNAs were rare, comprising 7.3% of the small RNAs in strain LT5a and 1% in strain 7498 (Fourounjian et al. 2019). While the 9509 genome had the lowest DNA methylation rate of any plant sequenced at 9%, the TEs had an average methylation rate of 20% (Michael et al. 2017). Furthermore, older TEs were annotated based on the mapping of 22–24nt siRNAs, suggesting that they were expressed and active. The *Spirodela* genome also revealed a low number of old TEs suggesting that it has been very successful at halting their proliferation (Wang et al. 2014; Michael et al. 2017). Taken together it looks like the RdDM pathway is working with little to no 24nt het-siRNAs. This could be similar to the results seen in Norway spruce where 24nt het-siRNAs are localized to flowers, which are very rare in *Spirodela*, or perhaps other mechanisms may be at play. The mystery of how the Lemnaceae, particularly *Spirodela*, regulate their TEs is an exciting field of research that is still currently unfolding.

Acknowledgements Thank you to Dr. Alex Harkess for reviewing this chapter to confirm its accurate, but not complete description of the RdDM pathway.

References

Biscotti MA, Olmo E, Heslop-Harrison JS (2015) Repetitive DNA in eukaryotic genomes. Chromosom Res 23:415–420. https://doi.org/10.1007/s10577-015-9499-z

Cao HX, Vu GTH, Wang W et al (2016) The map-based genome sequence of Spirodela polyrhiza aligned with its chromosomes, a reference for karyotype evolution. New Phytol 209:354–363. https://doi.org/10.1111/nph.13592

Ernst E, Martienssen R (2016) Status of the Lemna gibba 7742a and Lemna minor 8627 genomes. Duckweed Forum, 12

Fourounjian P, Tang J, Tanyolac B, Feng Y, Gelfand B, Kakrana A, Tu M, Wakim C, Meyers BC, Ma J, Messing J (2019) Post-transcriptional adaptation of the aquatic plant under stress and hormonal stimuli. Plant J

Matzke MA, Kanno T, Matzke AJM (2015) RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. Annu Rev Plant Biol 66:243–267. https://doi.org/10.1146/annurev-arplant-043014-114633

Melters DP, Bradnam KR, Young HA et al (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10. https://doi.org/10.1186/gb-2013-14-1-r10

Michael TP, Bryant D, Gutierrez R et al (2017) Comprehensive definition of genome features in Spirodela polyrhiza by high-depth physical mapping and short-read DNA sequencing strategies. Plant J 89:617–635. https://doi.org/10.1111/tpj.13400

Van Hoeck A, Horemans N, Monsieurs P et al (2015) The first draft genome of the aquatic model plant Lemna minor opens the route for future stress physiology research and biotechnological applications. Biotechnol Biofuels 8:188. https://doi.org/10.1186/s13068-015-0381-1

Wang W, Haberer G, Gundlach H et al (2014) The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun 5:1–13. https://doi.org/10.1038/ncomms4311

Wang W, Kerstetter RA, Michael TP (2011) Evolution of genome size in Duckweeds (Lemnaceae). J Bot 2011:1–9. https://doi.org/10.1155/2011/570319

Wendel JF, Jackson SA, Meyers BC, Wing RA (2016) Evolution of plant genome architecture. Genome Biol 17:1–14. https://doi.org/10.1186/s13059-016-0908-1