METHOD ARTICLE

Identification and analysis of functional associations among natural eukaryotic genome editing components [version 1; peer review: 1 approved, 1 approved with reservations]

Estienne C. Swart ¹, Cyril Denby Wilkes², Pamela Y. Sandoval¹, Cristina Hoehener¹, Aditi Singh ¹, Dominique I. Furrer¹, Miroslav Arambasic¹, Michael Ignarski¹, Mariusz Nowacki ¹

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

²Centre de génétique moléculaire - CGM, Gif-sur-Yvette, France

V1 First published: 08 Aug 2017, 6:1374
<https://doi.org/10.12688/f1000research.12121.1>
 Latest published: 08 Aug 2017, 6:1374
<https://doi.org/10.12688/f1000research.12121.1>

Abstract

During development in the ciliate *Paramecium*, excess DNA interspersed throughout the germline genome is deleted to generate a new somatic genome. In this process, most of the intervening DNA is excised by a Piggybac-derived transposase, assisted by small RNAs (scnRNAs and iesRNAs) and chromatin remodelling. As the list of genes involved in DNA elimination has been growing, a need for a general approach to discover functional relationships among these genes now exists. We show that deep sequencing-based comparisons of experimentally-induced DNA retention provide a sensitive, quantitative approach to identify and analyze functional associations among genes involved in native genome editing. This reveals two functional molecular groups: (i) iesRNAs/scnRNAs, the putative Piwi- and RNA-binding Nowa1/2 proteins, and the transcription elongation factor TFIIS4; and (ii) PtCAF1 and Ezl1, two proteins involved in chromatin remodelling. Comparative analyses of silencing effects upon the largely unstudied regions comprising most developmentally eliminated DNA in *Paramecium* suggests a continuum between precise and imprecise DNA elimination. These findings show there is now a way forward to systematically elucidate the main components of natural eukaryotic genome editing systems.

Keywords

genome editing, DNA excision, sRNA, transposase, chromatin remodelling

Open Peer Review

Approval Status ? ✓

	1	2
version 1	?	✓
08 Aug 2017	view	view

1. Kazufumi Mochizuki , CNRS, University of Montpellier, Montpellier, France
2. Martin Simon, Saarland University, Saarbrücken, Germany
 Sivarajan Karunanithi, Saarland University, Saarbrücken, Germany
 Max Planck Institute for Informatics, Saarbrücken, Germany

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Estienne C. Swart (estienne.swart@gmail.com)

Author roles: **Swart EC:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Denby Wilkes C:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Sandoval PY:** Investigation, Methodology, Writing – Review & Editing; **Hoehener C:** Investigation, Methodology, Writing – Review & Editing; **Singh A:** Investigation, Methodology, Writing – Review & Editing; **Furrer DI:** Investigation, Methodology, Writing – Review & Editing; **Arambasic M:** Investigation, Methodology, Writing – Review & Editing; **Ignarski M:** Investigation, Methodology, Writing – Review & Editing; **Nowacki M:** Funding Acquisition, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Swiss National Science Foundation (31003A_129957, 31003A_146257 and 31003A_166407 and NCCR RNA & Disease to MN); the European Research Council (EPIGENOME and G-EDIT to MN); the European Cooperation in Science and Technology (Action BM1102 "Ciliates as model systems to study genome evolution, mechanisms of non-Mendelian inheritance, and their roles in environmental adaptation" to MN). CDW was supported by a Ph.D. fellowship from the French Ministère de l'Enseignement Supérieur et de la Recherche and the French National Research Agency ("Inferno").

Copyright: © 2017 Swart EC *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Swart EC, Denby Wilkes C, Sandoval PY *et al.* **Identification and analysis of functional associations among natural eukaryotic genome editing components [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2017, 6:1374 <https://doi.org/10.12688/f1000research.12121.1>

First published: 08 Aug 2017, 6:1374 <https://doi.org/10.12688/f1000research.12121.1>

Introduction

Paramecium, like other ciliates, contains both a large, somatic, polyploid macronucleus (MAC), and a relatively tiny, gametic, diploid micronucleus (MIC), within the same cell. During sexual or asexual “reproduction” (conjugation or autogamy), a copy of *Paramecium*’s micronucleus develops into a new macronucleus, and at the same time the micronuclear genome it contains is transformed into a new macronuclear genome by extensive, targeted DNA deletion. Concomitant with DNA deletion, the developing MAC genome is amplified from 2N to ~800N. In *Paramecium*, the best studied germline DNA sequences destined to be deleted, known as internally eliminated sequences (IESs), are interspersed among the macronuclear-destined sequences (MDSs) (Figure 1). Characterized *Paramecium* IESs are typically short (> 90% are < 150 bp), TA-flanked, and unique (~45,000 IESs per *Paramecium tetraurelia* MIC genome¹). These IESs are recognized and generally precisely removed by the PiggyMac transposase complex². For some IESs, development-specific small RNAs (sRNAs) are required for the recognition and efficient removal of all their copies³.

Paramecium’s current annotated IESs comprise just a small fraction (~13%) of all the micronuclear-specific DNA that is eliminated during development¹. Though much of the remaining eliminated DNA has been assembled during IES annotation¹, it has yet to be assembled as a complete micronuclear genome, and comparatively little is known about it. For the sake of explanation, and because some of this DNA may reside at chromosome ends, and hence may not be internally eliminated, we classify this DNA as “other eliminated sequences” (OESs; see Figure 1).

Developmental elimination of the few OESs investigated in detail^{4–6} is associated with two possible outcomes: either breakage of the MIC chromosomes into smaller MAC chromosomes, accompanied by end capping with telomeres, or fusion of the regions flanking the removed region^{5–7}. The lengths of studied OESs are on the order of kilobases to tens of kilobases^{4–6}. The boundaries of MAC chromosome telomere addition sites occur at similar locations to the boundaries of the internally removed sequences⁶. In contrast to classical IESs, the cut sites of these internally removed sequences are imprecise⁶. During chromosome breakage, the exact locations

of the telomere addition sites vary over a span of at least hundreds of base pairs, but whether this is due to resectioning of DNA before telomere addition or due to initial imprecise DNA cleavage, or both, is unknown⁶.

An important question has therefore been whether the mechanisms underlying imprecise DNA elimination are distinct from those employed in precise IES excision, as previously proposed⁶, or whether some of the same developmental DNA deletion pathway is shared, as previously proposed⁴. Suggesting the latter may be true, substantial retention of OESs occurs after silencing the gene encoding PiggyMac (PGM)¹, and OESs and transposons embedded within them are typically retained when gene silencing leads to IES retention^{3,8–12}.

In ciliates, small RNAs (> 21 nt and < 32 nt) involved in MAC genome development have proposed roles either in DNA targeting and excision (as in *Tetrahymena* and *Paramecium*^{13–16}) or in marking/protecting DNA for retention (as in *Oxytricha*^{17,18}). *Paramecium tetraurelia* has two classes of sRNAs involved in DNA elimination: scnRNAs and iesRNAs^{3,13}. scnRNAs are produced in gametic MICs during micronuclear meiosis, early during sexual development. Newly formed scnRNAs are transported to the existing, old MAC where MDS-matching scnRNAs are removed (i.e. removed by “RNA scanning”, as originally proposed in the ciliate *Tetrahymena*¹⁹), after which the non-MDS-matching scnRNAs (including IES-matching scnRNAs) are transported to the developing new MAC where they target DNA deletion^{3,13}. *Paramecium*’s scnRNAs have been proposed to bind to longer RNA transcripts, rather than directly to DNA, in both the old and new MAC^{12,20}. iesRNAs are produced in the developing MAC, where, in principle, they can immediately target unexcised IESs, and their production peaks later during development³. iesRNAs are hypothesized to be produced from transcripts of excised IESs³.

scnRNAs and RNA scanning were discovered in the ciliate *Tetrahymena* (belonging to class Oligohymenophorea, like *Paramecium*)^{14,16,21}. Other than the use of scnRNAs to target excision, *Tetrahymena*’s macronuclear genome development resembles that of *Paramecium* in a number of other key ways,

Figure 1. *Paramecium* genome development schematic. A representative micronuclear chromosome is transformed into macronuclear chromosomes via precise excision of internally eliminated sequences (IESs) and imprecise elimination of other eliminated sequences (OESs). Macronuclear-destined sequences (MDSs) inbetween are joined and capped with new telomeres to give rise to the final macronuclear chromosomes. When DNA is imprecisely eliminated, the telomere-capped ends of the macronuclear chromosomes occur at different sites. In some cases, this process also produces alternative macronuclear chromosome forms. Note that the distinction between OESs and IESs, though useful for the purposes of description and analysis here, is an artificial one, and that there may only be one main class of eliminated DNA in *Paramecium* (see later in Discussion).

most notably in employing a domesticated PiggyBac in DNA elimination²², and that both species have two phases of development-specific sRNA production¹⁶. There are also important differences between *Paramecium* and *Tetrahymena* genome development, in particular, unlike *Paramecium*'s classical short, precisely excised, single copy IESs, *Tetrahymena*'s conventional IESs tend to be longer, imprecisely excised, multicopy elements^{23,24}. In contrast to *Paramecium*, which produces scnRNAs and iesRNAs with distinct Dicer-like proteins (Dcl2/3 for scnRNAs¹³; Dcl5 for iesRNAs³), *Tetrahymena* produces two types of scnRNAs (early-scnRNAs and late-scnRNAs¹⁶) with a single Dicer-like protein (Dcl1)^{25,26}. These scnRNAs are differentiated by the Piwi proteins (Twi1 and Twi11) processing Dcl1-generated dsRNA precursors¹⁶. At present, it is not known how the development-specific sRNAs of *Paramecium* and *Tetrahymena* guide the domesticated PiggyBacs to their DNA targets, nor what role proteins implicated in their genome editing have in mediating the intervening interactions.

In the past, the effect on the excision of *Paramecium* IESs following microinjection of IES DNA copies into the old MAC during autogamy was investigated for a small set of IESs. IESs were classified either as maternally controlled (i.e. governed by the old or maternal MAC) when the DNA injection led to IES retention in the new MAC, and non-maternally controlled when there was no observable IES retention^{27,28}. In principle, since *DCL2/3* co-silencing stops transmission of epigenetic information from the old macronucleus to the new one by eliminating the scnRNAs bearing this information, we previously classified IESs affected by *DCL2/3* silencing as epigenetically controlled, and those that were not, as non-epigenetically controlled^{3,15}. This classification is a simplification, since there is no clear boundary between epigenetically and non-epigenetically controlled IESs (and between maternally and non-maternally controlled IESs), and because it may be difficult to detect low levels of epigenetic control. After *DCL2/3* silencing, some non-maternally controlled IESs are weakly retained, below the detection level of conventional PCR analyses³, and hence are weakly epigenetically controlled. IESs affected by *DCL5* silencing do not correlate well with those affected by *DCL2/3* silencing, suggesting that a substantial amount of *Paramecium*'s sRNA-dependent IES excision is not epigenetically controlled^{3,15}.

In the last few years, additional proteins influencing IES removal in *Paramecium* have been reported, including homologs of proteins with known functions in other eukaryotes, such as those involved in histone modification (PtCAF1, Ezl1) and RNA transcription (TFIIS4)¹⁰⁻¹² (see also Table 1 and Table 2). From these studies, following existing models in *Tetrahymena*^{29,30}, this has led to the proposal that scnRNAs in *Paramecium* direct certain chromatin modifications that may facilitate DNA excision^{10,11}. Proteins involved in IES removal with no readily detectable homologs in other eukaryotes have also been investigated⁸. Other than these proteins, two closely related protein paralogs, Nowa1 and Nowa2 (Nowa1/2), were previously implicated in IES elimination and proposed to be involved in scnRNA-mediated, trans-nuclear crosstalk³¹. Identification and establishment of the relationships between all these proteins and the development-specific sRNAs is necessary to determine how genome editing functions in *Paramecium*, and the main objective of the research presented here.

In this paper, highlighting the utility of DNA retention analyses in discovering associations between genome editing proteins, we present what we have learnt about the role of Nowa1/2 proteins and their relation to other *Paramecium* genome editing components. We report strong correlations between the IESs affected by *NOWA1/2* co-silencing (*NOWA1/2*-KD; KD=knockdown), *DCL2/3/5* triple silencing (*DCL2/3/5*-KD) and *TFIIS4* silencing (*TFIIS4*-KD), suggesting that Dcl2/3/5, Nowa1/2 and TFIIS4 are all components of an sRNA-guided DNA excision subsystem. Strong correlations also exist between *PTCAF1*-KD and *EZL1*-KD IES retention, whereas the correlations between the IES retention of these gene knockdowns and those of RNA-related genes are somewhat weaker. We also observe more IES retention due to the silencing of chromatin modifying genes (*PTCAF1* and *EZL1*) than due to depletion of most IES-matching sRNAs by *DCL2/3/5*-KD, suggesting that chromatin modification due to PtCaf1 and Ezl1 also influences genome editing in an sRNA-independent manner. The correlations in OES retention following gene knockdown suggest most OESs are effectively the longest IESs.

Methods

Paramecium strains, cultivation, and autogamy

All experiments were carried out with *Paramecium tetraurelia* stock 51, which may be obtained from the ATCC as ATCC30567. Cells were grown in a wheat grass powder (WGP; Pines International, Lawrence, KS) infusion medium bacterized the previous day with *Klebsiella pneumoniae* supplemented with 0.8 mg/l β -sitosterol. Cultivation and autogamy were carried out at 27°C as previously described²⁸.

Gene silencing

Gene silencing (knockdown) was carried out by RNA interference by feeding of *Paramecium* cells with *E. coli* expressing double-stranded RNA from a portion of the target gene cloned into the L4440 vector, as previously described^{3,8,10,31}. Sequence regions used for each of the genes silenced in *DCL2/3/5*-KD were as previously described^{3,13}. Silencing of *PTCAF1* and *ND7* were as previously described^{10,31}.

Three knockdowns of *NOWA1/2* were performed; where necessary we distinguish between these knockdowns (i.e. *NOWA1/2_a*-KD, *NOWA1/2_b*-KD, *NOWA1/2_c*-KD). Macronuclear DNA from postautogamous cells from the first and third knockdowns was used to produce Illumina paired-end sequence libraries. Illumina sRNA libraries were produced from *NOWA1/2_b*-KD at the same time as those produced for sRNAs from *DCL2/3*-KD, *DCL5*-KD and control cells - the latter knockdowns were described in "Experiment 1" of 3.

The following controls were used in this manuscript:

Control 1 = feeding with an empty vector (EV) construct. This was the control for *DCL2/3/5*-KD and *PDSG2*-KD;

Control 2 = *ND7*-KD. This was the control for *PGM*-KD¹,

Control 3 = *ND7*-KD. This was the control for an unpublished experiment (Singh *et al.*, in preparation);

Control 4 = *ND7*-KD. This was the control for *EZL1*-KD¹¹.

ND7 is a trichocyst discharge gene which is not expected to affect IES retention³².

Table 1. Gene knockdown effects during development. Effects of knockdowns as described in this and existing studies. See Results for a description of the functional group assignment.

Knockdown	IES excision affected	Long IESs strongly affected	Transposon excision affected	OES elimination affected	Effect on scnRNAs	Effect on iesRNAs	Effect upon H3K27me3 and H3K9me3	Functional group
<i>PGM</i>	All known ¹	Yes	Yes ¹	Yes	Reduction ³	Reduction ³	None ¹¹	DNA excision
<i>DCL2/3</i>	scnRNA- and indirectly, some iesRNA-dependent ³	Yes ³	Yes ³	Yes ¹²	Elimination ³	Reduction ³	Reduction in old and new MAC ¹¹	RNA-associated
<i>DCL5</i>	iesRNA-dependent ³	No ³	Very little ³	Very little	None ³	Elimination ³	None ¹¹	RNA-associated
<i>DCL2/3/5</i>	scnRNA- and iesRNA-dependent	Yes	Yes	Yes	Elimination	Elimination	Not determined	RNA-associated
<i>NOWA1/2</i>	scnRNA- and iesRNA-dependent	Yes	Yes ³¹	Yes	Nonremoval of MAC genome-matching scnRNAs	Reduction	Reduction in old and new MAC ³	RNA-associated
<i>TFIIS4</i>	scnRNA- and iesRNA-dependent ¹²	Yes ¹²	Yes ¹²	Yes ¹²	Little reduction, if any ¹²	Elimination ¹²	Not determined ¹²	RNA-associated
<i>EZL1</i>	scnRNA-, iesRNA-dependent and other ¹¹	Yes ¹¹	Yes ¹¹	Yes ¹¹	Not determined ¹¹	Not determined ¹¹	Reduction in old and new MAC ¹¹	Histone modification
<i>PTCAF1</i>	scnRNA-, iesRNA-dependent and other ¹⁰	Yes	Yes ¹⁰	Yes	Nonremoval of MAC genome-matching scnRNAs ¹⁰	Reduction ¹⁰	Reduction in old and new MAC ¹⁰	Histone modification
<i>PDG2</i>	scnRNA- and iesRNA-dependent	Yes	Yes ⁸	Yes	Little reduction, if any ⁸	Elimination ⁸	Increase in new MAC ⁸	

Table 2. Properties of proteins implicated in *Paramecium* genome editing. Proteins are those considered in this study. Timing of expression during development is gauged from microarray data³³. Localization is that determined from GFP fusion proteins. Putative function indicates the clearest known function.

Protein	Timing of expression during development	Localization	Putative function
Pgm ^{1,2}	Early ²	New MAC ²	DNA elimination
Dcl2/Dcl3 ¹³	Early ¹³	MIC ¹³	scnRNA production
Dcl5 ³	Late ³	New MAC ³	iesRNA production
Nowa1/2 ³¹	Early to Late ³¹	Old and new MAC ³¹	Linking Piwi-bound sRNAs to RNA transcripts
TFIIS4 ¹²	Late ¹²	New MAC ¹²	Transcription of IESs in developing MAC
Ezl1 ¹¹	Early ¹¹	MIC, old and new MAC ¹¹	Trimethylation of histone H3K9 and H3K27
PtCAF1 ¹⁰	Early ¹⁰	Old and new MAC ¹⁰	Histone H4 binding; Trimethylation of histone H3K9 and H3K27
Pdsg2 ⁸	Late ⁸	New MAC ⁸	Unknown

Survival tests and cytological staging

Tests of survival of sexual progeny cells were performed by isolating and refeeding 30 individual cells. *NOWA1/2* knockdown is generally lethal, i.e., typically all cells die within at most the time taken for 3 divisions of control cells following autogamy; for *DCL2/3/5-KD* 90% of cells were dead 3 days after autogamy; for *PDSG2-KD* after 3 days 50% of cells were dead and 47% appeared to be “sick” (i.e. dividing slower than usual); for *PTCAF1-KD* after 3 days 80% of cells were dead and 17% were sick.

Morphological characteristics of MAC development were used to assign the developmental stages of the control and knockdown samples: “Early”, “Middle” and “Late”, correspond to 50% of cells with fragmented MACs; 100% of cells with fragmented MACs; and 100% cells with fragmented MACs and a new MAC; [Figure S3A,C](#)). Note that because *Paramecium* cells are not completely synchronous during induction of autogamy³³ and because the underlying assessments are somewhat subjective, due caution should be exercised when comparing developmentally staged analyses between gene knockdowns.

Northern blotting

To assess the effectiveness of *DCL2/3/5-KD*, northern blotting was performed as in [3](#). The *DCL2* probe was from gene position 857-1810 (ParameciumDB: GSPATG00003051001); the *DCL3* probe was from gene position 518-1354 (ParameciumDB: GSPATG00027456001); the *DCL5* probe was from gene position 2074-2461 (ParameciumDB: GSPATG00012932001). For *PTCAF1-KD* and control (*ND7-KD*) cells, northern blotting was performed as in^{10,31}.

Nowa1 antibody production and use in western blotting

A portion of Nowa1's C-terminal domain (amino acids 744-1024) fused to an N-terminal Strep-Tag was expressed in *E. coli* BL21-CodonPlus(DE3)-RIL cells (formerly Stratagene, currently Agilent, USA). The protein was column purified, followed by removal of the tag by HRV 3C protease cleavage. Immunization of a rabbit with this protein was performed by Eurogentec, Belgium, following their 28-day polyclonal antibody protocol, then affinity purified using the matrix bound C-terminal domain portion of Nowa1, before performing western blotting using a 1:500 antibody dilution ([Figure S3B](#)). As a loading control, western blotting was performed using 1:5000 dilution of monoclonal anti- α -tubulin (Sigma T8203 lot 111M4828) and 2% BSA for blocking.

Small RNA electrophoresis

For small RNA electrophoretic analysis, total RNA samples (5.0 μ g) were 5'-end-labelled with [γ -³²P] ATP (5000 Ci/mmol, Amersham) by the exchange reaction of T4 polynucleotide kinase (Fermentas), then denatured and run on a 15% polyacrylamide, 7 M urea gel.

Electrophoretic mobility shift assay (EMSA)

Synthetic Nowa1 codon adjusted for *E. coli* translation (GENART synthesis) was cloned into the pMAL-c2x vector and expressed in Rosetta2(DE3)p-LysS (formerly Novagen, currently Merck) cells. Protein translation was performed at 37°C; cells were supplied with 1 mM IPTG and 0.2% glucose. Protein was purified according to the pMAL™ Protein Fusion and Purification System (NEB e8200s). EMSA was performed with the LightShift® Chemiluminescent RNA EMSA Kit (Thermo scientific, 201558) using 6.25 μ M of RNA or DNA, 4 μ g of MBP, and 12 μ g of MBP-Nowa1.

DNA and RNA extraction, library preparation and sequencing

Macronuclear DNA was isolated from post-autogamous cells as previously described¹. For PCR analyses of IES retention, DNA was extracted from approximately 800 cells with a Nucleospin Tissue kit (Macherey-Nagel).

Total RNA was extracted from 400 ml of culture per the TRI Reagent BD protocol (Sigma) and was resuspended in RNase free water. A MirVana miRNA isolation kit (Ambion) was used for sRNA enrichment.

MAC genomic DNA and small RNA libraries were produced and sequenced per standard Illumina protocols. Paired-end Illumina DNA-seq libraries were constructed and sequenced on either the Illumina GAIIX, HiSeq 2000 or HiSeq 3000 sequencers.

Sequence data corresponding to the MAC DNA from *NOWA1/2_a-KD* can be obtained from the GenBank Sequence Read Archive (SRA) under the accession [SRR646462](#). MAC DNA sequence data for *NOWA1/2_c-KD*, *DCL2/3/5-KD*, *PTCAF1-KD*, *PDSG2-KD*, *ND7-KD* and control 1 (empty vector) is available from the European Nucleotide Archive, respectively under the accession numbers [ERS1033672](#), [ERS1033670](#), [ERS1033674](#), [ERS1033673](#), [ERS1033676](#), [ERS1033671](#). sRNA sequence data for *NOWA1/2_b-KD* and empty vector control can be obtained from the European Nucleotide Archive accessions [ERS1455004](#), [ERS1455005](#), [ERS1455006](#), [ERS1455007](#), [ERS1455008](#), [ERS1455009](#), and for *DCL2/3/5-KD*, from [ERS1455358](#), [ERS1455359](#), [ERS1455360](#).

PCR analysis of IES retention

PCR conditions and primers used to test IES retention following *NOWA1/2_a-KD* were the same as those in [3](#).

Reference genomes and IES retention scores

The following reference genomes for *Paramecium tetraurelia* stock 51, obtained from ParameciumDB³⁴, were used in the IES analyses and read mapping:

MAC: http://paramecium.cgm.cnrs-gif.fr/download/fasta/assemblies/ptetraurelia_mac_51.fa

MAC+IES: http://paramecium.cgm.cnrs-gif.fr/download/fasta/assemblies/ptetraurelia_mac_51_with_ies.fa (md5 checksums: dbcc54fb2987c8f60f8e765db7ed274c and 3e5b3fa65ebfaa484566a1ffddf20239, respectively).

IES retention scores (IRSs) were determined by ParTIES³⁵. For each IES, ParTIES counts mapped reads with unexcised IESs (IES+) and excised IESs (IES-) to determine an IES retention score, $IRS = IES^+ \div (IES^+ + IES^-)$. A table of all the IES retention scores used in this manuscript is included as **Dataset 1** (see Data and software availability). IES retention scores from previous knockdown experiments, i.e. *PGM-KD*¹, *TFIIS4-KD*¹², *EZL1-KD* (*EZL1-2* silencing fragment)¹¹, *DCL2/3-KD*³ and *DCL5-KD*³, were downloaded from ParametriumDB³⁴. Unless otherwise indicated, *NOWA1/2-KD* IRSs are for *NOWA1/2_a-KD*.

IES and OES retention score regressions

To generate the IES and OES score regression matrices, shown in **Figures 3A–C**, **Figure 4E** and **Figures S1A–B**, we provide the program “after_ParTIES.py” (see Data and software availability), which can be run on **Dataset 1** and **Dataset 2** (IES and OES retention scores, respectively). This was tested using Python3.6 and relevant modules for scientific computing from the Anaconda software distribution from Continuum Analytics (version: Anaconda3-4.3.1).

Three regressions are graphically shown for IRS pairs: an ordinary least squares (OLS) regression, a LOWESS regression (non-parametric) and an orthogonal distance regression (ODR) with a linear function initialized by the slope and intercept from the OLS. OLS and ODR used standard SciPy functions. LOWESS regressions used the Python statsmodels library. ODR was performed since errors in both the “dependent” and “independent” regression variables are expected (both IRSs), not just the dependent variable as assumed by OLS. The combined effect of background IES retention and systematic IES retention errors on the regressions, and the effect of subtracting the mean of multiple control IES retention scores is shown in **Figure S1A**. For each pair of knockdown IES retention scores, Spearman’s rank correlation coefficient (r_s) was calculated instead of Pearson’s correlation coefficient to mitigate the effect of outliers. Reproducibility of IES retention scores between experiments is demonstrated by the strongly correlated scores of different *NOWA1/2* silencing experiments (**Figure S1B**).

Two-sided p-values, calculated with SciPy’s `spearmanr` function, were used to test the null hypothesis of no correlation (for $p \leq \alpha/m \leq 0.01/45$; i.e. with a Bonferroni correction for multiple hypothesis testing; uncorrected p-values are given in **Supplementary Data S1–4**), which was rejected in all the displayed correlations in the main figures, except where indicated by a circumflex accent. To facilitate assessment of the statistical significance of differences between r_s values, tables containing the two-tailed p-values are provided in **Supplementary Data S5–7**. p-values were calculated using methods from the R statistical package `cocor` version 1.1-3³⁶ that employ Fisher’s r-to-Z transformation (which are also applicable to r_s ³⁷). Appropriate methods for calculating test statistics for correlations with independent groups or dependent groups with either overlapping or non-overlapping variables were used from `cocor`

(noted in the supplementary data files). All the comparisons of r_s differences described in the text are statistically significant at $\alpha=0.01$ with a Bonferroni correction for multiple hypothesis testing.

Assembly and selection of OESs

Genomic regions that are not part of the current reference *Paramecium* MAC+IES assembly were previously assembled¹, but we decided to use an alternative, metagenome assembler to attempt to produce a better assembly, and to minimize bacterial contamination. Reads from *PGM-KD*¹ not mapping to the current *Paramecium tetraurelia* strain 51 MAC genome and 51 MAC genome with IESs (bwa³⁸ version 0.7.12-r1039; default parameters) were assembled with IDBA-UD³⁹ version 1.0.9 (default parameters). To remove bacterial contaminants assembled during the assembly, only scaffolds with GC < 40% were selected (bacterial scaffolds peak around 58% GC; **Figure S5A**). This procedure resulted in a ~20 Mb assembly consisting of 7266 scaffolds. As a small fraction of the *P. tetraurelia* MAC genome is currently missing from the reference assembly, this fraction was present in the ~20 Mb assembly. To exclude the missing MAC genome regions, empty vector control (control 1) reads were mapped to the 7266 scaffolds with HISAT2⁴⁰ version 2.0.4 (parameters: “-X 600 --pen-noncansplice 0 --mp 24,8 --max-intronlen 1000”), and then, using a custom Python script, only those regions with little or no sequence coverage ($\leq 2.0\times$; compared to a mean MAC coverage of $\sim 100\times$) were extracted (**Figure S5C** shows an illustrative scaffold from which OESs were extracted). Coverage was smoothed using a bi-directional average of two exponentially weighted moving averages (EWMA) with a 200 bp span. This procedure extracted 4799 sequence regions (12.3 Mb).

To minimize the effects of potential genome assembly chaff (short assembled regions, esp. those that might arise from sequence errors), analyses were continued with only the low control sequence coverage regions extracted from scaffolds longer than 300 bp and with a *PGM-KD* sequence coverage of between $3\times$ and $70\times$ (**Figure S5B**). These 3882 regions, obtained from 5236 scaffolds, corresponding to 11.8 Mb of sequence data, are classified as putative OESs (**Supplementary Data S8**). Note that OES boundaries are a first approximation, since imprecise DNA elimination generates variable boundaries (as illustrated in **Figure 1** and shown in **Figure S5C**), and many OESs situated at the ends of the scaffolds are likely incomplete due to the limitations of genome assembly (esp. since the assembly may end when untraversable repetitive regions are encountered).

Estimation of OES retention

For each knockdown, to quantify the effects of gene knockdown on OES retention, reads were mapped to the OESs (identified in the previous methods section) with HISAT2⁴⁰ version 2.0.4 (parameters: “-X 600 --pen-noncansplice 0 --mp 24,8 --max-intronlen 1000”) before determining coverage, and normalized to the median coverage of the knockdown reads mapped with HISAT2 to the *Paramecium tetraurelia* strain 51 MAC genome plus IES assembly, i.e. $OES\ retention\ score = OES_{coverage} / median(MAC-with-IES-genome_{coverage})$. OES retention scores may not be as accurate as IES retention scores, both because their denominators use a measure of global, rather than local, sequence coverage and because

only reads that map within OES boundaries are counted. In the latter case, read mapping may fail to detect short matches to an OES when only a short part of an OES/MDS-derived read matches, and so the shorter the OES, the greater the underestimation of its retention. Shorter OESs are more weakly retained in most knockdowns affecting genome reorganization (Figure S5D) in *Paramecium*, but at least some of this effect may be due to boundary-related underestimation of retention. In future, when a complete *Paramecium* MIC genome is assembled, the relationship between OES length and retention can be fully assessed. OES retention estimates are provided in Dataset 2. Control OES retention was subtracted from the knockdown OES retention for each OES used in the regression analyses.

sRNA mapping

sRNAs were mapped and histograms of their genomic distribution were produced as previously³ with the addition of OESs.

Sequence logos

Sequence logos were created with WebLogo⁴¹, version 3.3, with redundant sRNAs and base frequencies A=U=0.4 and C=G=0.1.

Results

Examination of basic IES retention properties

Previously, each newly silenced *Paramecium* gene involved in IES excision appeared to generate a different distribution of IES retention^{3,11,12}. From additional knockdowns we have subsequently performed, it can now be seen that the IES retention distributions for *NOWA1/2* knockdown, *TFIIS4* knockdown and *DCL2/3/5* triple knockdown are very similar (Figure 2A; see Methods for the IES retention score calculation). Given the role of *TFIIS4* and *Dcl2/3/5* in producing and processing of RNAs, and the suggested role of *Nowa1* in RNA interactions (see “*Nowa1/2* is involved in RNA scanning, may mediate scnRNA- and iesRNA-targeted DNA cutting and excision” later in the results), we classify these proteins as “RNA-associated” proteins involved in *Paramecium* MAC genome development. The IES retention score distributions of the *EZL1* knockdown and *PTCAF1* knockdown also have similar forms, though the effect of the former knockdown is slightly stronger (Figure 2B). Since both *Ez1* and *PtCAF1* affect histone modifications^{10,11}, we classify these proteins as those involved in “histone modification”.

As we previously reported, the effects of either *DCL2/3* knockdown or *DCL5* knockdown on IES retention alone are quite modest^{3,15}. Even when their IES retention scores are summed, it is apparent that the knockdowns of *DCL2/3* and *DCL5* influence considerably fewer IESs, and to a lesser degree than the *PGM* knockdown. For *DCL2/3/5*, *NOWA1/2* and *TFIIS4*, given that the gene knockdowns are efficient (as judged by depletion of their mRNAs or proteins; Figure S3B,D and¹²), we think the similarity of their IES retention score distributions indicates that the upper bounds of IES retention for these knockdowns have been approached. Consequently, these knockdowns may indicate most of the *Paramecium* IESs requiring scnRNAs and iesRNAs for their excision.

Though an interesting idea, we find no evidence of a proposed missing class of sRNAs^{11,12} responsible for targeting and excision

of additional IESs. Instead, we observe that after *DCL2/3/5*-KD the levels of IES-matching sRNAs are negligible (Figure 5C). We therefore infer that most IES excision in *Paramecium* does not require IES-targeting sRNAs. This is consistent with past studies reporting that most tested IESs were non-maternally controlled^{27,28}, i.e. that scnRNAs alone are insufficient for most IES excision^{3,15}. On the other hand, from IES retention following *DCL2/3/5*-KD we infer that many IESs require some scnRNAs or iesRNAs to cleanly remove all their copies (the exact number of affected IESs depending on the chosen retention score threshold, e.g. 19043 IESs are affected by *DCL2/3/5*-KD above a threshold of 0.1 retention).

An important observation from the *DCL2/3/5* knockdown is the considerably stronger IES retention compared to the separate *DCL2/3* and *DCL5* knockdowns (Figure 2C); for example, at an IES retention score > 0.1, for *DCL2/3/5*-KD 19043 IESs are affected vs. 4815 IESs for *DCL2/3*-KD, and 5384 IESs for *DCL5*-KD. This suggests some co-operation between scnRNAs and iesRNAs in IES excision. From the *DCL2/3/5* triple silencing, it is clear that multiple gene co-silencing may be necessary to gauge the full contributions of individual genes to IES excision. This is a critical consideration in gene function investigations in *Paramecium tetraurelia* with its abundance of paralogs that arose from successive whole genome duplications⁴².

As we previously proposed, IES excision/recognition efficiency and the need for particular genome reorganization proteins vary as a function of IES end bases and lengths¹⁵. A trend of increasing IES retention with IES length was shown for the knockdowns of *DCL2/3*, *TFIIS4* and *EZL1*^{3,11,12,15}. It can now be seen that, with the exceptions of *PGM*-KD and *DCL5*-KD, most knockdowns leading to IES retention show this trend (Figure 2E,F). A new observation with respect to IES length is that IES retention for the knockdowns of *DCL2/3/5*, *TFIIS4* and *NOWA1/2* also has an obvious 10-11 base periodicity for ~45-130 bp IESs (Figure 2E), mirroring that of the IES length peaks (which are proposed to reflect the constraints of DNA twist on transposase excision¹). This signal is not apparent in knockdowns with weaker effects (*DCL2/3*-KD, *DCL5*-KD and *PDSG2*-KD), and only weakly visible in the knockdowns with stronger effects (*PTCAF1* and *EZL1*). We previously also observed that the terminal IES base frequencies relative to IES length show a periodic, 10-11 bp signal (Figure 2A of 15). As in knockdowns of *DCL2/3*, *DCL5* and *PGM*¹⁵, for all the new knockdowns we examined, the end base frequencies of longer IESs (e.g. 45-55 bp) vary in a similar manner with respect to their IES retention scores (Figure 2G-H; Figure S2), consistent with the idea that some IESs have a greater requirement for the products of these genes than others¹⁵.

Relationships among genes involved in genome editing inferred from IES retention

While similar IES retention distributions suggest different gene knockdowns might affect retention of particular IESs in a similar manner, examination of pairwise correlations is necessary to clearly establish this. We therefore determined all the possible pairwise correlations between IES retention scores of different gene knockdowns (Figure 3). From this figure, it can be seen that the *DCL2/3/5*, *NOWA1/2*, *TFIIS4* (group 1), and histone modification

Figure 2. IES retention scores and their relation to basic IES properties. (A–D) For the sake of clarity, histogram bars with more than 10000 IESs have been truncated. The number of IESs binned in the left-most bars are as follows: (A) *NOWA1/2-KD* - 17468, *TFIIS4-KD* - 14152, *DCL2/3/5-KD* 18915 (Figure S3B and S3D provide Western and Northern blots, for *NOWA1/2-KD* and *DCL2/3/5-KD*, respectively); (B) *PTCAF1-KD* - 12614, *EZL1-KD* - 8047; (C) *DCL2/3-KD* - 37410, *DCL5-KD* - 35019, *DCL2/3/5-KD* - 22136; (D) Control - 43294; *PDSG2-KD* - 20730. Note that, as judged by survival tests in comparison to those previously published 8, *PDSG2-KD* may have been of reduced efficiency. (E, F) IES retention scores (scales on right y-axes) vs. IES length over shorter (E) (≤ 200 bp) and longer (F) (≤ 1000 bp) IES length scales. The IES length distribution is shown in grey in the background (scale on left y-axes). Lines are the exponentially-weighted mean averages (EWMA) with spans of 5 (E) or 50 (F) bp. (G–H) Mean base frequencies of positions 1–3 after the TA repeat relative are plotted relative to IES retention score for IESs from the third (45–55 bp) length peak (only the highest frequency bases are shown — see (H) for the base color scheme; similar trends are visible for ~ 10 bp windows surrounding longer IES length peaks). IES length is restricted to control for IES end base frequencies variation associated with large scale (26–1000 bp) IES length variation. EWMA lines for spans of 10 data points (intervals of 0.01) are plotted for data within 2 standard deviations (dotted vertical lines) of the mean retention score (dashed vertical line). Mean IES lengths (bp; counting only one of the two TA pointers) vs. IES retention score are plotted in the bottom subgraphs (navy diamonds). See Figure S2 for end base frequency graphs of *EZL1-KD* and *TFIIS4-KD* and for end base frequency graphs of shorter IESs (26–36 bp). Equivalent graphs for *DCL2/3-KD*, *DCL5-KD* and *PGM-KD* can be seen in 15.

Figure 3. IES retention score correlations and regressions. (A) For the purposes of regression, for each IES, negative control IES retention scores have been subtracted and zeroed when the subtraction yields a negative value, i.e. $IRS_{\text{regression}} = \max(IRS_{\text{experiment}} - \text{mean}(IRS_{\text{controls}}), 0)$ (controls=controls 1-4). This calculation was performed to reduce the effects of natural, background IES retention and systematic errors in IES retention score estimation as much as possible (see Figure S1, which examines these effects on regression). Hexagonal binning of IES retention scores was used to generate the plots in the lower triangular matrix. Red lines are for OLS regression, orange lines for LOWESS, and grey lines for ODR. IES retention score distributions given along the diagonal are shown to a maximum of 4000 IESs except for *PGM*-KD (maximum 8500 IESs). For each plot in the lower triangular matrix (M_{ij}), Spearman's rank correlation coefficient (r_s) is given in the corresponding position diagonally opposite in the upper triangular matrix (M_{ji}). Note $r_s=0.91$ for biological replicates of *NOWA1/2*-KD (Figure S1B). (B) as in (A) but for short IESs (≤ 35 bp). IES retention score distributions given along the diagonal are shown to a maximum of 1500 IESs except for *PGM*-KD (maximum 3200 IESs). (C) as in (A) but for long IESs (≥ 500 bp). IES retention score distributions given along the diagonal are shown to a maximum of 50 IESs except for *PGM*-KD (maximum 75 IESs).

Figure 4. OES retention and OES retention correlations. (A–D) For the sake of clarity, histogram bars with more than 2000 OESs have been truncated. The number of OESs binned in the left-most bars are as follows: (A) *NOWA1/2*-KD - 1238, *TFIIIS4*-KD - 1230, *DCL2/3/5*-KD 2812; (C) *DCL2/3*-KD - 473, *DCL5*-KD - 3788, *DCL2/3/5*-KD - 2812; (D) Control - 3882; *PDSG2*-KD - 3062; *PGM*-KD - 709. (E) OES retention score correlations and regressions. For the purposes of regression, for each IES the mean of two negative control OES retention scores (control 1 and control 2) have been subtracted, and zeroed when the subtraction yields a negative value, i.e. $ORS_{\text{regression}} = \max(ORS_{\text{experiment}} - \text{mean}(ORS_{\text{controls}}), 0)$. This calculation was performed to reduce the effects of natural, background OES retention and systematic errors in OES retention score estimation as much as possible (but note that when OES retention scores are very low, some residual background retention may lead to some correlation, e.g. between *DCL5*-KD and control 3). Hexagonal binning of OES retention scores was used to generate the plots in the lower triangular matrix. Red lines are for OLS regression; orange lines for LOWESS and grey lines for ODR. For each plot in the lower triangular matrix (M_{ij}), Spearman's rank correlation coefficient (r_s) is given in the corresponding position diagonally opposite in the upper triangular matrix (M_{ji}).

Figure 5. Analysis of the effects and properties of NOWA1/2-KD. (A) Length distribution of small RNAs from control cells and NOWA1/2-KD cells (see Methods for details of histogram construction and normalization). Vector=vector backbone only. Figure 2SA shows the cytological characterization of the cells these sRNAs were isolated from. (B) Radioactively end-labelled total RNA from NOWA1/2-KD cells and control cells (cells fed with *E. coli* producing dsRNAs corresponding to the plasmid L4440 with no sequence target in the *Paramecium* genome; the gel portion showing the control sRNAs was previously shown in Figure 2A of 3 as "Control 1; Experiment 1"). (C) sRNA length distribution for DCL2/3/5-KD; the peak at 33 nt corresponds to an unknown sRNA family (e.g. 5'-GGAUCUAUCGUAUAGUGGUAGUACCUGAGGCU-3') that matches a handful of locations in the MAC genome, which we have also observed in other controls and knockdowns (e.g. from 8). See Figure 2SE for cytological characterization. (D) To examine the relationship between maternal control and IES retention scores (values determined in Table S1), maternal control scores were obtained from the maximal IES retention observed in Figure 6 of 28. Linear regression lines are shown along with Pearson's r and the two-tailed p -values for hypothesis testing with a null hypothesis that the regression slope is zero. (E) Retention of IESs tested by PCR for control and NOWA1/2-KD cells. NOWA1/2-KD IES retention scores are given to the right. Additional bands indicate the presence of heterodimers of ssDNAs with excised and unexcised IESs. (F) Electrophoretic mobility shift assay with MBP (maltose-binding protein), MBP-Nowa1 (alternating lanes) and oligonucleotides, visualized on a native polyacrylamide gel.

genes (group 2) not only have the most similar IES retention score distributions within each of these two groups, but also the strongest IES retention score correlations. Though not as strong as those within the *DCL2/3/5*, *NOWA1/2*, *TFIIS4*, and histone modification groups, positive correlations are also evident between these groups (Figure 3; e.g. Spearman's rank correlation coefficient, $r_s=0.74$ between *DCL2/3/5*-KD and *PTCAF1*-KD). The correlations between *PGM* knockdown and other gene knockdowns are generally weak — not much greater than the correlations between the control (ctrl3) and other gene knockdowns. Such weak correlations are an expected consequence of the ability of PiggyMac to excise most IES copies without scnRNAs/iesRNAs and without chromatin state changes brought about by *Ez11* and *PtCAF1*.

Since IES retention increases with IES length for many of the knockdowns influencing IES excision (see Figure 2E,F), we examined the retention score correlations between the abundant IESs corresponding to the first length peak (e.g. ≤ 35 bp; ~15600 IESs; the IES length peaks can be seen in the background of Figure 2E) compared to longer IESs (e.g. ≥ 500 bp; ~380 IESs; patterns of correlations between these length intervals are intermediate to those in Figure 2E,F). For short IESs there are good correlations between all the possible pairs of IES retention scores from *DCL2/3/5*-KD, *NOWA1/2*-KD, *TFIIS4*-KD, *PTCAF1*-KD and *EZL1*-KD. This could be because the protein products of all the genes involved in these knockdowns are cooperating in the excision of short IESs. For the subset of short IESs, the right-skewed IES retention score distributions of these knockdowns also all resemble each other (diagonal of Figure 3B). We previously noted that *DCL5*-KD leads to greater retention of short IESs than *DCL2/3*-KD, i.e. iesRNAs are marginally more important for short IES excision than scnRNAs and short IESs are generally weakly epigenetically controlled^{3,15}. This is reflected by the stronger correlations in IES retention score for short IESs between *DCL5*-KD and the knockdowns of *NOWA1/2*, *TFIIS4*, *PTCAF1* and *EZL1*, relative to the correlations between *DCL2/3*-KD and the same genes.

When focusing on longer IESs (e.g., those > 500 bp) the IES retention score correlations between the knockdowns of histone-associated genes, and also those between *DCL2/3/5*, *NOWA1/2*, *TFIIS4* and histone-associated genes weaken substantially. Compared to their short counterparts the retention score distributions of longer IESs also look very different, with those of *DCL2/3*, *DCL2/3/5*, *NOWA1/2*, *TFIIS4* closer to uniform, and those of the histone modification group closer to normal. The forms of the histone modification group IES retention score distributions now also become more like that of *PGM*-KD. These patterns are consistent with the chromatin state changes brought about by *EZL1*-KD and *PTCAF1*-KD exerting more influence on long IESs than short ones, with a much lesser requirement for iesRNAs, and generally some requirement for scnRNAs (i.e. long IESs are generally more strongly epigenetically controlled). At longer scales, the effect of *PTCAF1*-KD and *EZL1*-KD on IES retention (e.g. at ≥ 500 bp, $r_s=0.58$) does not correlate as well as at the shorter scale (e.g. at ≤ 35 bp, $r_s=0.88$), likely reflecting that the products of these genes do not have exactly the same consequences on chromatin state at this scale. The decrease in IES retention score correlations between either *DCL2/3*-KD or *DCL2/3/5*-KD and histone modification

groups largely reflects that iesRNAs have little influence on long IES excision.

Relationships among genes involved in genome editing inferred from OES retention resemble those of long IESs

To date, information about the effects of gene silencing upon *Paramecium*'s other eliminated sequences (OESs), the remaining deleted DNA currently not annotated as IESs, has typically come from Southern blotting or dot blotting, e.g. 8,10,11,13,31, though more recent studies have also begun to investigate the overall effects of gene knockdown on these regions^{11,12}. Most gene knockdowns that lead to IES retention also lead to retention of transposons within OESs^{8,10-13,31}. For each of the knockdowns used to investigate IES retention, we examined OES retention by mapping the knockdown reads to genomic regions which do not match MDSs and IESs from the reference *P. tetraurelia* assemblies that were extracted from a *PGM*-KD DNA assembly (see Methods for more details). Since *PGM*-KD leads to the most IES retention, like previous reports^{11,12}, we assume that much of the remaining DNA eliminated from the MIC genome can be recovered from this knockdown.

Examination of the distributions of OES retention (Figure 4A-D), generated similarly to those for IES retention (Figure 2A-D), shows that while some features of OES retention resemble those of IES retention, e.g. the forms of the *NOWA1/2*-KD, *TFIIS4*-KD and *DCL2/3/5*-KD OES retention distributions are roughly similar, there are also important differences. OES retention due to *DCL5*-KD is very weak, hardly differing from that of a negative control, and *DCL2/3/5*-KD OES retention is weaker than that of *DCL2/3*-KD. Compared to the very strong IES retention following *PGM*-KD versus all other knockdowns (Figure 2D), *PGM*-KD OES retention is not as pronounced (Figure 4D). Like IESs, in part this may be because the strength of OES retention in other knockdowns is greater, but it is also possible that DNA cutting enzymes other than PiggyMac may be involved in OES elimination.

Examination of the correlations in OES retention in Figure 4E reveals a number of important effects: (i) between *PGM*-KD and other knockdowns, OES retention correlation is greater than general IES retention score correlations (Figure 3A); (ii) *NOWA1/2*-KD correlates most strongly with *TFIIS4*-KD and *DCL2/3*-KD; *PTCAF1*-KD and *EZL1*-KD correlate the best with each other, and moderately with *DCL2/3*-KD, *DCL2/3/5*-KD, *TFIIS4*-KD and *NOWA1/2*-KD; (iii) though a clear role for the *Pdsg2* protein has yet to be defined, OES retention following *PDSG2*-KD correlates most strongly with *PGM*-KD; (iv) as OES length increases (e.g. > 4 kb) the *PTCAF1*-KD and *EZL*-KD OES retention scores become very narrowly distributed in contrast to those of *DCL2/3*-KD, *NOWA1/2*-KD and *PGM*-KD, suggesting that chromatin state changes brought about by *PtCaf1* and *Ez11* may have a very general effect on OES retention. From these observations and those of the effects of previously reported analyses of the overall sequence complexity of retained OES DNA following gene silencing^{11,12}, it follows that OES elimination mostly does not need iesRNAs, whereas scnRNAs, together with *TFIIS4* and *Nowa1/2* are necessary for OES elimination. The stronger correlations in OES retention scores of other knockdowns relative to those of *PGM*-KD suggests greater cooperation between PiggyMac, scnRNAs and

other protein factors in OES elimination than in general IES excision. Overall, the patterns of OES retention correlation for the different knockdowns resemble those of retention correlations for long IESs, suggesting most OESs may be treated like long IESs.

Nowa1/2 is involved in RNA scanning, and may mediate scnRNA- and iesRNA- targeted DNA cutting and excision

As the *Nowa1/2* proteins have previously been implicated in removal of maternally controlled IESs³¹, but their precise role in *Paramecium* genome development has not yet been established, we wished to obtain additional insight into the function of these proteins, and their relation to other molecules involved in this process. *Nowa1/2* proteins contain two distinct domains: an unstructured N-terminal domain consisting of alternating blocks of ‘FRG’ repeats and ‘GGWG’ repeats, and a structured C-terminal domain which is sufficient to route *Nowa1* from the maternal MAC to the zygotic nuclei³¹. *Nowa1/2*’s ‘FRG’ repeats resemble the ‘RGG’ boxes present in many hnRNP and other RNA-binding proteins, raising the possibility that they could facilitate interaction between Piwi-bound sRNAs and nascent transcripts on the old and the new MAC. One way *Nowa1/2* may mediate these interactions is via its “GGWG” repeats, which resemble the “WG/GW” repeats found in Argonaute binding proteins^{43,44} (Argonaute proteins are members of the Piwi protein superfamily⁴⁵). “WG/GW” repeat-containing proteins in *Tetrahymena* (Wag1 and CnjB) have been shown to interact with its Twi1 Piwi protein⁴⁶. In principle, *Nowa1/2* could interact with *Paramecium*’s sRNA-bound Piwi proteins in both the old and the developing new MAC, and during the internuclear transport of the sRNAs. Like *PtCAF1*-KD, *NOWA1/2*-KD also substantially decreases H3K9me3 and H3K27me3 in the old and developing new macronuclei¹⁰.

To gain more insight into the role of *Nowa1/2* in *Paramecium* genome reorganization we analyzed the effects of *NOWA1/2*-KD upon IES retention and the sRNA population during macronuclear development (Figure 5). We previously showed that in control and *DCL5*-KD cells, MAC genome-matching scnRNAs decrease during development, consistent with RNA scanning, and that there is a progressive increase in iesRNAs³. It can now also be seen that OESs give rise mainly to 25 nt scnRNAs and, compared to IESs, relatively few iesRNAs (Figure 5A; see also Figure S3G,H).

Since it is known that smaller IESs can be nested within larger IESs⁴⁷, we searched for IESs embedded within OESs. Visual inspection of DNA reads mapping to the assembled OESs from different gene knockdowns confirms the presence of shorter IESs in OESs, often with matching iesRNAs. We previously noted that, though both scnRNAs and iesRNAs are enriched at IES ends, scnRNAs overlap IES ends, whereas iesRNAs map very specifically within the limits of IESs³. Given the patterns of OES retention shown in Figure 4E, we therefore suggest that iesRNAs primarily facilitate precise IES excision, whereas scnRNAs facilitate both precise IES excision and imprecise OES elimination.

Compared to control cells, scnRNAs accumulate and iesRNAs are reduced after *NOWA1/2* silencing (Figure 5A, B). scnRNA accumu-

lation is consistent with inhibition of RNA scanning when *NOWA1/2* is silenced, i.e. MDS scnRNAs are not removed. We do not think that *NOWA1/2*-KD affects RNA scanning by interfering with transport of scnRNA-bearing Piwis (Ptiwi01/09⁹), as we observed no differences in localization of GFP-tagged Ptiwi09 with and without silencing *NOWA1/2* (Figure S4). On the other hand, the pronounced mobility shift in an electrophoretic mobility shift assay upon addition of synthetic MBP-*Nowa1* protein (Figure 5F) indicates *Nowa1* may interact with RNA, RNA:DNA hybrids, and, to a lesser degree, DNA. Hence *NOWA1/2*-KD-induced interference in RNA scanning in the old MAC could be due to disruption of interactions normally mediated by *Nowa1/2* between nucleic acids and sRNA-carrying Ptiwi proteins.

From the electrophoretic visualization of sRNAs (Figure 5B) it can be seen that iesRNAs are reduced following *NOWA1/2*-KD (since Figure 5A shows only proportions, it is not possible to deduce whether absolute levels of sRNAs change from this figure). A decrease in iesRNAs was also observed following *DCL2/3* co-silencing, and, to a lesser degree for the independent *DCL2* and *DCL3* silencings³. For the lower quantity of iesRNAs in *NOWA1/2*-KD cells, there are two possible explanations, both of which may be valid: (i) *Nowa1/2* is directly involved in iesRNA production or stability; (ii) *Nowa1/2* depletion inhibits scnRNA-dependent IES excision due to the inhibition of RNA scanning, which in turn inhibits iesRNA production. Teasing apart these possibilities is a possible future experimental avenue to gain additional insight into the function of the *Nowa1/2* proteins.

NOWA1/2 silencing was previously reported to lead to retention of maternally controlled IESs, and indeed maternally controlled IES retention levels (following microinjection of complementary IESs into the old MAC) correlate well with the IES retention scores (Figure 5D). However, following *NOWA1/2* silencing we also observe low levels of retention of some non-maternally controlled IESs by PCR analyses and in DNA-seq data (Figure 5D, E; Table S1). As the non-maternally controlled IESs that are weakly retained following *NOWA1/2*-KD are also weakly retained following *DCL2/3*-KD, we infer that they are weakly epigenetically controlled (Table S1). Therefore, due to the sensitivity of IES retention detection in DNA-seq data, some “non-maternally controlled” IESs may actually be weakly epigenetically controlled (it is also possible that with more sensitive methods these IESs would be classified as weakly maternally controlled). From Figure 3, we infer that *Nowa1/2* is not only involved in scnRNA-dependent IES excision (epigenetically controlled IES excision), but also iesRNA-dependent IES excision, since IES retention scores of *NOWA1/2*-KD are positively correlated with those of both *DCL2/3*-KD and *DCL5*-KD, and strongly correlated with those of *DCL2/3/5*-KD.

From these observations of the effects of *NOWA1/2*-KD, we propose that *Nowa1/2*’s WG/GW (Argonaute-binding) repeats could mediate IES targeting by both Ptiwi01/09-bound scnRNAs (both old and new MAC) and iesRNAs bound to other Ptiwi proteins⁴⁸ in the new MAC. The apparent failure to remove MAC genome-matching scnRNAs after *NOWA1/2* silencing (Figure 5A) is consistent with the

involvement of Nowa1/2 in RNA scanning in the old macronucleus, as expected if these proteins mediate pairing of scnRNAs with nascent maternal non-coding transcripts. In principle, Nowa1/2 proteins could interact with such transcripts via their FRG (RNA-binding RG boxes) repeats. From the correlations in Figure 4, we infer that Nowa1/2 proteins may operate predominantly in concert with scnRNAs and not iesRNAs during OES elimination.

Discussion

The strong similarity of IES retention following silencing of *DCL2/3/5*, *NOWA1/2* and *TFIIS4* suggests that the protein products of these genes may work in the same sRNA-dependent DNA elimination pathway during short IES excision, and that *DCL2/3*, *NOWA1/2* and *TFIIS4* work in the predominantly scnRNA-dependent DNA pathway responsible for OES and long IES elimination. *PGM*-KD-induced DNA retention correlates very weakly with that of other knockdowns for short IESs, but moderately with longer IESs and OESs. This suggests that for most short DNA excision, interactions between PiggyMac and the other proteins and scnRNAs/iesRNAs are less important than for longer DNA excision. For comparison, the silencing effects upon these and related genes affecting *Paramecium* genome development, and the expression/localization patterns of their encoded proteins are summarized in Table 1 and Table 2, respectively.

Even though the knockdowns of *DCL2/3/5*, *NOWA1/2* and *TFIIS4* seem to affect short IESs in a very similar manner, they affect the development-specific sRNAs in quite different ways. *DCL2/3/5*-KD leads to elimination of both scnRNAs and iesRNAs; *NOWA1/2*-KD inhibits RNA scanning (MAC genome-matching scnRNAs are not properly removed) and iesRNA production; *TFIIS4*-KD has little, if any, effect upon RNA scanning, but does eliminate iesRNA production¹². Though limited to four analyzed IESs, it is thought that *TFIIS4* may be involved in general IES transcription in the developing new MAC¹², including that of an IES whose excision has no apparent requirement for scnRNAs, iesRNAs or *TFIIS4*¹² (51A4404; see Table S1). This is proposed to produce RNA to which the sRNAs bind, facilitating IES targeting¹². Since *TFIIS4*-KD also leads to OES retention, it is likely that *TFIIS4* is also involved in extensive transcription of these regions. It is also apparent that even when there is a good correlation in IES or OES retention scores, the normal expression and localization patterns of tagged proteins encoded by the genes tested in these knockdowns may be quite distinct, e.g. Nowa1/2 proteins are present in both old and new macronuclei, whereas *TFIIS4* is predominantly present in new macronuclei (Table 1).

At present, the simplest explanation which reconciles the observed correlations in IES and OES retention scores between gene knockdowns, the effects of these knockdowns on sRNAs, and the expression and localization of the proteins encoded by these genes is that: (i) *TFIIS4*-KD removes the RNA transcripts which both scnRNAs and iesRNAs need to bind to their target IESs (or OESs, in the case of scnRNAs), hence IES retention following *TFIIS4*-KD correlates well with that following *DCL2/3/5*-KD (and *TFIIS4*-KD OES retention correlates modestly with *DCL2/3*-KD); (ii) *NOWA1/2*-KD prevents binding of Piwi-bound scnRNAs and iesRNAs to their targets via the nascent RNAs

produced by *TFIIS4*, hence *NOWA1/2*-KD IES retention correlates well with both *TFIIS4*-KD and *DCL2/3/5*-KD IES retention (and *NOWA1/2*-KD OES retention correlates well with both *TFIIS4*-KD and *DCL2/3*-KD IES retention).

For the shortest IESs (those < 35 bp), we propose that the high IES retention score correlations between *DCL2/3/5*-KD, *NOWA1/2*-KD, *TFIIS4*-KD, *PTCAF1*-KD and *EZL1*-KD, may be a consequence of cooperation of scnRNAs, iesRNAs, RNA transcripts and chromatin state changes allowing these genomic regions to be recognized and targeted (Figure 3B). For longer IESs the correlations in IES retention score between the different possible pairs of *DCL2/3*-KD, *DCL2/3/5*-KD, *NOWA1/2*-KD, *TFIIS4*-KD are all strong, as are those between *PTCAF1*-KD and *EZL1*-KD, but those between the latter two knockdowns and the former ones are modest (Figure 3C). This reflects that long IESs have a greater requirement for scnRNAs, Nowa1/2 proteins and *TFIIS4*-dependent RNA transcripts, and very little requirement for iesRNAs, but that only some IES retention due to *PTCAF1*-KD and *EZL1*-KD also occurs when either scnRNAs, Nowa1/2 or RNA transcripts are depleted.

Though potential co-operation between chromatin modifying genes and scnRNA/iesRNA genes in IES excision is suggested by the modest correlations between their knockdown-induced IES retention scores, much of the excision requiring the Ezl1 and PtCAF1 chromatin-modifying proteins does not appear to need development-specific sRNAs, as there is generally much less IES retention after *DCL2/3/5*-KD (median 0.053) than after *EZL1*-KD or *PTCAF1*-KD (median 0.37 and 0.23, respectively) even though *DCL2/3/5*-KD effectively eliminates the IES-matching sRNAs (Figure 5C). *DCL2/3* knockdown abolishes most H3K27me3 and H3K9me3 during new MAC development¹¹, like *EZL1*-KD, *PTCAF1*-KD and *NOWA1/2*-KD^{10,11}, whereas *DCL5*-KD does not influence these modifications¹¹. Thus it has been suggested that scnRNAs alone may guide histone modifications that facilitate DNA elimination in *Paramecium*¹¹.

The idea that *Paramecium*'s scnRNAs guide chromatin state changes in the developing new MAC should, however, be interpreted cautiously: though both *PTCAF1*-KD and *NOWA1/2*-KD affect H3K27me3 and H3K9me3 histone modifications in the new MAC, both PtCAF1 and Nowa1/2 also localize in the actively expressed, old MAC, and both *PTCAF1*-KD and *NOWA1/2*-KD¹⁰ affect H3K9me3 and H3K27me3 and scanning of scnRNAs proposed to occur in this nucleus, showing that PtCAF1 and Nowa1/2 (and potentially also the chromatin modifications that they influence) operate upstream of any possible scnRNA-directed chromatin modifications in the new MAC. It will thus be important to determine whether, like *PTCAF1*-KD and *NOWA1/2*-KD, *EZL1*-KD influences RNA scanning and the scnRNA population. Assuming fairly accurate quantification, more IESs are very strongly retained following *DCL2/3/5*-KD than *EZL1*-KD or *PTCAF1*-KD, e.g. at a threshold of > 0.7, 4316 IESs for *DCL2/3/5*-KD, compared to 1077 and 649 IESs for *EZL1*-KD or *PTCAF1*-KD, suggesting there may be some scnRNA/iesRNA-dependent IES excision which does not require the chromatin state changes induced by Ezl1- and PtCAF1. In future, it will thus be necessary to determine how the silencing of chromatin modifying genes, such as *EZL1* and *PTCAF1*, and

the removal of histone modifications, prevent excision of scnRNA/iesRNA-independent IESs, and, reciprocally, how depletion of scnRNAs and iesRNAs by *DCL2/3/5* silencing prevents IES excision independently of Ezl1- or PtCAF1-dependent histone modifications.

Previously, it was proposed that the mechanisms involved in imprecise DNA elimination in *Paramecium* are distinct from those involved in IES excision⁶. For imprecise DNA elimination, it was suggested that the resultant boundaries are a mixture of AT-rich direct repeats⁶. Since these repeats are an arbitrary length and are in relatively AT rich genomic regions it is straightforward to find them purely by chance: e.g. 39 out of 40 of the proposed direct repeats bounding the imprecisely eliminated DNA contained a TA dinucleotide, with all 7 of the two nucleotide repeats being TA⁶. We examined additional junctions of imprecisely eliminated regions (OESs) in DNA from control cells and consistently found that where direct repeats were present and unambiguous, they invariably contained a TA dinucleotide. We therefore favor the simple explanation that, like IESs, the imprecisely eliminated OESs are excised by PiggyMac transposases, with TA dinucleotides as the invariant bases necessary for this cleavage⁴⁷. We previously noted that long IESs in *Paramecium* tend to be underrepresented in MIC genome regions that develop into coding sequences in the MAC genome, suggesting counterselection against such IESs⁴⁹, either due to their lower excision efficiency or higher excision imprecision. Supporting the latter, we find that IESs with evidence of alternative excision in their vicinity are longer than those that do not (Figure S5E).

For the correlations between pairs of gene knockdowns, the effects on short IESs can be seen to differ from those on long IESs, and OES retention correlations more closely resemble those of long IESs than those of short IESs. We therefore propose there is a continuum of excision precision, which is highest for the shortest IESs and lowest for the longest IESs and OESs. In this case, the abbreviation “IES” may be a misnomer, since excision of longer DNA may lead to chromosome fragmentation rather than internal elimination. In a model in which PiggyMac operates as a dimer¹ that recognizes and cuts at two sites, for short IESs it is most likely that both subunits recognize and cut where expected. As IES length increases, DNA elimination imprecision increases. When IESs/OESs become sufficiently long, the two PiggyMac proteins become increasingly unlikely to make contact, or the cut sites may be too far apart to be rejoined efficiently (the latter, proposed by 6), thus imprecise chromosome breakage results. Thus, consistent with our overall observations, it follows that the recognition and delimitation of the longest eliminated DNA sequences, including transposons, benefits most from additional factors, such as sRNAs or chromatin alterations.

In conclusion, we have shown that comparisons of DNA excision readout following gene silencing by RNAi now provide the means

to identify and study associations among genome editing components in *Paramecium* on a genomic scale. Using this approach, we have proposed putative associations between pairs of Dicer-like and Piwi proteins involved in DNA elimination, which were subsequently confirmed by experimental pulldowns of these proteins and their bound sRNAs⁴⁸. In principle, the application of similar methods would also be useful in identifying and analyzing the molecular components of pathways in forms of developmental DNA deletion found in diverse eukaryotes, including that occurring in parasitic nematodes⁵⁰, sea lamprey⁵¹, and copepods⁵². In the case of the parasitic *Ascaris* nematodes, this includes DNA elimination which is considered unlikely to have a requirement for development-specific sRNAs such as those in ciliates⁵³, but which could conceivably require specific chromatin modifications. As highlighted by the analysis of *NOWA1/2* knockdown effects in this paper, these methods, together with complementary experiments, set the stage to begin generally disentangling interactions between candidate genome editing genes.

Data and software availability

Datasets 1 and 2, IES and OES retention scores, respectively, are available from Zenodo: DOI, [10.5281/zenodo.823766](https://doi.org/10.5281/zenodo.823766)⁵⁴.

The latest source code of after_ParTIES is available from GitHub: https://github.com/gh-ecs/After_ParTIES.

The archived source code as at the time of publication is available from Zenodo: DOI, [10.5281/zenodo.832236](https://doi.org/10.5281/zenodo.832236)⁵⁵.

License: BSD 2-Clause License.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Swiss National Science Foundation (31003A_129957, 31003A_146257 and 31003A_166407 and NCCR RNA & Disease to MN); the European Research Council (EPiGENOME and G-EDIT to MN); the European Cooperation in Science and Technology (Action BM1102 “Ciliates as model systems to study genome evolution, mechanisms of non-Mendelian inheritance, and their roles in environmental adaptation” to MN). CDW was supported by a Ph.D. fellowship from the French Ministère de l’Enseignement Supérieur et de la Recherche and the French National Research Agency (“Inferno”).

Acknowledgements

The authors wish to thank Linda Sperling (Institute for Integrative Biology of the Cell (I2BC), CNRS, CEA, University of Paris Sud, France) for her feedback on this manuscript.

Supplementary material

Figure S1. IES retention score correlations among controls and evaluation of reproducibility of *NOWA1/2*-KD. (A) Where indicated, IES retention score subtractions are to a minimum value of zero. IES retention scores for CtrlM are the mean of the other controls. Randomly permuted CtrlI IES retention scores were produced with NumPy's random.permutation function. For each plot in the lower triangular matrix (M_{ij}), Spearman's rank correlation coefficient (r_s) is given in the corresponding position diagonally opposite in the upper triangular matrix (M_{ji}). (B) Correlation between *NOWA1/2_a*-KD and *NOWA1/2_c*-KD IES retention scores. Spearman's rank correlation coefficient (r_s) is given. As described in the Methods section, mean control retention scores were subtracted before performing the regressions.

[Click here to access the data.](#)

Figure S2. Relationships between knockdown IES retention scores and IES end base frequencies. Graphs constructed as in Figure 2G–H.

[Click here to access the data.](#)

Figure S3. Cytological characterization and verification of knockdown efficacy. (A) Cytological characterization of control and *NOWA1/2*-KD cells whose sRNAs were sequenced. Bar graphs represent the percentages of cells in the vegetative stage or during MAC development, coloured per the observed developmental stages: “Skein” - cells displaying maternal MACs stretched to a characteristic convoluted structure; “Frag” - cells with fragmentation of the maternal MAC with no evident new MAC; and “Frag + new MAC” indicates cells at later stages of development with a fragmented MAC and a new developing MAC. In the early stage (“E”) ~50% of cells have fragmented maternal MACs. The middle stage (“M”) and late stage (“L”) corresponds to succeeding stages with an increasing proportion of new MACs. (B) Western blot (6% SDS-PAGE) for control and *NOWA1/2*-KD cells. (C) Cytological characterization of *DCL2/3/5*-KD. Legend as in (A). Cytological characterization of *DCL2/3*-KD and *DCL5*-KD cells was given in 3. (D) Northern blot for *DCL2/3/5*-KD and control cells (E) Cytological characterization of *PTCAF1*-KD and control (*ND7*-KD) cells. (F) Northern blot for *PTCAF1*-KD and control (*ND7*-KD) cells. (G) Sequence logo of 25 nt scnRNAs from late development control cells. (H) Sequence logo of 27 nt iesRNAs from late development control cells.

[Click here to access the data.](#)

Figure S4. Localisation of Ptiwi09-GFP during sexual reproduction in control and *NOWA1/2*-KD cells. The GFP construct was obtained from 9. N-terminally GFP tagged Ptiwi09 localized in the maternal MAC upon meiosis and later in the macronuclear fragments and developing new MACs in control and *NOWA1/2*-KD. Scale bar: 10 μ m.

[Click here to access the data.](#)

Figure S5. Properties of non-reference *P. tetraurelia* MAC+IES scaffolds, OESs and IESs. (A) Base composition of 7266 IDBA-UD scaffolds assembled from *PGM*-KD reads not mapping to the *P. tetraurelia* MAC genome and MAC genome with IES assemblies. (B) *PGM*-KD sequence coverage vs. scaffold length of the 4799 scaffolds with little retention in an empty silencing vector control. (C) Illustrative read mapping by BWA with empty silencing vector control (EV) reads aligned to a scaffold assembled by IDBA-UD using *PGM*-KD reads not mapping to the current *P. tetraurelia* MAC+IES genome assembly. OESs between the MDSs are extracted by identifying genomic regions with low read coverage. The positions of two IESs between MDSs is shown. Note that the boundaries of MDSs and IESs are not precise – e.g. the coverage of the right-most MDS region has a low sequence coverage region preceding the main MDS region. (D) OES retention scores (scales on right y-axes) vs. OES length. Lines are the exponentially-weighted mean averages (EWMA) with spans of 50 bp.

[Click here to access the data.](#)

Table S1. Comparison of maternal control scores and IES retention scores. IESs are those described in (Duharcourt *et al.*, 1998) and maternal control scores were calculated as the maximal observed retention described previously (Duharcourt *et al.*, 1998, Sandoval *et al.*, 2014). Maternally controlled IESs are those whose maternal control score is > 0; non-maternally controlled IESs are those with maternal control score = 0. For all the IESs listed in this table the mean retention scores across the four controls used in this paper are 0.00.

[Click here to access the data.](#)

Supplementary Data S1 to S8 is available from Zenodo: DOI, [10.5281/zenodo.823766](https://doi.org/10.5281/zenodo.823766)⁵².

Supplementary Data S1: Statistics for Spearman's correlations for all IESs

Supplementary Data S2: Statistics for Spearman's correlations for IESs <= 35 bp

Supplementary Data S3: Statistics for Spearman's correlations for IESs >= 500 bp

Supplementary Data S4: Statistics for Spearman's correlations for OESs (>= 300 bp and 3x-70x *PGM*-KD coverage)

Supplementary Data S5: Statistics for comparisons of Spearman correlation differences for all IESs

Supplementary Data S6: Statistics for comparisons of Spearman correlation differences for all OESs

Supplementary Data S7: Statistics for comparisons of the differences in Spearman correlations between IESs and OESs

Supplementary Data S8: OES sequences

References

1. Arnaiz O, Mathy N, Baudry C, *et al.*: **The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences.** *PLoS Genet.* 2012; **8**(10): e1002984.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Baudry C, Malinsky S, Restituito M, *et al.*: **PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*.** *Genes Dev.* 2009; **23**(21): 2478–83.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Sandoval PY, Swart EC, Arambasic M, *et al.*: **Functional Diversification of Dicer-like Proteins and Small RNAs Required for Genome Sculpting.** *Dev Cell.* 2014; **28**(2): 174–88.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Amar L: **Chromosome end formation and internal sequence elimination as alternative genomic rearrangements in the ciliate *Paramecium*.** *J Mol Biol.* 1994; **236**(2): 421–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Forney JD, Blackburn EH: **Developmentally controlled telomere addition in wild-type and mutant *paramecia*.** *Mol Cell Biol.* 1988; **8**(1): 251–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Le Mouél A, Butler A, Caron F, *et al.*: **Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in *paramecia*.** *Eukaryot Cell.* 2003; **2**(5): 1076–90.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Caron F: **A high degree of macronuclear chromosome polymorphism is generated by variable DNA rearrangements in *Paramecium primaurelia* during macronuclear differentiation.** *J Mol Biol.* 1992; **225**(3): 661–78.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Arambasic M, Sandoval PY, Hoehener C, *et al.*: **Pdsg1 and Pdsg2, novel proteins involved in developmental genome remodelling in *Paramecium*.** *PLoS One.* 2014; **9**(11): e112899.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Bouhouche K, Gout JF, Kapusta A, *et al.*: **Functional specialization of Piwi proteins in *Paramecium tetraurelia* from post-transcriptional gene silencing to genome remodelling.** *Nucleic Acids Res.* 2011; **39**(10): 4249–64.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Ignarski M, Singh A, Swart EC, *et al.*: ***Paramecium tetraurelia* chromatin assembly factor-1-like protein P1CAF-1 is involved in RNA-mediated control of DNA elimination.** *Nucleic Acids Res.* 2014; **42**(19): 11952–64.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Lhullier-Akakpo M, Frapporti A, Denby Wilkes C, *et al.*: **Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements.** *PLoS Genet.* 2014; **10**(9): e1004665.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Maliszewska-Olejniczak K, Gruchota J, Gromadka R, *et al.*: **TFIIS-Dependent Non-coding Transcription Regulates Developmental Genome Rearrangements.** *PLoS Genet.* 2015; **11**(7): e1005383.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Lepère G, Nowacki M, Serrano V, *et al.*: **Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*.** *Nucleic Acids Res.* 2009; **37**(3): 903–15.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Mochizuki K, Fine NA, Fujisawa T, *et al.*: **Analysis of a *piwi*-related gene implicates small RNAs in genome rearrangement in *tetrahymena*.** *Cell.* 2002; **110**(6): 689–99.
[PubMed Abstract](#) | [Publisher Full Text](#)
15. Swart EC, Wilkes CD, Sandoval PY, *et al.*: **Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion.** *Nucleic Acids Res.* 2014; **42**(14): 8970–83.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Noto T, Kataoka K, Suhren JH, *et al.*: **Small-RNA-Mediated Genome-wide trans-Recognition Network in *Tetrahymena* DNA Elimination.** *Mol Cell.* 2015; **59**(2): 229–42.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Fang W, Wang X, Bracht JR, *et al.*: **Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement.** *Cell.* 2012; **151**(6): 1243–55.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Zahler AM, Neeb ZT, Lin A, *et al.*: **Mating of the stichotrichous ciliate *Oxytricha trifallax* induces production of a class of 27 nt small RNAs derived from the parental macronucleus.** *PLoS One.* 2012; **7**(8): e42371.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Mochizuki K, Gorovsky MA: **Conjugation-specific small RNAs in *Tetrahymena* have predicted properties of scan (scn) RNAs involved in genome rearrangement.** *Genes Dev.* 2004; **18**(17): 2068–73.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Lepère G, Bétermier M, Meyer E, *et al.*: **Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in *Paramecium tetraurelia*.** *Genes Dev.* 2008; **22**(11): 1501–12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Schoeberl UE, Kurth HM, Noto T, *et al.*: **Biased transcription and selective degradation of small RNAs shape the pattern of DNA elimination in *Tetrahymena*.** *Genes Dev.* 2012; **26**(15): 1729–42.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Cheng CY, Vogt A, Mochizuki K, *et al.*: **A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in *Tetrahymena thermophila*.** *Mol Biol Cell.* 2010; **21**(10): 1753–62.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Chalker DL, Yao MC: **DNA elimination in ciliates: transposon domestication and genome surveillance.** *Annu Rev Genet.* 2011; **45**: 227–46.
[PubMed Abstract](#) | [Publisher Full Text](#)
24. Fass JN, Joshi NA, Couvillion MT, *et al.*: **Genome-Scale Analysis of Programmed DNA Elimination Sites in *Tetrahymena thermophila*.** *G3 (Bethesda).* 2011; **1**(6): 515–22.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Malone CD, Anderson AM, Motl JA, *et al.*: **Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of *Tetrahymena thermophila*.** *Mol Cell Biol.* 2005; **25**(20): 9151–64.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Mochizuki K, Gorovsky MA: **A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase.** *Genes Dev.* 2005; **19**(1): 77–89.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Duhaucourt S, Butler A, Meyer E: **Epigenetic self-regulation of developmental excision of an internal eliminated sequence on *Paramecium tetraurelia*.** *Genes Dev.* 1995; **9**(16): 2065–77.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. Duhaucourt S, Keller AM, Meyer E: **Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in *Paramecium tetraurelia*.** *Mol Cell Biol.* 1998; **18**(12): 7075–85.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Liu Y, Taverna SD, Muratore TL, *et al.*: **RNA-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in *Tetrahymena*.** *Genes Dev.* 2007; **21**(12): 1530–45.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Taverna SD, Coyne RS, Allis CD: **Methylation of histone h3 at lysine 9 targets programmed DNA elimination in *tetrahymena*.** *Cell.* 2002; **110**(6): 701–11.
[PubMed Abstract](#) | [Publisher Full Text](#)
31. Nowacki M, Zagorski-Ostojka W, Meyer E: **Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in *Paramecium tetraurelia*.** *Curr Biol.* 2005; **15**(18): 1616–28.
[PubMed Abstract](#) | [Publisher Full Text](#)
32. Skouri F, Cohen J: **Genetic approach to regulated exocytosis using functional complementation in *Paramecium*: identification of the ND7 gene required for membrane fusion.** *Mol Biol Cell.* 1997; **8**(6): 1063–71.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Arnaiz O, Goût JF, Bétermier M, *et al.*: **Gene expression in a paleopolyploid: a transcriptome resource for the ciliate *Paramecium tetraurelia*.** *BMC Genomics.* 2010; **11**: 547.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Arnaiz O, Sperling L: ***ParameciumDB* in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*.** *Nucleic Acids Res.* 2011; **39**(Database issue): D632–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Denby Wilkes C, Arnaiz O, Sperling L: **ParTIES: a toolbox for *Paramecium* interspersed DNA elimination studies.** *Bioinformatics.* 2016; **32**(4): 599–601.
[PubMed Abstract](#) | [Publisher Full Text](#)
36. Diedenhofen B, Musch J: **cocor: a comprehensive solution for the statistical comparison of correlations.** *PLoS One.* 2015; **10**(3): e0121945.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
37. Myers L, Sirois MJ: **Spearman Correlation Coefficients, Differences between.** *Encyclopedia of Statistical Sciences.* John Wiley & Sons, Inc.; 2004.
[Publisher Full Text](#)
38. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–60.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Peng Y, Leung HC, Yiu SM, *et al.*: **IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth.** *Bioinformatics.* 2012; **28**(11): 1420–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
40. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Methods.* 2015; **12**(4): 357–60.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Crooks GE, Hon G, Chandonia JM, *et al.*: **WebLogo: a sequence logo generator.** *Genome Res.* 2004; **14**(6): 1188–90.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

42. Aury JM, Jaillon O, Duret L, *et al.*: **Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*.** *Nature*. 2006; **444**(7116): 171–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
43. El-Shami M, Pontier D, Lahmy S, *et al.*: **Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components.** *Genes Dev*. 2007; **21**(20): 2539–44.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
44. Partridge JF, DeBeauchamp JL, Kosinski AM, *et al.*: **Functional separation of the requirements for establishment and maintenance of centromeric heterochromatin.** *Mol Cell*. 2007; **26**(4): 593–602.
[PubMed Abstract](#) | [Publisher Full Text](#)
45. Swarts DC, Makarova K, Wang Y, *et al.*: **The evolutionary journey of Argonaute proteins.** *Nat Struct Mol Biol*. 2014; **21**(9): 743–53.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
46. Bednenko J, Noto T, DeSouza LV, *et al.*: **Two GW repeat proteins interact with *Tetrahymena thermophila* argonaute and promote genome rearrangement.** *Mol Cell Biol*. 2009; **29**(18): 5020–30.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
47. Gratiás A, Lepère G, Garnier O, *et al.*: **Developmentally programmed DNA splicing in *Paramecium* reveals short-distance crosstalk between DNA cleavage sites.** *Nucleic Acids Res*. 2008; **36**(10): 3244–51.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
48. Furrer DI, Swart EC, Kraft MF, *et al.*: **Two Sets of Piwi Proteins Are Involved in Distinct sRNA Pathways Leading to Elimination of Germline-Specific DNA.** *Cell Rep*. 2017; **20**(2): 505–20.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Swart EC, Nowacki M: **The eukaryotic way to defend and edit genomes by sRNA-targeted DNA deletion.** *Ann N Y Acad Sci*. 2015; **1341**: 106–14.
[PubMed Abstract](#) | [Publisher Full Text](#)
50. Wang J, Mitreva M, Berriman M, *et al.*: **Silencing of germline-expressed genes by DNA elimination in somatic cells.** *Dev Cell*. 2012; **23**(5): 1072–80.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
51. Smith JJ, Antonacci F, Eichler EE, *et al.*: **Programmed loss of millions of base pairs from a vertebrate genome.** *Proc Natl Acad Sci U S A*. 2009; **106**(27): 11212–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
52. Sun C, Wyngaard G, Walton DB, *et al.*: **Billions of basepairs of recently expanded, repetitive sequences are eliminated from the somatic genome during copepod development.** *BMC Genomics*. 2014; **15**: 186.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
53. Wang J, Czech B, Crunk A, *et al.*: **Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles.** *Genome Res*. 2011; **21**(9): 1462–77.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
54. Swart EC: **Identification and analysis of functional associations among natural eukaryotic genome editing components.** *Zenodo*. 2017.
[Data Source](#)
55. gh-ecs: **gh-ecs/After_ParTIES: First release.** *Zenodo*. 2017.
[Data Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 22 August 2017

<https://doi.org/10.5256/f1000research.13119.r24871>

© 2017 Simon M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Martin Simon

Molecular Cell Dynamics, Centre for Human and Molecular Biology, Saarland University, Saarbrücken, Germany

Sivarajan Karunanithi

¹ Cluster of Excellence "Multimodal Computing and Interaction", Saarland University, Saarbrücken, Germany

² Max Planck Institute for Informatics, Saarbrücken, Germany

In *Paramecium* as well as other ciliates, massive genome re-arrangements in form of elimination of thousands of transposon derived IES sequences occur after sexual fertilization. Some of these excision events have been described to be controlled by small RNAs acting in a genome comparison mechanism. Swart and colleagues compared IES retention and OES retention in F1 progeny after dsRNA induced knockdown of genes involved in these genome re-arrangements and aimed to classify functional relationship of genes by similarity of the individual IES/OES retention. They show that RNA associated and chromatin associated genes form individual groups. In addition, useful information is provided by the *de novo* assembly of eliminated sequences apart of IESSs. The paper represents an interesting analysis providing useful tools and solid data for future research applications.

Major points:

1. Although the described method to classify IES retention after silencing of different genes is innovative, the paper is a balancing act between methods and research paper. In its current shape, the conclusions drawn and the main results of this paper are of biological significance pushing the methodological aspect into the background. Right now, the paper is more a research paper. The authors need to decide for a strategy.

Minor Points:

1. The identification of the functional groups of Dcl2/3/5, TFIIS4, NowaX and Ezl1, PtCAF also reflects the expression behavior of these enzymes during development (with the summary of Dcl2/3 early - Dcl5 late). As it makes sense at first glance that cooperating enzymes need to become induced at the same time, the authors should also discuss whether the dsRNA

feeding efficiency could be different in early and late stages of development.

2. In Table 2, the induction time points differ for some enzymes to the microarray data in ParameciumDB described in Arnaiz *et al.* 2010. For instance, TFIIS4 is indicated to be expressed late in Table 2 but according to Maliszewska-Olejniczak *et al.* 2015 and ParameciumDB, this gene is induced early during autogamy.
3. As most of the analyses here relies on the calculation of IES retention scores, please give a more precise description of their definition in the text, as this term could be new for readers outside the field. Maybe also give more information in Methods, which reads are counted that map to an IES junction or to IES, in particular those, which only partially overlap to the IES. Also, are there any cutoffs set at any time of calculation?
4. In Methods: IES and OES retention score regressions: It's nice to see the effort put in creating different kinds of regression. However, in the paper, authors do not address about regression in the text (Also, not in Supp. Fig. S1). They only focus on the correlation coefficient and the IES retention score distributions.
5. Introduction: asexual reproduction would be cell division in *paramecia* which is not autogamy or conjugation
6. In Methods: Estimation of OES retention section, the first sentence is too long and difficult to follow. Kindly rephrase it. Authors do not explain how exactly OES scores are obtained.
7. I'm looking for how the authors mapped the 3882 OESs with variable boundaries (obtained from the 5236 scaffolds in the new assembly) to the already available 697 scaffolds (in MAC+IES assembly).
8. In Results: Examination of basic IES retention properties (last paragraph) Shouldn't it be Figure 2G-J.
9. Figure 2: In the legend H should be replaced with J. Also in the sentence "(G-H) Mean base frequencies of positions 1-3 after the TA repeat relative are plotted relative to IES retention score" there seems to be a typographical error with the use of the word relative. And, for G-J, along the Y-axis it would be nice to say "Bases after TA repeat" along the Y-axis for the base frequency plots.
10. Discussion: The idea that....This sentence is too long.
11. Authors say they have attempted to produce a better assembly. It would be helpful to provide this assembly.
12. For readers not used to *Paramecium* Mac development, the different classifications of IESs into epigenetic/not epigenetic, scnRNA dependent/non scnRNA dependent, long/short etc. can be confusing. Together with the classification of the genetic requirements, is it possible to sum the discussion up in kind of a venn diagram to compare IES properties and genetic requirements for excision? As a suggestion...

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 15 August 2017

<https://doi.org/10.5256/f1000research.13119.r24870>

© 2017 Mochizuki K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Kazufumi Mochizuki

Institute of Human Genetics, CNRS, University of Montpellier, Montpellier, France

In this manuscript, the authors aimed to classify genes reported to be involved in DNA elimination of *Paramecium* by comparing sensitivities of internal eliminated sequences (IESs) and other eliminated sequences (OESs) to RNAi knockdown (KD) of the genes. Their classification method provides quantitative scores for pairwise correlations of genes in DNA elimination pathway and they successfully separated two groups of genes: RNA-associated genes and histone modification genes. The authors also showed that shorter and longer IESs have different sensitivities to RNAi KD of different genes and the sensitivities of OESs were similar to those of longer IESs. Therefore, they suggested that the mechanisms for IES and OES eliminations are overlapping. I believe the presented data is compelling and the strategy described is useful to elucidate the function of uncharacterized genes in DNA elimination in future study.

Major points

1. Because this is a method article, I suggest the authors to minimize biological interpretations and rather extensively discuss methodological/strategical validities and technical/theoretical disadvantages, if any. Related to this point, it may be better to present the data shown in Figure 5 in another paper.
2. The authors' strategy does not distinguish direct roles of a genes in DNA elimination from its indirect effects in DNA elimination. I believe this point should be discussed. Some possibilities may be discusses are: a) histone modification genes may affect DNA elimination not only directly through histone modifications on IESs/OESs but also indirectly through regulating expressions of genes involved in DNA elimination; b) RNA-associated genes may act not only on chromatin but also in post-transcriptional regulations of genes involved in DNA elimination, like classical RNAi. In these cases, direct targets of RNA-associated genes and histone modification genes may be identical and the observed differences are due to indirect effects.
3. I am confused by the following descriptions in the 2nd and the 3rd paragraphs of Results (page 8), which need some clarification.
 - a) Though they first mentioned that "these knockdowns may indicate most of the Paramecium IESs requiring scnRNAs and iesRNAs for their excision", they then wrote in the next paragraph that "We therefore infer that most IES excision in Paramecium does not require IES-targeting sRNAs."
 - b) I do not see why "most IES excision in Paramecium does not require IES-targeting sRNAs" is "consistent" with the fact that "scnRNAs alone are insufficient for most IES excision". I think a requirement cannot be inferred from an insufficiency.
4. According to the authors' recent publication (Allen et al. 2017), the production of iesRNA requires circularization of IESs. Therefore, different sensitivities of short and long IESs (and OESs) to KDs of the DNA elimination genes are likely related to their ability to make circle when they are eliminated. I think this possibility should be discussed.
5. Very little discussion for PDSG2 in the manuscript. The IES retention scores of PDSG2 KD shows no strong correlation with those of any other KDs. Does this mean PDSG2 forms the third group? If so, what would be the function of the third group? If they do not provide any additional explanation, it may be better to consider removing PDSG2-related data from this study.

Other minor points

1. Abstract: I do not see "comparisons of experimentally-induced DNA retention" in this study.
2. Abstract: although I agree that the approach is quantitative but not clear how the authors can claim it is sensitive.
3. PtCAF1 encodes a protein similar to chromatin remodeling factor and thus classifying this gene as "histone modification" would be misnomer.
4. Figure 3A: The colors for "IESs per bin" are not visible in the individual plots (all dots look purple). Same for Figure 4E.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research