

Identification of novel, functional long non-coding RNAs involved in programmed, large-scale genome rearrangements

Sebastian T. Bechara^{1,2,†}, Lyna E. S. Kabbani^{1,2,†}, Xyrus X. Maurer-Alcalá^{1,3} and Mariusz Nowacki^{1*}

¹ Institute of Cell Biology, University of Bern, Bern, Bern, 3012, Switzerland

² Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Bern, 3012, Switzerland

³ Division of Invertebrate Zoology & Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York, 10024, United States of America

* Correspondence to: mariusz.nowacki@unibe.ch, +41 31 684 46 54

† Joint Authors

Present Address: Xyrus X. Maurer-Alcalá, Division of Invertebrate Zoology & Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York, 10024, United States of America

Short title: Long non-coding RNAs during genome rearrangements

Keywords: lncRNA, genome rearrangement, DNA elimination, sRNA, ciliate

ABSTRACT

Non-coding RNAs (ncRNAs) make up to ~98% percent of the transcriptome of a given organism. In recent years one relatively new class of ncRNAs, long non-coding RNAs (lncRNAs), were shown to be more than mere by-products of gene expression and regulation. The unicellular eukaryote *Paramecium tetraurelia* is a member of the ciliate phylum, an extremely heterogeneous group of organisms found in most bodies of water across the globe. A hallmark of ciliate genetics is nuclear dimorphism and programmed elimination of transposons and transposon-derived DNA elements, the latter of which is essential for the maintenance of the somatic genome. *Paramecium* and ciliates in general harbour a plethora of different ncRNA species, some of which drive the process of large-scale genome rearrangements, including DNA elimination, during sexual development. Here, we identify and validate the first known functional lncRNAs in ciliates to date. Using deep-sequencing and subsequent bioinformatic processing and experimental validation, we show that *Paramecium* expresses at least 15 lncRNAs. These candidates were predicted by a highly conservative pipeline and informatic analyses hint at differential expression during development. Depletion of two lncRNAs, Inc1 and Inc15, resulted in clear phenotypes, decreased survival, morphological impairment and a global effect on DNA elimination.

INTRODUCTION

In recent years, the advent of various next generation sequencing techniques, such as high throughput RNA sequencing, have revealed that the vast majority of eukaryotic genomes are transcribed into non-coding RNAs (Dunham et al. 2012). These non-coding transcripts can be broadly divided into the following categories: small non-coding RNAs (sRNAs), long non-coding RNAs (lncRNAs) and ribozymes such as ribosomal RNA. sRNAs comprise different species with at times strict classification characteristics, such as micro RNAs (miRNA) or small nuclear and small nucleolar RNAs (sn/snoRNAs); and at other times, more loose characteristics such as piwi-interacting RNAs (piRNAs) (Schmitz et al. 2016; Amaral and Mattick 2008). lncRNAs are generally defined by two characteristics. The first is their length, which is roughly defined by being long, meaning ≥ 200 bp. This cut-off was chosen arbitrarily to differentiate them from other known and well defined small RNA molecules such as tRNAs and their precursors (Schmitz et al. 2016; Amaral and Mattick 2008). The other characteristic, as the name suggests, is the lack of translation into a functional protein. The advances in ribosome profiling however revealed that a considerable amount of lncRNAs contain small open reading-frames (sORFs) that are translated into so-called “micropeptides” (Zampetaki et al. 2018; Rivas et al. 2016; Statello et al. 2021). The function of lncRNAs is highly variable and can range from gene regulation on transcriptional, post-transcriptional and even on post-translational levels, to scaffolding and the formation of nuclear condensates (reviewed in: Statello et al. 2021; Quinn and Chang 2016). Scaffolding is a diverse mode of action for functional lncRNAs. As a scaffold, the lncRNA binds effector molecules, bringing them together spatio-temporally and allowing them to fulfil their function. Examples for lncRNA scaffolds include TERC, the telomerase RNA mediating telomerase assembly (Lustig 2004), and HOTAIR, a lncRNA which binds two histone modifying complexes, promoting gene repression (Rinn et al. 2007; Tsai et al. 2010). Some, but not all, scaffolding lncRNAs mediate the formation of nuclear condensates, membrane-less compartments that exert a wide array of molecular functions such as pre-mRNA splicing and

gene expression regulation (reviewed in: Ramírez-Colmenero et al. 2020; Statello et al. 2021; Bhan and Mandal 2015). lncRNAs are well known to be expressed, exert a specific function, and even be conserved to a certain degree between different organisms (Sarropoulos et al. 2019). In unicellular eukaryotes, little is known about the existence and function of lncRNAs compared to their multicellular counterparts. Some studies have shed light on this field in recent years, most of which were conducted in yeast *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. These lncRNAs were found to be involved in transcriptional regulation and some can function as scaffolds (Niederer et al. 2017). In *Plasmodium falciparum*, lncRNAs were found to be involved in a multitude of cellular processes, including telomere maintenance and virulence gene regulation (Broadbent et al. 2011). Recent studies revealed differential expression of ~1,500 long intergenic non-coding RNAs (lincRNAs) and ~2,600 natural antisense transcripts (NATs) under various environmental constraints in the diatom *Phaeodactylum tricornutum* (Cruz de Carvalho et al. 2016; Cruz de Carvalho and Bowler 2020). Moreover, the ciliate *Pseudourostyla cristata* was found to express lncRNAs related to encystment (Pan et al. 2021). As is evident by these and numerous other studies, lncRNAs are well present and functional in unicellular eukaryotes divided by millennia of divergent evolution.

Ciliates comprise a large group of ciliated protozoans that are common in bodies of water all around the world. All ciliates share a common characteristic, termed “nuclear dimorphism” (Rzeszutek et al. 2020). They harbour two distinct nuclei, the macronucleus (MAC) and the micronucleus (MIC). The diploid MIC contains the cell’s germline genome and it is generally thought to be transcriptionally silent, although there are examples showing that MIC-limited genes exist and are actively transcribed (Neeb et al. 2017; Chen et al. 2014; Miller et al. 2021). The MAC on the other hand contains the somatic genome used for maintaining cellular functions. In contrast to the MIC, this genome variation is highly polyploid, ranging from ~45n in *Tetrahymena thermophila* to over ~75-140n in *Paramecium caudatum* and up to ~800n in the *Paramecium aurelia* complex (Aury et al. 2006; Eisen et al. 2006). During

sexual development, a new MAC genome is formed from the zygotic genome which undergoes complex DNA rearrangements. During this process chromosome fragmentation and DNA elimination occurs. The majority of the eliminated DNA consists of repetitive regions such as mini- and microsatellites, transposable elements (TEs) and transposon-derived single copy “internal eliminated sequences” (IESs)(Klobutcher et al. 1984; Rzeszutek et al. 2020; Maurer-Alcalá et al. 2018) IESs can be found intragenically, thus making their precise elimination crucial for the formation of a functional MAC genome. Different ciliates have evolved contrasting strategies to engage in large-scale genome reorganisation. Some excise IESs and ligate the MAC chromosomes back together (Marmignon et al. 2014; Sandoval et al. 2014; Mochizuki and Gorovsky 2004), others carry out an additional step, wherein they excise IESs and reshuffle the remaining DNA pieces through a process called unscrambling (Greslin et al. 1989; Chen et al. 2014). The resulting macronuclear genome becomes highly fragmented, usually carrying a single gene per chromosome (Nowacki et al. 2010). The common ground between all those approaches is that sRNAs drive the reorganisation process. Ciliates that unscramble their genome, like *Oxytricha trifallax*, utilise long non-coding transcripts of the parental somatic chromosomes to guide the DNA reorganisation (Nowacki et al. 2008; Lindblad et al. 2017), as well as parental sRNAs that protect somatic DNA from elimination (Fang et al. 2012; Zahler et al. 2012).

sRNA mediated epigenetic silencing of DNA elements is a feature also found in other single celled eukaryotes such as *S. pombe*. In this case, the process relies on a complex called RNA-induced initiation of transcriptional gene silencing (RITS), which mediates heterochromatin formation through sRNAs (Grewal and Jia 2007; Bhattacharjee et al. 2019). Ciliates however take this epigenetic silencing to the extreme, by eliminating DNA altogether from the somatic genome. The biological properties of ciliates and their peculiar genetics make them an ideal model organism to study epigenetics among others. Indeed, several key discoveries have been made in ciliates: The discovery of telomerase (Greider and Blackburn

1985), ribozymes (Cech 1985) and the first histone-modifying enzymes (Brownell et al. 1996) are just a few examples of ground breaking studies conducted in ciliates. Given the wide array of non-coding RNA species in *Paramecium*, we reasoned it is possible that lncRNAs exist and may be involved in genome rearrangements which take place during development. Hence, we investigated whether the ciliate *Paramecium tetraurelia* harbours functional lncRNAs. We collected RNA from different time points during a developmental time course, depleted rRNA as well as poly(A) transcripts and processed the data using a custom pipeline that combines reference based and *de novo* transcriptome assemblies, followed by various filtering steps for coding domains. We identified a lncRNA, lnc1, that is implicated globally in large-scale genome rearrangements. A second candidate, lnc15, is required for maintenance of cell morphology. Depletion of both candidates has detrimental effects on survival and, in the case of lnc1, DNA elimination during development.

RESULTS

Time course and sampling

To obtain the most comprehensive pool of candidates, we sampled one time point during vegetative growth and three time points during the sexual development of *Paramecium*. The three developmental time points represent different stages of the chromosomal rearrangement process. *Paramecium tetraurelia*, like other ciliates such as multiple marine *Euplotes* species, *Tetrahymena rostrata*, as well as several *Paramecia* in the aurelia clade, can undergo autogamy, a process of self-fertilisation undertaken by a single cell (Kaczanowski et al. 2016; Dini 1984; Diller 1934). During development, the old parental MAC fragments and the genome rearrangement takes place in the developing MAC. This process results in newly generated macro- and micronuclear genomes. Under laboratory conditions, autogamy can be induced by various stress conditions, including starvation (Beisson et al. 2010). Total RNA of the four samples were rRNA depleted and the early and post-autogamous samples were poly(A) depleted to detect non-polyadenylated molecules. These samples were subsequently sequenced employing stranded RNA sequencing (see Materials and Methods). The rRNA depletion was not complete but reduced rRNA levels enough for them not to be masking other transcripts (Fig. S1). We did not confirm the rRNA depletion by qRT-PCR or Northern blot, since we already observed a relative depletion of rRNA in our sequencing data.

The need for an alternative pipeline

Most lncRNA prediction pipelines or programmes employ either a combination of machine learning and filtering or rely on machine learning solely. We have tested different programmes and pipelines for their suitability in predicting lncRNAs in *P.tetraurelia*.

Programmes like CPAT that infer coding probability over k-mer usage utilizing a coding and

non-coding training set unfortunately classified known coding transcripts as non-coding. Other programmes utilising similar k-mer inference also resulted in the misclassification of genes. Pipelines such as FEELnc were able to perform up until the filter module but were not able to continue with the coding prediction because of a lack of candidates. This problem is most likely due to the fact that the macronuclear genome is highly gene rich. As such, machine learning approaches proved to be incompatible with *Paramecium* biology. Additionally, closely related organisms did not provide data that is compatible with machine learning approaches either. We therefore had to devise an alternative approach with more conventional methods relying on filtering the candidate pool while applying strict parameters. A similar approach was already applied to identify ncRNAs in *Oxytricha trifallax* (Jung et al. 2011). To this end, we constructed a modular pipeline based on a reference guided transcriptome assembly and a *de novo* assembly, followed by filtering for transcripts lacking any coding annotations and predicted coding domains (Fig. 1). The single modules can be viewed as steps in predicting or filtering non-coding RNA species, for instance 1) Transcriptome assembly, 2) Coding domain prediction and filtering, 3) Refinement by discarding known non-coding RNA species. The first step is the assembly of a transcriptome from RNA deep-sequencing by one of the first two modules, comprising a reference guided assembly and/or a *de novo* assembly followed by an initial filtering by using the first module of the FEELnc-pipeline and/or a CDS filtering step using BLAST respectively. Both modules output a preliminary candidate list. The *de novo* assembly module has the potential to output more candidates, as it does not account for directionality and abundance of the assembled transcripts. The third module comprises the translation of the candidates in all six reading frames and filters out those with coding sequences by searching for coding domains against a protein database like Pfam. Up to this point in the pipeline, the candidates of both assembly modules were kept separate. Candidates from both assembly approaches that show a major overlap with each other are merged to form a more contiguous candidate list. Finally, the merged candidates are further filtered by discarding candidates that are

predicted by Infernal to be unannotated non-coding RNA species. Subsequently, the candidates are screened for coding potential by CPC2. Although not applying machine learning *per-se*, CPC2 compares certain criteria to a set of values obtained from training a support vector machine with data sets from several model-organisms, making it suboptimal for non-standard model organisms (Kang et al. 2017).

Pipeline predicts 15 shared lncRNA candidates

Running the pipeline on all four time points resulted in multiple putative transcripts that were predicted by both branches. During sexual development, *Paramecium tetraurelia* generates non-coding transcripts in its macronuclear and micronuclear genomes, which are used for the so-called “RNA scanning”. Scanning is performed by “scan RNAs” (scnRNAs), the first of the two small RNA classes driving DNA rearrangements. During this process, scnRNAs are compared to the MAC genome in order to select sequences to be excised (Lepère et al. 2009). Because the long transcripts generated during development may be heavily fragmented, it is probable that they show up as false positive in the results of the pipeline. One sample was harvested during vegetative growth; therefore, these macro- and micronuclear non-coding transcripts should not be present there. To reduce the risk for false-positives and to reduce heterogeneity within the candidates, we regarded transcripts as putative lncRNAs if they were present in at least two of the four samples. As expected, the *de novo* assembly produced more candidates than the reference-based assembly followed by the FEELnc filter module. In comparison, the *de novo*-based workflow resulted in ~5-7 fold more candidates. We found 15 shared putative lncRNAs (Table S1).

Transcripts per million (TPMs) were calculated for all samples and z-transformed values were visualised in a heat-map (Fig. 2A). These results hint at a differential expression pattern during development. Clustering reveals that the lncRNA candidates are distinctly expressed in certain stages of *Paramecium* development. Generally, expression of most candidates seems to peak mainly in early development suggesting a role during early stages

of the RNA-guided genome reorganization (Sandoval et al. 2014; Swart et al. 2014). Because we could only calculate TPMs and we manipulated the samples prior to sequencing, these results can only be used with low confidence, and further studies need to be conducted to assess the proper differential expression of the predicted candidates. Similar to mRNAs, most lncRNAs are single stranded transcripts and often convey their function over specific secondary structures and/or sequence guided interactions (Statello et al. 2021; Quinn and Chang 2016). Reports have shown that some lncRNAs like enhancer RNAs (eRNAs) or enhancer associated lncRNAs (elncRNAs) are bidirectionally transcribed (Andersson et al. 2014; Hon et al. 2017). Since the data used here was obtained from a stranded RNA sequencing, we are able to specifically identify the orientation of the sequenced fragment. If the lncRNA candidates at hand are bidirectionally transcribed, we are able to detect this. Figure 2B shows the directionality of reads mapped to all lncRNA candidates. As evident, the reads primarily map in one direction, highlighting the single stranded nature of our lncRNA candidates.

Knock-down of the candidates lnc1 and lnc15 affects survival and morphology

To investigate whether the predicted candidates are true positives and not just artefacts, we verified their presence via RT-PCR. We tested the six largest predicted candidates (Table S1). As shown in figure 3A, we were able to amplify four of the six tested candidates with their predicted full length or near full length. lnc1 has a predicted size of over 6 kbp (Table S1), but we were able to amplify a 2.5 – 3 kbp fragment, suggesting that it is present in the cell. Similarly, lncRNA candidates 13 – 15 were amplified in their near full length of approximately 1 kbp. We were unable to amplify lnc3 and lnc7 in their full length, but fragments at either end of these predicted transcripts were amplifiable. It is possible that our pipeline predicted larger sizes than are present in the cell, which could be the reason why we were unable to amplify some candidates in their full length. Another possibility might be the presence of stable secondary structures. All the candidates except for lnc1 were

amplified from RNA taken at a vegetative time point. Lnc1 was amplified from RNA taken at an early stage during development. This is consistent with the observed expression pattern from the previously described RNA sequencing (Fig 2A) and suggests that the pipeline predicted RNA molecules that are present in the cells.

To test whether the (near) full length candidates exert a function, we performed silencing by feeding (see Materials and Methods) of all four candidates and screened for an effect on survival of progeny cells after autogamy. Of the tested candidates, Inc1 and Inc15 showed an effect on survival (Fig. 3B). This effect was reproducible. Both silencings were efficient judging by RT-qPCR and the phenotypes arising from the depletion in subsequent replicates (Fig. S2 & S3). Since most of the lncRNAs already showed an expression during the vegetative time point, we induced silencing during vegetative growth and allowed the cells to have several fissions (~12 fissions) in the silencing media. We observed that the cells were growing slower than the typical division-rate of 4/24h (Beisson et al. 2010) before inducing autogamy. This indicates an effect on cellular fission, which could be attributed to light mortality, impaired cell division or possibly a slower metabolism. The sexual progeny of Inc1 and Inc15 silenced cells showed a visible decrease in survival. 77% of the Inc1 silenced cells died in our survival tests over a period of 3 days after the progeny cells were refed and 23% did not divide at the usual rate. The control progeny cells showed normal growth during the same time frame. The vast majority of cells (~90%) in the Inc15 silencing culture developed a morphological abnormality during vegetative growth (Fig. S3A). The cells were still able to go through autogamy, which was evident because we observed fragmentation of the old MAC, a sign of progression through development. Although the cells subjected to Inc15 silencing underwent development, they did so at a decreased division rate, i.e., the cells needed ~2 additional days to complete sexual development compared to the EV control culture. We attribute this delay to an increased mortality and defective division due to the morphological effect.

Once the Inc15-KD cells were finally able to undergo development, we could observe that from the initial ~90% showing the morphological abnormality, the percentage was reduced to nearly 50%, with the other half appearing “normal”, which we also attribute to a certain degree of mortality during vegetative growth. Because of the morphological phenotype not affecting all cells at the onset of development, or some cells being able to survive without apparent defects, we conducted survival tests on cells showing morphological abnormalities and on cells that appear normal (Fig. 3B). 27% of the cells showing the phenotype died, 30% divided at vastly reduced rates (sometimes only once) and 43% showed normal growth. Out of the Inc15 knock-down cells with an apparent normal morphology only 7% died and 13% showed a decreased division rate. This result suggests that the phenomenon observed for Inc15 silencing might be linked to a dilution of either Inc15 or an increase of cells that are able to withstand the loss of the aforementioned lncRNA.

Inc1 knock-down affects IES excision on a global scale

Since the knock-down of Inc1 and Inc15 influenced survival, we tested whether both candidates affect IES excision. To this end, we tested IES retention by PCR, using 8 primer pairs flanking known IESs (Fig. 4A). If an IES is excised correctly, a shorter fragment will be amplified, representing the genomic region lacking the IES. If the IES in question is retained, a longer fragment will be amplified. The silencing of Inc1 affects the retention of IES 5 reproducibly. The silencing of Inc15 showed no effect on IES retention of the tested IESs; however, we observed a smaller than expected PCR product for IES10 which is shorter than the regular macronuclear sequence devoid of the IES. This may be due to the use of alternate TA boundaries outside of the original IES boundaries, leading to a larger deletion. Because Inc1 showed IES retention of one IES in the initial PCR experiments, we analysed whether Inc1 has a global effect on genome rearrangement or whether it was solely affecting a small subset of IESs. To this end, we sequenced DNA isolated from newly developed macronuclei following Inc1 silencing, which was used to calculate IES retention scores

(IRSs) (see Materials and Methods). The IRSs are consistent with our IES retention PCR analysis. A silencing of PiggyMac (PGM), the domesticated PiggyBac transposase responsible for IES excision, was initially used to identify all known ~45,000 *Paramecium* IESs and shows a mean IRS of 0.77 (77%) (Arnaiz et al. 2012). IESs can be classified into sRNA dependent and independent, meaning they either require or do not require scnRNAs for their excision by PGM. This classification is dependent on the IRS from knock-downs of Dicer-like (Dcl) enzymes, which produce the sRNAs required for IES excision. ScnRNAs, which are derived from the MIC genome, are produced by Dcl2/3 while iesRNAs, derived from excised IESs, are produced by Dcl5. The initial excision of sRNA dependent IESs is mediated by scnRNAs while iesRNAs ensure complete excision of the remaining copies of excised IESs, thus acting as a positive feedback loop (Sandoval et al. 2014; Furrer et al. 2017). Most IESs in the *Inc1* silencing are weakly or not retained compared to the original PGM silencing, but show a comparable IRS distribution to other key players of IES elimination such as the Dcls, Ptiwis and PDSG2 (Fig. 4B-D). IRS distribution is indicative of the general function of a gene involved in the rearrangement mechanism relative to the function of known genes that impact the process. All depicted key players in Figure 4B-D show a large number of IESs with a relatively low IRS, which is typical for genes involved in the sRNA guided excision pathway. The *Inc1* knock-down has a mean IRS of 0.06 (6%). In comparison, knock-down of Dcl2/3 and Dcl5 lead to a mean IRS of 3 and 2,6 % respectively (Sandoval et al. 2014). Out of the ~45,000 IESs, 12,431 IESs were retained with an IRS higher than 0.05 (5%), 8,548 IESs were retained with an IRS higher than 0.1 (10%). Depletion of *Inc1* seems to affect IESs from all sizes similarly (Fig. 5C-D). There is no apparent bias towards smaller or larger IES judging by the mean IRS. In order to rule out a strict *cis* interaction of *Inc1* with the affected IESs, we quantified IESs with an IRS higher than 0.1 per each scaffold (Fig. 5B). Scaffolds with a number higher than 200 show 100% retention. This due to the fact that they are rather small in comparison, often only 20 kbp (Fig. S4). As evident, the *Inc1* silencing affects IESs on all scaffolds with a similar severity.

To estimate whether a gene is involved in a similar pathway/molecular function to any other gene, global IRSs can be correlated to one another, which hint at the function of the gene in question, e.g. if IRSs of a gene silencing strongly correlate to those of a Dcl2/3/5 and Ptiwi01/09 knock-down, it will probably be involved in the scanning process, i.e. the scnRNAs pathway (Swart et al. 2017). Correlating the IRSs generated from a *Inc1* silencing reveals a moderate correlation with the Dcl enzymes and the Ptiwi-enzymes (~0.5 each, Fig. 5A) (Swart et al. 2017). *Inc1* shares half of the Dcl5-sensitive IESs and Dcl2/3/5-sensitive IESs (Fig. 4B&E). Dcl2/3 and Ptiwi01/09 are responsible for the biogenesis and transport of the scnRNAs respectively during the early stages of the programming of the DNA elimination process. Dcl5 and Ptiwi10/11 are enzymes expressed during late stages of development and mediate iesRNAs and transport respectively (Sandoval et al. 2014; Swart et al. 2017; Furrer et al. 2017; Bouhouche et al. 2011). *Inc1* shows the highest correlation coefficients with 0.7 for PDSG2, a protein involved in iesRNA processing (Arambasic et al. 2014). This correlation hints at an involvement in the iesRNA pathway. *Inc1* knock-down covers ~63% of all IESs impacted by PDSG2 knock-down (Fig.4D&F). Although the precise function of PDSG2 is unknown, its depletion was shown to abolish detectable iesRNAs, suggesting an impairment in their production or stability (Arambasic et al. 2014). In addition to *Inc1*, we have also sequenced genomic DNA from a *Inc15* knock-down (data not shown). *Inc15* does not correlate with any of the tested gene knock-downs, indicating that the *Inc1* correlations are not merely coincidental.

***Inc1* knock-down affects iesRNA levels during DNA elimination**

Given the correlation coefficients of the *Inc1* knock-down with PDSG2 knock-down, we investigated whether depletion of *Inc1* affects sRNA levels, specifically iesRNAs levels, as it is the case for PDSG2. We isolated total RNA from cells following *Inc1* knock-down or EV control at an early and late developmental time point and performed sRNA sequencing. The obtained sRNA sequences were mapped to the *Paramecium* genome. Both classes of

sRNAs that mediate IES excision have distinct properties when it comes to size and temporal expression pattern: scnRNAs are exclusively 25bp long, are produced during the early stages of development and map to macronuclear destined sequences (MDSs) i.e., genomic sequences to be retained, as well as IESs and other eliminated sequences (OESs) (Swart et al. 2017). OESs are DNA sequences which are germline-specific and cannot be merged with the existing genome assembly. iesRNAs have a size range of 25-35bp, are produced during the late stages of development and map exclusively to IESs, i.e., sequences to be excised from the genome (Sandoval et al. 2014; Lepere et al. 2009). We routinely map and analyse sRNAs by size for given time points by plotting relative abundance in a histogram, which can give additional information about a phenotype. Disruption or delay of sRNA production is usually reflected in a visible change of abundance or timing of occurrence of those sRNAs. Mapping and quantification of sRNAs from Inc1 knock-down and EV shows a visible reduction of iesRNAs upon Inc1 knock-down (Fig.6). Timing of expression of both iesRNAs and scnRNAs as well as abundance of scnRNAs does not seem to be affected. iesRNAs production is not completely abolished upon Inc1 depletion, in contrast to PDSG2, whose knock-down yields no detectable iesRNAs sequences. The effect of Inc1 knock-down on iesRNAs abundance suggests that Inc1 may be involved in sRNA processing and further illustrates its involvement in IES elimination during late developmental stages.

Inc1 depletion affects nuclear localisation and distribution of Dcl5-GFP in developing macronuclei

Depletion of PDSG2 affects the localization of Dcl5, the enzyme responsible for iesRNAs production, in the developing MACs (Sandoval et al. 2014; Arambasic et al. 2014). Dcl5 localizes to the developing MACs as nuclear foci (Sandoval et al. 2014). PDSG2 depletion leads to the disruption of said Dcl5 foci (Arambasic et al. 2014). It is hypothesized that Dcl5 operates within those foci, possibly together with other factors, and that those foci act as

processing centres for iesRNAs production. Considering this and the fact that nuclear condensates are a common mode of action for the functionality of lncRNAs, we tested whether depletion of Inc1 leads to a similar effect as PDSG2 depletion. We tagged Dcl5 with GFP on its N-terminus (Sandoval et al. 2014) and a silencing of Inc1, PDSG2 (Arambasic et al. 2014) as well as a co-silencing of both and an EV control was performed as described above. The cells were monitored throughout autogamy and imaged during development of the new macronuclei (Fig. 7). Depletion of Inc1 leads to the disruption of Dcl5-GFP foci in a similar fashion as PDSG2. Interestingly, co-depletion of both Inc1 and PDSG2 leads to a notable decrease of Dcl5-GFP foci, suggesting an additive effect. Consistent with the previous study on PDSG2, disruption of Dcl5-GFP was not observed for the EV control. This result suggests a function for Inc1 within the foci formed by Dcl5 during IES excision.

DISCUSSION

In the present study, we investigated whether *Paramecium tetraurelia* harbours functional lncRNAs. To this end, we predict lncRNA candidates and show that depleting two of them has a detrimental effect on survival (lnc1 & lnc15), morphology (lnc15) and large-scale genome rearrangements (lnc1). We applied a custom pipeline to predict these lncRNA candidates. Since it does not apply any machine learning and solely works via filtering steps, the magnitude of different lncRNA classes is limited. All of the predicted candidates are found intergenically to their neighbouring coding gene, technically classifying them as lincRNAs (Table S1). Because hits to any CDSs are filtered out in the early steps of the pipeline, all overlapping species, which make up a substantial amount of lncRNAs, are not caught by the pipeline, highlighting its strict nature. NATs also will not appear in the output of the pipeline as long as they overlap with coding transcripts, although they can be found in the intermediate output of the first two modules by custom scripts. Previously predicted putative lncRNAs such as MS2 (Tanabe and Le 2006; Tanabe and Mori 2003) in *P.tetraurelia* were found by the pipeline in the first two modules, suggesting that it functions under agreeable parameters. In summary, the pipeline is by no means an alternative for use in model organisms where enough data is available to train machine learning algorithms. It is meant as a first step to the discovery of novel non-coding transcripts in non-model organisms or organisms with divergent biology where standard approaches are not applicable.

Knock-down of lnc1 leads to the retention of several thousands of IESs. The global retention is relatively low with a mean IRS of 6% but comparable to the mean IRS of 3% and 2.6% in a Dcl2/3 and Dcl5 silencing respectively (Sandoval et al. 2014). Dcl2/3 are responsible for the generation of scnRNAs and Dcl5 generates iesRNAs. Their silencing therefore leads to IES retention in an indirect manner, which might also be the case for lnc1. The high correlation to the IES retention of PDSG2 and moderate correlation to key proteins also

expressed during late development, such as Ptiwi10/11 and Dcl5 silencing, suggest that Inc1 may work together with these proteins to facilitate IES excision during the late stages of development. PDSG2 is suggested to be involved in the production of iesRNAs. A previous study found that PDSG2 depletion leads to near complete lack of iesRNAs, which was attributed to PDSG2 directly, because the scanning process appeared to function normally and scnRNAs were not retained in later stages of development (Arambasic et al. 2014). In contrast to PDSG2, we found that Inc1 does not abolish iesRNAs production but instead leads to a notable reduction in their abundance, suggesting that Inc1 is involved in iesRNA processing. This result, as well as the correlation of IRSs between PDSG2 and Inc1 further suggests that Inc1 functions in the same pathway during late stages of development, thus facilitating DNA elimination.

Dcl5 localizes as foci in the developing MACs. These foci were seen to be disrupted when observing GFP tagged Dcl5 in a PDSG2 depleted background (Sandoval et al. 2014; Arambasic et al. 2014). Nuclear condensates are focused processing centres for various cellular processes such as splicing and translational regulation, often associated with lncRNAs. In light of this, we analysed Dcl5-GFP localization in a Inc1, PDSG2 and Inc1/PDSG2 depleted background. We observed the previously characterized disruption of foci upon PDSG2 knock-down (Arambasic et al. 2014) and found that depletion of Inc1 has a similar effect on Dcl5-GFP foci in the new developing MACs. Furthermore, co-depletion of Inc1 and PDSG2 leads to a strong decrease of the number of Dcl5-GFP foci, suggesting an interplay between Dcl5, PDSG2 and Inc1 within said foci. Interaction between Inc1, Dcl5, PDSG2 and perhaps other proteins or ncRNAs could mediate the function or maintenance of those foci, which in turn could participate in the IES excision process. An example of a lncRNA operating through nuclear speckles is NEAT1, which plays an integral part in the function and formation of nuclear speckles by functioning as a scaffold. Among numerous binding partners, it interacts with MALAT1, another lncRNA of crucial importance for the

function of nuclear speckles (West et al. 2016; Tripathi et al. 2010; Cai et al.; Fei et al.). Depletion of NEAT1 leads to a disruption of nuclear speckles, however in the case of MALAT1, depletion does not affect the formation of speckles; rather it was found that their composition is impacted (Tripathi et al. 2010, 1). It is entirely conceivable that depletion of Inc1 affects the composition, and therefore the function, of the putative nuclear condensates formed in developing MACs during IES excision, leading to the observed IES retention. In light of the presented results, we suggest Inc1 may act as a scaffolding molecule for proteins involved in IES excision. It is conceivable that one of the roles Inc1 fulfills involves facilitation of iesRNAs production, which could be achieved by scaffolding proteins necessary for this process such as PDSG2. Ranging from transcription and pre-mRNA splicing to deposition of epigenetic marks, scaffolding lncRNAs participate in various cellular processes and contribute largely to nuclear architecture by providing membrane-less compartments, which concentrate specific proteins and nucleic acids (Rinn et al. 2007; Tsai et al. 2010; reviewed in: Banani et al. 2017) .

Interestingly, multiple other key effectors of DNA elimination in *Paramecium* besides Dcl5 localize to the developing macronuclei as foci. This includes PGM, the domesticated transposase responsible for the excision of DNA elements and one of the earliest known proteins involved in IES excision (Baudry et al. 2009). Other examples of proteins with similar localization patterns include Ezl1 (Lhuillier-Akakpo et al. 2014) and PtCAF-1 (Ignarski et al. 2014, 1), both of which are associated with the H3K9me3 and H3K27me3 histone modifications. PGM, Ezl1 and PtCAF-1 are all required for proper H3K9me3 and H3K27me3 localization, two histone modifications needed for IES excision. A recent study conducted in *Tetrahymena* presented evidence for the fact that Ezl1 and other members of the PRC complex form nuclear condensates, which they termed “Polycomb bodies” (Xu et al. 2021). Given the link between lncRNAs, nuclear condensates and the localization pattern of key effectors of IES excision, it is possible that the IES elimination process is mediated through

nuclear condensates. As mentioned for other examples, nuclear condensates act as focussed processing centres for various biological pathways. It is possible that DNA elimination itself and/or processes indirectly contributing to DNA elimination, such as iesRNAs production, are carried out within those environments. The spatially confined nature of condensates and their capability to retain or recruit specific factors could contribute to the efficiency of DNA elimination, especially considering the vast number of effectors needed as well as the large ploidy of the *Paramecium* genome. Curiously, Dcl5 shows a correlation of only 0.56 with Inc1 and only roughly half of its IESs overlap with Inc1 affected IESs. Inc1 affected IESs are not explained by combining the PDSG2 and Dcl5 affected IESs, suggesting that Inc1 might have an additional function. This fact, combined with the observed clustering of Inc1 expression during vegetative growth and early developmental stages (Fig.2A), as well as the mild correlation observed with Dcl2/3 and Ptiwi1/9 silencing (Fig.5) suggest an involvement in the early stages of genome rearrangements as well. It is possible that there are multiple functions at different stages of genome rearrangements for Inc1. Because the specific molecular function of PDSG2 is unknown, we can only speculate on what interaction Inc1 might facilitate during late development. Further study is needed to elucidate the exact mode of action of Inc1.

The morphological effect of Inc15 depletion hints at a functional involvement of this candidate in the cytoskeleton or the cortical body of *Paramecia* cells. The fungal pathogen *Cryptococcus neoformans* utilises a lncRNA to facilitate its transition from yeast to hypha, by regulating the key player in hypha formation Znf2 in *cis* (Chacko et al. 2015). Another lncRNA termed *Tug1* was shown to be responsible for male fertility. Knock-out mice showed a low sperm count as well as abnormal sperm morphology (Lewandowski et al. 2020). Similar to those two lncRNAs, we speculate that Inc15 may be involved in proper cell formation and morphology, by either directly controlling cytoskeletal elements or indirectly by controlling their expression. These cells also have difficulties dividing properly, further hinting

at an involvement in cytoskeletal function (Fig. S3B). Some cells surviving Inc15 depletion and morphologically reverting to seemingly wildtype cells may indicate a dilution effect. Inc15 may be needed at a set equilibrium for the cell to maintain proper morphology. RNAi by feeding may be insufficient to disrupt Inc15 function, since it will be expressed at normal levels once the siRNAs are completely digested.

MATERIAL AND METHODS

***Paramecium* cultivation**

Paramecium tetraurelia strain 51 of mating-type 7 was used in this study. Cultivation and autogamy were carried out at 27°C as previously described (Beisson et al. 2010). Cells were grown in wheat grass powder (WGP; Pines International, Lawrence, KS) infusion medium bacterised with *Klebsiella pneumonia*, supplemented with 0.8mg/l of β -sitosterol (Calbiochem, Millipore).

Total RNA extraction, rRNA/mRNA depletion and sequencing

Total RNA was extracted from 200-400mL of a *Paramecium* culture during the vegetative growth state, an early developmental time point (15% of cells with fragmented old MAC), a late time point (40% of cells have visible anlagen; ~12 hours after all cells are fragmented) and a post-autogamous time point (2 days after sampling of the late stage), using TRI reagent (Sigma-Aldrich) according to the manufacturer's protocol. Ribosomal RNA was depleted using the Ribo-Zero Gold rRNA Removal Kit (Yeast; Illumina) following the manufacturer's protocol. This kit has been previously used for studies conducted in *Paramecium* (Gotz et al. 2016; Pirritano et al. 2020). In order to eliminate the majority of mRNA transcripts, we performed poly(A) depletion using the Dynabeads mRNA Purification Kit (Invitrogen, ThermoFisher Scientific). In contrast to the manufacturer's protocol we, discarded the pulled-down mRNA and purified the supernatant using the RNA Clean & Concentrator-25 Kit (Zymo Research). An Illumina TruSeq, Stranded mRNA library was

prepared according to standard Illumina protocols and sequenced with 100 cycles single-end at the NGS platform at the University of Bern. For small RNA sequencing, RNA was extracted as described above and sequenced by the NGS facility at FASTERIS SA (Geneva, Switzerland). An Illumina small RNA-seq library was prepared according to standard Illumina protocols and sequenced with 50 cycles single-end.

RNAi by feeding

Knock-down (KD) of lncRNA candidates was performed using RNAi by feeding as previously described (Beisson et al. 2010). Lnc1/15 fragments were cloned into L4440 vector and transformed into HT1115 feeding bacteria. Pre-cultures of feeding bacteria were inoculated overnight with shaking at 37°C in LB media supplemented with 0.0125mg/mL tetracycline and 0.1mg/mL ampicillin. The pre-culture was diluted 1:100 in WGP medium containing 0.1mg/mL ampicillin and expanded overnight at 37°C. The following day, the bacterial culture was diluted 1:4 in WGP medium containing 0.1mg/mL ampicillin and incubated at 37°C with shaking until it reached the log growth phase (OD between 0.07 and 0.1). OD was assessed using LLG-uniSPEC 2 spectrophotometer (Lab Logistics Group) according to the manufacturer's instructions. Double strand RNA production was induced by the addition of 0.4mM IPTG and incubation of the culture at 37°C for at least 4 hours with shaking. After induction, the silencing medium was cooled down to 27°C and supplemented with 0.8mg/l of β -sitosterol. *Paramecium* cells were seeded at a concentration of a 100 cells per mL into the silencing medium and continuously diluted with additional silencing medium over the next few days in order to allow robust silencing during vegetative growth. "Empty Vector" (EV) silencing, RNAi using L4440 plasmid without an insert, was used for subsequent analysis as a negative control.

Post autogamous assessment of survival after RNA

Viability of progeny *Paramecium* cells following vegetative silencing of lncRNA candidates was assessed by refeeding the cells post autogamy. 30 post-autogamous cells per condition were monitored over the span of three days and survival was quantified..

IES retention PCR

IES retention PCRs were performed as previously described, using genomic DNA from post-autogamous cells and standard primers (Sandoval et al. 2014).

Macronuclear DNA extraction and Illumina Sequencing

Macronuclear DNA was extracted from a lnc1-KD cell culture a few days after completing autogamy as previously described (Arnaiz et al. 2012). An Illumina TruSeq, PCR-free DNA library was prepared according to standard Illumina protocols and sequenced with 150 cycles paired-end at the NGS platform at the University of Bern.

RT-qPCR

Total RNA from vegetative cells in silencing medium was extracted as described above and reverse transcribed into cDNA using the GoScript RT system (Promega) and primers containing random hexamers. qPCR on EV, lnc1-KD and lnc15-KD was performed using MESA Green qPCRTM Mastermix Plus for SYBR[®] Assay on an ABI Prism 7000 Sequence Detection System (7000 SDS instrument) according to the manufacturer's protocol. GAPDH was used to normalize the expression levels of lnc1 and lnc15 to the EV sample with the $\Delta\Delta C_t$ method (Primers are listed in table S2).

Confocal microscopy

Cells were collected at different developmental time points and stored in 70% EtOH at 4°C. For imaging, cells were washed twice with PBS and incubated in staining solution (0.5% Triton; 0.002% 4,6-diamidino-2-phenylindole (DAPI) in PBS) for 1 hour at room temperature. Following staining, cells were mounted onto microscopy slides using

ProLong™ Glass Antifade Mountant with NucBlue™ (Invitrogen) and imaged on a Leica SP8 STED confocal microscope using the 63x oil objective.

Modular pipeline to predict lncRNAs without machine learning algorithms

The first two modules comprise *de novo* and reference-based transcriptome assemblies and their initial filtering. Both modules take a FastQ file as input. The resulting candidates are processed by the third module, which filters for coding domains. Transcripts that pass this filter are further evaluated by the fourth module, which collapses duplicated candidates, filters for non-lnc ncRNA species and adds another coding potential check. Transcripts that pass the pipeline are considered putative lncRNAs.

Reference based module: the FastQ file is if necessary, trimmed with `bbduk.sh` version 38.98 (`ktrim=l` `mink=11` `qtrim=r` `trimq=15`) and mapped to a given genome with HiSat2 version 2.1.0 with the `--dta` flag active. The generated SAM file is passed to StringTie version 2.1.1 together with a GFF/GTF file to assemble transcripts. The GTF file is passed to the first module of FEELnc pipeline, which identifies non-lncRNA transcripts and applies a size filter of 200 bps (Wucher et al. 2017). The resulting GTF file is converted to a FastA using `gffread` version 0.11.7. and a reference genome.

De novo-based module: The FastQ is, if necessary, trimmed with `bbduk.sh`, and passed to SPAdes version 3.13.1 which is executed with standard parameters adding the `--rna` flag. The resulting transcripts are filtered by size, with a cut-off of 200 bps. These transcripts are aligned to the reference genome using `blastn` version 2.9.0+ and filtered for overlaps with annotated coding genes using custom scripts.

Coding domain filter: The putative transcripts from both modules are translated in all six reading frames and the peptide sequences are passed to HMMer (`hmmsearch`) version 3.3. which searches a provided Pfam database for coding domains. The HMMer output is filtered by the provided full sequence and best domain E-values. If both E-values are smaller than

1e-3 the transcript was discarded as possibly protein coding. The putative candidates are written to a FastA.

Last filter: Until now the putative transcripts were kept separate. They are checked for overlapping transcripts by blastn with the -ungapped flag set and using custom scripts. The larger transcript is kept if one transcript spanned the other. If transcripts are 80-85 % identical the larger transcript is kept. The collapsed putative transcripts are filtered for possible sn/sno/tRNAs that may have been missed by the annotation of the organism in question using Infernal version 1.1.3, applying filters at levels used by Rfam, using clan competition, which scores the best hit in relation to hits in the same clan and removing redundant hits (the following flags were set: --rfam --cut_ga --nohmmonly --oclan --oskip --clanin <input.clanin>). Hits with an E-value lower than 1e-3 were discarded. CPC2 version 1.0.1 was run with default parameters and transcripts with the coding flag were discarded.

Genome-wide analysis and calculation of IES retention scores (IRSs)

IRSs were calculated using ParTIES as described in (Denby Wilkes et al. 2016). For each IES, reads with excised IESs (IES⁻) and with unexcised IESs (IES⁺) were counted and IRSs were calculated ($IRS = IES^+ / (IES^+ + IES^-)$).

Small RNA mapping and quantification

Small RNAs were binned into different size classes (15 – 35 nts) and then mapped with HiSat2 (version 2.1.0) using default parameters. Mapped reads were filtered to specific features including MAC, IES and OES sequences, the mitochondrial DNA, DNA sequence of the feeding bacteria *Klebsiella pneumoniae* and the L4440 vector (Addgene) backbone. We normalised the mapped reads with the total number of reads.

Reference genomes used for read mapping

The following sequences were used for the analysis and mapping of sequencing data: the *Paramecium tetraurelia* strain 51 MAC genome (Aury et al. 2006) (https://paramecium.i2bc.paris-saclay.fr/files/Paramecium/tetraurelia/51/sequences/ptetraurelia_mac_51.fa), the MAC+IES genome (Arnaiz et al. 2012) (https://paramecium.i2bc.paris-saclay.fr/files/Paramecium/tetraurelia/51/sequences/ptetraurelia_mac_51_with_ies.fa), the FACS sorted MIC genome (Guerin et al. 2017) (https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/tetraurelia/51/sequences/ptetraurelia_mic2.fa), the mitochondrial DNA (Pritchard et al. 1990) (https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/all/all/sequences/paramecium_mitochondrial_genomes_v1.1.fa) and the *Klebsiella pneumoniae* genome (Liu et al. 2012) (https://www.ncbi.nlm.nih.gov/assembly/GCF_000240185.1/).

GFP localization experiments with Dcl5

Dcl5 tagged with GFP on its N-terminus (Sandoval et al. 2014) was used for the localization experiments. *Paramecium* cells were microinjected with the Dcl5-GFP linearized plasmid (Beisson et al. 2010). The transformed cells were subjected to Inc1, PDSG2, and EV silencing as described above and observed throughout their development. This was followed by imaging on a Leica microscope.

Data availability

All sequencing data sets are available in the NCBI BioProject database under accession number PRJNA789403. The hard coded script of the pipeline is available at GitHub (<https://github.com/SebastianBechara>). IncRNA sequences have been deposited at GenBank under the accession numbers OL962699-OL962713 (individual accession numbers are listed in table S1).

Supplemental material

The supplement includes the following figures and tables:

- **Figure S1 Coverage of obtained reads:** histogram depicting the different proportions of reads from RNA sequencing used to identify our lncRNA candidates.
- **Figure S2 RT-qPCR analysis of the Inc1 and Inc15 knock-downs:** Histograms with plotted relative expression levels of lncRNAs knocked-down in our study.
- **Figure S3 Inc15 is involved in maintaining the cell's morphology:** Microscopy pictures of cells showcasing the morphological phenotype of Inc15-KD.
- **Figure S4 Size distribution of Paramecium MAC genome scaffolds:** Graph depicting the size distribution of *Paramecium* macronuclear genome scaffolds by length.
- **Table S1 General information on the predicted lncRNAs:** List of all identified lncRNAs, which includes information such as size.
- **Table S2 Primers used in this study:** List of oligonucleotide sequences used for our study.

Funding

This research was supported by European Research Council Grants (ERC) [260358](#) "EPIGENOME" and [681178](#) "G-EDIT", Swiss National Science Foundation Grants [31003A_146257](#) and [31003A_166407](#), and grants from the National Center of Competence in Research (NCCR) RNA and Disease.

ACKNOWLEDGMENTS

We would like to thank Nasikhat Stahlberger for technical support and members of the Nowacki laboratory for discussion.

REFERENCES

- Amaral PP, Mattick JS. 2008. Noncoding RNA in development. *Mammalian Genome* **19**: 454–492.
- Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.
- Arambasic M, Sandoval PY, Hoehener C, Singh A, Swart EC, Nowacki M. 2014. Pds_{g1} and Pds_{g2}, novel proteins involved in developmental genome remodelling in *Paramecium*. *PLoS one* **9**: e112899.
- Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury JM, Denby Wilkes C, Garnier O, Labadie K, Lauderdale BE, Le Mouel A, et al. 2012. The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS genetics* **8**: e1002984.
- Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**: 171–178.
- Banani SF, Lee HO, Hyman AA, Rosen MK. 2017. Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology* **18**: 285–298.
- Baudry C, Malinsky S, Restituto M, Kapusta A, Rosa S, Meyer E, Betermier M. 2009. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev* **23**: 2478–83.
- Beisson J, Bétermier M, Bré MH, Cohen J, Duharcourt S, Duret L, Kung C, Malinsky S, Meyer E, Preer JR, et al. 2010. *Paramecium tetraurelia*: The renaissance of an early unicellular model. *Cold Spring Harbor Protocols* **5**.
- Bhan A, Mandal SS. 2015. LncRNA HOTAIR: A master regulator of chromatin dynamics and cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1856**: 151–164.
- Bhattacharjee S, Roche B, Martienssen RA. 2019. RNA-induced initiation of transcriptional silencing (RITS) complex structure and function. *RNA Biol* **16**: 1133–1146.
- Bouhouche K, Gout JF, Kapusta A, Betermier M, Meyer E. 2011. Functional specialization of Piwi proteins in *Paramecium tetraurelia* from post-transcriptional gene silencing to genome remodelling. *Nucleic Acids Res* **39**: 4249–64.
- Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, Ribacke U, Volkman S, Duraisingh M, Wirth D, Sabeti PC, et al. 2011. A global transcriptional analysis of *Plasmodium falciparum* malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biology* **12**: 1–15.
- Brownell JE, Zhou J, Ranalli T, Kobayashi R, Edmondson DG, Roth SY, Allis CD. 1996. Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* **84**: 843–51.
- Cai Z, Cao C, Ji L, Ye R, Wang D, Xia C, Wang S, Nature ZD-, 2020 undefined. RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *nature.com*.
- Cech TR. 1985. Self-splicing RNA: implications for evolution. *International review of cytology* **93**: 3–22.

- Chacko N, Zhao Y, Yang E, Wang L, Cai JJ, Lin X. 2015. The lncRNA RZE1 Controls Cryptococcal Morphological Transition. *PLoS Genetics* **11**: e1005692.
- Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, Swart EC, Perlman DH, Doak TG, Stuart A, Amemiya CT, et al. 2014. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell* **158**: 1187–98.
- Cruz de Carvalho MH, Bowler C. 2020. Global identification of a marine diatom long noncoding natural antisense transcripts (NATs) and their response to phosphate fluctuations. *Scientific Reports* **10**: 14110.
- Cruz de Carvalho MH, Sun HX, Bowler C, Chua NH. 2016. Noncoding and coding transcriptome responses of a marine diatom to phosphate fluctuations. *New Phytologist* **210**: 497–510.
- Denby Wilkes C, Arnaiz O, Sperling L. 2016. ParTIES: A toolbox for Paramecium interspersed DNA elimination studies. *Bioinformatics* **32**: 599–601.
- Diller WF. 1934. AUTOGAMY IN PARAMECIUM AURELIA. *Science (New York, NY)* **79**: 57.
- Dini F. 1984. On the Evolutionary Significance of Autogamy in the Marine Euplotes (Ciliophora: Hypotrichida). *The American Naturalist* **123**: 151–162.
- Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.
- Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biology* **4**: 1620–1642.
- Fang W, Wang X, Bracht JR, Nowacki M, Landweber LF. 2012. Piwi-Interacting RNAs Protect DNA against Loss during *Oxytricha* Genome Rearrangement. *Cell* **151**: 1243–1255.
- Fei J, Jadalaha M, Harmon T, Li I, ... BH-J of cell, 2017 undefined. Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *jcs.biologists.org*.
- Furrer DI, Swart EC, Kraft MF, Sandoval PY, Nowacki M. 2017. Two Sets of Piwi Proteins Are Involved in Distinct sRNA Pathways Leading to Elimination of Germline-Specific DNA. *Cell Reports* **20**: 505–520.
- Gotz U, Marker S, Cheaib M, Andresen K, Shrestha S, Durai DA, Nordstrom KJ, Schulz MH, Simon M. 2016. Two sets of RNAi components are required for heterochromatin formation in trans triggered by truncated transgenes. *Nucleic Acids Res* **44**: 5908–5923.
- Greider CW, Blackburn EH. 1985. Identification of a specific telomere terminal transferase activity in *tetrahymena* extracts. *Cell* **43**: 405–413.
- Greslin AF, Prescott DM, Oka Y, Loukin SH, Chappell JC. 1989. Reordering of nine exons is necessary to form a functional actin gene in *Oxytricha nova*. *Proc Natl Acad Sci U S A* **86**: 6264–6268.
- Grewal SIS, Jia S. 2007. Heterochromatin revisited. *Nature Reviews Genetics* **8**: 35–46.
- Guerin F, Arnaiz O, Boggetto N, Denby Wilkes C, Meyer E, Sperling L, Duharcourt S. 2017. Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements. *BMC Genomics* **18**: 327.

- Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. 2017. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**: 199–204.
- Hug IS, Maurer-Alcalá XX, Bechara ST, Nowacki M. 2021. RNA Polymerase III Transcribes Small Extra Chromosomal Circular DNA. In preparation.
- Ignarski M, Singh A, Swart EC, Arambasic M, Sandoval PY, Nowacki M. 2014. Paramecium tetraurelia chromatin assembly factor-1-like protein PtCAF-1 is involved in RNA-mediated control of DNA elimination. *Nucleic Acids Res* **42**: 11952–11964.
- Jung S, Swart EC, Minx PJ, Magrini V, Mardis ER, Landweber LF, Eddy SR. 2011. Exploiting Oxytricha trifallax nanochromosomes to screen for non-coding RNA genes. *Nucleic Acids Research* **39**: 7529–7547.
- Kaczanowski A, Brunk CF, Kazubski SL. 2016. Cohesion of Clonal Life History, Senescence and Rejuvenation Induced by Autogamy of the Histophagous Ciliate Tetrahymena rostrata. *Protist* **167**: 490–510.
- Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, Gao G. 2017. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research* **45**: W12–W16.
- Klobutcher LA, Jahn CL, Prescott DM. 1984. Internal sequences are eliminated from genes during macronuclear development in the ciliated protozoan oxytricha nova. *Cell* **36**: 1045–1055.
- Lepère G, Nowacki M, Serrano V, Gout JF, Guglielmi G, Duharcourt S, Meyer E. 2009. Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. *Nucleic Acids Research* **37**: 903–915.
- Lepere G, Nowacki M, Serrano V, Gout J-F, Guglielmi G, Duharcourt S, Meyer E. 2009. Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. *Nucleic Acids Res* **37**: 903–915.
- Lewandowski JP, Dumbović G, Watson AR, Hwang T, Jacobs-Palmer E, Chang N, Much C, Turner KM, Kirby C, Rubinstein ND, et al. 2020. The Tug1 lncRNA locus is essential for male fertility. *Genome Biology* **21**: 1–35.
- Lhuillier-Akakpo M, Frapporti A, Denby Wilkes C, Matelot M, Vervoort M, Sperling L, Duharcourt S. 2014. Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements. *PLoS Genet* **10**: e1004665.
- Lindblad KA, Bracht JR, Williams AE, Landweber LF. 2017. Thousands of RNA-cached copies of whole chromosomes are present in the ciliate Oxytricha during development. *RNA* **23**: 1200–1208.
- Liu P, Li P, Jiang X, Bi D, Xie Y, Tai C, Deng Z, Rajakumar K, Ou H-Y. 2012. Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. *J Bacteriol* **194**: 1841–1842.
- Lustig AJ. 2004. Telomerase RNA: A Flexible RNA Scaffold for Telomerase Biosynthesis. *Current Biology* **14**: R565–R567.
- Marmignon A, Bischerour J, Silve A, Fojcik C, Dubois E, Arnaiz O, Kapusta A, Malinsky S, Bétermier M. 2014. Ku-Mediated Coupling of DNA Cleavage and Repair during Programmed Genome Rearrangements in the Ciliate Paramecium tetraurelia. *PLoS Genetics* **10**.

- Maurer-Alcalá XX, Yan Y, Pilling OA, Knight R, Katz LA. 2018. Twisted Tales: Insights into Genome Diversity of Ciliates Using Single-Cell 'Omics. *Genome Biol Evol* **10**: 1927–1939.
- Miller RV, Neme R, Clay DM, Pathmanathan JS, Lu MW, Yerlici VT, Khurana JS, Landweber LF. 2021. Transcribed germline-limited coding sequences in *Oxytricha trifallax*. *G3 (Bethesda)* **11**: jkab092.
- Mochizuki K, Gorovsky MA. 2004. Small RNAs in genome rearrangement in Tetrahymena. *Current Opinion in Genetics and Development* **14**: 181–187.
- Neeb ZT, Hogan DJ, Katzman S, Zahler AM. 2017. Preferential expression of scores of functionally and evolutionarily diverse DNA and RNA-binding proteins during *Oxytricha trifallax* macronuclear development. *PLoS ONE* **12**: e0170870.
- Niederer RO, Hass EP, Zappulla DC. 2017. Long noncoding RNAs in the yeast *S. Cerevisiae*. In *Advances in Experimental Medicine and Biology*, Vol. 1008 of, pp. 119–132, Springer New York LLC.
- Nowacki M, Haye JE, Fang W, Vijayan V, Landweber LF. 2010. RNA-mediated epigenetic regulation of DNA copy number. **107**: 22140–22144.
- Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, Landweber LF. 2008. RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* **451**: 153–158.
- Pan N, Bhatti MZ, Zhang W, Ni B, Fan X, Chen J. 2021. Transcriptome analysis reveals the encystment-related lncRNA expression profile and coexpressed mRNAs in *Pseudourostyla cristata*. *Scientific Reports* **11**: 8274.
- Pirritano M, Zaburanyi N, Grosser K, Gasparoni G, Müller R, Simon M, Schrollhammer M. 2020. Dual-Seq reveals genome and transcriptome of *Caedibacter taeniospiralis*, obligate endosymbiont of *Paramecium*. *Scientific Reports* **10**: 9727.
- Pritchard AE, Seilhamer JJ, Mahalingam R, Sable CL, Venuti SE, Cummings DJ. 1990. Nucleotide sequence of the mitochondrial genome of *Paramecium*. *Nucleic Acids Res* **18**: 173–180.
- Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics* **17**: 47–62.
- Ramírez-Colmenero A, Oktaba K, Fernandez-Valverde SL. 2020. Evolution of Genome-Organizing Long Non-coding RNAs in Metazoans. *Frontiers in Genetics* **11**. <https://www.frontiersin.org/article/10.3389/fgene.2020.589697>.
- Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. 2007. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* **129**: 1311–1323.
- Rivas E, Clements J, Eddy SR. 2016. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods* **14**: 45–48.
- Rzeszutek I, Maurer-Alcalá XX, Nowacki M. 2020. Programmed genome rearrangements in ciliates. *Cellular and Molecular Life Sciences* **77**: 4615–4629.
- Sandoval PY, Swart EC, Arambasic M, Nowacki M. 2014. Functional Diversification of Dicer-like Proteins and Small RNAs Required for Genome Sculpting. *Developmental Cell* **28**: 174–188.
- Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. 2019. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**: 510–514.

- Schmitz SU, Grote P, Herrmann BG. 2016. Mechanisms of long noncoding RNA function in development and disease. *Cellular and Molecular Life Sciences* **73**: 2491–2509.
- Statello L, Guo CJ, Chen LL, Huarte M. 2021. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology* **22**: 96–118.
- Swart E, Denby Wilkes C, Sandoval P, Hoehener C, Singh A, Furrer D, Arambasic M, Ignarski M, Nowacki M. 2017. Identification and analysis of functional associations among natural eukaryotic genome editing components. <http://europepmc.org/abstract/PPR/PPR41103>.
- Swart EC, Wilkes CD, Sandoval PY, Arambasic M, Sperling L, Nowacki M. 2014. Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion. *Nucleic Acids Research* **42**: 8970–8983.
- Tanabe H, Le S. 2006. Short Communication Prediction of structural homologs to functional RNAs involved in determination of life span of *Paramecium tetraurelia*. **39**: 151–156.
- Tanabe H, Mori M. 2003. Transcriptional Regulation of the MS2 Gene of *Paramecium tetraurelia*. *Japanese Journal of Protozoology* **36**: 97–104.
- Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, et al. 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell* **39**: 925–938.
- Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**: 689–693.
- West JA, Mito M, Kurosaka S, Takumi T, Tanegashima C, Chujo T, Yanaka K, Kingston RE, Hirose T, Bond C, et al. 2016. Structural, super-resolution microscopy analysis of paraspeckle nuclear body organization. *Journal of Cell Biology* **214**: 817–830.
- Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, Jagannathan V, Cadieu E, David A, Lohi H, et al. 2017. FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research* **45**: 57.
- Xu J, Zhao X, Mao F, Basrur V, Ueberheide B, Chait BT, Allis CD, Taverna SD, Gao S, Wang W, et al. 2021. A Polycomb repressive complex is required for RNAi-mediated heterochromatin formation and dynamic distribution of nuclear bodies. *Nucleic Acids Research* **49**: 5407–5425.
- Zahler AM, Neeb ZT, Lin A, Katzman S. 2012. Mating of the Stichotrichous Ciliate *Oxytricha trifallax* Induces Production of a Class of 27 nt Small RNAs Derived from the Parental Macronucleus ed. L.-H. Qu. *PLoS ONE* **7**: e42371.
- Zampetaki A, Albrecht A, Steinhofel K. 2018. Long non-coding RNA structure and function: Is there a link? *Frontiers in Physiology* **9**.

FIGURE LEGENDS

Figure 1. Assembly dependent pipeline to predict non-coding transcripts.

The pipeline takes a FastQ file as input and assembles a reference based and a *de novo* transcriptome. The reference based transcripts are filtered by the first module of FEELnc, which generates putative ncRNAs. The *de novo* transcripts are BLASTed against a

reference genome and filtered for CDSs. Pre-filtered candidates are then translated in all six read frames and a coding domain (CD) search is conducted with HMMer. Transcripts with CD hits are discarded and kept candidates are further evaluated by Infernal and CPC2 (more information in materials and methods).

Figure 2. Bioinformatic prediction of 15 shared candidates. (A) Heat map generated with the Z-transformed TPMs of each candidate in each sample. Distinct clustering can be seen at different time points. (B) Relative per base pair coverage of all lncRNA candidates. Orange depicts the negative / reverse strand and blue the positive / forward strand. As evident, all candidates show a heavy strand bias.

Figure 3. Validation of selected candidates. (A) Total RNA was reverse transcribed using primers containing random hexamers. The resulting cDNA was subjected to PCRs, in order to test the presence of the predicted candidates in the samples. RNA from an early developmental time point was used to detect lnc1. The remaining candidates were amplified from RNA taken during vegetative growth. 1 kb Plus DNA Ladder (Thermo Scientific) was used as size marker. (B) Survival tests of sexual progeny (30 cells per condition) for three consecutive days in a lnc1 and lnc15 knock-down background. Cells were divided into three groups: dead in black; sick (cells showing abnormalities in division rate) in grey, and healthy in light grey. Shown are the empty vector (EV) control, the lnc1 knock-down and the lnc15 knock-down divided into cells with abnormal (P) and normal (N) morphology.

Figure 4. lnc1 is involved in large-scale genome rearrangements. (A) Effect of lnc1 and lnc15 knock-down on IES excision. Retention of different sRNA dependent (IESs 4, 5, 7 and 9) and independent (IESs 6, 8 and 10) IESs was analysed by PCR using primers flanking each IES in question. Top band represents the retained IES, whereas the bottom band corresponds to properly processed MAC DNA. (B - D) IRS retention score distributions of several key effectors in sexual development including the Dcl (B) and the Ptiwi (C) proteins

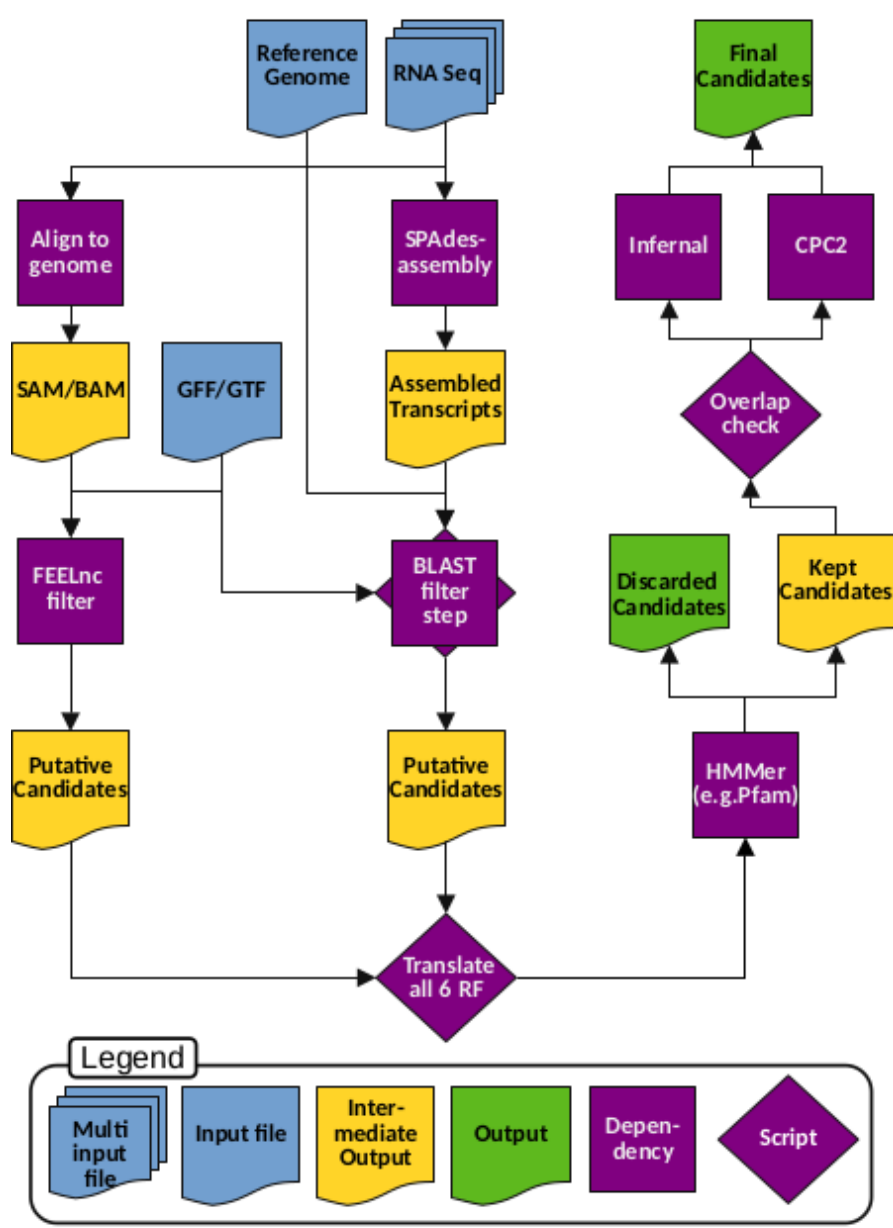
in comparison to Inc1. Dcls and Ptiwis are responsible for producing and shuttling the sRNAs driving the rearrangement process respectively. A comparison in retention score distribution between Inc1, PDSG2 and Dcl5 are given (D). All shown silencings only affect a small subset of IESs. Retention scores range from 0 (no retention) to 1 (IES is retained in all 800 genome copies). (E & F) Venn diagrams depicting the overlap in IESs with an retention score higher than 0,1. Given are the overlaps between Inc1 and Dcl5 (Sandoval et al. 2014) / Dcl2/3/5 (Sandoval et al. 2014) and PDSG2 (Arambasic et al. 2014) / Dcl5 respectively.

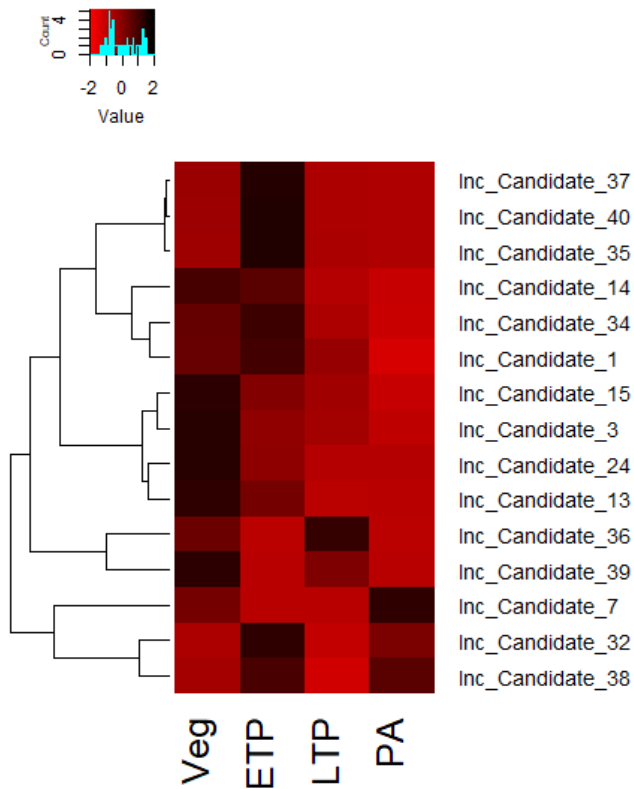
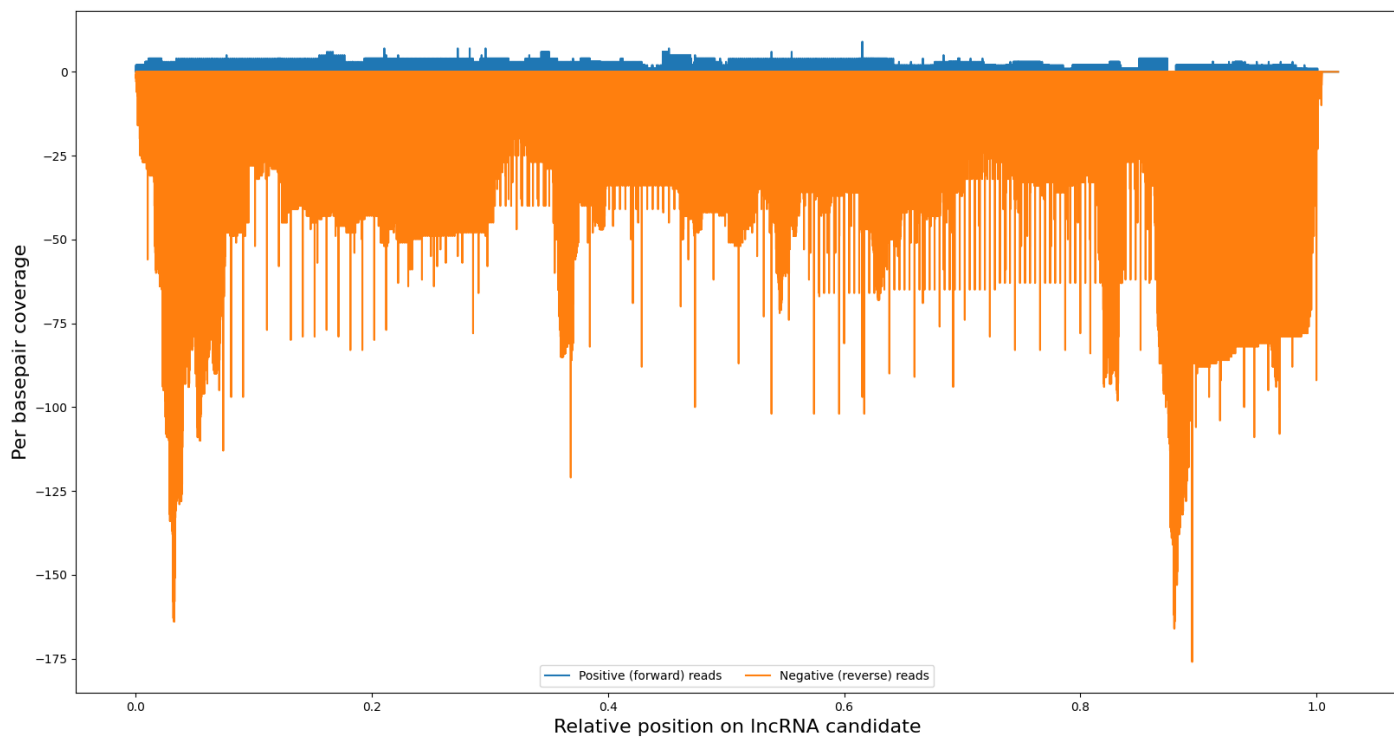
Figure 5. IESs affected by Inc1 knock-down do not show size bias. (A) Correlation matrix between several key players in the rearrangement process and Inc1. PDSG2 is included in the matrix because of the observed correlation with Inc1 in our initial analysis. Correlations were calculated using the Pearson method. IES retention scores for each knock-down were correlated using the correlation function in the pandas Python library implementing the Pearson method (B) Relative abundance of retained IESs in a Inc1 depletion background per scaffold. Most IESs can be found on the first ~200 scaffolds, leaving scaffolds with a higher number with only very few IESs (Fig. S4). 1 represents complete retention of all IESs on a given scaffold; 0 is equivalent to no retention, i.e. complete excision of IESs on a given scaffold. (C & D) The relationship between IRSs and IES length for short (≤ 200 bps; C) and long (≤ 1000 bps; D) IES. IES length distribution is given in the background as grey histogram. Lines represent a mean IRS in a 5 bp (C) and 50 bp (D) window. IRSs for single IESs are given as scatterplot in the appropriate colour.

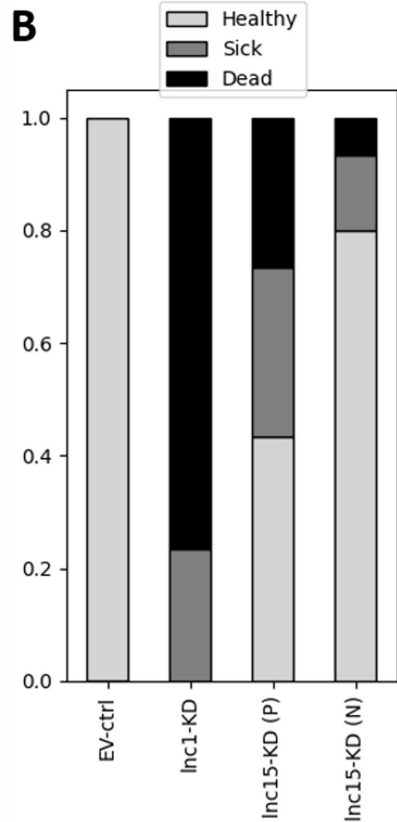
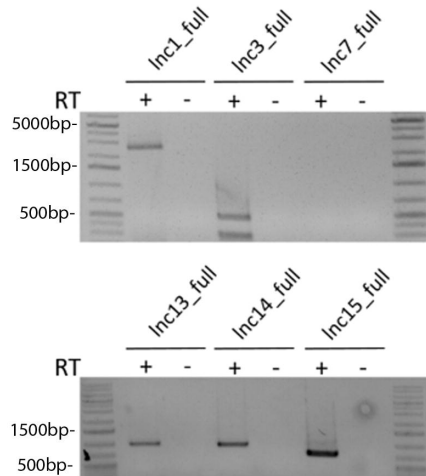
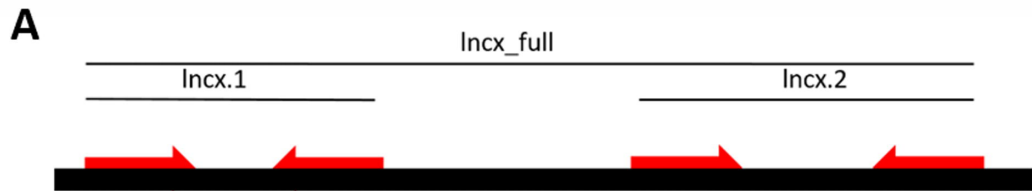
Figure 6. Inc1 depletion affects sRNA levels. Histograms of sRNAs binned by length. The top panels show the sRNA distribution during an early developmental time point, while the bottom panels show a late time point. Small RNAs from a Inc1 knock-down and an EV control were sequenced from an early and late developmental time point and mapped to the *Paramecium* genome. The proportion of reads mapping to different features such as the MAC genome (green) and IESs (red) is shown as different colours. A notable decrease of

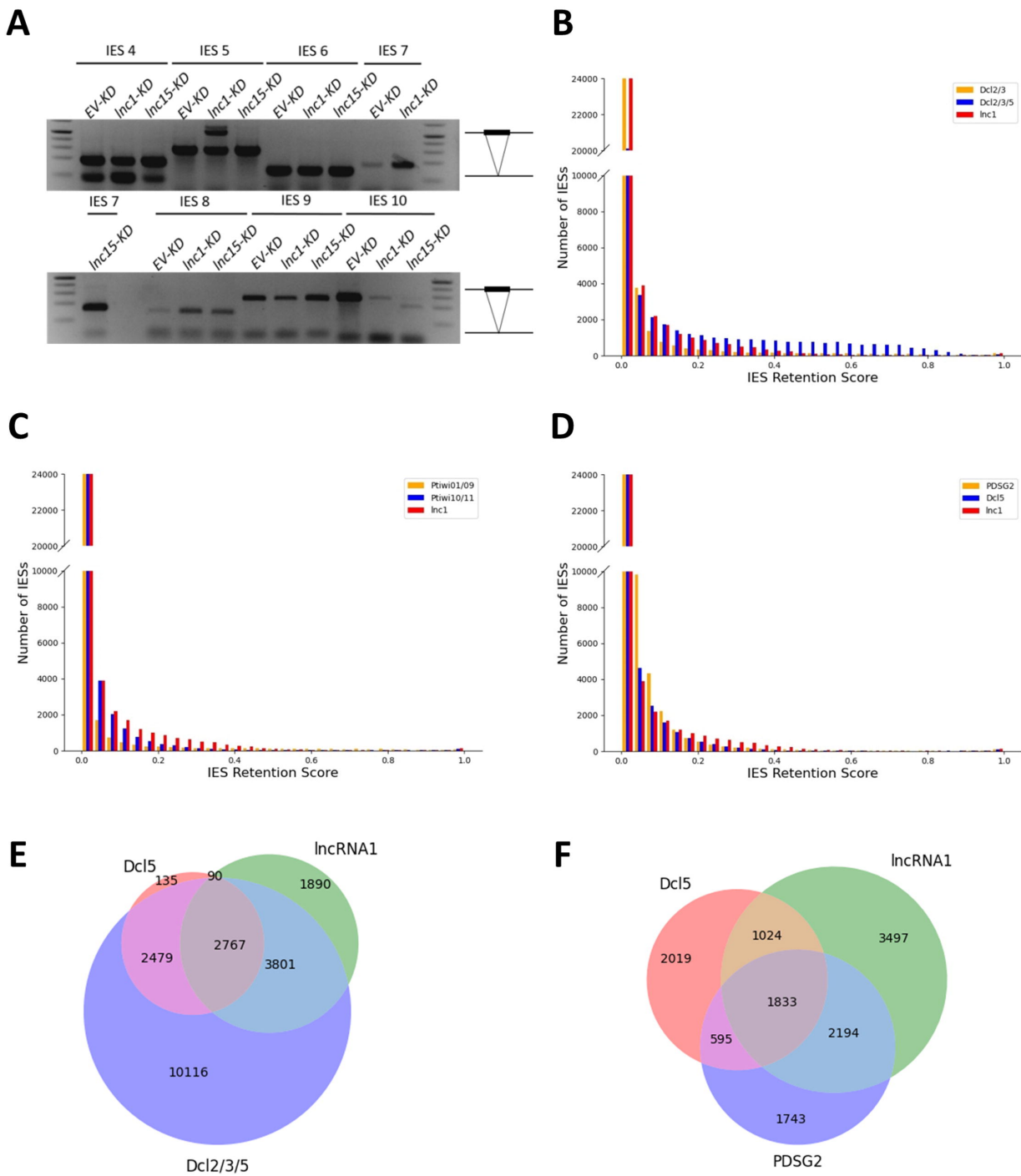
26-30bp IES-matching RNAs (iesRNAs) in the late time point can be observed in the *Inc1* knock-down.

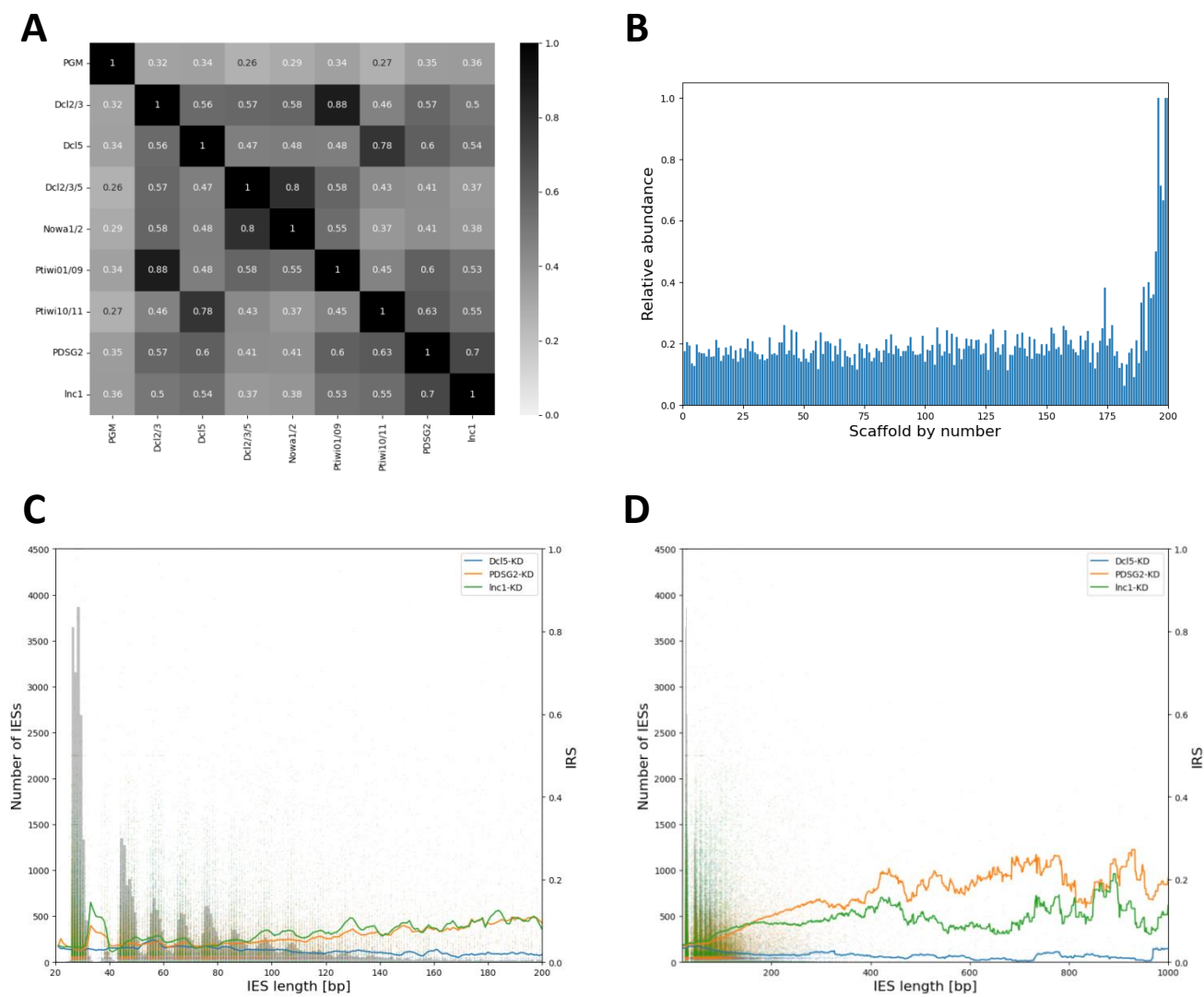
Figure 7. *Inc1* depletion affects Dcl5-GFP foci. Localization of Dcl5 tagged with GFP in the developing MACs in the EV control as well as the *Inc1* knock-down (*Inc1*-KD), PDSG2 knock-down (PDSG2-KD) and co-silencing of *Inc1* and PDSG2 (*Inc1*/PDSG2-KD). The top panels show DAPI staining in blue, which visualizes DNA, while the panels in the middle show GFP signal in green. Developing MACs are highlighted with arrows. The bottom panel shows a merge of DAPI and GFP; one of the developing MACs per cell is highlighted in detail. A similar disruption of foci can be observed between *Inc1*-KD and PDSG2-KD. *Inc1*/PDSG2-KD shows a visible decrease of Dcl5-GFP foci suggesting an additive effect.

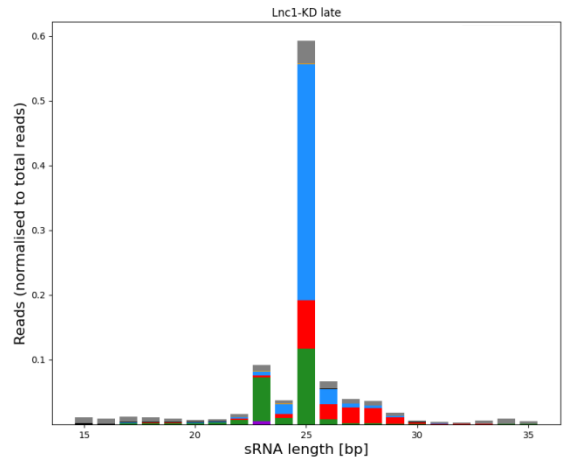
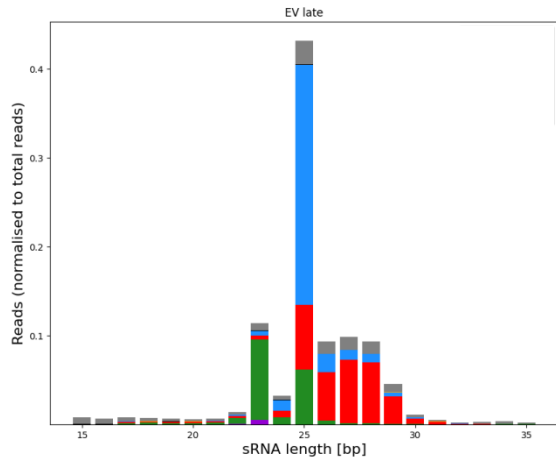
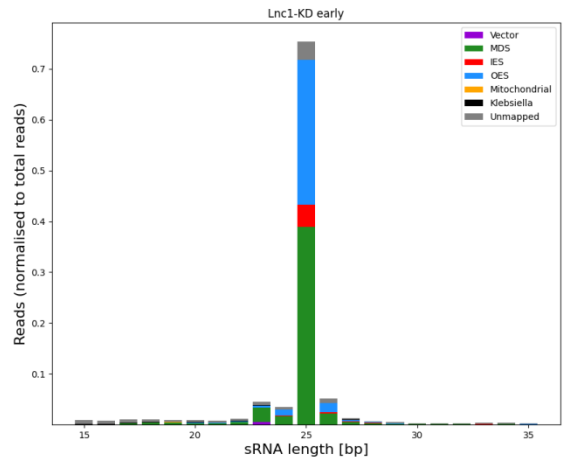
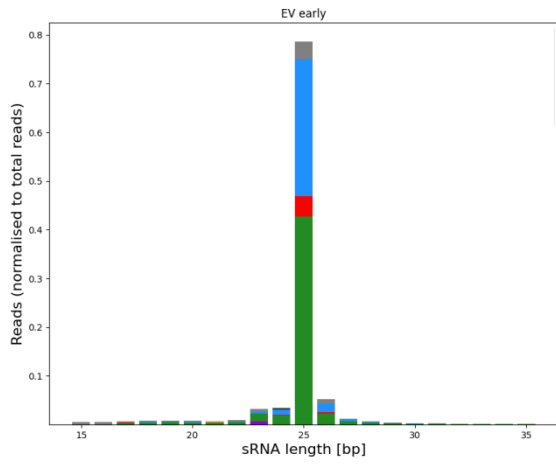


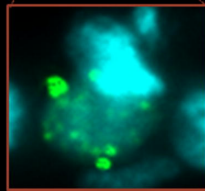
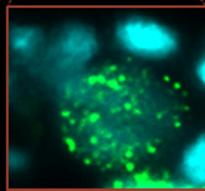
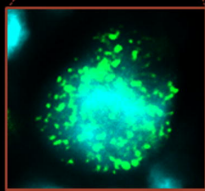
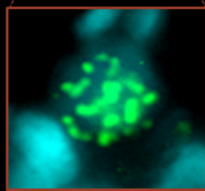
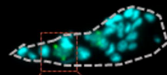
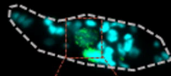
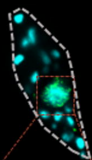
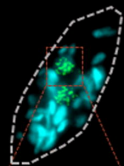
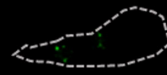
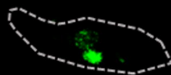
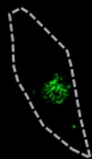
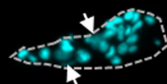
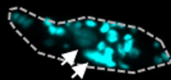
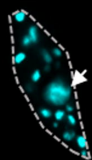
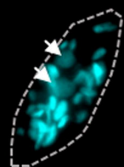
A**B**













RNA

A PUBLICATION OF THE RNA SOCIETY

Identification of novel, functional long non-coding RNAs involved in programmed, large scale genome rearrangements.

Sebastian Bechara, Lyna Kabbani, Xyrus Maurer Alcalá, et al.

RNA published online June 9, 2022

Supplemental Material <http://rnajournal.cshlp.org/content/suppl/2022/06/09/rna.079134.122.DC1>

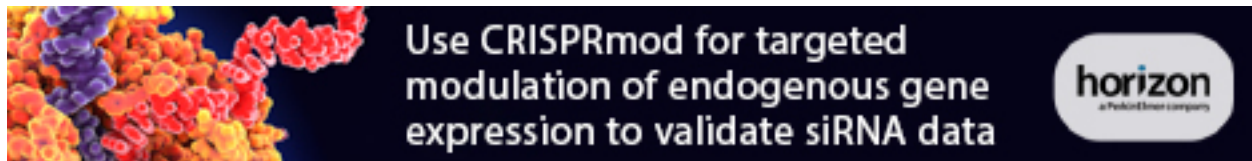
P<P Published online June 9, 2022 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *RNA* Open Access option.

Creative Commons License This article, published in *RNA*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

An advertisement banner for CRISPRmod. On the left, there is a colorful 3D molecular model of a protein complex. The text in the center reads "Use CRISPRmod for targeted modulation of endogenous gene expression to validate siRNA data". On the right, there is a logo for "horizon" with the tagline "a PerkinElmer company" below it.

Use CRISPRmod for targeted modulation of endogenous gene expression to validate siRNA data

horizon
a PerkinElmer company

To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>