Mapping the Algal Secret Genome: The small RNA Locus Map for *Chlamydomonas reinhardtii*

Sebastian Y. Müller[1,*], Nicholas E. Matthews[1,2,*], Adrian A. Valli[1,3], David C. Baulcombe[1]

1 *Department of Plant Sciences, University of Cambridge, Cambridge, CB2 3EA, UK*
2 *Present address: The University of Manchester, Oxford Road, Manchester, M13 9PL, UK*
3 *Present address: Spanish National Centre for Biotechnology (CNB-CSIC), 28049, Madrid, Spain.*
***** *These authors contributed equally to this work.*

+ Correspondence and requests for materials should be addressed to D.C.B. (email: dcb40@cam.ac.uk)

Abstract

Small (s)RNAs play crucial roles in the regulation of gene expression and genome stability across eukaryotes where they direct epigenetic modifications, post-transcriptional gene silencing, and defense against both endogenous and exogenous viruses. The green alga *Chlamydomonas reinhardtii* is a well-studied unicellular alga species with sRNA-based mechanisms that are distinct from those of land plants. It is, therefore, a good model to study sRNA evolution but a systematic classification of sRNA mechanisms is lacking in this and any other algae. Here, using data-driven machine learning approaches including Multiple Correspondence Analysis (MCA) and clustering, we have generated a comprehensively annotated and classified sRNA locus map for C. reinhardtii. This map shows some common characteristics with higher plants and animals, but it also reveals distinct features. These results are consistent with the idea that there was diversification in sRNA mechanisms after the evolutionary divergence of algae from higher plant lineages.

Introduction

Small (s)RNAs in many organisms are involved in regulation of gene expression at the transcriptional (epigenetic marks) and post-transcriptional (RNA degradation/translational repression) levels, as well as in host defenses against viruses [25]. In eukaryotes, their double stranded or highly structured RNA precursors are processed by Dicer-like (DCL) endonucleases into short 20-25nt RNA duplexes that are bound by Argonaute (AGO) proteins. One of the sRNA strands guides an AGO-containing effector complex toward complementary DNA/RNA to mediate the various mechanisms of RNA silencing [23]. Based on their origin and biogenesis, sRNAs can be classified in diverse classes including small-interfering (si)RNAs, micro(mi)RNA, and piwi-interacting (pi)RNAs [10], but there are many different sRNA subtypes [2] and diffuse boundaries between sRNA classes, as demonstrated for *Arabidopsis thaliana* [13]. Consequently, development and improvement of comprehensive classification methods is required for understanding of sRNA-based regulation networks.

Due to its small genome, vegetative/sexual reproduction, fast growth, motility and capacity to use acetate as carbon source, the photosynthetic green alga *Chlamydomonas reinhardtii* (hereafter referred to as Chlamydomonas) has been an important model organism for decades [31] and it was the first unicellular organism in which miRNAs were described [26, 44]. These Chlamydomonas miRNAs, however, are distinct from those of land plants: they have certain animal-like features including their biogenesis and mode of action [6, 7, 36, 42]. The Chlamydomonas key proteins in RNA silencing pathways (three AGO and three DCL proteins) are also distinct from homologues in land plants. Phylogenetic, structural and functional analyses indicate a divergence of both protein families since the common ancestor of algae and land plant about 1 billion years ago [4, 7, 32, 36]. There is also doubt about whether DCLs in Chlamydomonas contain PAZ domains [32,36], which are known to be important in cleavage of precursors for sRNAs of specific lengths [17]. Other differences with land plants include (i) the absence in Chlamydomonas of RNA-dependent RNA polymerases (RDRs) that generate dsRNAs from single stranded RNA [5] and (ii) the almost complete absence of non-CG methylation in transposons that is a hallmark of sRNA-directed DNA methylation [9,19].

These previously described characteristics of Chlamydomonas sRNA pathways suggest potential divergence from land plants. However, to-date, there has been no comprehensive characterization of the sRNA species found in in this alga. To address this issue we examined Chlamydomonas sRNAs, including miRNAs and siRNAs, based on the distinct loci from which they are produced. We used a Bayesian approach to generate the first comprehensive sRNA locus map for Chlamydomonas. Annotation of the loci based on intrinsic and extrinsic features allowed us to carry out a Multiple Correspondence Analysis (MCA) followed by clustering. We identified 6 classes of sRNAs, which may correspond to distinct RNA silencing pathways. Through comparison of the results with those previously reported for Arabidopsis [13], distinct features to those found in higher plants were uncovered, such as a particular sRNA loci distribution across the genome and their association with the epigenetic landscape. Together, these results help to understand the function of sRNAs in this single-celled alga Chlamydomonas and the evolution of sRNA-related pathways in green algae and land plants.

Results

RNAs characteristics in Chlamydomonas

In order to construct a complete locus map we first obtained a comprehensive collection of 145 sRNA libraries encompassing 54 replicate groups (Supplementary Table S1). To capture a maximum diversity of sRNA the libraries represent a wide range of conditions, strains and stages of the life cycle. After initial trimming and filtering (see Methods), the libraries contained a total of 22.3 million non-redundant (336 million redundant) sRNA reads mapping to the Chlamydomonas reference assembly genome [24].

To obtain a general overview of the sRNA population within our datasets, we first computed the same intrinsic key determinants found to be informative in the sRNA locus classification of Arabidopsis [13]: locus size, 5' nucleotide, repetitiveness of genomic mapping locations and abundance of individual sRNA species. To avoid a bias towards non-representative conditions, for this first part of the analysis we selected only libraries made from wild type strains.

The size distribution of sRNAs (Figure 1A) is dominated by 21nt species comprising 28% of all sRNAs, in line with previous studies [6, 26, 36, 38, 44] and unlike Arabidopsis and other land plants in which the 24nt sRNA fraction predominates (Figure 1A). Also, the 21nt species are the only class exhibiting counts with more than 1 million copies per sRNA species which makes up about 20% within that class(Figure 1B). Many reads

(43% redundant and 70% non-redundant) map uniquely to the reference genome, which is comparable to Arabidopsis (52% and 80%), albeit more evenly distributed across size range as shown in Figure 1C. This difference is most likely due to the absence of a RdDM pathway in Chlamydomonas. We did not find an abundant 23nt class found in Chlamydomonas from northern blot experiments [4].

Studies have shown the importance of 5' nucleotide for AGO binding specificity in Chlamydomonas [38] and Arabidopsis [34]. For Chlamydomonas, and in line with previous reports, we found that most sRNAs have adenine (A) and uracil (U) as 5' nucleotide. As in Arabidopsis, however, the 5'-end nucleotide varies greatly for different size classes (Figure 1D). For the predominant 21nt fraction, there was a preference for A and U (26% and 53%) with a higher proportion of 5' A sRNAs in the larger fractions up to 27nt. Overall, we found general agreement of our data with other datasets and we concluded that they are suitable for subsequent locus map generation and classification.

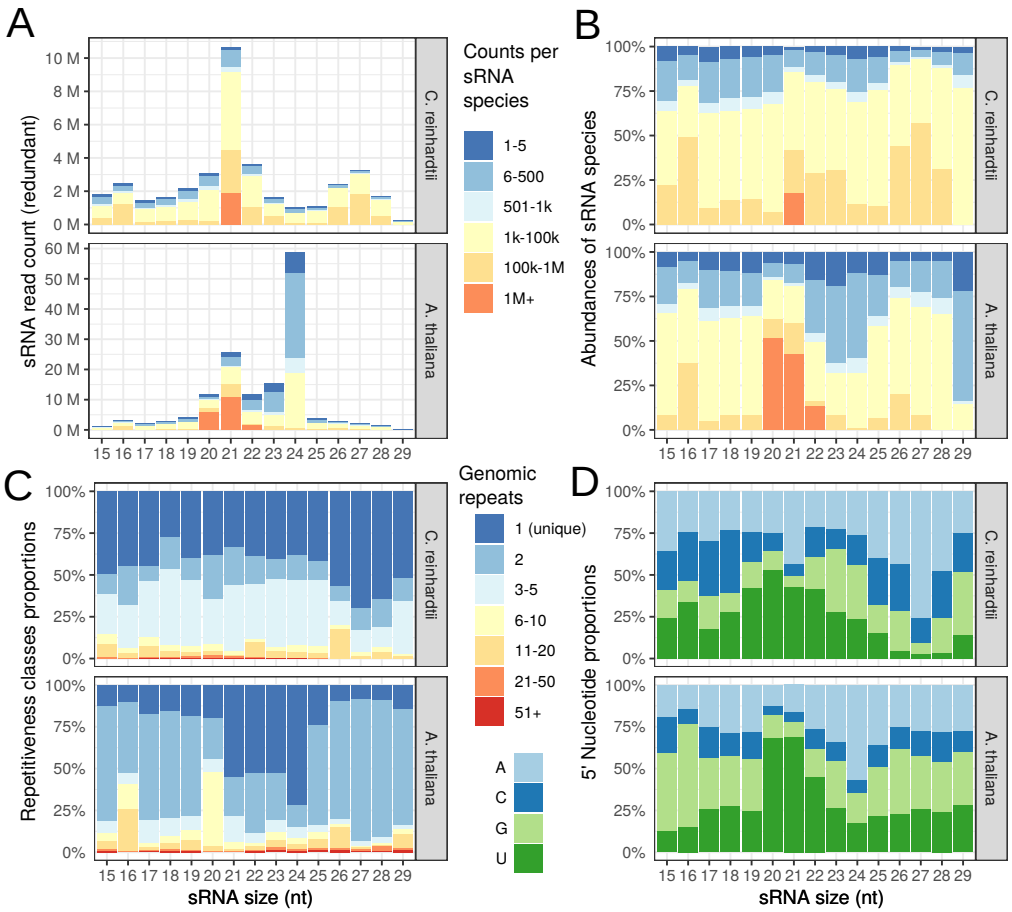

Figure 1. Characterisation of endogenous sRNA derived from wild-type libraries. (A) Size distribution of Chlamydomonas and Arabidopsis sRNAs. Redundant read counts were used to quantify the number of sequences obtained for each size class. Color reflects distribution of individual sRNA species abundance. (B) Similar to (A), but fraction of counts instead of total count. (C) Composition of reads mapping to various locations on the genome. Dark blue corresponds to fraction of sRNAs mapping to only one location in the genome, red to 51 or more locations. (D) Sequence composition of the 5' nucleotide of the sRNAs. (sRNAs with multireads greater than 20 were removed)

Defining a comprehensive small RNA locus map

To generate a locus map with these datasets we used the R package SegmentSeq. It employs a heuristic approach based on sRNA densities to derive an initial locus map which is then refined using Bayesian methods to take into account replicate groups [12]. The locus map based on a false discovery rate of less than 0.05 (FDR) had 6164 loci (Figure 2A) and covering 4.1% (4.57Mb) of the reference genome (110Mb). While the size of the libraries varies markedly, the number of loci per library scales roughly with library size (Figure 2B) and a cumulative analysis indicates that very few extra loci are likely to have been identified with further sequencing (Figure 2C). Figure 2D shows a degree of conservation of loci across replicate groups, although it should be noted that the libraries consist of a variety of strains, mutants and growth conditions, so complete conservation is not expected. For more stringent FDRs we saw a greater conservation of loci across replicate groups (Figure 3).

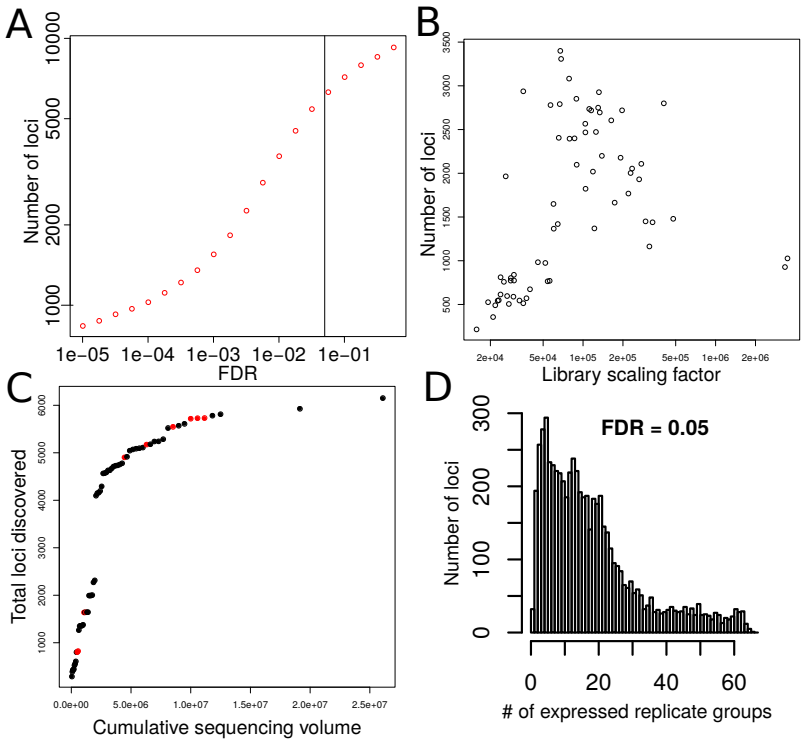

Figure 2. Diagnostic plots for validation of sRNA locus map. (A) Plot of the number of loci for different FDR levels. Vertical black line corresponds to cutoff used (FDR=0.05) (B) Number of loci discovered per replicate group plotted against library size. (C) Scatter-plot of number of loci discovered as cumulative sequencing depth increases. Red dots represent WT libraries. (D) Number of loci expressed in a given number of replicate groups.

Small RNA loci annotation

To gain insight into potential function of individual loci, we annotated them based on intrinsic loci features, such as locus size and repetitiveness, and based on features associated with the sRNAs, such as sRNA size, 5'-nucleotide, strand bias and phasing pattern (Supplementary Table S3 and Methods). In addition, extrinsic annotation features of sRNA loci included sRNA expression in specific conditions, genotype or

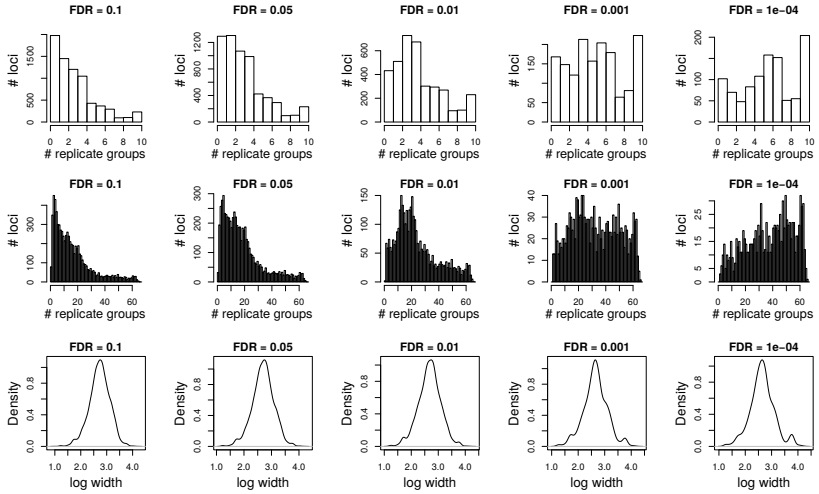

Figure 3. Locus conservation and width for loci identified at different FDR thresholds from 0.1 to 1e-4. The top and middle rows show the frequency (y-axis) of loci found in a given number of replicate groups (x-axis) for the 10 wild type control replicate groups (top row) and all replicate groups (middle row). The bottom row shows locus size density distributions with the log of locus width on the x axis.

overlap with genomic features (e.g. genes, transposons, methylation level). In a strong validation of the locus map, all 42 miRNAs previously identified in Chlamydomonas appear as defined sRNA loci [36]. An example is depicted in Figure 4 showing the loci CRSL0041450 which represents miR1157 and miR1157* as part of the 22th intron of the gene Cre12ġ537671.

Most annotation features are categorical in nature (i.e. overlap with genomic features is either true or false), but others are quantitative (i.e. size and phasing score). In preparation for MCA (see below) the quantitative features were classified into discrete groups according to the modality of the density distributions (Figure 5). Figure 5A shows a clear bimodal distribution of high or low locus repetitiveness and so we annotated loci in three groups corresponding to the two modes and the intervening section. Strand bias (Figure 5B) shows multiple modal peaks which can be neatly divided into strong bias ($0.2 < x < 0.8$) and medium bias (0.2-0.4 and 0.6-0.8). Figure 5C shows phasing to have just a single peak with a cut-off of 60 capturing the modal peak (< 60) and the long-tail (> 60). Finally, locus size cut-offs were chosen based on marked changes in gradient (Figure 5D).

The overall annotation results are shown in Table 1 and Supplementary Table S2. The majority of loci are between 400 and 1500nt long (50.4%), followed by 100-400 long loci (34%). Each of the remaining three classes make up less than 10%. Most loci have a predominant population of 21nt sRNAs (53%), followed by loci with <20nt sRNAs (22%), 20nt (18%) and <21nt (17%). Notably, 3924 out of 6164 loci (64%) overlap with genes making this feature as the most represented. These loci overlap (not necessarily exclusively) exons (55%), introns (39%), 3' untranslated region (UTR, 26%) promoters (19%) and 5' UTR (14%). There were also loci overlapping with transposons including both L1 (20%) and other LINE elements (21%), which is consistent with other studies [38]. Most other transposons do not overlap with sRNA loci (Supplementary Table S2). Furthermore, only few loci showed evidence of AGO3 dependency (6%), DCL3 dependency (11%) and sRNA phasing (2%). Many loci (41%) show a strong stand bias, where sRNA are found to be predominantly from either the + or - strand, but not both. 35% of loci had a medium bias. In addition loci were generally found to be very

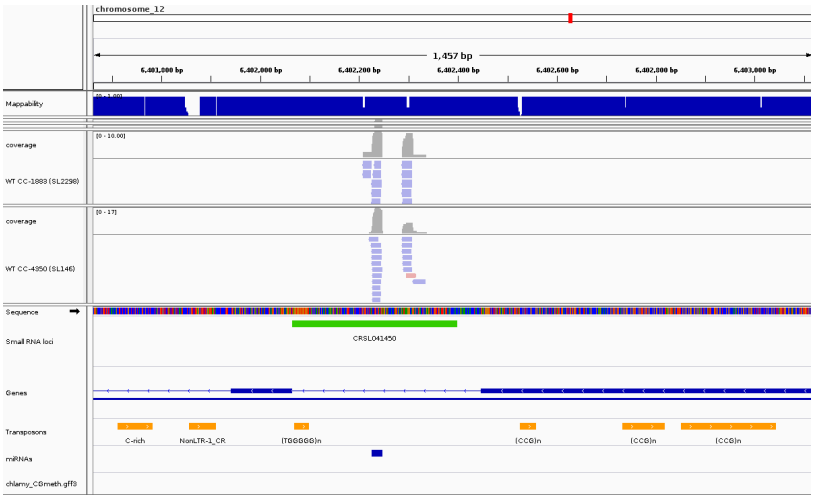

Figure 4. Genome browser view of a LC3 paragon. Upper panel shows location of the depicted section within the *C. reinhardtii* genome. Mappability track is shown below ranging between 0 (low mappability) and 1 (high). Individual mapped small RNAs are shown as red (mapped on + strand) and blue (mapped on - strand) bars for 2 wild type libraries (CC-1883 and CC-4350) along with coverage on top. Genome sequence of the section is shown with a color for each nucleotide (C=blue, G=orange, A=green, T=red). Green bars correspond to individual loci with CRSL prefix followed by a running number. Bottom panel show Genes (blue bars) and Transposons (orange bars).

repetitive in their constituent sRNAs with 56% having a high repetitiveness score.

Multiple Correspondence Analysis

Having assigned intrinsic and extrinsic features to each locus (a detailed breakdown of intrinsic and extrinsic features are listed in Supplementary Table S3), we used the MCA function from the FactoMineR R package [16] to search for underlying patterns and feature associations. Following dimensional reduction with MCA, k-means clustering was used to group loci according to the annotation patterns (see Methods for more details). We hypothesized that such grouping might identify distinct sRNA types and therefore potentially reveal distinct biogenesis or effector pathways.

To optimize the number of clusters and dimensions to be used we followed an approach used in a similar analysis of Arabidopsis [13]. We first evaluated the stability of the clustering using different combinations of clusters and dimensions through random sub-sampling (Figure 6A). We also calculated the additional variance explained by inclusion of an additional dimension (Figure 6B). Taken together, this analysis indicated that seven dimensions is the optimal number. In addition to the stability tests, computation of the gap statistic [35] and the normalised mutual information (NMI) between the clusters and annotation features both suggested six clusters to be optimal (Figure 6C-E). The presence of robust clustering for two and three clusters was noted and used to generate a hierarchy plot to demonstrate how loci from clusters grouped together for lower values of k (Figure 7).

The resulting six clusters, referred to as locus class (LC) 1-6, have relatively similar sizes and their association with different annotation features is shown in Figure 8 (also see Supplementary Table S4). The cluster hierarchy (see Figure 7) suggests that the primary division in clusters is between LC1-2 and LC3-6. LC1-2 have a high levels of

Table 1. Summary of selected annotation features. The table shows the number and percentage of loci with a particular annotation. Annotations include overlap with particular genome annotations, dependency on specific sRNA pathway machinery (from mutant libraries) or intrinsic locus features (e.g. strand bias).

Overlap	Yes	No
Genes	3924	2240
	64%	36%
Exons	3373	2791
	55%	45%
Introns	2385	3779
	39%	61%
3'UTR	1610	4554
	26%	74%
5'UTR	882	5282
	14%	86%
Promoter	1175	4989
	19%	81%
Intergenic	2240	3924
	36%	64%
miRNAs	42	6122
	1%	99%
IRs	1846	4318
	30%	70%
TE L1	1213	4951
	20%	80%
TE Order LINE	1325	4839
	21%	79%

Dependency		
AGO3 dependency	368	5796
	6%	94%
DCL3 dependency	700	5464
	11%	89%

Phasing Class	none	median	high		
	6020	123	21		
	98%	2%	0%		
Expression Class	specific	inbetween	common		
	1515	2984	1665		
	25%	48%	27%		
Strand bias class	strong	medium	none		
	2536	2182	829		
	41%	35%	13%		
Repetitiveness Class	low	med	high		
	518	1777	3477		
	8%	29%	56%		
Loci size	(0,100]	(100,400]	(400,1.5e+03]	(1.5e+03,3e+03]	(3e+03,Inf]
	272	2113	3109	552	118
	4%	34%	50%	9%	2%

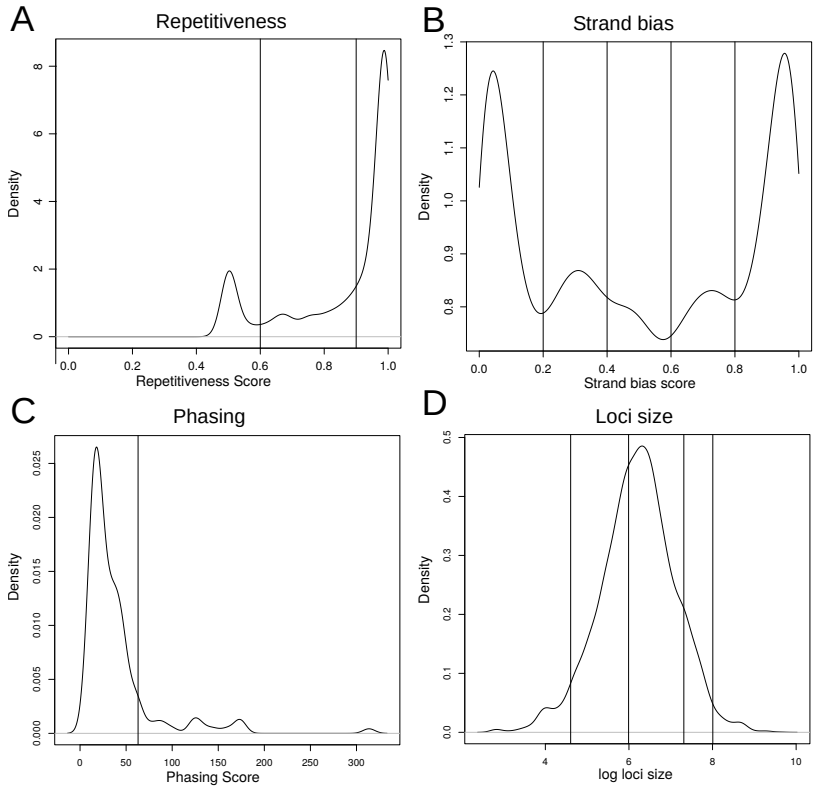

Figure 5. Density plots used to determine cut-offs for locus classification. Density plots shown for (A) Repetitiveness score, (B) Strand bias score, (C) Phasing score and (C) Log of loci size. The vertical lines indicating the cut-offs used to classify the loci into discrete classes.

repetitiveness and a stronger association with genomic methylated regions than LC3-6.

As observed in Arabidopsis [21], the repetitiveness of loci in the genome may correlate with different sRNA silencing mechanisms. LC1 is more associated with transposons (specifically retrotransposons) than LC2. Figure S1 illustrates the overlap of LC1 loci with transposons, low mappability (which corresponds to high repetitiveness) and absence of overlapping genes. Loci were split up mostly due to varying coverage.

LC2 is associated with genic regions, exemplified by a LC2 paragon (CRSL003890) in Figure S2 further can be seen in Supplementary Figures S2-5. Interestingly, a CRTOC1 transposon is superimposed on both the loci and the gene. LC2 contains the largest loci of all classes. Indeed, the shown sRNA locus is 6kb long, which is well outside the normal locus size range.

All miRNA containing loci are in LC3 along with the majority of DCL3- and AGO3-dependent loci. Of the AGO3-dependent loci, 73% have a predominance for U at the 5' end, consistent with AGO3's strong U preference [38]. LC3 loci also demonstrated common expression across wild type libraries and enrichment for 21 nt sRNAs. Interestingly, LC3 contained 130 out of a total of 149 loci that exhibit evidence of phasing.

LC4 sRNAs were similarly represented in most wild type libraries suggesting constitutive or housekeeping roles. However, unlike LC3, there was a lack of sRNA size specificity with 95% having a bias for sRNAs larger or smaller than the modal 20 and 21 nt sRNAs. DNA transposons are associated with 11% of LC5 for which there is a strong bias for sRNAs with a C at the 5'-end potentially indicative of different AGO protein

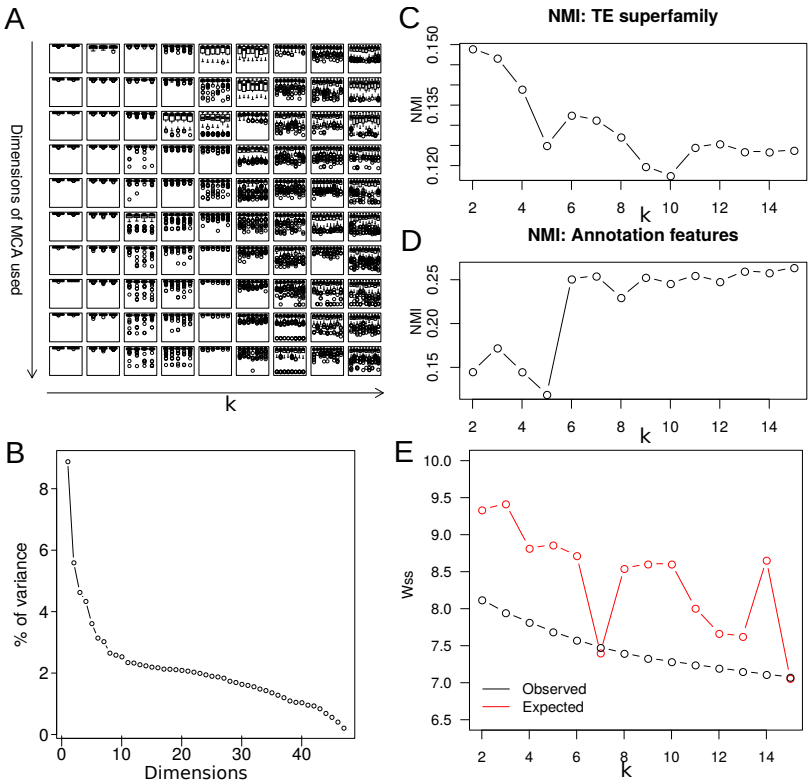

Figure 6. Diagnostic plots used to determine the number of dimensions and clusters to use in the analysis. For panels A, C-E, the x-axis correspond to cluster result running k-means with (k) clusters. (A) Stability of the cluster results (under bootstrapped sampling) achieved for all combinations of dimension selection from 1-8 and all numbers of clusters from 2-10. (B) Screeplot: Ranked percentage of variance explained by each dimension of the MCA transformed data. (C-D) Non mutual information (NMI) comparing clustering partitioning with (C) Transposon superfamily overlap and (D) annotation feature overlap. (E) Observed and expected sum-of-squares within each cluster relative to the cluster means.X-axis correspond to cluster result running k-means with (k) cluster for panels A,C-E.

association [8]. There was also a much higher level of LC5 loci specifically expressed during the zygote stage (5%) perhaps indicating roles in silencing DNA transposons at specific points in the life-cycle. LC6 had typically smaller loci (average size 386 nt) as well as most loci (85%) having an enrichment for sRNAs shorter than 20 bp.

Chromosome tracks (Figures 9), demonstrate distinct genomic location patterns for the different LC. LC1-2 are concentrated at the centromere along with higher levels of DNA methylation and a concentration of retrotransposons whereas LC3-6 meanwhile are more evenly spread along the chromosome arms. These patterns are a validation of the locus clusters as chromosomal location was not included as a feature in the MCA.

Discussion

Chlamydomonas has a silencing machinery more complex than might be expected for a unicellular organism. This complexity precludes a comprehensive characterisation of sRNA-related pathways simply by investigating individual sRNA loci in detail or by studying individual genome-wide features. To address this complexity in this study we

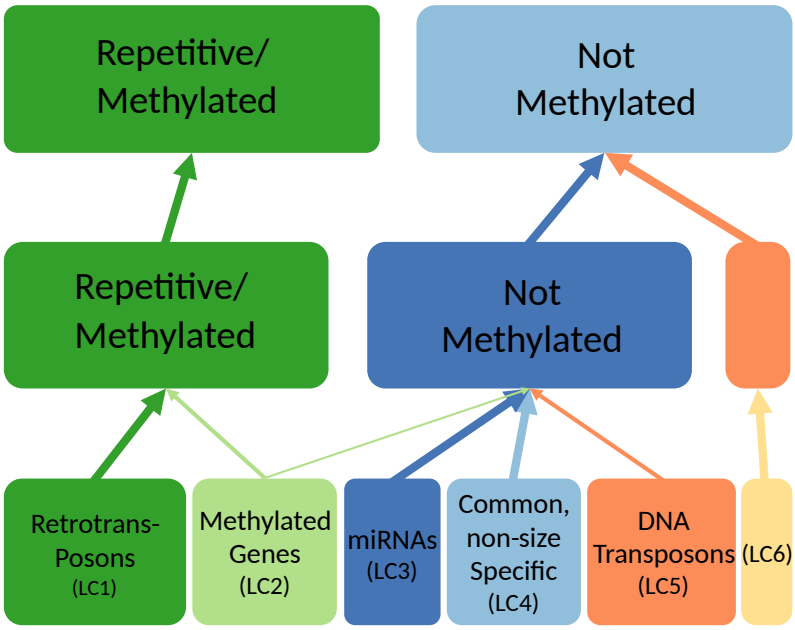

Figure 7. Cluster hierarchy plot. Hierarchy of clusters for k=2,3 and 6. Clusters are annotated with their main distinguishing features. Width of the arrow denotes the proportion of the loci contained within the "higher" cluster.

followed an approach that allows the data-driven identification of distinct types of sRNA loci [12, 13]. Importantly, the MCA uses a wide range of features as inputs to enable the robust identification of clusters, which could have not been derived by using individual features. Inspection of the feature associations for each cluster enables the validation of the clustering, as well as the dissection of important and unimportant features.

By reporting the first comprehensive map of sRNA loci in a unicellular organism we demonstrate the multi-applicability of our pipeline, which was previously used for Arabidopsis [12, 13], for locus map generation, annotation and clustering. Our results confirm (i) the overall size bias for 21nt sRNAs, (ii) the lack of enrichment in the 24nt fraction associated with the RdDM in higher plants, and (iii) the bias for U and A at the 5'-end of sRNAs [26, 36, 38, 44]. Importantly, our analyses groups known Chlamydomonas miRNAs into the same cluster, LC3. Together, these results indicate the robustness of our approach and the overall validity of our findings.

The characteristics of LC3 loci indicate that this cluster includes canonical miRNAs along with other sRNAs, all produced by DCL3-mediated cleavage of precursors and then bound by the AGO3 effector protein. As these non-miRNAs have most, but not all, features corresponding to *bona fide* miRNAs, we propose that they derive from immature miRNA precursors, which will evolve to either become canonical pre-miRNAs or are on a pathway for potential elimination from the genome.

A potential novel class of Chlamydomonas sRNAs, in LC4, is independent of DCL3 and AGO3 and, among other peculiarities, they lack bias for U/A at the 5'-end and they are variable in size. The variability in size may indicate imprecise processing by DCL1/2 possibly due to their lack of PAZ domain, which is thought to confer measuring specificity [20]. Importantly, loci in LC4 tend to be large, commonly expressed, and have low repetitiveness, suggesting that they do not represent noise. Further analyses of sRNA species from dcl1/2 and ago1/2 mutants will provide key insight on this matter.

It is likely that LC1-3 are also processed by DCL1/2 because they have a high 21nt bias. In that scenario a PAZ domain-like measuring function may be performed by other

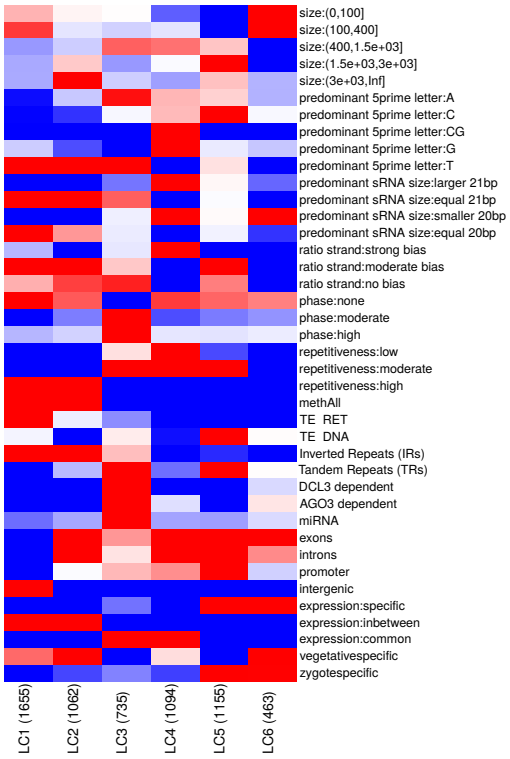

Figure 8. Heatmap showing association with annotations for the six loci clusters. Red colours indicate association while blue colours represent disassociation. The size of the clusters is shown in brackets along the x axis.

dsRNA-binding proteins, as with DGCR8 in the microprocessor complex of animals [28]. An RNA-binding protein DUS16, similarly partners with DCL3 for the proper processing of miRNAs [40, 41].

Our findings also raise questions about DNA methylation and genomic defense from transposons in Chlamydomonas. In Arabidopsis, the RdDM pathway is well characterized [22] and, in the Arabidopsis map, a subset of loci show a combination of RDR2 dependence, bias for 24nt sRNAs (the size class which directs RdDM), and overlap with methylated DNA regions including transposons (the primary targets for RdDM pathway) [13]. The crucial RdDM machinery has not been identified in Chlamydomonas and, in this study consistent with previous findings, no enrichment in the 24nt size fraction was found in any of the sRNA types. However, the presence of loci overlapping with methylated retrotransposons (LC1) and with methylated genic regions (LC2) suggests a possible role of sRNAs during establishment and/or maintenance of methylation states in Chlamydomonas genome. If this connection between sRNA and methylation does exist, then there is a possibility that these loci may represent a distinct form of RdDM in Chlamydomonas that is highly divergent from that of higher plants. Moreover, LC5 loci, with their zygotic-specific expression and DNA transposon overlap, could possibly represent a transposon silencing system activated specifically during the zygotic stage.

Our data-driven approach to identify and classify sRNA loci is intended primarily for the purpose of hypothesis generation, giving possible insights into the biosynthesis and function of sRNAs in Chlamydomonas and, potentially, in other unicellular eukaryotes. The results have allowed the identification of a number of areas for further exploration, as discussed above. Overall, when compared to the previously presented Arabidopsis locus map [13], the results indicate both convergence with higher plants (e.g. LC3)

as well as diversification (e.g. LC4). These findings support a view that significant diversification in sRNA pathways in Chlamydomonas occurred after division from higher plants. Further studies with mutant strains will enable deeper characterisation aiming to elucidate the functional significance of sRNAs in the unicellular algae Chlamydomonas as well as the evolution of RNA silencing pathways in diverse linages.

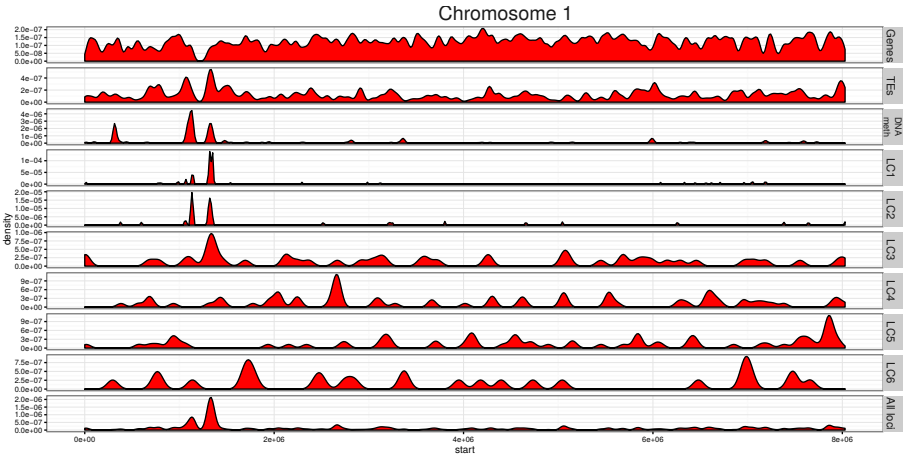

Figure 9. Chromosome tracks for Chromosome 1. Density of gene, transposon and methylation loci are shown on the first three panels. The "DNA meth" track represents combined data for the three methylation contexts from the bisulphite sequencing data. LC1-6 show the locus density of the six clusters while "all loci" plots the density of all loci.

Materials and Methods

Chlamydomonas strains and culture conditions

C. reinhardtii strains were obtained from the Chlamydomonas Resource Center (University of Minnesota) and maintained by passing cells into new fresh solid TAP media $[2-amino-2-(hydroxymethyl)-1,3-propanediol$ (TRIS)-acetate-phosphate [14] in the presence of 1,5% agar] every two months, in constant light, at 21 degree celcius.

Preparation and sequencing of sRNA libraries

Chlamydomonas cells were grown in liquid TAP media with constant shaking at 25 degree celcius under continuous illumination until cultures reached saturation. Total RNA was extracted from cell pellets with TRIzol reagent (ThermoFisher) by following a protocol previously described [26]. sRNA libraries were prepared directly from total RNAs by using the TruSeq v2 RNA Sample Preparation Kit (Illumina) following the manufacturer instructions, and then they were further sequenced on a HiSeq 2000 sequencer. Sequencing data were preprocessed using the ADDAPTS pipeline and tracking system [13, 27]. After 3' adaptor removal, all sequences <15 nt in length were discarded, and the remaining sequences were aligned against *C. reinhardtii* genome using the bowtie alignment program tolerating zero mismatches [15, 24]. Only sequences with at least one perfect match were included in further analyses. The *C. reinhardtii* reference genome and transcriptome used were Phytozome v5.0 and version 281, respectively.

sRNA Locus Map

145 sRNA libraries consisting of 54 replicate groups were used as the basis for analysis. 142 libraries were internally generated laboratory datasets [26, 37] while 3 were from [18]. A locus map for Chlamydomonas sRNAs was produced using the Bioconductor (www.bioconductor.org) package segmentSeq [12]. This package uses a heuristic approach based on sRNA densities to establish an initial locus map which is then refined using Bayesian methods to take into account the separate replicate groups. Sequences that aligned to the genome more than 200 times were excluded from the segmentation. Any gap of greater than 100nt with no reads was sufficient to split a locus. The quality of the segmentation was analysed using a series of diagnostic plots (Figure 2). The locus map was formed from all loci with a false discovery rate (FDR) of less than 0.05.

Locus Map Annotation

The locus map was annotated with intrinsic locus characteristics and publicly available annotations using functions mainly run in the R programming language [30]. The full annotated locus map can be found in Supplementary File S1.

Loci sizes

Chlamydomonas loci were classified into five discrete size classes with a plot of locus size distribution (Figure 5D) used to determine the appropriate class divisions of 100, 400, 1500 and 3000 (corresponding to log 4.6, 6.0, 7.3 and 8.0 respectively) nucleotide width.

Predominant 5' Nucleotide

To investigate whether there was a predominance for particular 5' nucleotides of the sRNAs originating from each locus, each of the four nucleotides were tested as to whether levels differed significantly from the normal ratio across all loci assuming a binomial distribution [3]. The Benjamini-Hochberg procedure was used to minimise the FDR [1].

Repetitiveness

Repetitiveness is a measure of the extent to which small RNAs that align to a given location may also align to other genomic locations. We assessed this at each locus using the following equation:

$$R = 1 - \sum_i \frac{x_i}{m_i} / \sum_i x_i$$

where x_i is the number of times the ith small RNA within the locus is sequenced and m_i is the number of genomic locations to which that small RNA aligns. This gave a score between 1 and 0 (1 being highly repetitive, 0 not at all repetitive) which was divided into three groups (low $R < 0.6$, median $0.6 < R < 0.9$, and high with $R > 0.9$) corresponding to the peaks of the distribution shown in 5A.

sRNA Strand Bias

Strand bias ratios were calculated from loci in wild type samples with more than five reads. Confidence intervals for strand bias at each locus were calculated assuming a binomial distribution and using a modified Jeffreys interval [3]. Loci were then classified as having a strong (< 0.2 or > 0.8), medium (between 0.2 and 0.4 or between 0.6 and 0.8) or no bias (between 0.4 and 0.6) 5B.

sRNA Size Ratios

Since both 20nt and 21nt long sRNAs were shown to be predominant sRNA classes in Chlamydomonas [26] and Figure 1 we calculated the ratio between them. Furthermore, there are potential physiological roles proposed for larger [43] and smaller [33] sRNAs in Arabidopsis. We therefore assigned each locus according its predominant sRNA population namely using 20nt and 21nt sizes as thresholds.

Phasing

Secondary phased (pha)siRNAs production was detected using PhaseTank, a perl based tool which searches for regions of a minimum of four 21nt sRNAs and computes a phasing score for each one [11]. Overlap between sRNA loci and phased regions were then used to annotate loci as phased or not. The phasing score was reported as part of the annotation in Supplementary File S1 ranging between 0 (no phasing) and 313. This allowed the identification of loci overlapping with a medium phasing region ($0 < \text{score} < 60$) or with a high phasing region ($\text{score} > 60$).

Expression Type

Loci were classified as to how ubiquitous their expression was. 10 wild type replicate groups were identified and loci present in more than 5 defined as common (expression:common in Fig 5), between 1 and 5 as inbetween and only 1 as specific.

Mutant, Strain and Developmental Stage Annotation

DCL3 and AGO3 dependence was calculated by determining loci present in at least one wild type replicate group but not in any of the mutant libraries. We also determined loci found specifically in one of the three used strains (CC4350, CC1883, and J3). Presence of loci specifically in libraries of either vegetative or zygotic developmental stages was also calculated.

Genome Annotations

Overlap with various genome features was calculated including predictions for genes, 5' and 3' UTRs, exons and coding sequences obtained from the Phytozome genomics portal (phytozome.jgi.doe.gov) using the most recent Chlamydomonas genome annotation (v5.5) [24]. Promoter regions were calculated as the 500 bp flanking each gene. Transposon locations were established by processing the Chlamydomonas repeat masker file to remove any sequence not explicitly identified as known transposons. Transposons were then classified (Supplementary Table S5) according to the unified system proposed by Wicker et al. and using the extensive literature concerning transposon identities [39]. Predictions of miRNAs, inverted repeats (IRs) and tandem repeats (TRs) were sourced from internal lab data with the IRF and TRF algorithms used to identify IRs and TRs respectively [36].

DNA Methylation

Bisulphite sequencing data generated was processed using yama (https://github.com/tjh48/YAMA) with the heuristics based functionality of segmentSeq used to determine loci enriched in CG, CHH or CHG methylation. sRNA Loci were then probed for overlap with these methylation enriched loci.

Multiple Correspondance Analysis

MCA was used to cluster the loci according to their annotations using the CRAN (cran.r-project.org) package FactoMineR with the HCPC function adapted to enable K-means clustering [16]. Some annotations were used as supplementary where they were not predictive of the clustering but their correlations were calculated (Supplementary Table S3).

The numbers of dimensions and clusters to select was determined by integrating information from a number of analyses consistent with that applied by Hardcastle et al. [13]. Dimensional reduction techniques like MCA are designed to concentrate the variance explained in the lower dimensions. Thus, at a certain cut-off, higher dimensions can be excluded as not being particularly significant for explaining overall variation. The graphical elbow-method is a common means to do this by considering the % of variation explained for each dimension, with a clear elbox displayed suggesting 6-10 dimensions would be appropriate 6C-E. Using random sampling from the sRNA loci with replacement to generate a consistently sized sample and re-calculating the clustering results we can observe the stability of the clustering for different combinations of dimensions (1-10) and clusters (2-10) (Figure 6A). For between six and ten dimensions clustering was generally only stable for two, three or six clusters. Based on the combination of these two analyses we determined seven dimensions to be appropriate.

To determine whether two, three or six clusters would be most appropriate, we calculated the gap statistic for seven dimensions seeing that two three and six clusters [35]. All showed large differences between the observed and expected sum-of-squares within the clusters compared to the cluster means (Wss) (Figure 6E). Computing the normalised mutual information (NMI) compared the clustering to annotation feature overlap showed a large increase in NMI for six clusters (Figure 6C and D). Thus six clusters was selected for primary analyses.

A cluster hierarchy was also generated by carrying out clustering for two, three and six clusters each time using seven dimensions which allowed the determination of how loci clustered together for lower values of k 7.

Acknowledgments

We would like to thank Attila Molnar, Betty Chung, Daisy Hessenberger and Andrew Bassett for preparing sRNA libraries, and to Thomas Hardcastle for his help during the initiation of this study. N.E.M. acknowledges the support of the Engineering and Physical Sciences Research Council (Grant numbers: EP/M506436/1, EP/ M507969/1, and EP/N509565/1). A.A.V. was supported by grant BIO2015-73900-JIN from the Spanish Ministry of Science and Innovation. S.Y.M. was supported by European Research Council Advanced Investigator Grant ERC-2013-AdG 340642 - TRIBE. D.C.B. is the Royal Society Edward Penley Abraham Research Professor.

Author contributions statement

N.E.M. and S.Y.M. conceived the experiment, S.Y.M and N.E.M. analysed the results with help of A.A.V and D.C.B, N.E.M., S.Y.M. and A.A.V. wrote the manuscript. All authors reviewed the manuscript.

Additional information

Data Availability: High throughput sequencing data have been deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-8526. All code used for bioinformatic analysis was deposited on zenodo [29].

Competing Interests

The authors declare no competing interests.

Supplemental Table

Supplementary Table S1. Overview of sRNA libraries used to determine the locus map, including details of replicate groups, genotype, lift cycle stage, and mutant libraries.

Supplementary Table S2. Summary of all annotation features showing the number and percentage of loci with a particular annotation.

Supplementary Table S3. Overview of annotations used for MCA. The "type" column indicates whether the annotation is intrinsic (a characteristic of the locus itself) or extrinsic (showing overlap with other genome features of appearing in specific strains/mutants). The final column states whether the annotation primary (i.e. used predictively in the MCA) or supplementary (i.e. not predictive for the MCA but correlations to clusters calculated).

Supplementary Table S4. Locus annotations seperated by locus cluster. This table show the number and percentage of loci in each cluster which correspond to a particular annotation. Only annotations with a binary true/false distinction are show.

Supplementary Table S5. Transposable element classification schema demonstrating the search-terms used to classify repetative sequences from the repeatmasker output into transposon superfamilies, orders and classes.

Supplementary Files

Supplementary File S1. GFF (General Feature Format) file containing all derived loci along with annotations derived in this study.

Supplemental Figures

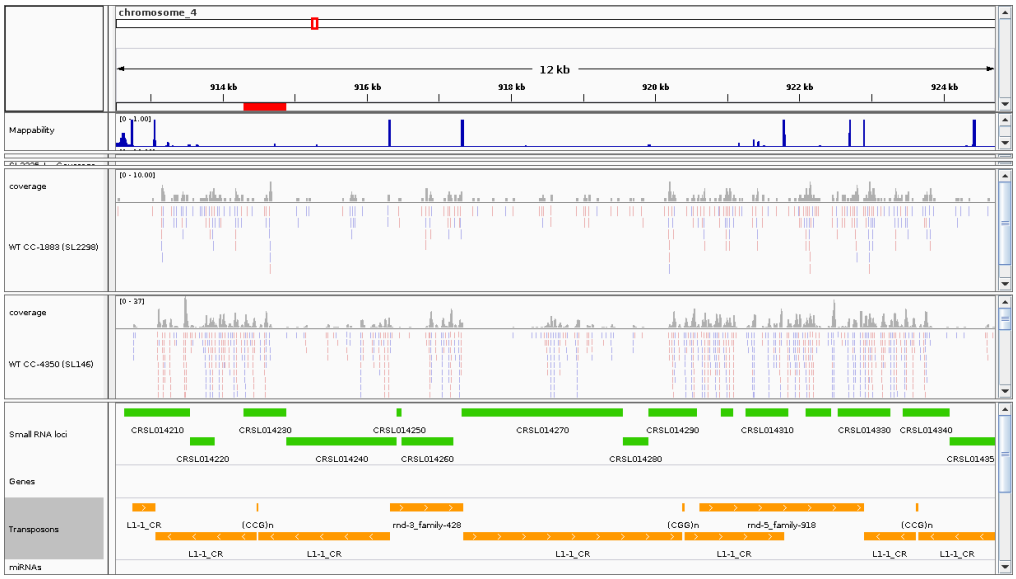

Figure S1. Genome browser view of example LC1 paragon loci (CRSL014210-350). Tracks are annotated as in 4

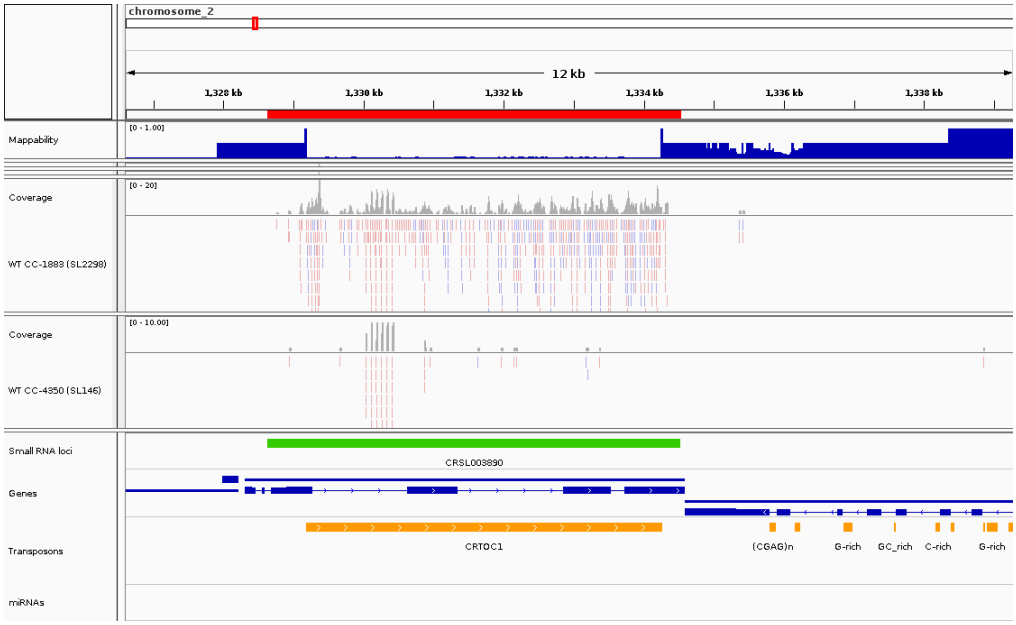

Figure S2. Genome browser view of an example LC5 paragon loci (CRSL000070). Tracks are annotated as in Figure 4.

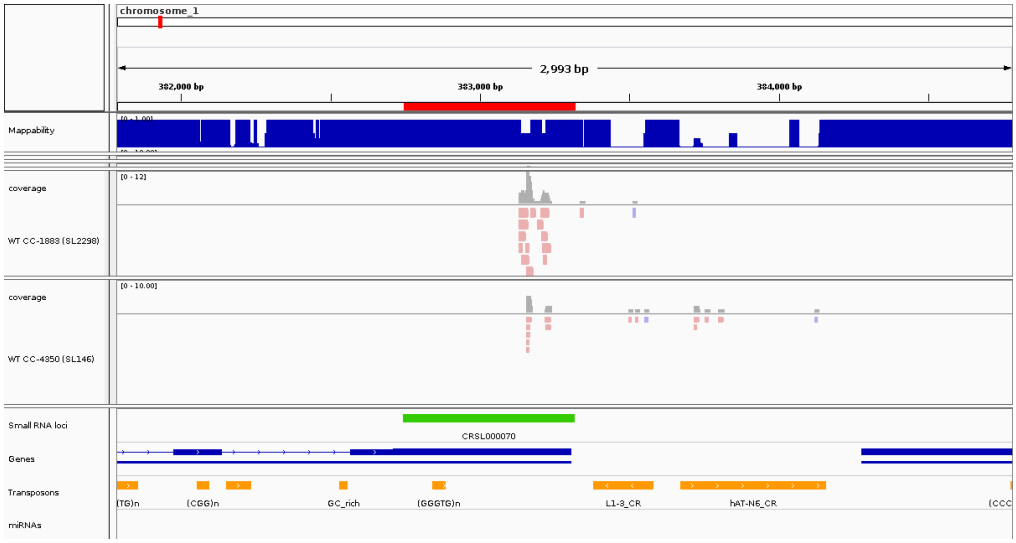

Figure S3. Genome browser view of an example LC5 paragon loci (CRSL000070). Tracks are annotated as in Figure 4.

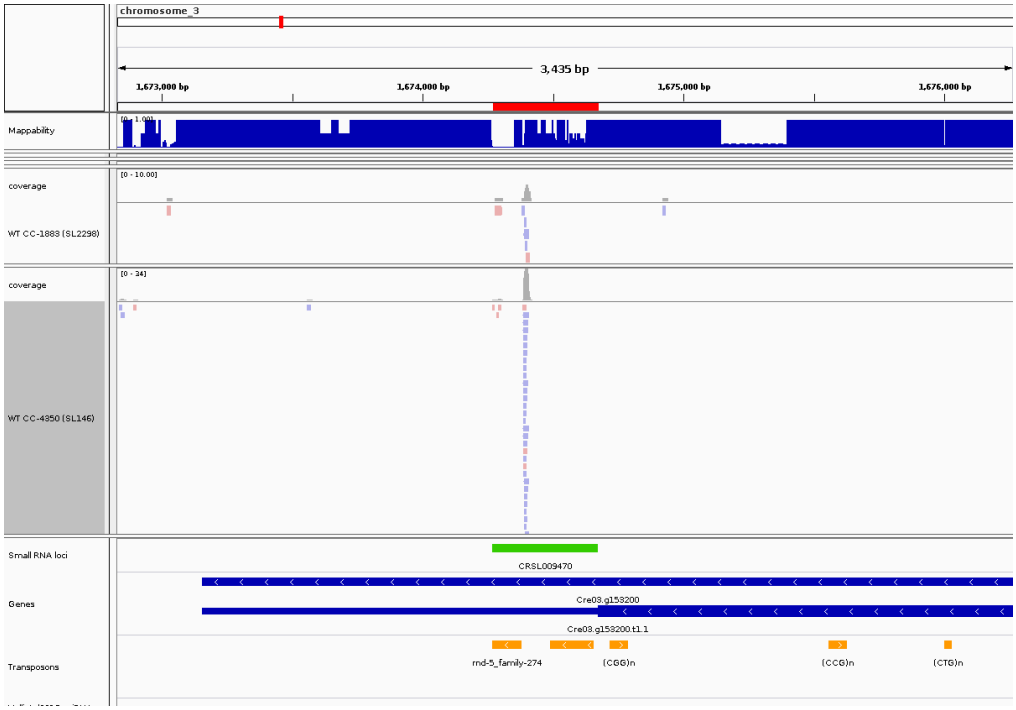

Figure S4. Genome browser view of an example LC5 paragon loci (CRSL000070). Tracks are annotated as in Figure 4.

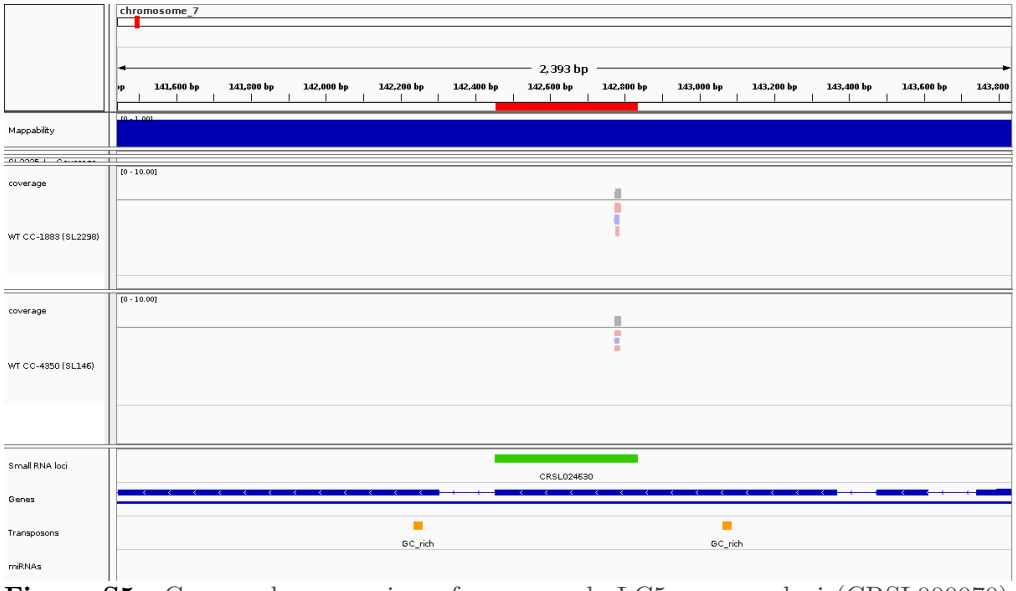

Figure S5. Genome browser view of an example LC5 paragon loci (CRSL000070). Tracks are annotated as in Figure 4.

References

1. Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

2. F. Borges and R. A. Martienssen. The expanding world of small RNAs in plants. *Nature Publishing Group*, 16(12):727–741, 2015.

3. L. D. Brown, T. T. Cai, and A. Dasgupta. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2):101–133, 2001.

4. J. A. Casas-Mollano, J. Rohr, E.-J. Kim, E. Balassa, K. van Dijk, and H. Cerutti. Diversification of the Core RNA Interference Machinery in Chlamydomonas reinhardtii and the Role of DCL1 in Transposon Silencing. *Genetics*, 179(1):69–81, may 2008.

5. H. Cerutti and J. A. Casas-Mollano. On the origin and functions of RNA-mediated silencing: from protists to man. *Current Genetics*, 50(2):81–99, aug 2006.

6. B. Y. Chung, M. J. Deery, A. J. Groen, J. Howard, and D. C. Baulcombe. Endogenous miRNA in the green alga Chlamydomonas regulates gene expression through CDS-targeting. *Nature Plants*, 3(10):787–794, oct 2017.

7. B. Y. Chung, A. Valli, M. J. Deery, F. J. Navarro, K. Brown, S. Hnatova, J. Howard, A. Molnar, and D. C. Baulcombe. Distinct roles of Argonaute in the green alga Chlamydomonas reveal evolutionary conserved mode of miRNA-mediated gene expression. *Scientific Reports*, 9(1):11091, dec 2019.

8. B. Czech and G. J. Hannon. Small RNA sorting: matchmaking for Argonautes. *Nature Reviews Genetics*, 12(1):19–31, jan 2011.

9. S. Feng, S. J. Cokus, X. Zhang, P.-Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, C. Ukomadu, K. C. Sadler, S. Pradhan, M. Pellegrini, and S. E. Jacobsen. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19):8689–8694, 2010.

10. M. Ghildiyal and P. D. Zamore. Small silencing RNAs: an expanding universe. *Nature Reviews Genetics*, 10(2):94–108, feb 2009.

11. Q. Guo, X. Qu, and W. Jin. PhaseTank: genome-wide computational identification of phasiRNAs and their regulatory cascades. *Bioinformatics*, 31(2):284–286, jan 2015.

12. T. J. Hardcastle, K. A. Kelly, and D. C. Baulcombe. Identifying small interfering RNA loci from high-throughput sequencing data. *BIOINFORMATICS ORIGINAL PAPER*, 28(4):457–46310, 2012.

13. T. J. Hardcastle, S. Y. Müller, and D. C. Baulcombe. Towards annotating the plant epigenome: The Arabidopsis thaliana small RNA locus map. *Scientific Reports*, 8(1):1–15, 2018.

14. E. H. Harris. Preface to Volume 1. page ix. Academic Press, London, 2009.

15. B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

16. S. Lê, J. Josse, and F. Husson. FactoMineR : An R package for multivariate analysis. *J Stat Softw*, 25(1):1–18, 2008.

17. Q. Liu, Y. Feng, and Z. Zhu. Dicer-like (DCL) proteins in plants. *Functional and Integrative Genomics*, 9(3):277–286, 2009.

18. K. Loizeau, Y. Qu, S. Depp, V. Fiechter, H. Ruwe, L. Lefebvre-Legendre, C. Schmitz-Linneweber, and M. Goldschmidt-Clermont. Small RNAs reveal two target sites of the RNA-maturation factor Mbb1 in the chloroplast of Chlamydomonas. *Nucleic acids research*, 42(5):3286–97, 2014.

19. D. A. Lopez, T. Hamaji, J. Kropat, P. De Hoff, M. Morselli, L. Rubbi, S. T. Fitz-Gibbon, S. D. Gallaher, S. S. Merchant, J. G. Umen, and M. Pellegrini. Dynamic changes in the transcriptome and methylome of Chlamydomonas reinhardtii throughout its life cycle. *Plant Physiology*, 169(December):pp.00861.2015, 2015.

20. I. J. Macrae, K. Zhou, and J. A. Doudna. Structural determinants of RNA recognition and cleavage by Dicer. *Nature Structural and Molecular Biology*, 14(10):934–40, 2007.

21. A. Marí-ordóñez, A. Marchais, M. Etcheverry, A. Martin, V. Colot, and O. Voinnet. Reconstructing de novo silencing of an active plant retrotransposon. *Nature Publishing Group*, 45(9):1029–1039, 2013.

22. M. a. Matzke and R. a. Mosher. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature reviews. Genetics*, 15(6):394–408, jun 2014.

23. G. Meister. Argonaute proteins: functional insights and emerging roles. *Nature Reviews Genetics*, 14(7):447–459, 2013.

24. S. S. Merchant, S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin, L. Marechal-Drouard, W. F. Marshall, L.-H. Qu, D. R. Nelson, A. A. Sanderfoot, M. H. Spalding, V. V. Kapitonov, Q. Ren, P. Ferris, E. Lindquist, H. Shapiro, S. M. Lucas, J. Grimwood, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C.-L. Chen, V. Cognat, M. T. Croft, R. Dent, S. Dutcher, E. Fernandez, H. Fukuzawa, D. Gonzalez-Ballester, D. Gonzalez-Halphen, A. Hallmann, M. Hanikenne, M. Hippler, W. Inwood, K. Jabbari, M. Kalanon, R. Kuras, P. A. Lefebvre, S. D. Lemaire, A. V. Lobanov, M. Lohr, A. Manuell, I. Meier, L. Mets, M. Mittag, T. Mittelmeier, J. V. Moroney, J. Moseley, C. Napoli, A. M. Nedelcu, K. Niyogi, S. V. Novoselov, I. T. Paulsen, G. Pazour, S. Purton, J.-P. Ral, D. M. Riano-Pachon, W. Riekhof, L. Rymarquis, M. Schroda, D. Stern, J. Umen, R. Willows, N. Wilson, S. L. Zimmer, J. Allmer, J. Balk, K. Bisova, C.-J. Chen, M. Elias, K. Gendler, C. Hauser, M. R. Lamb, H. Ledford, J. C. Long, J. Minagawa, M. D. Page, J. Pan, W. Pootakham, S. Roje, A. Rose, E. Stahlberg, A. M. Terauchi, P. Yang, S. Ball, C. Bowler, C. L. Dieckmann, V. N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roeber, S. Rajamani, R. T. Sayre, P. Brokstein, I. Dubchak, D. Goodstein, L. Hornick, Y. W. Huang, J. Jhaveri, Y. Luo, D. Martinez, W. C. A. Ngau, B. Otillar, A. Poliakov, A. Porter, L. Szajkowski, G. Werner, K. Zhou, I. V. Grigoriev, D. S. Rokhsar, and A. R. Grossman. The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. *Science*, 318(5848):245–250, 10 2007.

25. D. Moazed. Small RNAs in transcriptional gene silencing and genome defence. *Nature*, 457(January):413–420, 2009.

26. A. Molnár, F. Schwach, D. J. Studholme, E. C. Thuenemann, and D. C. Baulcombe. miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, 447(7148):1126–1129, 2007.

27. R. Narayan, K. Rutherford, R. Nag, and K. Kelly. ADDAPTS: A Data-Driven Automated Pipeline and Tracking System. *F1000Posters*, 2011.

28. T. A. Nguyen, M. H. Jo, Y.-G. Choi, J. Park, S. C. Kwon, S. Hohng, V. N. Kim, and J.-S. Woo. Functional Anatomy of the Human Microprocessor. *Cell*, 161(6):1374–1387, jun 2015.

29. nmatthews323. nmatthews323/chlamy_locus_map: chlamy_locus_map, May 2020.

30. R Core Team. R: A Language and Environment for Statistical Computing, 2018.

31. S. Sasso, H. Stibor, M. Mittag, and A. R. Grossman. From molecular manipulation of domesticated Chlamydomonas reinhardtii to survival in nature. *eLife*, 7:1–14, nov 2018.

32. M. Schroda. RNA silencing in Chlamydomonas : mechanisms and tools. *Current Genetics*, 49(2):69–84, 2006.

33. R. J. Taft, E. a. Glazov, N. Cloonan, C. Simons, S. Stephen, G. J. Faulkner, T. Lassmann, A. R. R. Forrest, S. M. Grimmond, K. Schroder, K. Irvine, T. Arakawa, M. Nakamura, A. Kubosaki, K. Hayashida, C. Kawazu, M. Murata, H. Nishiyori, S. Fukuda, J. Kawai, C. O. Daub, D. a. Hume, H. Suzuki, V. Orlando, P. Carninci, Y. Hayashizaki, and J. S. Mattick. Tiny RNAs associated with transcription start sites in animals. *Nature genetics*, 41(5):572–8, 2009.

34. C. J. Thieme, C. Schudoma, P. May, and D. Walther. Give It AGO: The Search for miRNA-Argonaute Sorting Signals in Arabidopsis thaliana Indicates a Relevance of Sequence Positions Other than the 5'-Position Alone. *Frontiers in plant science*, 3(December):272, 2012.

35. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, may 2001.

36. A. A. Valli, B. A. Santos, S. Hnatova, A. R. Bassett, A. Molnar, B. Y. Chung, and D. C. Baulcombe. Most microRNAs in the single-cell alga Chlamydomonas reinhardtii are produced by Dicer-like 3-mediated cleavage of introns and untranslated regions of coding RNAs. *Genome Research*, 26(4):519–529, 4 2016.

37. A. A. Valli, B. A. C. M. Santos, S. Hnatova, A. R. Bassett, A. Molnar, Y. Betty, D. C. Baulcombe, B. Y. Chung, and D. C. Baulcombe. Most microRNAs in the single-cell alga Chlamydomonas reinhardtii are produced by Dicer-like 3-mediated cleavage of introns and untranslated regions of coding RNAs. *Genome Research*, pages 1–11, 2016.

38. A. Voshall, E.-J. Kim, X. Ma, E. N. Moriyama, and H. Cerutti. Identification of AGO3-Associated miRNAs and Computational Prediction of Their Targets in the Green Alga Chlamydomonas reinhardtii. *Genetics*, 200(1):105–121, 2015.

39. T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman. A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics*, 8(12):973–982, 2007.

40. T. Yamasaki and H. Cerutti. Cooperative processing of primary miRNAs by DUS16 and DCL3 in the unicellular green alga Chlamydomonas reinhardtii. *Communicative & Integrative Biology*, 10(1):e1280208, jan 2017.

41. T. Yamasaki, M. Onishi, E.-J. Kim, H. Cerutti, and T. Ohama. RNA-binding protein DUS16 plays an essential role in primary miRNA processing in the unicellular alga Chlamydomonas reinhardtii. *Proceedings of the National Academy of Sciences*, 113(38):10720–10725, sep 2016.

42. T. Yamasaki, A. Voshall, E.-J. Kim, E. Moriyama, H. Cerutti, and T. Ohama. Complementarity to an miRNA seed region is sufficient to induce moderate repression of a target transcript in the unicellular green alga Chlamydomonas reinhardtii. *The Plant Journal*, 76(6):1045–1056, dec 2013.

43. R. Ye, Z. Chen, B. Lian, M. J. Rowley, N. Xia, J. Chai, Y. Li, X.-J. He, A. T. Wierzbicki, and Y. Qi. A Dicer-Independent Route for Biogenesis of siRNAs that Direct DNA Methylation in Arabidopsis. *Molecular Cell*, 61(2):1–14, 2015.

44. T. Zhao, G. Li, S. Mi, S. Li, G. J. Hannon, X.-j. J. Wang, and Y. Qi. A complex system of small RNAs in the unicellular green alga. *Genes & Development*, pages 1190–1203, 2007.