

UNIVERSITÄT DES SAARLANDES

**A highly condensed genome without
heterochromatin: orchestration of gene
expression and epigenomics in
*Paramecium tetraurelia***

Dissertation

zur Erlangung des Grades des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät
der Universität des Saarlandes

von

Franziska Drews (M.Sc.)

Saarbrücken

Juli, 2022

Tag des Kolloquiums	09. Dezember 2022
Dekan	Prof. Dr. Ludger Santen
Berichterstatter	Prof. Dr. Martin Simon Prof. Dr. Jörn E. Walter Prof. Dr. Douglas L. Chalker
Akad. Mitglied	Dr. Anna Welle
Vorsitz	Prof. Dr. Alexandra K. Kiemer

Einmal trafen sie eine Krähe. "Vögel sind nicht dumm", sagte der kleine Bär und er fragte die Krähe nach dem Weg. "Welchen Weg?", fragte die Krähe. "Es gibt hundert und tausend Wege."

Janosch

Abstract

Epigenetic regulation in unicellular ciliates can be as complex as in metazoans and is well described regarding small RNA (sRNA) mediated effects. The ciliate *Paramecium* harbors several copies of sRNA-biogenesis related proteins involved in genome rearrangements resulting in chromatin alterations. The global chromatin organization thereby is poorly understood, and unusual characteristics of the somatic nucleus, like high polyploidy, high genome coding density, and absence of heterochromatin, ought to call for complex regulation to orchestrate gene expression.

The present study characterized the nucleosomal organization required for gene regulation and proper Polymerase II activity. Histone marks reveal broad domains in gene bodies, whereas intergenic regions are nucleosome free. Low occupancy in silent genes suggests that gene inactivation does not involve nucleosome recruitment. Thus, *Paramecium* gene regulation counteracts the current understanding of chromatin biology.

Apart from global nucleosome studies, two sRNA binding proteins (Ptiwis) classically associated with transposon silencing were investigated in the background of transgene-induced silencing. Surprisingly, both Ptiwis also load sRNAs from endogenous loci in vegetative growth, revealing a broad diversity of Ptiwi functions. Together, the studies enlighten epigenetic mechanisms that regulate gene expression in a condensed genome, with Ptiwis contributing to transcriptome and chromatin dynamics.

Epigenetische Regulation kann in einzelligen Ciliaten so komplex sein wie in Vielzellern und wurde umfassend angesichts kleiner RNA (sRNA)-vermittelter Effekte untersucht. Der Ciliat *Paramecium* besitzt mehrere Kopien sRNA-Biogenese assoziierter Proteine, die an Genomprozessierungen und resultierenden Chromatinänderungen beteiligt sind.

Die globale Organisation des Chromatins ist dabei kaum verstanden und obskure Eigenschaften des somatischen Kerns, wie hohe Polyploidie, Kodierungsdichte und Fehlen von Heterochromatin, sollten eine komplexe Regulation zur Steuerung der Genexpression erfordern.

Die vorliegende Studie charakterisiert die Chromatinorganisation, die für die Genregulation und Polymerase II Aktivität notwendig ist. Histonmodifikationen zeigen breite Verteilungen in Genen, während intergenische Regionen Nucleosomen-frei sind. Ein Stilllegen von Genen scheint ohne die Rekrutierung von Nucleosomen zu erfolgen, womit die Genregulation in *Paramecium* dem aktuellen Verständnis der Chromatinbiologie widerspricht.

Neben Nucleosomenstudien wurden zwei sRNA-bindende Proteine (Ptiwis), die klassisch mit Transposon-Silencing assoziiert sind, im Hintergrund des Transgen-induzierten Silencings untersucht. Überraschenderweise laden Ptiwis sRNAs von endogenen Loci im vegetativen Wachstum, was vielfältige Ptiwi-Funktionen offenbart. Die Studien zeigen epigenetische Mechanismen zur Genregulation in einem kompakten Genom, wobei Ptiwis zur Transkriptom- und Chromatindynamik beitragen.

Publications resulting from this work

Drews, F., Boenigk, J., Simon, M. (2022). Paramecium epigenetics in development and proliferation. *Journal of Eukaryotic Microbiology*, 00, e12914.

Drews, F., Salhab, A., Karunanithi, S., Cheaib, M., Jung, M., Schulz, M., Simon, M. (2022). Broad domains of histone marks in the highly compact Paramecium macronuclear genome. *Genome Research*, gr.276126.121. Advance online publication.

Drews, F., Karunanithi, S., Götz, U., Marker, S., deWijn, R., Pirritano, M., Jung, M., Gasparoni, G., Schulz, M. H. Simon, M. (2021). Two Piwis with Ago-like functions silence somatic genes at the chromatin level. *RNA Biology*, 18(sup2), 757-769.

Karunanithi, S., Oruganti, V., de Wijn, R., **Drews, F.**, Cheaib, M., Nordström, K., Simon, M., Schulz, M. H. (2020). Feeding exogenous dsRNA interferes with endogenous sRNA accumulation in Paramecium. *DNA Research*, 27(1), dsaa005.

Karunanithi, S., Oruganti, V., Marker, S., Rodriguez-Viana, A. M., **Drews, F.**, Pirritano, M., Nordström, K., Simon, M. Schulz, M. H. (2019). Exogenous RNAi mechanisms contribute to transcriptome adaptation by phased siRNA clusters in Paramecium. *Nucleic Acids Research*, 47(15), 8036-8049.

Contents

Abstract	v
1 Background	1
1.1 Epigenetics	2
1.1.1 Chromatin	2
1.1.2 Nucleosome Organization	4
Excursion: How to Find Nucleosome Positions	5
1.1.3 Histone Modifications	6
Histone Modifications in Gene Expression	7
Polymerase II Recruitment and Transcriptional Activation	8
Transcriptional Repression	10
Histone Modifications in Disease and Epigenetic Inheritance	11
Ongoing Debates on the Function of Histone Modifications	12
1.1.4 Cross-talk of Regulators: RNA and Chromatin	12
RNA Interference	13
RNAi on the Co-Transcriptional Level	16
Silencing of Loci in <i>Trans</i> and Systemic RNAi	18
1.2 Ciliates as Models in Epigenetics	19
1.2.1 Nuclear Dimorphism in <i>Paramecium tetraurelia</i>	22
<i>Paramecium</i> MAC Chromosomes Have Extraordinary Features	23
1.2.2 Building a Functional MAC Genome Involves RNAi Components	24
IES Excision Upon Heterochromatin Formation	26
1.2.3 RNAi Machinery Apart from Development	28
1.2.4 Studying Ciliates is Fun	29
1.3 Aim and Outline of the Thesis	30
2 Material and Methods	33
2.1 Organisms and Cultivation Conditions	34
2.1.1 <i>Paramecium tetraurelia</i>	34
2.1.2 Bacteria	34
2.2 Standard Molecular Biology Techniques	35
2.2.1 Polymerase Chain Reaction	35
2.2.2 Transformation of <i>E. coli</i>	37
2.2.3 Plasmid Isolation from <i>E. coli</i> by Alkaline Lysis	37
2.2.4 Plasmid Isolation in Large Scale (Midi-Prep)	38
2.2.5 Agarose Gel Electrophoresis	38
2.2.6 Re-Isolation of DNA from Agarose Gels	38
2.2.7 Ligation Procedure and Restriction Enzyme Analyses	38
2.2.8 Sanger Sequencing	39
2.2.9 Preparation of Electrocompetent Bacteria	39
2.3 Handling of <i>Paramecium tetraurelia</i> and <i>Paramecium</i> Specific Methods	39
2.3.1 Growing of Clonal Cell Lines of Defined Age in Mass Cultures	39

2.3.2	Staining of Nuclei	40
2.3.3	Isolating Serotype Pure Cell Lines	40
2.3.4	Trichocyst Discharge	41
2.3.5	RNAi by Feeding	41
2.3.6	Microinjection	43
2.4	Protein Specific Methods	43
2.4.1	Total Protein Isolation and Macronuclei Enrichment	43
2.4.2	SDS-Gels for Western Blot	45
2.4.3	Western Blot	46
2.4.4	Affinity Purification of Polyclonal Peptide Antibodies	47
2.4.5	Competition Assay and Dot Blot	48
2.4.6	Immunostaining	49
2.4.7	Expression of Tagged Proteins and Immunoprecipitation	51
2.5	RNA Specific Methods	52
2.5.1	RNA Isolation	52
2.5.2	RNA Integrity Check	53
2.5.3	DNase I Treatment	54
2.5.4	Gel Purification of Small RNAs	54
2.5.5	Dissection of 3'-Modifications by Periodate Oxidation	55
2.6	Chromatin Specific Methods	56
2.6.1	Fixation of Cells	56
2.6.2	Isolation of Macronuclei from Fixed Material	57
2.6.3	Shearing of Chromatin	57
2.6.4	Quality Control for Chromatin Shearing	57
2.6.5	Immunoprecipitation from Chromatin	58
2.6.6	MNase Treatment	58
2.7	Library Preparation and Sequencing	59
2.7.1	Transcriptome Library Preparation	59
2.7.2	Small RNA Library Preparation	60
2.7.3	DNA Library Preparation	60
2.7.4	Library Quantification and Quality Control	60
2.7.5	Sequencing	60
2.8	Processing and Analyses of Sequencing Data	61
2.8.1	small RNA Analyses	61
	sRNA Normalization Using RAPID	61
	sRNA Sequence Logos and Nucleotide Count	62
	Overlapping read pairs	62
2.8.2	Calculation of Gene Expression and Plasticity from mRNA Data	62
2.8.3	ChIP-seq and MNase-seq Analyses	62
	Segmentation by ChromHMMM	63
2.8.4	Phylogenetic Analyses and Protein Sequences Alignments	64
2.9	Devices, Chemicals, Kits	64
2.10	Software, Packages, Web pages	64
3	The <i>Paramecium</i> Macronuclear Epigenome	67
3.1	Background	68
3.2	Methods	70
3.3	Results	71
3.3.1	Histone H3 Modifications in the Vegetative <i>Paramecium</i> MAC	71
3.3.2	Nucleosome Patterns Unveiled by MNase-seq	72
3.3.3	Coupling of Nucleosome Occupancy and Gene Expression	74

3.3.4	<i>Paramecium's</i> Extraordinary -1 Nucleosome	76
3.3.5	Combinatorial Patterns of Histone Marks	77
3.3.6	Gene Expression Regulation by Polymerase II Occupancy	80
3.3.7	Pausing Regulation with a Highly Divergent Polymerase II CTD	80
3.3.8	How do Epigenetic Marks Orchestrate Gene Expression?	82
3.3.9	Correlation of Epigenetic Features	84
3.4	Discussion	85
3.4.1	Outlook	89
4	<i>Paramecium tetraurelia</i> Piwi Proteins Silence on the Chromatin Level	93
4.1	Background	94
4.2	Methods	96
4.3	Results	98
4.3.1	Ptiwi Phylogeny and Localization	98
4.3.2	sRNA Loading Preferences	100
4.3.3	Transgene-Induced Silencing: Loading of 1° and 2° siRNAs	101
4.3.4	Ptiwis Load sRNAs from Endogenous sRNA Producing Clusters	105
4.4	Discussion	108
4.4.1	Outlook	111
5	General Discussion and Future Perspective	115
5.1	Interaction of Epigenetic Key Players in a Crowded Nucleus	116
5.1.1	High Coding Hardware Is Unprotected	118
5.1.2	Recruitment of Epigenetic Regulators to Genes Drive Expression	121
5.1.3	Processivity of Polymerase II: Beyond the CTD	123
5.1.4	Small RNAs Shape Gene Expression	123
5.1.5	Future Perspective: Surface Antigen Expression Regulation by Epigenetic Marks	126
	Bibliography	129
	A Supplementary Material Chapter 3	145
	B Supplementary Material Chapter 4	151
	C Manuscripts	159
	Acknowledgments	217
	Declaration of Authorship	219

List of Figures

1.1	Layers of epigenetic control	2
1.2	Nucleosome structure	3
1.3	Histone H3 modifications	7
1.4	RNA interference	14
1.5	Co-transcriptional gene silencing	17
1.6	Phylogeny of the SAR clade	20
1.7	Nuclei features in <i>Paramecium tetraurelia</i>	22
1.8	Genome rearrangements in <i>Paramecium tetraurelia</i> 's sexual development	27
2.1	Map of the double T7 L4440 vector	41
3.1	<i>Paramecium</i> histone H3 modifications	71
3.2	Localization of H3 modifications in <i>Paramecium</i> cells	72
3.3	Nucleosome signals from MNase-seq	73
3.4	Correlation of nucleosome signals to gene expression	75
3.5	Categories of gene configuration	76
3.6	Segmentation analyses of ChIP-seq data	78
3.7	<i>Paramecium</i> 's divergent Po II	81
3.8	Poll II pausing analysis	82
3.9	Heatmaps of epigenetic marks	83
3.10	States of plastic gene groups	84
4.1	Argonaute protein phylogeny	99
4.2	sRNAs from Ptiwi IP experiments	100
4.3	Transgene-associated siRNAs	102
4.4	Antisense ratio of transgene-associated siRNAs	103
4.5	Sequence logos of transgene-associated siRNAs	104
4.6	Uridine-content of sRNAs	105
4.7	sRNAs from endogenous clusters	106
4.8	sRNAs accumulation in genes from various expression groups	107
5.1	Model	126
A.1	MNase digest on isolated nuclei	146
A.2	Nucleosome profile at the TSS in different species	147
A.3	Length of genes in different intron frequency groups	147
A.4	Comparison of MNase-seq pipelines	148
A.5	Neighboring genes expression	148
A.6	Expression levels of genes with different length	149
A.7	Partial correlations of epigenetic marks	149
B.1	Alignment of Ptiwi protein sequences	152
B.2	Alignment of Ptiwi protein sequences (continued)	153
B.3	Antibody specificity	154

B.4	Logos of sense and antisense sRNAs from Ptiwi IPs	155
B.5	Logos of periodate treated sRNAs	155
B.6	SRCs in Ptiwi IPs	156

List of Tables

1.1	Genome features of unicellulars and metazoans	24
2.1	Bacterial Strains	35
2.2	Q5 High-Fidelity polymerase PCR reaction and program setup	36
2.3	Taq polymerase PCR reaction and program setup	36
2.4	List of Oligonucleotides	36
2.5	Feeding fragments cloned into L4440 vector to target genes by dsRNA feeding.	42
2.6	List of injected transgenes (TG).	44
2.7	SDS-gel composition	45
2.8	Peptides for competition assays	49
2.9	Antibodies	50
5.1	Characteristics of sRNA species	125
A.1	External datasets	146

List of Abbreviations

bp	base pair	TG	transgene
cDNA	complementary DNA	TGS	transcriptional gene silencing
cds	coding sequence	TPM	transcripts per million
ChIP	chromatin immunoprecipitation	tRNA	transfer RNA
CTD	carboxy terminal domain	TSS	transcription start site
5'U	5' uridine	TTS	transcription termination site
DNA	deoxy-ribonucleic acid	UTR	untranslated region
dsRNA	double-stranded RNA	WGD	whole genome duplication
GSRC	genes associated with small RNA cluster	WT	wildtype
HAT	histone acetyltransferase		
HM	histone modification		
IES	internal eliminated sequence		
IGV	Integrative genomics viewer		
IP	immunoprecipitation		
LECA	last eukaryotic common ancestor		
MAC	macronucleus		
mcIES	maternally controlled IES		
MIC	micronucleus		
miRNA	micro RNA		
mRNA	messenger RNA		
ncRNA	non-coding RNA		
NDR	nucleosome depleted region		
NGS	next generation sequencing		
NHEJ	non-homologous end joining		
nt	nucleotide		
ORF	open reading frame		
PCR	polymerase chain reaction		
PIC	preinitiation complex		
piRNA	piwi-interacting RNA		
PTM	post-translational modifications		
PTGS	post-transcriptional gene silencing		
RDRP	RNA-dependent RNA polymerase		
RNA	ribonucleic acid		
rRNA	ribosomal RNA		
RNAi	RNA interference		
SE	start-to-end		
siRNA	small interfering RNA		
sRNA	small RNA		
SS	start-to-start		
ssRNA	single stranded RNA		
SRC	small RNA cluster		
TE	transposable element		

Dedicated to Opa Günter

Chapter 1

Background

1.1 Epigenetics

80 years ago, Conrad Waddington postulated, "We certainly need to remember that between genotype and phenotype, and connecting them to each other, there lies a whole complex of developmental processes" and thus introduced the term epigenetics into the field of molecular biology (Waddington, 1942). Today, textbooks describe epigenetics as mechanisms that provide an additional layer of regulation to DNA encoded information (*epi*, Greek, 'upon') without changing the nucleotide sequence but by modifying gene expression and even cell fate and differentiation. It is broadly accepted that the outcome of epigenetic mechanisms is reversible, which is in contrast to the irreversible effects of changes in the DNA sequence itself (Allis and Jenuwein, 2016). To take this to an even more fascinating level, the term 'epigenetic' can be extended to the transmission of epigenetic information to the next generation, a mechanism termed as transgenerational inheritance, expanding the Mendelian gene definition of being "more than just a DNA moiety" (Klar, 1998; Jenuwein and Allis, 2001).

Figure 1.1 summarizes the key players of epigenetic mechanisms, with post-translational histone modifications, DNA methylation, and non-coding RNAs (ncRNAs) being the three major epigenetic modulators. While most epigenetic mechanisms such as DNA methylation have been extensively studied in multicellular organisms, being involved in the definition of varying cellular fates from one single zygote (Reik, Dean, and Walter, 2001), the first histone modifying enzymes were identified in the nuclei of unicellular organisms, showing no cellular differentiation but reacting on external stimuli by rapidly changing their gene expression using epigenetic mechanisms. There is accumulating evidence for cross-talks among components of the epigenetic machinery, e.g. DNA methylation patterns can be directed by histone methylation (Cedar and Bergman, 2009), and small RNA molecules can guide histone enzymes and remodelers to locally induce changes in DNA accessibility (Gutbrod and Martienssen, 2020).

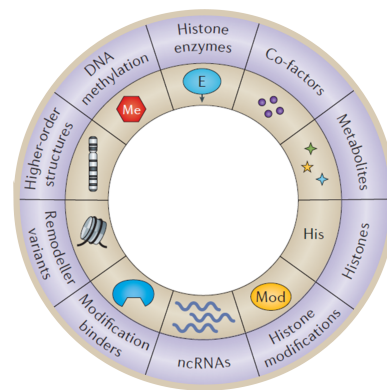


Figure 1.1 Summary of players in epigenetics that orchestrate gene expression in addition to the DNA encoded information. Modified from Allis and Jenuwein, 2016.

1.1.1 Chromatin

The eukaryotic DNA is organized by an orchestra of proteins, forming the chromatin, a structure that helps to nest long nucleotide stretches into the nucleus while regulating accessibility to specific genes and simultaneously protecting others from being exposed to the transcription machinery. The smallest units of chromatin are nucleosomes, of which the first X-ray crystal structure was published by Luger et al., 1997, showing 146-147bp of DNA being wrapped around a protein spool, creating the nucleosome core particle.

Figure 1.2 shows the organization of approximately 147bp DNA wrapped in almost two helical turns around eight histone proteins, a structure that seems to be

highly conserved amongst eukaryotes. Two copies of H2A-H2B and H3-H4 dimers form the nucleosome core, while the linker histone H1 (not shown) is attached to the DNA at the nucleosome entry/exit sites. Two neighboring nucleosomes are separated by a stretch of DNA called *linker*. Proteins with *histone fold* structures can be found in all domains of life, but heterodimerization of histones into the octamer form is found only in eukaryotes (Talbert and Henikoff, 2021a). Canonical histone proteins are quite small ($\approx 11-15$ kDa, 120 amino acids) and are incorporated into the chromatin during the S-phase/replication. They are encoded by multiple genes that produce mRNAs lacking introns and polyA tails. Each histone harbors an amino-terminal residue segment that extends from the surface of the nucleosome by 25-30 (mostly basic), amino acids. Furthermore histone H2A is unique in having an additional ≈ 37 amino acid carboxy-terminal domain that protrudes from the nucleosome (Mannironi, Bonner, and Hatch, 1989).

The nucleosome contains 14 non-covalent contacts to the DNA and the nucleotide sequence, which affects DNA bendability and influences assembly of nucleosomes: A sequence pattern of AA/TT and GC dinucleotides in a 10bp phasing favors positioning while poly(dA:dT) stretches are likely to be less bendable thus disfavoring DNA wrapping around the nucleosome core particle (Segal and Widom, 2009). In principle, the positively charged basic patches of the globular histone domains and negative charge of the DNA phosphate backbone balance themselves, an interaction involved in inter-nucleosome interactions and intra-nucleosomal histone tail-DNA interactions (Peppenella, Murphy, and Hayes, 2014).

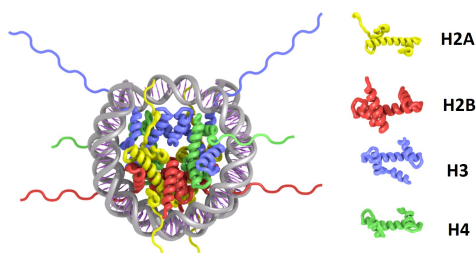


Figure 1.2 Schematic view of the nucleosome structure composed of the histone proteins H2A,H2B,H3,H4, and the DNA wrapped around the histone octamer. Adapted from Draizen et al., 2016.

Nucleosomes can be organized in a periodic manner along the DNA to form a 'bead on a string' fiber that can reach a higher level of compactness by the recruitment of associated proteins such as the heterochromatin protein 1 (HP1), intra-nucleosomal contacts and contacts of histone tails with the DNA. Thereby, the chromatin fiber of 10 nm diameter folds, twists, and coils to reach a degree of compact chromatin domains, probably by transitionally forming a 30 nm fiber, that has not yet been detected *in vivo* but seems to be a higher order secondary chromatin structure as it was shown in *in vitro* experiments (Tremethick, 2007; Hansen et al., 2018) (Figure 1.3).

Specific sites important for genome integrity, such as centromeres and telomeres, show a highly compact chromatin organization in all stages of the cell cycle and are usually associated with low levels of transcription, consequently termed constitutive **heterochromatin**. Additionally, specific sites can show facultative heterochromatinization in terms of cell development and gene expression regulation (Trojer and Reinberg, 2007). Regions more open and accessible for the transcription machinery, called **euchromatin**, are associated with higher gene expression.

Since nucleosomes cover DNA and consequently hide binding sites for regulatory components of the RNA and DNA polymerase machinery, their characteristics were extensively studied in terms of their impact on guidance of transcription, replication, and cell development (Brahma and Henikoff, 2020). Regulation on the chromatin

level is not only orchestrated by the amount or organization of nucleosomes along the chromatin fiber, but also by the incorporation of histone variants and modifications (HMs) of the histone N-terminal tail, protruding the nucleosome core.

1.1.2 Nucleosome Organization

The chromatin landscape is highly dynamic and can be followed by the analysis of changing single nucleosome positions along the genome. The spacing of nucleosomes can be measured by the nucleosome repeat length (NRL), defined as the length of nucleosomal and linker DNA that varies between species in a range of 150bp to 260bp (Szerlong and Hansen, 2011). Thereby, the length of the linker DNA can highly vary between different species and also between tissues at the same time (Szerlong and Hansen, 2011).

Changes in nucleosome positioning are achieved by activation of nucleosome remodelers, transcription machinery and replication fork. Thus, two terms need to be clearly separated when it comes to the biological interpretation of nucleosome array data: occupancy and positioning. Positioned nucleosomes seem to be less mobile and resistant to being removed by remodelers and transcription machinery thereby showing a robust, reproducible profile covering the same DNA stretch in a population of cells. Occupancy, in contrast, describes high turn over of nucleosomes at a specific site of a gene region, with high occupancy simply referring to a high percentage of cells from a population that contain a nucleosome at a given position (Chereji, Bryson, and Henikoff, 2019).

Based on groundbreaking studies in budding yeast *Saccharomyces cerevisiae* and the increasing availability of high-resolution nucleosome positioning maps, a widely accepted dogma was established: promoters showing high nucleosome occupancy result in an *off* state in gene expression, while eviction of those nucleosome patterns can induce gene expression (Henikoff and Shilatifard, 2011). This dogma waters down with accumulating data from different species and single-cell analysis, giving insights into the organization of nucleosomes in regulatory regions and their regular spacing in transcribed and untranscribed regions. It has long been thought that transcription factor (TF) binding sites upstream of the gene to be expressed need to be accessible and therefore located in a nucleosome depleted region (NDR), a pattern that is widely conserved amongst eukaryotes (Talbert, Meers, and Henikoff, 2019). This black-and-white assumption does not seem to properly reflect reality, since pioneer TFs are capable of targeting the nucleosome surface, recruiting chromatin remodelers, achieving partial nucleosome unwrapping from DNA, and facilitate binding of TFs (Brahma and Henikoff, 2020). Subsequent recruitment of remodelers, which themselves push away nucleosomes in an ATP-dependent manner, creates a profile of an NDR flanked by a downstream -1 nucleosome and a +1 nucleosome in the 5' region upstream of the transcription start site (TSS), with following phased (regularly spaced) nucleosomes along the gene body, a pattern that seems to be well conserved from fungi to plants (Dion et al., 2007; Baldi, Korber, and Becker, 2020).

ATP-dependent chromatin remodeling complexes overcome the intrinsic favoring of nucleosome positioning encoded in the DNA sequence itself (Struhl and Segal, 2013). As a consequence of nucleosome movement, the +1 nucleosome is precisely positioned, partially or fully covering the TSS, which is crucial for recruitment of the transcription machinery and TSS selection for gene transcription (Baldi, Korber, and Becker, 2020). ATP-dependent remodelers thereby can regulate nucleosome spacing (ISWI, CHD and INO80), exchange of the histone dimer (INO80), and contribute to octamer eviction (SWI/SNF) with proteins of the SWI/SNF family being already

identified in the last eukaryotic common ancestor (LECA) (Iyer et al., 2008; Talbert and Henikoff, 2021a).

Excursion: How to Find Nucleosome Positions

In general, nucleosome maps, or chromatin landscapes, are profiled by physical accessibility methods using enzymes with specific preferences. The pioneering studies in 1970 (Weintraub and Groudine, 1976) revealed that open, accessible chromatin regions are sensitive to deoxyribonuclease I (DNase I), and till this day, this enzyme is used to separate regions that are protected by TFs from unprotected DNA. With diverse emerging methodical setups, chromatin nucleosome landscapes are described for a wide range of species and even different cell types such as embryonic stem cells, hematopoietic cells, or brain tumor tissue (Wu et al., 2021). Therefore, the methods benefit from high-throughput sequencing approaches, allowing the analysis of entire genomes and the associated proteins (epigenomes) in several days. Limitations occur only for species that are not extensively studied and that lack full genome annotations. The first genome-wide profiling of chromatin in combination with next generation sequencing (NGS) (DNase I-seq) was performed in 2008 (Boyle et al., 2008), whereby the NGS approach allows the rapid readout of each base of millions of DNA sequences in parallel.

Aside from DNase I digestion, the most common approach is the digestion of chromatin, either in native conditions or upon fixation with formaldehyde, with micrococcal nuclease (MNase). This endo- and exonuclease preferentially cuts in between the linker DNA of two adjacent nucleosomes with preferences for AT-rich DNA stretches and can digest up to the size of mononucleotides to the center of the nucleosome core particle. Resulting DNA fragments corresponding to the size of mononucleosomal DNA can be isolated and prepared for NGS approaches (Cuatrecasas, Fuchs, and Anfinsen, 1967; Oberbeckmann et al., 2019). When the obtained reads are aligned back to a reference genome, positions that were covered with nucleosomes can be identified. The limit hereby is the amount of DNA that is needed for the sequencing-readout, sometimes demanding chromatin isolation from millions of cells. In consequence, nucleosome landscapes mostly describe the profile of a cell population rather than from one single cells.

Within the last years, adopted protocols with low input amounts allow for the analysis of single-cell chromatin, probably shedding more light on the dynamic process of gene regulation by chromatin changes simply by subjecting single cells to different conditions. Since there is an ongoing debate about enzyme sequence preferences biasing toward AT-rich DNA, additional methods were developed: using a highly reactive transposase (Tn5), which cuts in between nucleosomes and inserts all vehicle-sequences necessary for NGS, nucleosome-free regions can be detected (Assay for Transposase-Accessible Chromatin using sequencing, ATAC) (Buenrostro et al., 2013). Using Tn5 also allows for the reduction of input material because DNA fragment enrichment is no longer needed prior to sequencing. Therefore, the most popular approach to single-cell chromatin is probably scATAC-seq. The elegant sc-CUT&Tag method uses an antibody-loaded Tn5 direct to specific histone marks. Tn5s loaded with different antibodies can be mixed to allow for the detection of several modifications in one cell, although the subsequently needed bioinformatic power is huge (Janssens et al., 2022). CUT&Tag excludes shearing of chromatin, which is crucial for classical assays using antibodies like Chromatin immunoprecipitation (ChIP).

The number of methods for studying chromatin and generating high-resolution chromatin maps is rising since the 1970s, and methods were extensively reviewed by Minnoye et al., 2021. E.g. NoME-seq or single DNA-molecule foot printing combine de-novo methylation at GpC sites with subsequent bisulfite conversion for detection of unprotected chromatin regions on the single-molecule level (Kelly et al., 2012; Kleinendorst et al., 2021). Methods for chromosome conformation capture such as Hi-C allow for the detection of interactions between single chromosomes upon cross-linking.

Not only DNA can be analyzed by NGS, but also RNA-species of different sizes and biochemical properties. By selectively purifying mRNAs using their polyA tail, transcripts of single cells and populations can be analyzed (transcriptomics) with respect to varying conditions, linking transcriptome changes to adapted nucleosome landscapes (Wang, Gerstein, and Snyder, 2009). The holistic multi-omics approach allows for the analysis of multiple layers of the same population or even single cells. Aside from extensive wet-lab approaches and time-consuming adjustments of protocols, the data obtained from the methods described above are highly complex and must be treated with caution. 'Sequencing reads', basically millions of short nucleotide strings obtained by the sequencing device, have to be aligned precisely to a reference genome. The interpretation of patterns, such as read accumulation in a given region, must be performed in a biological context and assigned e.g. to regulatory regions such as the TSS, enhancers, or promoters. Just to give one example, the DANPOS2 pipeline, designed to precisely determine nucleosome positions and occupancy from sequencing data, performs at least four subsequent steps until one can start interpreting profiles, not including a comparison of different samples, for example, from silencing experiments with each other (Chen et al., 2013).

1.1.3 Histone Modifications

For all histones shown in Figure 1.2, short peptide sequences in the long N-terminal tail of circa 20-35 amino acids that protrude from the octamer core can be chemically modified to achieve different levels of regulation, with modifications such as acetylation, methylation, and phosphorylation already present in the LECA (Iyer et al., 2008; Talbert and Henikoff, 2021b). The combinatorial pattern of those modifications, which can be 'read' by enzymes recruiting downstream effector proteins, is called the histone code. The degree of complexity of this code reaches beyond imagination since several amino acids in each histone tail of all eight histones in one nucleosome can be modified, with those post-translational modifications (PTMs) being newly introduced or erased.

To date, 20 different, covalent histone PTMs have been identified (Huang et al., 2015), with numerous of them being linked to transcription activation, repression, DNA damage response, cell cycle regulation, and DNA replication. These modifications on the one hand, can directly achieve changes in DNA packing by altering the charge of histones, thereby weakening the interaction with negatively charged DNA or, on the other hand, by recruiting specific binding proteins (metaphorically termed *readers*) and associated downstream binding partners (effector proteins) (Talbert and Henikoff, 2021a).

PTMs well studied and extensively documented in the literature are found mainly on histone H3 and H4 tails, which could be due to the lower turnover rate of the H3-H4 dimers compared to H2A-H2B dimers (Talbert and Henikoff, 2017) (Figure 1.3). Huang et al., 2015 list covalent modifications appearing at a minimum of 25 different amino acids, with several of them capable of carrying different modifications,

depending on the regulatory context. While acetylation seems to be exclusively related to transcriptional activation, methylation can be read as an activating or repressive mark: methylation at H3 lysine 4 (H3K4) is usually linked to transcriptional activation in humans, while adjacent methylation of lysine 9 (H3K9) is generally involved in transcriptional repression. The distribution and combinatorial pattern of these marks, together with information on transcription factors and DNA methylation along regulatory regions, can be translated into a road map for cell identity and gene expression (Kundaje et al., 2015).

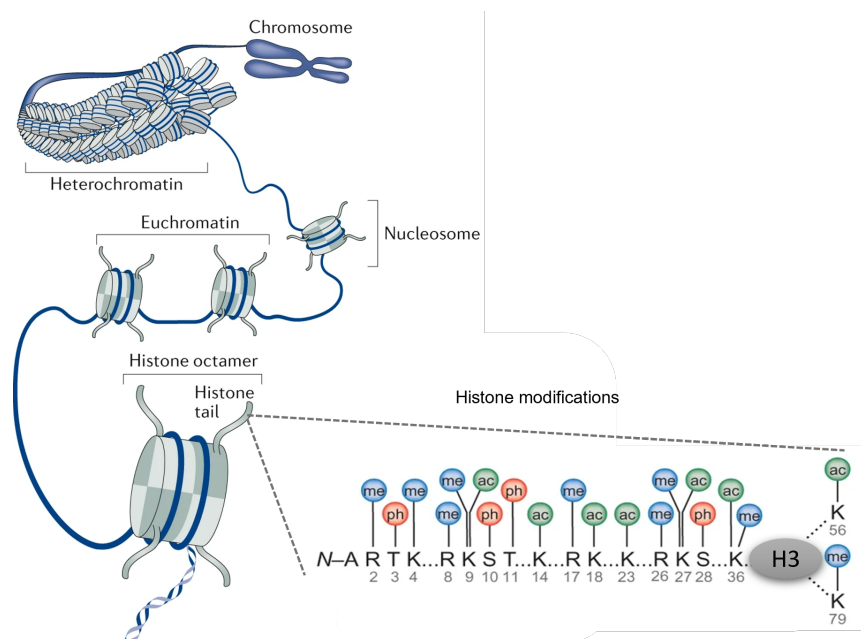


Figure 1.3 Sketch of chromatin organization to the degree of heterochromatin formation, centering on the nucleosome N-terminal modifications at distinct sites. Adapted from Hogg et al., 2020 and Bhaumik, Smith, and Shilatifard, 2007.

From integrating the combinatorial patterns, it must be taken into account that genes can be marked by active and repressive modifications at the same time; a pattern that is termed bistable chromatin, which allows rapid switch from repression to activation during cell development (Sneppen and Ringrose, 2019). This conformation was first described in embryonic stem cells and is often referred to as a bivalent domain or poised state, including information on Polymerase II positioning.

PTMs have been described for both the canonical histones in numerous species, as well as for histone variants, which are encoded by single genes that have introns and a polyA tail and can be incorporated independent of replication. The variants H2A.1/2, H2A.Z, H2A.X, H3.1/2, and H3.3 carry only minor differences in their amino acid sequences to the canonical histones, but their incorporation independent of replication during the cell cycle adds another level of chromatin variability (Huang et al., 2015). The most divergent and universal H3 variant is cenH3 (CENP-A in animals) which is centromere-specifically incorporated and has a longer N-terminal tail than other H3 variants (Talbert and Henikoff, 2021a).

Histone Modifications in Gene Expression

An organism's reaction to stimuli such as changes in temperature and food availability is realized at the level of changes in gene expression, which means switching

between transcriptional activation or repression of genes. Gene expression is regulated at the level of mRNA synthesis, stability, and translation into functional proteins.

Cramer et al., 2000 published a model of the key player in the synthesis of eukaryotic mRNA, RNA Polymerase II (Pol II), a complex comprising 12 subunits, which shows a high degree of conservation at the amino acid level from yeast to humans. Pol II unwinds DNA, synthesizes RNA transcripts, and performs proofreading-processes that are highly orchestrated by the association with general transcription and elongation factors and the writing and reading of histone marks in regulatory regions and gene bodies. Although many studies within the last decades focused on the Pol II transcription machinery in yeast, it becomes apparent that most of these principles are also applicable to multicellular organisms.

Rpb1 and Rpb2, the two largest subunits of the complex, form a cleft for DNA entry and interaction with TFs for transcription start site selection, while Rpb3,10,11 and Rpb12 have anchoring functions. One of the most studied structures in terms of transcriptional regulation is the carboxy terminus of Rpb1, which consists of a flexible linker that is followed by a structure of seven amino acids that are repetitively organized (Spåhr et al., 2009).

The number of heptad repeats of the YSPTSPS consensus sequence in the carboxy terminal domain (CTD) of Rpb1 is quite different between species: while the human CTD is composed of 52 repeats, *Plasmodium yoelii* has only five of them. Breaking the consensus sequence into submotifs like SPxY or YSPx, even more species could be identified, having a repeated structure in their CTD (Chapman et al., 2008).

Although the CTD is not necessary for the catalytic activity of Pol II, it is involved in transcriptional regulation. Almost all residues in the heptad repeats become phosphorylated by different kinases at some points during transcription and the abundance of those modifications changes during Pol II transition on the DNA (Harlen and Churchman, 2017). While deletion of the whole CTD is lethal in many species, cells are able to survive with reduced numbers in heptad repeats (Nonet, Sweetser, and Young, 1987).

Polymerase II Recruitment and Transcriptional Activation

On naked DNA, gene transcription starts by binding of activators upstream of the promoter and TSS, followed by the recruitment of coactivators such as the Mediator complex and chromatin remodelers which promote attachment of general transcription factors (GTFs). Since the Mediator complex has a high affinity for unmodified Pol II CTD, unmodified CTD seems to be involved in transcription initiation, while the introduction of post-translational modifications is linked to productive elongation. Pol II is guided by TFIID, TFIIA, TFIIF, TFIIE, and TFIIB to its binding site to form the preinitiation complex (PIC), and RNA synthesis is initiated once 10-15bp of the DNA is separated into single strands allowing Pol II to pass through (Li, Carey, and Workman, 2007).

During the first 30bp of transcription, the CTD of Pol II is phosphorylated at the Ser5 and Ser7 by a cyclin-dependent kinase that is part of the general transcription factor TFIIH subunit, and Pol II proceeds onto the elongation stage by losing its contacts to general transcription factors (GTFs). The introduced pattern of serine phosphorylation seems to be conserved amongst multiple species, with Ser5 and Ser7 phosphorylation peaking at the TSS and increasing levels of Ser2 phosphorylation along the gene body. Ser5 phosphorylation is thereby essential for the successful recruitment of the capping enzyme. In metazoans, Pol II shows pausing 20-100bp downstream

of the TSS, which is introduced by the negative elongation factor (NELF) and DRB sensitivity-inducing factor (DSIF) comprised of SPT4 and SPT5, with SPT5 being conserved across all kingdoms (Guo et al., 2008). NELF and DSIF, the latter interacting with nascent RNA, DNA, and Pol II, hinder the incorporation of nucleotide tri-phosphates and thereby block Pol II in effective elongation. To release Pol II from promoter-proximal pausing, NELF is phosphorylated by the positive elongation factor B (P-TEFb), resulting in NELF dissociation and by phosphorylation of DSIF, the factor switches into a positive elongation factor for Pol II release. P-TEFb additionally introduces Ser2 phosphorylation, a modification that recruits elongation factors and chromatin modifiers such as the Paf1 complex and histone chaperone SPT6 and FACT complex for promoting elongation.

Elongation of transcription is orchestrated by several conserved factors among eukaryotes, with TFIIS being the first one to be described. TFIIS helps to cleave nascent transcripts from backtracked Pol II, a state where the transcript is mislocated in the Pol II complex and needs to be cleaved. Thereby, TFIIS promotes elongation. Transcription is finally terminated by binding of the cleavage and polyadenylation machinery on phosphorylated Ser2 and Tyr1 to create a native mRNA. Strikingly, most transcription initiation events fail to be productive, as it was shown by photo bleaching experiments in human cell lines. Only 10% of Pol II molecules that load at a promoter successfully initiate transcription, and only 10% of those initiation events convert to elongation (Steurer et al., 2018).

Different from what was mentioned above, transcription does not take place on naked DNA. The nucleosome is a physical barrier that needs to be overcome by the PIC for elongation of transcription, which is promoted by partial eviction of histones that reconstitute on the DNA again with the aid of histone chaperons once the Pol II passed through. For successful transcription initiation, nucleosomes in regions upstream of the TSS are H3 acetylated for high expressed genes as a result of recruitment of histone acetyltransferases (HATs) prior to full PIC assembly. Thereby, acetylation is thought to neutralize positive charges on lysines and reduce histone-DNA interactions *in cis*, while on the other hand this mark can be read by bromodomain-containing factors to recruit effector proteins helping to mobilize nucleosomes. Such effectors could be remodelers like SWI/SNF to promote nucleosome movement and create regions of accessible DNA (Chen, Koutelou, and Dent, 2022).

The elongation of transcription is coordinated by the incorporation of histone variants and histone modifications along the open reading frame (ORF). For most studied organisms, genes show a 5' to 3' gradient of histone mark distribution with changing marks along with an ongoing transcription. Histone H3K4 methylation is introduced by proteins of the Set1 methyltransferase family which are recruited, together with SPT6 and FACT by the PAF1 complex. FACT has been shown to function in the disassembly and reassembly of H2A/H2B dimers in ongoing transcription.

Set1 introduced H3K4 trimethylation (H3K4me3) tends to have a signaling function for recruiting complexes for further transcription, such as the NURF complex (another chromatin remodeler). H3K4 methylation levels are strongly correlated with transcription, with H3K4me3 peaking at the +1 nucleosome. However, H3K4 modifications appear to have little direct effects on gene expression, and rather serve as a scaffold for localization of other proteins to aid gene expression.

Another histone mark positively correlated with transcription is H3K36 trimethylation which is enriched towards the 3' end of the gene body and introduced by Set2 proteins. H3K36me3 prevents intragenic transcription by activating histone deacetylases (HDACS, Rpd3p in yeast), so the interplay of acetylation and deacetylation

seems to be necessary for successful elongation. *Sir2* (Silent Information Regulator; *Sirt2* in humans) was one of the first HDACS identified in yeast. Another protein of this family, *Sirt6*, removes H3K56ac which consequences in NELF stabilization and Pol II pausing at +1 nucleosome. Effector histone acetylases, such as NUA4 and SAGA, are recruited by the Pol II CTD, highlighting the regulatory function of this domain. Patterns of Pol II pausing are also seen at splice sites and nucleosomes in gene bodies, obstacles that are overcome by CTD phosphorylation, and probably also the involvement of NELF. Thus, Pol II pausing helps not only to control elongation itself but also probably maintains a more open chromatin state for effective, robust transcription (Price, 2018).

Transcriptional Repression

Histone modifications and the positioning of nucleosomes are not only involved in the activation of gene expression but also in silencing of not only genes but transposable and repetitive elements and the establishment of constitutive heterochromatin. They hinder spurious transcription at cryptic promoters and prevent activation of transposons, of which subsequent integration into genes would be harmful for the organism.

As mentioned, the balance of histone acetylation and deacetylation is essential for the regulation of the *on* and *off* states of genes, as postulated by Allfrey in 1964 (Allis and Jenuwein, 2016). The first mammalian histone deacetylase, named HD1, was identified in 1996, and HDAC superfamilies were present in the LECA (Talbert, Meers, and Henikoff, 2019), including Silent Information Regulator (Sir) in yeast and sirtuins (Sirt) in humans. One of the most famous examples of the regulatory function of histone acetylation levels is described for X-chromosome inactivation by hypoacetylation in mammalian females. In *Drosophila*, on the contrary, the only X-chromosome is hyperacetylated for transcriptional activation in male *Drosophila* cells, a mechanism called dosage compensation (Talbert and Henikoff, 2021a). This event also involves a long, non-coding RNA, linking two epigenetic key players in one mechanism.

The two histone marks best described in terms of silencing are H3K9 trimethylation, established by the SET domain methyltransferase termed Suppressor of Variegation (SUV39H1/2), and the H3K27 trimethylation, respectively introduced by Enhancer of Zeste (EZH1/2). Both marks are shown to be involved in silencing of repetitive elements in unicellular organisms, probably invented as defense mechanism, and are co-opted for developmental silencing in multicellular organisms. The human SUV39H1 is the first lysine methyltransferase being identified which is homologous to Su(var)3-9 in *Drosophila*.

H3K9me3 is further recognized by the chromodomain protein HP1, which specifically binds methylated histones, bridges nucleosomes, and recruits more methyltransferases for heterochromatin spreading. This kind of spreading can be seen at constitutive pericentromeric chromatin domains, repetitive elements, and also in the regulation of genes, such as in the regulation of the mating-type determination in yeast (Zhang et al., 2008).

H3K27me3 is well described not only for silencing of repetitive sequences originating from transposable elements (TEs) discovered by Barbara McClintock in the 1950s (McClintock, 1950), but especially for silencing in development in multicellular organisms. This histone mark is introduced by the Polycomb repressive complexes PRC1 and PRC2, especially by the methyltransferase Enhancer of Zeste (E(z))

in *Drosophila* (EZH1/2 in mammals) of the PRC2. Polycomb silencing, acting antagonistically to gene activation by Trithorax group proteins, is a system well described in *Drosophila* embryogenesis, maintaining the correct spatial and temporal expression pattern of Hox genes through transcriptional repression, resulting in building of correctly orientated body compartments. In humans, Polycomb silencing has been described for stem cell maintenance and cancer development, including regulation of tumor-suppressor genes (Sparmann and Lohuizen, 2006). H3K27me3 is usually present in large contiguous domains over genes, which form when the human PRC2 member EED binds H3K27me3 and positions the E(z) homolog EZH2 to methylate an adjacent nucleosome, facilitating heterochromatin spreading.

Histone Modifications in Disease and Epigenetic Inheritance

In 2006, the first epigenetic drugs for human cancer therapy became available, such as vorinostat, a histone deacetylase inhibitor that can reactivate aberrantly silenced tumor-suppressor genes. Probably, multiple writers, readers, and erasers can be used as targets for cancer therapy. Varying expression levels of writers are shown to be positively related to cell proliferation in cancer, and changes in readers' expression, such as HP1, lead to chromosome instability and cancer development. Additionally, mutations in H3.3, called onco-histone mutations, appear to alter the binding of specific marks, probably leading to cancer as well (Wang, Allis, and Chi, 2007; Allis and Jenuwein, 2016). However, in drug development, pleiotropic effects on the whole chromatin landscape, not only in cancer-related genes, need to be considered. One has to mention that the mechanism of histone segregation during replication and HM inheritance is still obscure. In S-Phase, replication-dependent de-novo nucleosome deposition occurs subsequently to nucleosome disruption. How PTMs are transmitted to newly synthesized histones is not understood, which is in contrast to the well-described distribution and de-novo methylation of cytosines on the new DNA strands upon replication (Almouzni and Cedar, 2016). It has been shown that H3K9me3 can be transmitted for many generations in the absence of TFs initially guiding the methylation, simply by H3K9me3 and a writer-reader balance in yeast (Ragunathan, Jih, and Moazed, 2015).

PTMs on histones, and other epigenetic features like DNA methylation, have not only been linked to cancer but also to diseases such as obesity, especially in future generations. Studies in mice revealed that the parental diet could reprogram an offspring metabolism by altering DNA methylation levels in the zygote and impacting histone modification patterns. There is an ongoing debate over viral and bacterial infections in mammals regarding alterations of epigenetic patterns in sperm and fetuses, resulting in F1 and F2 altered phenotypes (Katzmarski et al., 2022; Kleeman, Gubert, and Hannan, 2022). This pattern of epigenetic inheritance can also involve RNA species which might be transmitted via sperm, as it was shown in mice. The mechanism of mobile RNA has long been discovered in plants, regulating gene expression and transposon silencing throughout the entire plant life cycle up to generating seeds, but the function of shuttling RNA species is less understood in animals. In mice, the paternal diet affected the offspring's metabolism by altering cytosine methylation patterns on genes crucial for lipid metabolism (Carone et al., 2010). However, the DNA methylation patterns in the sperm itself were not altered across the entire genome, but in relatively few loci resulting in developmental effects in the animal. Additionally, small RNAs traveling with the sperm influence expression patterns in the offspring, and these small RNAs can probably be transmitted

from the parental epididymis (Sharma et al., 2018). In discussion of multigenerational epigenetic transfer, such as F2 effects, it is important to rule out simple plastic responses of the offspring to the maternal uterus environment. Conclusions from paternal effects avoid this issue as fathers often contribute little more than sperm, making studies in mice extraordinary models to examine transgenerational inheritance (Carone et al., 2010). 'You are what your grandparents ate' summarizes that organisms can inherit characters induced by ancestral environments, arguing for a Lamarckian inheritance, which contrasts classical Darwinian evolution and natural selection.

Ongoing Debates on the Function of Histone Modifications

It seems like the description of the interplay of histone modifications involved not only in regulation of transcription, but also DNA replication and repair gets out of hand. But, it is still under debate whether the distribution of marks is the cause or consequence of individual processes. Kornberg and Lorch, 2020, being pioneers in the field of nucleosome chemistry, claim the nucleosome itself has a primary role in gene regulation rather than the introduction of histone modifications. This is based on the central dogma that eukaryotic transcription is shut down by repressive nucleosome positioning and is only activated by the recruitment of positive regulators. Naked DNA must be exposed for gene activation, which is achieved by SWI/SNF proteins that remodel chromatin and the closely related RSC complex, which is necessary for nucleosome removal in activated genes, whereby the RSC creates a NDR and is associated with the -1 and +1 nucleosome (Lorch et al., 2011).

Nevertheless, the existence of a well-positioned +1 nucleosome contributes to gene expression and therefore, the nucleosome is also part of the transcription machinery. A promoter with a +1 nucleosome is a far better template for transcription *in vitro* than the corresponding naked DNA, and assembly of a PIC occurs most efficiently in the presence of a promoter nucleosome (Nagai et al., 2017).

The idea of DNA accessibility in a chromatin landscape is also discussed by Henikoff and Shilatifard, 2011, claiming that histone modifications have to be seen as cogs in a global machinery for fine-tuning of transcription and the overall nucleosome positioning allows for modifications and maintenance of a transcriptional state. The question raises whether PTMs are responsible for differences in chromatin states or differences in changes are simply consequences of a dynamic process. Probably, the truth lies somewhere in the middle.

1.1.4 Cross-talk of Regulators: RNA and Chromatin

The establishment of chromatin domains owing to distinct properties is highly orchestrated not only by the writers, readers, and erasers themselves, but also especially by their guidance to specific sites of action. While the CTD of the Pol II Rpb1 subunit functions as an assembly platform, long and small RNA species guide multi-subunit complexes to their target sites and fine-tune gene activation and repression. The regulatory function of short RNA molecules (sRNAs) of varying lengths among species have been underestimated for a long time since RNA was thought to be unstable and that short RNAs are simply junk. This view has changed with the discovery of sRNA-regulated genome integrity in yeast pericentromeric regions, which are established by a co-transcriptional gene silencing (CTGS) process involving sRNA-guided recruitment of heterochromatin forming complexes. In addition to CTGS, sRNAs can also target mRNAs directly for their degradation, which then happens

on the post-transcriptional level (PTGS). Either CTGS and PTGS are induced by the processing of small RNA molecules from double-stranded RNAs (dsRNA), and the basic mechanisms will be explained below.

RNA Interference

The phenomenon of sequence-specific mRNA degradation or inhibition of its translation induced by long double-stranded RNA, first described in *C. elegans* by Fire et al., 1998, is coined by the general term of RNA interference (RNAi). The authors demonstrated that injection of long RNA as a mixture of complementary sense and antisense strands could induce silencing of the *nuc-22* gene, resulting in a so-called twitching phenotype. In contrast, injection of sense or antisense siRNA alone resulted in only a modest phenotype at high RNA concentrations.

Functions of the RNAi machinery are conserved from unicellular to higher eukaryotes, and RNAi is an ancient pathway which was present in the LECA (Cerutti and Casas-Mollano, 2006), probably equipped with the three key players of the pathway responsible for processing long RNAs: an Ago-like protein/Piwi-like protein, Dicer-like protein and an RNA-dependent RNA polymerase, with the active domains of those proteins being highly conserved across pro- and eukaryotes. Proteins of these families are found in varying stoichiometric amounts in various species. The ancestral function of the RNAi machinery was probably to repress transposons and viruses that produce dsRNA at both the transcriptional and post-transcriptional levels. The machinery was later potentially co-opted for regulation of developmental processes, chromosome and even genome integrity.

Figure 1.4 shows the basic principle of the three major pathways of RNAi with the initial substrate and the processing into small RNA (sRNA) molecules responsible for final targeting of mRNA. The pathways have apparent mechanistic overlaps, can be interconnected, and are rather complex, with several proteins from each family adopted for a specific process, such as *Arabidopsis thaliana* shows both the miRNA and siRNA pathway. Nevertheless, the pathway can also be relatively simple in terms of the miRNA pathway in animals, since the evolution of the acquired immune system for pathogen defense, inducing the interferon response upon a dsRNA trigger, probably limits the need for a somatic RNAi machinery. RNAi begins with the production of small RNA molecules, whether siRNAs, miRNAs, or piRNAs, from distinct long RNA precursors.

Small interfering RNAs (siRNAs) are generated from dsRNA precursors, introduced by bidirectional transcription, hairpin formation with internal dsRNA stretches from hairpin repeat sequences, exogenous dsRNA or dsRNA that is synthesized by a RNA-dependent RNA polymerase (RDRP) using single-stranded RNA as template (Figure 1.4A). dsRNA is then cleaved by a protein of the Dicer family, which cleaves long dsRNA via its RNase III domain into small RNA duplexes of approximately 20-28nt length with a 2nt 3'OH overhang and a 5' monophosphate. The dsRNA terminus is bound by Dicers PAZ domain while two RNase III domains cleave the double-strand; the distance between the PAZ and RNase domain determines the length of the sRNA duplex, varying from \approx 20-28nt among different species. When starting from blunt end dsRNA at one terminus, Dicer cuts along its template, generating small RNAs in a regular manner (phased siRNAs), which induce mRNA cleavage and transcriptional silencing in plants. In both the siRNA and miRNA pathway, Dicer enzymes show a preference for dsRNA structures, although the enzymes show varying nucleotide sequence preferences amongst different species.

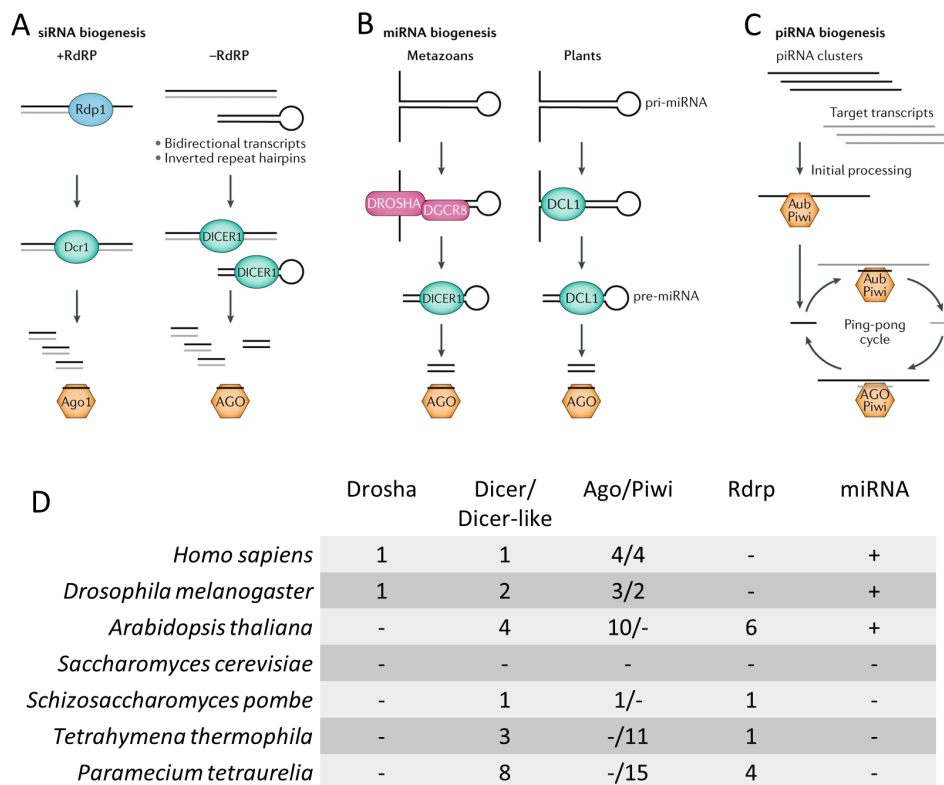


Figure 1.4 (A) Biogenesis of siRNAs (B) miRNAs and (C) piRNAs, including the key enzymes of each pathway. (D) Summary of RNAi components conservation amongst different model organisms. Adapted from Gutbrod and Martienssen, 2020. *Paramecium* data from Götz et al., 2016.

siRNA duplexes are loaded into Argonaute proteins, thus forming the active RNA-induced silencing complex (RISC). One strand - the passenger - is degraded while the other guides the Argonaute protein to its target mRNA in a sequence-dependent manner, resulting in cleavage of the mRNA in the siRNA-mRNA duplex. The place of action of both dicing and loading is believed to occur in the nucleus of *Schizosaccharomyces pombe*, while it was thought to occur in the cytoplasm of mammals. However, Ago shuttling from the cytoplasm to the nucleus along with nuclear Dicer localization have been shown for *Drosophila melanogaster* (Grimaud et al., 2006). In *A. thaliana*, one of the ten Ago paralogs, Ago4, shuttles from the cytoplasm to the nucleus after cytoplasmatic siRNA loading.

Argonaute proteins are present in all domains of life, and the Argonaute protein family consists of two clades: AGO proteins that bind miRNA and siRNAs, and PIWI proteins that bind single-stranded piRNAs and were originally identified in the germline. However, there is a high degree of diversification amongst Argonaute proteins, and a clear separation of Agos from PIWIs is not always possible since proteins seem to lack a high degree of functional conservation. As shown in Figure 1.4D, the level of complexity is quite diverse, with *A. thaliana* having only Agos but no PIWIs, which is the opposite in ciliates such as *Tetrahymena thermophila* or *Paramecium tetraurelia*, the latter having 15 PIWI proteins. It has been observed that different Argonautes can act in distinct silencing pathways by preferentially loading and sorting siRNAs from bulk siRNA pools. Argonaute proteins bind both ends of the siRNA: the 5' end by its MID and PIWI domain, and the 3' end by the PAZ

domain. The selection of the guide strand is believed to be based on the sequence preferences of different Argonautes as well as thermodynamics, whereas the less thermodynamically stable strand is preferentially loaded as the guide strand (Svoboda, 2020). Slicer activity on the targeted mRNA is performed by an RNase-H-like motif of the PIWI domain with a conserved active site of aspartate-aspartate-glutamate (DDE) residues (Song et al., 2004). Target recognition is highly specific, but allows mismatches outside of nucleotides 2-17 in *Drosophila* and at position 1, 14-21 in plants, while mismatches in the middle of a siRNA sequence are not tolerated (Svoboda, 2020).

In contrast to the conserved RNAi pathway, the microRNA pathway (miRNA) was not present in the LECA and evolved independently in animals and plants; and processors of the miRNA pathway in metazoans lack conservation. As indicated in Figure 1.4D, unicellulars lack the miRNA machinery, and most interestingly, *Saccharomyces cerevisiae* completely lacks enzymes of the RNAi pathway, while another yeast, *S. pombe* possesses at least one Dicer, one Ago, and one Rdrp homolog. Despite a lack of conservation, hairpin structures of the miRNA precursors arise from transposons that invaded the genome, highlighting the conserved function of genome defense by RNAi, which is now being co-opted for gene regulation. Thus, miRNAs can regulate $\approx 30\%$ of human protein-coding genes and act on cancer development, as miRNAs have been shown to exhibit lower expression in cancer cells and suppress oncogene expression, thus controlling cell differentiation (Vishnoi and Rani, 2017). In recent years, many breast oncomirs have been detected, which exert their oncogenic activity by targeting tumor-suppressor genes and activating oncogenic TFs (Schooneveld et al., 2015). Targeting mRNAs by miRNAs results in translation repression, degradation, or deadenylation of the mRNA.

miRNAs are cleaved from primary hairpin structures (pri-miRNA) (Figure 1.4B), which are encoded in the genome, e.g., in humans, there are ≈ 500 distinct loci (Griffiths-Jones et al., 2007). The miRNA pathway differs between species: in metazoans, Drosha-Dgcr8 recognizes a single-stranded RNA in a stem-loop structure of the pri-miRNA and generates the initial pre-miRNA that is further processed by Dicer1 into a mature miRNA complex. In plants, the miRNA duplex is generated by two successive Dicer-like 1 (Dcl1) cleavages. The small RNA duplexes are then, similar to the siRNA pathway, loaded into an Argonaute protein. miRNAs have imperfect complementarity to the target transcript, requiring homologous pairing in a so-called *seed* region at nucleotide 2-8 in human Ago2; imperfect pairing leads to transcriptional repression of multiple target mRNAs, while siRNAs cause transcript degradation by most perfect pairing. Mature miRNAs show a 5' U/A bias, which is in close relation to the loading preference of the downstream acting Argonaute protein.

PIWI-interacting RNAs (piRNAs) (Figure 1.4C), in contrast, are generated from single-stranded RNAs produced from genomically encoded piRNA clusters, firstly described in the germline of *D. melanogaster* to silence transposable elements. Although the piRNA biogenesis pathway is less understood than the si- and miRNA pathway, the RISC could be identified as the core component.

Precursor piRNA cluster transcripts are processed into ≈ 23 -29nt antisense piRNAs by a to date not fully understood mechanism and piRNAs are subsequently loaded by PIWI proteins. PiRNAs can enter a complex amplification pathway, called the ping-pong cycle, by which secondary sRNAs target the initial piRNA transcript. The 5' end of a piRNA is generated by the action of an endonucleolytic cleavage, whereas

the 3' end is generated by a second cleavage or by exonucleolytic trimming. The resulting piRNAs are highly diverse in size and sequence specificity but harbor a 5' U preference which is conserved amongst species. piRNAs target a transposon mRNA in the cytoplasm producing sense siRNAs, that show an internal A preference at position 10. Those sense piRNAs target piRNA precursors, resulting in a subsequent release of antisense piRNAs, again promoting further amplification. The ping-pong signature is not observed in *Caenorhabditis elegans*, where germline transposon silencing occurs via piRNAs called 21U RNAs, originating from more than 5,000 loci. If successful induction of transposon slicing needs a seed-like pairing mechanism or full complementarity cannot be answered to this day (Stein et al., 2019). Inhibition of transposon transcription is also achieved by introducing the histone modification H3K9me3 to prevent their spreading/mobilization in the germline, linking piRNAs to changes in the chromatin landscape. PIWI has been shown to interact with HP1A to enforce heterochromatin formation in *Drosophila*, where another variant of HP1 counteractively binds to the piRNA cluster and allows active transcription, since impairing of initially triggering the ping-pong cycle by silencing these clusters would be harmful (Klattenhoff et al., 2009) In animals, the piRNA pathway is crucial for transposon silencing, while plants use the siRNA pathway.

Amplification of the silencing trigger in the siRNA pathway is mediated by RNA-dependent RNA polymerases, which are directed by siRNAs to produce more dsRNA as a template for Dicers. Rdrps are not involved in the miRNA pathway and vertebrates seem to lack Rdrps, although it cannot be ruled out that dsRNA synthesis upon an initial trigger can be performed by other polymerases (Martienssen and Moazed, 2015). The RdrP reaction can extend beyond the sequence complementary to the initial dsRNA (transitivity), such as into upstream regions of the target mRNA.

Furthermore, a new population of siRNAs, called secondary siRNAs acting as an amplification mechanism comparable to the piRNA pathway, might be generated from the extended dsRNA (Sijen et al., 2007). In *A. thaliana*, transitivity occurs in both directions outside the targeted region since long dsRNA is produced by an Rdrp from mRNA cleavage products and newly synthesized dsRNA is diced, while in *C. elegans*, 2° siRNAs are produced Dicer-independent and are a product of an unprimed Rdrp activity. *C. elegans* secondary siRNAs are only found to match upstream targeted regions (Sijen et al., 2007).

RNAi on the Co-Transcriptional Level

The described RNAi mechanism so far appears to act at the post-transcriptional level, repressing genes either by inhibition of mRNA translation or degradation of the targeted mRNA by slicer activities. However, as indicated from piRNA transposon silencing, sRNAs can also induce changes in the chromatin landscape, a process that involves attacking nascent transcripts at distinct loci in the nucleus, thereby acting co-transcriptionally (co-transcriptional silencing, CTGS) (Bühler, Verdel, and Moazed, 2006). CTGS is well described in yeast and plants, but RNAi-induced gene silencing on the chromatin level appears to be challenging to detect in animals and is poorly understood (Woolcock et al., 2011).

By CTGS, RNAi reduces transcription at a given locus by introducing heterochromatin formation. Sequence specificity is thereby introduced by small RNAs that

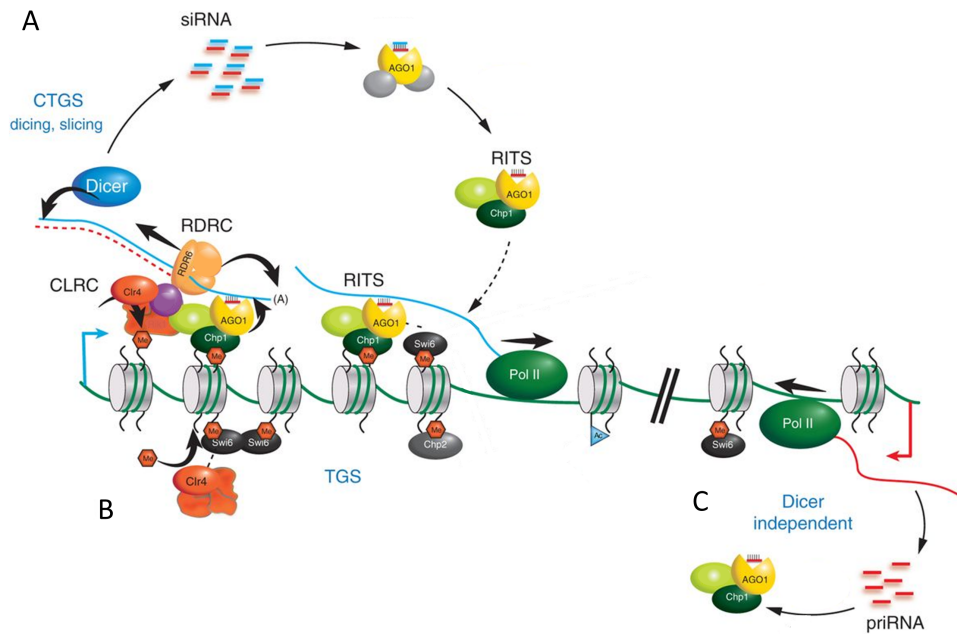


Figure 1.5 Sketch of the CTGS pathway in *Schizosaccharomyces pombe* leading to heterochromatin formation. (A) siRNAs are loaded into an Argonaute protein, forming the RITS that targets nascent transcripts, leading to the recruitment of the RDRC and dsRNA synthesis with subsequent dicing. (B) The RITS complex also recruits H3K9 methyltransferase complex (CLRC) to chromatin and H3K9 methylation and also provides binding sites for Swi6 and Chp1, stabilizing the association of RITS with chromatin. (C) Dicer-independent priRNAs contribute to low levels of H3K9 methylation and may trigger siRNA amplification starting from single strands. See text for more information. Adapted from Martienssen and Moazed, 2015.

target nascent Pol II transcripts at a given locus, recruiting chromatin effector complexes. So at first glance, this mechanism seems contradictory, with introduction of silencing calling for active transcription in the first place.

CTGS is best described in *Schizosaccharomyces pombe* and displayed in Figure 1.5. *S. pombe* has one homolog of the RNAi machinery components each - Dicer, Argonaute, and Rdrp - being essentially involved in the generation of centromeric heterochromatin, which is important for the correct attachment of the kinetochore and chromosome segregation during mitosis. The centromeric and pericentromeric regions are also introduced by an initial dsRNA trigger, which is produced by RNA Pol II bidirectionally transcribing centromeric repeats. As outlined above, Dcr1 generates siRNA duplexes loaded into Ago1, assembling the RNAi-induced transcriptional silencing complex (RITSC), targeting nascent, non-coding transcripts (Bühler, Verdel, and Moazed, 2006).

Tethering the RITSC to a nascent transcript recruits the histone methyltransferase Clr4, which introduces H3K9me3, a mark that spreads along the pericentromeric region and is bound by the HP1 homolog Swi6, resulting in a centromeric heterochromatin formation and the deposition of the centromere-specific histone variant Cnp1. Spreading hereby means the distribution of repressive chromatin state into neighboring regions, resulting in silencing of adjacent genes. This can result in expansion from pericentromeric regions and position effect variegation (PEV), where gene expression patterns variegate due to their positioning next to a heterochromatic region upon genome rearrangements, as it has been extensively studied in *D. melanogaster*.

The initial siRNA trigger is additionally re-inforced by the activity of an RNA dependent RNA polymerase complex (RDRC), which again produces dsRNA on ssRNA transcripts with further cleavage by Dcr1 and siRNA loading by Ago1, which interact with the RDRC. CTGS in *S. pombe* is also crucial for telomere integrity since subtelomeric regions contain regions homologous to the pericentromeric repeats, which facilitates RNAi-dependent heterochromatin formation. Apart from *cis*-silencing, it has been shown, that long ncRNA genes in euchromatic regions and terminal repeats are associated with low amounts of siRNAs and enzymes of the RNAi machinery, probably resulting in gene regulation independent of heterochromatin formation (Woolcock et al., 2011).

The mechanism of CTGS in *S. pombe* shows many parallels to RNA-mediated DNA methylation (RdDM), firstly seen by transgene DNA methylation in *A. thaliana*. Both RdDM and CTGS need active transcription of the locus to be silenced. In *A. thaliana*, RNAi transcripts are generated by RNA Pol IV, which further attacks Pol V-generated transcripts. Again, RNA Pol IV interacts with a Rdrp, which synthesizes dsRNAs, that are diced into 24nt siRNAs by Dicer-like 3 (Dcl3), and exported to the cytoplasm where they are loaded into an Argonaute protein. *A. thaliana* has ten Agos, with Ago4 studied in nucleus shuttling after siRNA loading to target nascent Pol V transcripts. This results in the deposition of DNA cytosine methylation by the DNA methyltransferase DRM2. There appears to be a crosstalk between histone and DNA methylation, as the latter is required for the recruitment of the SUVH4 H3K9 methyltransferase in *Arabidopsis* (Holoch and Moazed, 2015).

RNAi in the nucleus, apart from CTGS and RdDM, is mostly described for silencing processes in the germline. But several studies in *D. melanogaster* also revealed interaction of the RNAi and transcription machinery to repress somatic heat shock genes in non-stress conditions by maintaining Pol II in a paused state dependent on Ago2 and Rdr2. In *C. elegans*, cytoplasmatic loading of Ago NDRE-3 and shuttling to nucleus results in transcriptional gene silencing and H3K9me3 deposition, but the function of this endogenous siRNA pathway is unknown (Burkhart et al., 2011).

Silencing of Loci in *Trans* and Systemic RNAi

As discussed above, it seems like CTGS can occur only in *cis*, meaning at the position where the nascent transcript is generated. In contrast, in *Zea mays*, a phenomenon termed paramutation has been described, resulting in silencing of a locus in *trans*. Thereby, one allele is silenced by the presence of another silent heterochromatic allele in *trans*, a condition that can be transmitted to the next generation. Similar phenomena like quelling or co-suppression have the same outcome, but the mechanism beyond is poorly understood.

Heritable heterochromatin formation can also be induced by systemic RNAi, meaning the movement of sRNAs between adjacent cells and tissues, as has been shown in *C. elegans*, where sRNAs probably also enter the germline to induce heritable epigenetic modifications (Mao et al., 2015). In plants, systemic silencing involves spreading of the silencing trigger, e.g. Ago9-bound siRNAs, between tissues followed by induction of RdDM. sRNA traveling in the plant vascular system resulting in systemic spreading of RdDM probably provides a mechanism for the transmission of stress responses to the germline, affecting stress responses in subsequent generations (Matzke and Mosher, 2014).

1.2 Ciliates as Models in Epigenetics

The principles of epigenetics, such as the involvement of histone code readers and erasers, as well as chromatin conformation changes regulated by RNAi, have been extensively studied in unicellular organisms such as *S. pombe* to extrapolate dogmas that can be applied to complex multicellular organisms. Additionally, studies on morphologically more complex unicellular organisms contribute to our understanding of chromatin biology and clinical research applications. The first gene encoding for a histone acetyltransferase was identified by Allis and colleagues in 1996 (Allis and Jenuwein, 2016), studying the dynamic histone landscape in *Tetrahymena thermophila*. This unicellular organism belongs to the SAR clade, which unites Stramenophiles, Alveolates and Rhizaria and contains an immense diversity of lineages that live in soil, marine and freshwater. The SAR clade is equidistant from plants, animals, and fungi and comprises several important animal and plant parasites (Figure 1.6A) (Grattepanche et al., 2018).

Ciliates, belonging to the Alveolates, have diverged into approximately 8,000 species, some of them with extreme evolutionary distances, such as the distance between *Euplotes* (Spirotrichea) and *Tetrahymena* (Oligohymenophorea) being the same as between rat and corn (Prescott, 1994; Grattepanche et al., 2018). However, ciliates can be characterized by having cilia on their surface for food uptake and motility and by their nuclear dimorphism, which is in principle the differentiation between germline and soma in one single cell. Most species studied on the molecular level to date belong to two classes of Intramacronucleata: Oligohymenophorea (*Paramecium*, *Tetrahymena*) and Spirotrichea (*Oxytricha*, *Stylonychia*, *Euplotes*) (Figure 1.6A) (Drews, Boenigk, and Simon, 2022).

The dimension of nuclear dimorphism can vary drastically in nuclei number and shape, from two micronuclei (MICs) and one macronucleus (MAC) in *Paramecium tetraurelia* to several MACs in Karyorelictea and up to 20 MICs and hundreds of MACs in some Heterotrichea (Prescott, 1994) (Figure 1.6B). The MAC is the transcriptionally active nucleus, comparable to a somatic nucleus, expressing all genes necessary for cell viability and metabolism, while the MIC is transcriptionally silent, comparable to a backup version of the germline genome.

In a vegetative situation, without induction of a sexual event, cells divide their MICs by mitosis, whereas the MAC divides amitotically, by simply expanding and stretching and random distribution of the MAC chromosomes between daughter cells. Intranuclear microtubules assist in MAC shape transformation in Intramacronucleata while microtubules control amitosis outside the MAC in Heterotrichea (Figure 1.6B). In both classes, classical spindle apparatus for chromosome segregation by attachment to kinetochores is not detected. Amitosis is not seen in all ciliate species, a strong exception are Karyorelictea, for example, which build a new MAC upon each cell division (not included in Figure 1.6A).

Paramecia can vegetatively divide up to ≈ 300 times, and it is believed that imbalances in chromosome numbers by amitotic division of MACs seem to limit cell viability, which resembles senescence (Preer, 1976; Sonneborn, 1954). At this time point, as well as other unfavorable conditions such as starvation, *Paramecium* undergoes sexual development, building new MACs from zygotic MICs. This development can involve self-fertilization, called autogamy, or mating between different mating types. While the autogamy process is regularly seen in *Paramecium*, *Tetrahymena* can

divide at least 1,000 times asexually (Long et al., 2013).

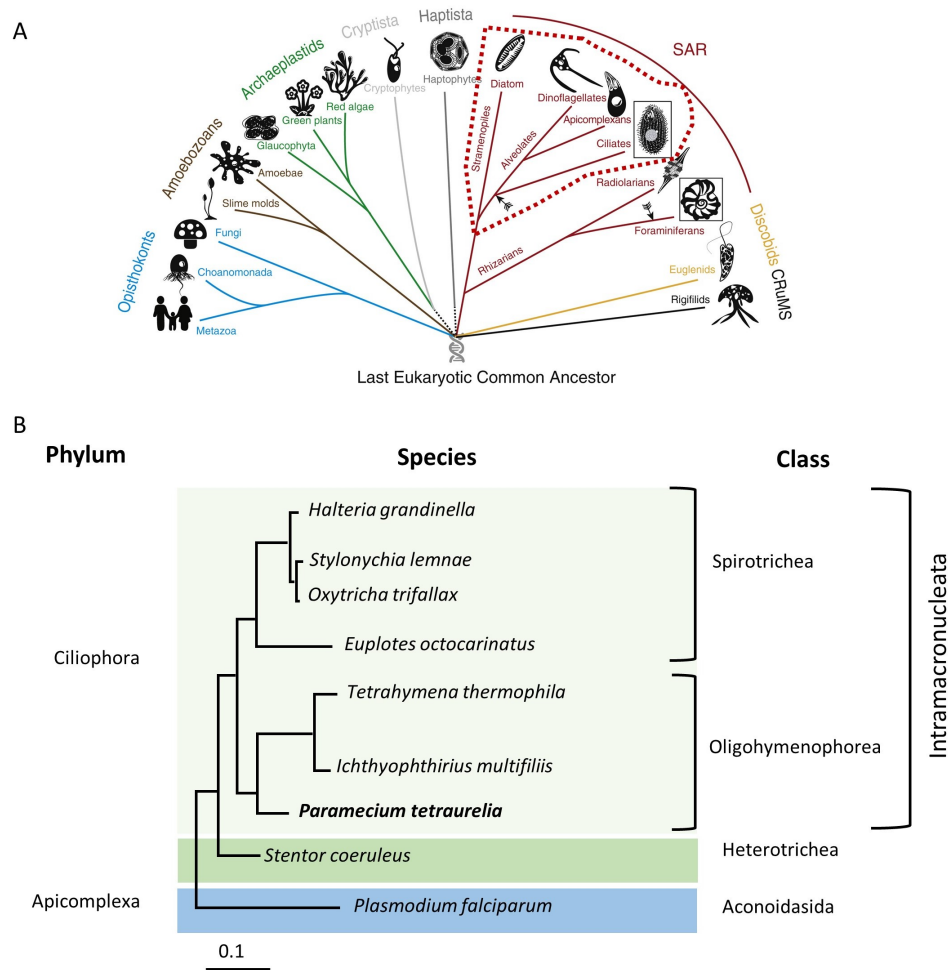


Figure 1.6 (A) Phylogenetic tree of eukaryotes includes the monophyletic SAR clade (uniting Stramenophiles, Alveolates and Rhizarians) with a great variety in diversity of lineages. Phylogenetic tree of eukaryotes based on SSU rDNA sequences. Alveolates, accommodating the lineages Ciliophora, Apicomplexan and Dinoflagellata are highlighted in dashed lines. Arrows indicating the origination of germline-soma differentiation (modified from Cheng et al., 2020). **(B)** Phylogenetic relationship of Ciliophora with highlighting *Paramecium tetraurelia* as the model organism used for studies presented in this work. Apicomplexa are included as an outgroup. Tree was built on multiple sequences alignment of 18S rRNA genes. Adapted from Zheng et al., 2021; Katz, 2001.

Paramecium tetraurelia (Intramacronucleata), being the subject of the following presented studies, is a member of the *Paramecium aurelia* complex, which contains 15 sibling species that are morphologically identical but genetically isolated, excluding interspecies mating. Mating in *Paramecium* depends on the two mating types, even (E) and odd (O), and mating occurs only between two cells of different mating types by conjugation. Conjugations at low levels between siblings in the aurelia complex result in death or sterility of the F1 (Sonneborn, 1975; Catania et al., 2009). Conjugation includes agglutination of cells and exchange of gametic nuclei, but with little exchange of cytoplasm. The inheritance of mating types is epigenetically controlled for some *Paramecium* species, where the new developing MACs in each ex-conjugant almost always become determined for the same mating type as

the parental MAC in that cell and such that the determination of mating type follows the maternal, cytoplasmic pattern (Orias, Singh, and Meyer, 2017). Moreover, mating type determination depends on genome rearrangements in loci responsible for mating-type specific genes, a mechanism that involves components of the RNAi machinery (Singh et al., 2014) and will be illustrated in the following sections. Due to his observations on mating-type inheritance, Sonneborn postulated in 1937 the existence of cytogenes, cytoplasmic determinants involved in heredity, already surmising the molecular mechanism of information shuttling between the different nuclei (Sonneborn, 1937).

Several other phenomena observed in *Paramecium* follow non-mendelian inheritance and involve key regulators of the epigenetic machinery, with the phenomenon of surface antigen expression being extensively studied to this day.

Experiments at the beginning of the twentieth century revealed that injection of *Paramecium* into rabbits results in serum that can immobilize *Paramecium* cells, an effect that was later shown to depend on immobilization by i-antigens or surface proteins (Rössle, 1905). Surface antigens (SAg) cover mainly the cilia of a *Paramecium* cell (Sommerville, 1969), and the observed immobilization reaction is based on a clumping of ciliary membranes and subsequent fusion of plasma membranes at the tips (Barnett and Steers, 1984). Cells from the same homozygous strains are capable of expressing dozens of surface antigens of this protein family, but only one is detectable at the given time (Sommerville, 1969); it has long been postulated that an extranuclear component is involved in expression regulation.

Surface antigen presentation on the outer membrane of unicellular organisms allows for a rapid reaction to external stimuli, such as environmental changes or activation of the host's immune system. By variation of the outer membrane protein surface, pathogens can escape the immune system, as has been shown for *Plasmodium falciparum*, an Apicomplexan that causes malaria in humans. In different unicellular species, antigen variation ('shift') can be regulated on multiple levels, such as co- or post-transcriptionally, involving chromatin remodeling or genome rearrangements. The latter has been shown for *Trypanosoma brucei* (Apicomplexa), when positioning of an actively expressed surface antigen in a telomere region results in gene slicing and antigen shift. *Paramecium* is free-living, non-pathogenic, and the function of antigen presentation and shift cannot be precluded to this day, although surface antigens are certainly involved in sensing external stimuli as well, since paramecia without surface antigens have never been detected (Sommerville, 1969; Simon, Marker, and Schmidt, 2006a).

Paramecium aurelia stock 51 serotypes, which means paramecia with specific i-antigens, were described in 1950 (Sonneborn et al., 1950; Schmidt, 1988), and it was shown that only one immobilizing antigen is present at a given time, implying mutual exclusive expression that follows the cytoplasm of the cell after division and/or conjugation (Beale, 1952). Stock 51 of *P. tetraurelia* is able to express at least 11 serotypes named A, B, C, etc., while alleles are named by combining the serotype letter with the stock number, e.g. 51A (Simon and Schmidt, 2005). The presentation of surface proteins depends on temperature and other cultivation conditions, while one antigen has a range in which it shows the highest stability, and the antigen shift does not occur randomly (Beale, 1952; Sommerville, 1969; Cheaib et al., 2015).

Genes encoding SAgS in *Paramecium* comprise a multigene family of 65 members with eight classical serotype genes (alphas) and six isogenes (Baranasic et al., 2014).

The codon usage allows for high gene expression, and the genes are devoid of introns (Meyer, Caron, and Baroin, 1985; Nielsen, You, and Forney, 1991). Studies in *Paramecium primaurelia* revealed that the central part of the coding sequence consists of tandem repeats, with the element of periodicity occurring every eight cysteines, forming a repeated immunogenic domain in the protein. Thus, the N- and C-terminus of these large proteins (250-300 kDa (Reisner, Rowe, and Sleight, 1969)) is more conserved than the internal sequences. Similar to genes encoding surface antigen in *T. brucei*, subtelomeric localization of surface antigen genes has also been shown for multiple surface antigen genes in *Paramecium* (Meyer, Caron, and Baroin, 1985; Baranasic et al., 2014).

Regulation of expression was initially thought to occur at the transcriptional level, since only mRNA for one i-Antigen could be detected by mercury gels (Preer, Preer, and Rudman, 1981; Meyer, Caron, and Guiard, 1984), although the molecular mechanism of SAg expression is not fully understood until this day. It has been postulated that the variation in serotype expression is dependent on gene rearrangements as it has been shown in somatic recombination in mammalian immunoglobulin genes (VDJ-recombination) or Trypanosoma antigen expression, depending on changes in the 3' end of genes. This model does not seem to hold true for *Paramecium* (Forney et al., 1983). Quite the contrary, it has been shown that the regulation of expression is also post-transcriptionally regulated (Simon, Marker, and Schmidt, 2006b) and further dependent on RNAi components like (i) Dicer, (ii) an RNA-dependent RNA polymerase, (Marker et al., 2010; Baranasic et al., 2014) and (iii) accumulating small RNAs, involving small RNA induced chromatin regulation.

Non-mendelian inheritance, as in mating-type determination, was also firstly described for serotypes in *Paramecium* by Epstein and Forney in 1984. The authors investigated a mutant that did not contain the A i-antigen gene in its MAC, but a complete copy of the gene in its MIC. Nevertheless, the absence of the gene in the paternal MAC resulted in impaired incorporation of the A i-antigen gene into the new macronucleus. This was the first evidence that a mechanism is available in ciliates to control the expression of a gene by regulating its incorporation into a newly formed MAC (Epstein and Forney, 1984; Scott et al., 1994). By manifestation and heredity of serotypes and mating types, paramecia of specific serotypes can be characterized as differentiated cells (Simon and Schmidt, 2007; Drews, Boenigk, and Simon, 2022).

1.2.1 Nuclear Dimorphism in *Paramecium tetraurelia*

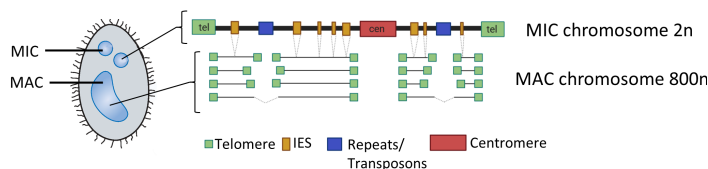


Figure 1.7 Sketch of a *Paramecium* cell with distinct nuclei and a summary of micronuclear (MIC) and macronuclear (MAC) chromosomes features.

2017) but thought to be at least 50 in number (Aury et al., 2006).

P. tetraurelia nuclei show some distinct features displayed in Figure 1.7. The two small ($\approx 3 \mu\text{m}$), genetically identical diploid micronuclei comprise circa 98 Mb of DNA in long chromosomes not fully annotated yet (Guérin et al.,

These MIC chromosomes contain common characteristics of metazoan chromosomes: they possess telomeres and centromeres, the latter probably being associated with the MIC-specific histones CenH3 (Lhuillier-Akakpo et al., 2016).

Further, they harbor ≈ 3 Mb transposable elements (TEs), ≈ 1.3 Mb repeats as well as 45,000 transposon remnants, called internal eliminated sequences (IES) (Guérin et al., 2017; Arnaiz et al., 2012). These sequences are removed when a functional MAC genome is built from a zygote upon sexual development. Since IES are located in genes, their excision must be precise, which is different from *T. thermophila*, where IES are located in intergenic regions enabling imprecise elimination (Chalker and Yao, 2011).

In addition to the consequences of eliminating transposons, repetitive sequences, and IES, newly built MAC chromosomes contain some heterogeneity since, upon elimination, chromosome breakage occurs by de-novo telomere addition at sites that can occur at various positions. These positions are under epigenetic control and stabilized across sexual generations (Meyer, 1992).

While the MIC is diploid, the MAC shows not only heterogeneity but also the extreme feature of polyploidy: prior to excision, the ≈ 200 chromosomes (Arnaiz et al., 2012) are pre-amplified to some extent, and after the chromosomes are processed, the number will increase to up to $\approx 800n$. *T. thermophila*, in contrast, shows only a polyploidy level of ≈ 45 to $90n$, whereas some Spirotrichs possess a polyploidy grade of $\approx 15,000n$. Chromosome copy numbers are kept equal to some extent for Oligohymenophoreans by an unknown mechanism, while Spirotrichs have extreme diverse copy numbers for distinct chromosomes, while numbers positively correlate with gene expression levels across chromosomes (Zhou et al., 2022b).

Amongst ciliates, the degree of chromosome processing from MIC (zygote) to MAC is quite diverse, with *Stylonychia* and *Oxytricha* (Spirotrichs) processing MIC chromosomes to the size of nanochromosomes, with some of them being just size the of one gene. Developmental genome rearrangements cannot only involve excision of sequences but also placing the remaining fragments in a new order, a process called unscrambling. This is seen for multiple fragments in *Oxytricha* and to some extent in *Tetrahymena*, but not in *Paramecium* (Sheng et al., 2020). The information of unscrambled gene organization is thereby again transmitted to the next generation, a process that is epigenetically controlled by RNA species shuttling between distinct nuclei (Nowacki et al., 2008).

Programmed DNA elimination upon genome rearrangements is not only seen in ciliates, but also in early embryogenesis of nematodes to suppress transposon activity and repeats and sea lamprey. The conserved aim seems to be a genome defense against TEs and offers an extreme form of modularity in genome architecture. Of the eliminated genes that have been identified, many are explicitly linked to reproduction (Drotos et al., 2022).

***Paramecium* MAC Chromosomes Have Extraordinary Features**

In *P. tetraurelia* sexual development, $\approx 25\%$ of the MIC genome sequences are removed to build a functional MAC genome, thus constructing a highly condensed genome that is further amplified to a high chromosome copy number (Guérin et al., 2017; Arnaiz et al., 2012). In comparison to other unicellular organisms such as *Tetrahymena* and yeast, *P. tetraurelia* has short intergenic regions of 352bp and the smallest intron size with an average of 25bp (Table 1.1). Of the $\approx 95,000$ introns, only 720 are longer than 40nt (Arnaiz et al., 2017).

The high coding density and high gene number with $\approx 42,000$ genes (Arnaiz et al., 2017) is the result of three successive whole genome duplications (WGD): one that took place prior to the separation of *Tetrahymena* from *Paramecium* lineages, one intermediate, and one that took place before the speciation of the *aurelia* complex. Interestingly, from the recent WGD, many genes remain in duplicate with high protein sequence conservation, and only 32% of genes lost their paralog, which is in contrast to the outcome of WGDs in yeast and plants (Aury et al., 2006).

Coding density is also reduced in humans due to long intergenic regions being involved in gene regulation by harboring functional elements that can even give rise to long ncRNAs (Hangauer, Vaughn, and McManus, 2013). Table 1.1 summarizes the unique features and highlights *Paramecium*'s idiosyncrasy.

Table 1.1 Comparison of ciliates MAC genome features with other unicellulars and metazoans. Adapted from Drews et al., 2022.

	<i>P. tetraurelia</i>	<i>T. thermophila</i>	<i>S. pombe</i>	<i>D. melanogaster</i>	<i>H. sapiens</i>
genome size	72 Mb	103.34 Mb	13.8 Mb	137 Mb	3.1 Gb
protein coding genes	40,460	26,258	4,824	13,947	22,802
mean gene size	1,084 bp	2,451 bp	1,407 bp	6,953 bp	62,825 bp
mean intron size	25 bp	80 bp	81 bp	1,648 bp	3,365 bp
coding density	80%	62%	53%	46%	3.3%
mean size of intergenic region	352 bp	1,403 bp	952 bp	5,548 bp	1,500 bp

1.2.2 Building a Functional MAC Genome Involves RNAi Components

How is the construction of a highly compact MAC genome from a zygote realized, including the excision of thousands of sequences in a precise manner? This process involves the core components of the RNAi machinery, Dicers and Argonautes, and to some extent, the introduction of histone modifications.

In *P. tetraurelia* (Figure 1.8), sexual development starts with meiosis of both diploid MICs, resulting in eight haploid nuclei, whereas seven of them are subsequently degraded. In MIC meiosis, all MIC chromosomes are transcribed bidirectionally, involving transcription factors SPT4 and SPT5 (Gruchota et al., 2017b; Owsian et al., 2022). In parallel, long ncRNA transcripts are generated in the MAC (Lepere et al., 2008) while general MAC transcription continues, although the MAC starts to disintegrate into fragments.

In the MIC, long dsRNA is cleaved by two Dicers, Dcl2 and Dcl3, producing small dsRNA duplexes with a strong 5' UNG signature, 3' 2nt overhang, and the precise length of 25nt (Lepere et al., 2009; Sandoval et al., 2014). Thereby, Dcl2 shows a sequence preference to cleave at the conserved ends of IES, thus enriching the sRNA pool for IES targeting (Hoehener, Hug, and Nowacki, 2018). These sRNAs are further termed scanRNAs, since they perform scanning of the old MAC transcripts in a homology-dependent manner.

Selection of single-stranded RNA is performed by *Paramecium*'s Argonaute proteins: *P. tetraurelia* harbors 15 PIWI proteins (Ptiwi 1-15) but no Agos. Some of these Ptiwis are paralogs from the WGD, and they are not only involved in developmental RNAi but also in silencing upon an exogenous dsRNA trigger. In developmental genome rearrangements, the two paralogs Ptiwi 01 and Ptiwi 09 load ssRNAs. In which

compartment loading occurs has not been shown yet, but shuttling of Ptiwis to load sRNAs in the MIC seems likely, and shuttling to the MIC has at least been shown for the two RNA binding proteins Nowa1/2 (Nowacki, Zagorski-Ostojka, and Meyer, 2005). Subsequent shuttling of sRNA-loaded Ptiwis to the maternal MAC fragments has been shown indeed by GFP-localization studies (Furrer et al., 2017).

In the old MAC fragments, scanning of MAC transcripts occurs: scnRNAs that show homology to the long transcript are degraded while scnRNAs that do not share homology, the ones that are MIC specific, remain Ptiwi-bound. In parallel to scnRNA biogenesis, the one remaining haploid MIC undergoes one mitotic division. A zygote is formed either by exchanging one haploid product with a mating partner and following fusion of nuclei or by fusion of the haploid nuclei in the maternal cell without exchange (autogamy). The zygote divides twice mitotically, rising four nuclei, two of them being the new developing MACs, called Anlagen.

In these Anlagen, after some pre-amplification of MAC chromosomes, MIC-specific scnRNAs target homologous sequences arising from MIC chromosomes that need to be excised. How targeting is achieved is not fully understood, but since RNA:DNA hybridization prior to genome rearrangements has not been shown, it is likely, that nascent transcripts from the developing MAC chromosomes are targeted by scnRNAs (Miró-Pina et al., 2022; Singh et al., 2022). The *Tetrahymena* PIWI protein Twi1 was co-immunoprecipitated with Rpb3, supporting the role of Pol II in ncRNA production and targeting of Twi1 to nascent transcripts (Zhao et al., 2019).

Transcripts are probably synthesized by RNA Pol II depending and elongation factor TFIIS4 (Maliszewska-Olejniczak et al., 2015b). Targeted sequences are excised by a domesticated piggyBac transposase called piggyMac (Pgm) and five associated transposases, termed piggyMac likes (PgmLs) (Bischerour et al., 2018). The Pgm-complex is anchored by the Ku70/80 heterodimer (Marmignon et al., 2014) and induces dsDNA breaks (DSB) at conserved TA dinucleotides at IES boundaries, resulting in overhangs centered around the TA dinucleotide sequence and proteins of the non-homologous end joining (NHEJ) pathway are recruited (Abello et al., 2020). Upon IES excision, remaining MAC-destined sequences (MDS) are ligated by Ligase IV and Xrcc4 in a streamlined fashion while one of the two TA dinucleotides remains in the final MAC genome sequence (Kapusta et al., 2011).

The excised IES are not degraded but circularized, either by forming concatemers of smaller IES or direct circularization of longer ones. These circles are transcribed, and long RNA is processed by Dcl5 into secondary, IES-targeting sRNAs, termed iesRNAs (Sandoval et al., 2014). These sRNAs of ≈ 25 -30nt have a 5' UAG signature and are thought to be an amplification mechanism for precise IES excision, since they target IES in the new developing MAC upon binding by Ptiwi 10/11 (Furrer et al., 2017).

The mechanism of genome rearrangements in *Tetrahymena* is somewhat similar, including the production of early scnRNAs of ≈ 28 nt that can target IES that previously did not contribute to scnRNA production, resulting in the accumulation of 2° sRNAs (late scnRNAs) which in turn can target IES in *trans*, resembling an amplification mechanism (Noto et al., 2015). In addition, *Tetrahymena* has twelve Piwi proteins, but only Twi1 and 11 are involved in genome rearrangements, although it is not clear if they have distinct functions (Bastiaanssen and Joo, 2021).

In contrast to *Paramecium* and *Tetrahymena*, ciliates with gene-sized nanochromosomes, like *Oxytricha* or *Stylonychia*, evolved an opposite mechanism with the same outcome. These species protect sequences that should remain in the new developing MAC by small RNAs. Similar to the *Paramecium* mechanism, *Oxytricha* sRNAs of precise 27nt length are shuttling from the parental MAC to the zygotic MAC, bound

to Otiwi1, one of 13 *Oxytricha* PIWI proteins. Otiwi1 bound sRNAs protect MDS while IES remain unprotected, consequently leading to their excision. Once the IES is excised, the remaining MDS must be sorted to generate a functional MAC genome, involving unscrambling, which is achieved by long RNAs from MIC chromosomes and pointer sequences (Chen et al., 2014).

In *P. tetraurelia*, the excision of IES is driven in a somewhat hierarchical manner: Pgm is required to excise all IES and transposable elements (Arnaiz et al., 2012), while only a subset of IES, termed maternally controlled IES (mcIES), is dependent on the described subset of Ptiwis and Dicers.

IES Excision Upon Heterochromatin Formation

Probably all transposons and a subset of IES ($\approx 70\%$) are dependent of histone modifications and nucleosome remodeling, comparable to transcriptional gene silencing as it is performed by Ptiwi proteins targeting nascent transcripts in the nucleus (Czech et al., 2018), with the extreme outcome of DNA excision and genome rearrangements.

Upon targeting the mcIES by Ptiwi-bound scnRNAs, trimethylation at the histone H3 at lysine K9 and K27 co-occurs. Both modifications are introduced by the histone methyltransferase Enhancer-of-zeste-like (Ezl1), which is probably associated with the chromatin assembly factor 1 (PtCaf1) that guides Ezl1 for methylation by its histone binding domain. Since PtCaf1 is also involved in the upstream scanning process in the maternal MAC, there is accumulating evidence for the RNA-guided DNA elimination linked to changes in chromatin conformation (Ignarski et al., 2014). Pull-down experiments show an interaction of Ptiwi 09 with the PRC2-complex, indicating a sRNA-guided deposition of histone modifications (Miró-Pina et al., 2022). Nevertheless, how Ptiwi finds homologous sequences in the Anlagen is still obscure. In *Tetrahymena*, it has been shown that nascent transcripts are targeted by the RNA helicase EMA1 and probably the involvement of Twi1, which further recruits EZL1 to induce heterochromatin formation (Aronica et al., 2008; Miró-Pina et al., 2022). Introduction of chromatin modification needs to occur in a strict local manner since the majority of IES is shorter than the DNA wrapped around a nucleosome (<150 bp), and only nucleosomes that cover IES should be targeted (Lhuillier-Akakpo et al., 2014). H3K9 and H3K27 trimethylation further recruits or activates an excisase complex comprised of the piggyBac transposase and a histone chaperone of the FACT complex, Spt16-1, that probably mediates chromatin rearrangements, allowing the Pgm-PgmL complex to access the DNA for excision (Vanssay et al., 2020). Opening the chromatin at IES sites is probably also guided by the ISWI chromatin remodeler (Singh et al., 2022).

Almost all transposons, which are excised imprecisely, and $\approx 70\%$ of all IES are dependent on Ezl1, with especially larger IES being dependent on the chromatin conformation changes (Lhuillier-Akakpo et al., 2014; Frapporti et al., 2019). Only $\approx 7\%$ of IES seem to need Dcl2/3 produced scanRNAs and an even a smaller fraction of IES is dependent on Dcl5 generated iesRNAs. Especially transposon elimination seems not to be dependent on Dcl5.

Overall, IES and transposon seem to differ in their recognition mechanism, and transposon elimination seems to be more dependent on chromatin-remodeling as described above. The short IES, being the oldest ones, do not seem to depend on the RNAi machinery in contrast to the younger, longer IES which are transposon

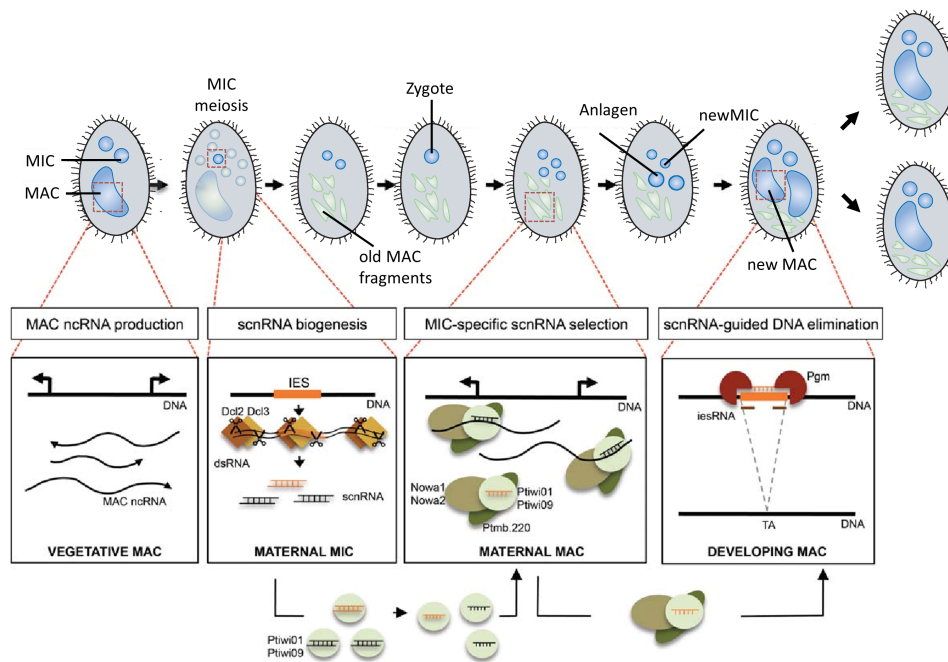


Figure 1.8 Top: Sketch of *Paramecium* developmental stages starting from a vegetative cell (left). Adopted from Drews, Boenigk and Simon, 2022. Bottom: Scanning model for RNA-guided DNA elimination, excluding the RNA helicase Ptmb.220. Adopted from Bétermier and Duharcourt, 2014.

remnants that reside in genic and intergenic regions (Sellis et al., 2021). Whole-genome sequencing revealed that also Ptiwi 01/09 silencing affects only the excision of the Dcl2/3-dependent mcIES and not all IES as previously thought (Sandoval et al., 2014) and only Ptiwi 01/09-silencing, but not Ptiwi 10/11, affects the excision of transposons, coherent with the dependence of transposon elimination on scn- but not on iesRNAs (Furrer et al., 2017; Sandoval et al., 2014).

The scnRNA pathway appears to be a genome defense mechanism that allows the removal of transposons and their relics (IES) from the somatic genome, by which the pathway can regulate cellular genes and also mediate epigenetic inheritance. Although the IES excision machinery is precise, $\approx 7,000$ sites with excision variability have been identified: events like occasional IES retention, excision of IES with alternative boundaries, and cryptic IES, i.e., the excision of MAC destined sequences at TA dinucleotides, contribute to the variability of the MAC genome (Duret et al., 2008; Swart et al., 2014). Coming back to mating-type inheritance, an interesting example has been described for the epigenetic regulation: promoter of the mtA gene becomes excised, similar to the excision of a cryptic IES, during MAC development by scnRNAs and Pgm, thus preventing the mtA expression resulting in the production of mating-type O cells. By this mechanism, the mating type can be inherited by the co-option of the genome rearrangement machinery to regulate gene expression (Singh et al., 2014; Sawka-Gądek et al., 2021). By a coordinated comparison between old and new MACs, *Paramecium* has the chance to control the fitness of its offspring and establish beneficial chromosome conformations, which follows the Lamarckian-based evolution theory (Drews, Boenigk, and Simon, 2022).

1.2.3 RNAi Machinery Apart from Development

Two distinct RNAi pathways have been identified in *Paramecium*, with proteins being exclusively involved in one or the other, but this does not rule out possible overlaps. The pathway described above involves the developmental RNAi machinery, but some RNAi components are also involved in processing exogenously introduced dsRNA and silencing of endogenous genes, a mechanism thought to be originally evolved for virus defense and transposon silencing, mimicking a genetic immune system. Specialization of proteins of the same class to different RNAi pathways is not unique to ciliates but has been described for the mi/siRNA pathways in plants and worms as well (Ketting, 2011).

Exogenous RNAi can be triggered by applying the *feeding* protocol initially described for homologous gene silencing in *C. elegans* (Timmons and Fire, 1998). Thereby, paramecia ingest dsRNA-producing bacteria, and the dsRNA escapes from the food vacuole by an unknown mechanism. The dsRNA is further passed through the endogenous RNAi machinery, leading to gene silencing by homology of the dsRNA trigger to the targeted gene. In *Paramecium*, Dcr1 produces 23nt 1° siRNAs matching the targeted region, which results in a silencing phenotype and the accumulation of 2° siRNAs including 5' to 3' transitivity (Carradec et al., 2015). Accumulating 2° siRNAs show an antisense bias and are mainly the product of Rdrp activity on a nascent transcript without further Dicer dependency. In *A. thaliana* and *C. elegans* 2° siRNAs seem to be the main contributor to a silencing effect, which is in contrast to *Paramecium*, where reduction of 2° siRNAs only had mild effects on silencing phenotype.

Interestingly, *Paramecium*'s 1° siRNAs also seem to be dependent of Rdrps, which seems contradictory because Dicer substrates for 1° siRNA production are already offered as double-stranded substrates. Maybe the initial silencing trigger is imported as ssRNA (Carradec et al., 2015). The feeding pathway also involves three Ptiwi proteins, Ptiwi 13, 12 and 15, with the two latter ones probably performing redundant functionalities (Bouhouche et al., 2011).

Paramecium genes can be either silenced by feeding of dsRNA-producing bacteria or by transformation of the *Paramecium* MAC with specifically designed transgenes. Therefore, truncated transgenes are injected in high copy numbers into the MAC, undergo de-novo telomere addition and are replicated and distributed stably during amitosis similar to pseudochromosomes, until the induction of developmental genome rearrangements (Gilley et al., 1988; Bourgain and Katinka, 1991). Injection of truncated transgenes, lacking the 3' UTR, results in production of aberrant transcripts from both strands, being longer or shorter than the targeted mRNA (Ruiz et al., 1998; Galvani and Sperling, 2001). Here, a set of Rdr3, Ptiwi 13 and Ptiwi 14, as well as Dcr1 is involved in the biogenesis of 1° and 2° siRNAs with the precise length of 23nt (Marker et al., 2010). It was thought that silencing occurs on the transcriptional level, but Götz et al., 2016 showed that a truncated transgene could induce heterochromatin formation at the endogenous locus in dependence of Dcr1, Rdr2 and Rdr3. Specifically, the endogenous 5' coding sequence was depleted for the activating histone marks H3K4me3 upon silencing, while the 3' coding region, not part of the transgene, shows enrichment for the repressive mark H3K27me3.

The *Paramecium* sRNA world seems to be even more complex apart from developmental RNAi or the reaction to exogenous triggers or transgene injection. Recently,

studies of small RNA isolates from wildtype, vegetative cultures revealed 2,602 clusters, that produce small RNAs of precise length that are dependent on two of the four Rdrps. Interestingly, no miRNA producing loci were identified. In an experimental setup, where the RNAi by feeding protocol was applied, subsequent changes on the transcriptome level have been observed. The RNAi machinery appears, therefore, to be involved in both endogenous and exogenous pathways, probably involving a competition between environmental and endogenous RNAi (Karunanithi et al., 2019; Karunanithi et al., 2020).

Endogenous clusters have also been identified in *T. thermophila* generating ≈ 23 –24nt sRNAs that accumulate throughout the life cycle and these sRNAs can be aligned to a number of discrete sites in the genome as well. As in *Paramecium*, these sRNAs also depend on one or more of three distinct Rdrp complexes. Their function seems to be in maintaining chromosome integrity and regulating DNA damage response (Lee and Collins, 2006; Lee et al., 2021).

1.2.4 Studying Ciliates is Fun

Ciliates sometimes seem to be the jack of all trades if one is looking for a favorite model organism: due to their broad range in habitats, they can be studied in terms of ecology and serve as indicators for the health of ecosystems. Several species harbor symbionts or are symbionts themselves, allowing to study host-symbiont or parasitic interactions on the molecular level efficiently.

Most of the model work in ciliate biology is limited to studies using *Tetrahymena* and *Paramecium*, which can be easily cultivated in the laboratory under reproducible conditions. The organisms can be easily kept at several temperatures, can be stressed by heat, starvation, and change of food supplements. In combination with whole genome sequencing approaches and transcriptome alteration analysis, organisms reaction to external stimuli can be studied on the molecular level as well. It is worth mentioning that complete MAC genomes are available for both *Tetrahymena* and *Paramecium*, the latter one still lacking a complete MIC genome sequence.

In both models, some ground-breaking findings were made, cheering for ciliates as model organisms in addition to studying mice, zebrafish or plants. Probably discovery of the telomerase in *Tetrahymena* by Blackburn, Greider, and Szostak, a finding honored by the Nobel prize in 2009, is just one of the most famous examples. *Paramecium* served as a model for genetics and especially epigenetics, the latter already being described in mating-type inheritance, patterns of cilia organization, surface antigen variation, and genome rearrangements, where sequences in the old, maternal MAC control building of a new functional genome. Ciliates have the advantage of studying the transgenerational RNA transfer as parental and zygotic nuclei are in the same cytoplasm. Therefore, ciliates remain the preferred organisms for analyzing the mechanism of RNA transfer, which is the general challenge of transgenerational epigenetics. Specificity, timing, and transport mechanisms for RNA transfer are generally poorly understood, although being a common and essential key player of epigenetics across kingdoms. Using high-resolution techniques on the molecular level will help to answer open questions on the unique and fascinating features of ciliates.

However, especially for *Paramecium*, its genome features are quite different from the features of other model organisms. Indeed, it is important to investigate and understand the regulation of unusual genomes, but sometimes there is no need to connect all findings to other eukaryotes. From time to time, one need to appreciate

that molecular biology dogmas are not set in stone for all the fascinating organisms out there.

1.3 Aim and Outline of the Thesis

The present studies aim to illuminate the organization of the *Paramecium tetraurelia* macronuclear chromatin during vegetative growth. Changes on the chromatin landscape upon sexual development have been described in recent years for several ciliated species, but a description of the *Paramecium* MAC chromatin in vegetative growth is missing. Therefore, methods to isolate MAC chromatin for subsequent antibody pull-downs (ChIP) and the isolation of mononucleosomal DNA by micrococcal nuclease (MNase) digest were established. Crucial parameters for robust, reproducible chromatin isolations were determined.

Protocols were applied to whole-cell culture chromatin and provided information on the global nucleosome landscape in the somatic MAC. Thereby, the isolation of mononucleosomal DNA in combination with library preparation and subsequent NGS approaches allowed for a holistic analysis of chromatin organization in inter- and intragenic regions. By targeting the histone modifications H3K4me3, H3K9ac, and H3K27me3 in pull-down assays, combinatorial patterns of marks regulating gene expression were examined. With regard to the absence of canonical heterochromatin in the MAC, it was questioned how gene silencing and activation are governed. Thus, bioinformatic pipelines were established to analyze nucleosome positioning, transcriptome dynamics, and combinatorial patterns of marks for future global chromatin studies.

Due to characteristics of the RNA Polymerase II CTD, the present study aimed to shed further light on the regulation of Pol II along the condensed MAC genome. Therefore, an antibody against the divergent CTD was produced and applied in immunolocalization assays and pull-downs from chromatin samples. chapter 3 summarizes the identified characteristics of the MAC epigenome.

Furthermore, the vegetative gene expression was manipulated by silencing of an endogenous gene via injection of a truncated transgene. The purpose of this study was to elucidate siRNA-mediated gene regulatory mechanisms shown to result in chromatin conformation changes at the silenced endogenous locus. Two Argonaute proteins involved in transgene-induced silencing, Ptiwi 13 and Ptiwi 14, were characterized in terms of their siRNA loading preferences and localization in wildtype and transgene mechanisms. Tagged Ptiwi proteins were used for pull-downs (IPs) of Ptiwi-bound sRNAs and chapter 4 summarizes the results of sRNA deep sequencing analyses, including the description of sRNA biochemical properties and discussion of diversity in sRNA-mediated silencing mechanisms.

Moreover, studies on vegetative small RNAs originating from distinct loci will be discussed in terms of proteins involved in sRNA biogenesis and Ptiwi-loading. The functionality of these sRNAs in dynamic transcriptome regulation will be discussed in context with chromatin regulation in chapter 5.

Chapter 2

Material and Methods

2.1 Organisms and Cultivation Conditions

2.1.1 *Paramecium tetraurelia*

Paramecium tetraurelia strains (51 and d4-2) were grown at temperatures from 4°C to 31°C in wheat grass powder (WGP) cultivation media, which was freshly bacterized with *Klebsiella pneumoniae* at 31°C the day before use and supplemented with 1x β -sitosterol to promote high cell division rates. *Paramecium* cells were grown in three well depression slides, test tubes or flasks, whereby it was crucial to permit gas exchange.

WGP Extract Stock Solution (20x)

Wheat grass powder	16.6% (w/v)
--------------------	-------------

The solution was prepared with H₂O_{dest} and boiled at 103 °C for 20 min, filtered through multiple layers of gaze and run through a cream separator (Westfalia Separator AG) to remove remaining particles and subsequently autoclaved at 121°C for 20 min.

WGP Buffer Stock Solution (20x)

Tris	95.8 mM
Na ₂ HPO ₄	105.6 mM
NaH ₂ PO ₄	33.3 mM

The pH was adjusted to 7.0 with HCl and the solution was autoclaved at 121 °C for 20 min.

WGP Cultivation Medium (1x)

WGP extract stock solution (20x)	5% (v/v)
WGP buffer stock solution (20x)	5% (v/v)

The medium was autoclaved at 121 °C for 20 min and stored at 4°C until further use.

β -sitosterol Stock Solution (5,000x)

β -sitosterol	0.4% (w/v) in 100% ethanol
---------------------	----------------------------

2.1.2 Bacteria

Unless otherwise stated, *Escherichia coli* were cultivated either in liquid LB medium at 37°C while shaking at 200 rpm or on solid LB agar plates. Selection of strains carrying plasmids was performed by adding antibiotics to the medium. *Klebsiella pneumoniae* were grown from glycerol stocks on angular agar in 15 mL tubes at 31°C over-night and stored at 4°C until the bacteria were used for bacterization of WGP cultivation media.

Table 2.1 Bacterial Strains

Strain	Characteristics	Origin	Application
<i>Escherichia coli</i> Top 10	electrocompetent, tetracycline resistant	Invitrogen	cloning strategy
<i>Escherichia coli</i> HT115(DE3)	electrocompetent, tetracycline resistant	A. Fire laboratory	dsRNA synthesis
<i>Klebsiella pneumoniae</i>	-	-	food bacteria for <i>P. tetraurelia</i>

LB Medium

Peptone	1% (w/v)
Yeast extract	0.5% (w/v)
Sodium chloride	171.1 mM

LB medium was autoclaved and stored at 4°C. LB agar plates were prepared by supplementing media with 1.5% (w/v) Agar-Agar (Carl Roth), boiling and, if desired, adding antibiotics after the medium was cooled down.

Angular Agar for *K. pneumoniae*

Nähragar (Carl Roth)	6.6% (w/v)
----------------------	------------

The agar was dissolved by boiling in water, aliquoted to 5 mL in 15 mL tubes and autoclaved.

Ampicillin Stock Solution (1,000x)

Ampicillin sodium salt	0.01% (w/v)
------------------------	-------------

The antibiotic solution was sterilized by filtration and stored at -20 °C.

Tetracycline Stock Solution (1,000x)

Tetracycline	1.25% (w/v) in 100% ethanol
--------------	-----------------------------

2.2 Standard Molecular Biology Techniques

2.2.1 Polymerase Chain Reaction

The Polymerase chain reaction is used to amplify specific DNA sequences from templates as genomic DNA or plasmids. Short DNA oligonucleotides (Primer) bind to the ends of the targeted DNA sequence and within a defined number of PCR cycles, the DNA Polymerase synthesizes new DNA fragments leading to an exponential amplification. For cloning strategies, Q5 High-Fidelity Polymerase (2 units/ μ L,

NEB) with a low error rate was used, whereas Taq Polymerase (kind gift of Konstantin Lepikhov) was used to check for successful injection of transgenes into the macronucleus (2.3.6). Either purified genomic DNA, colonies of bacteria or single *Paramecium* cells served as templates for PCR reactions.

Table 2.2 Q5 High-Fidelity polymerase PCR reaction and program setup

5X Q5 Reaction Buffer	5 µL	Initial denaturation	98°C	30s
10 mM dNTPs	0.5 µL	Denaturation	98°C	5-10 s
10 µM fwd Primer	1.25 µL	Annealing	50-72°C	10-30 s
10 µM rev Primer	1.25 µL	Elongation	72 °C	20-30 s per kb
DNA Polymerase	0.25 µL	Final extension	72°C	2 min
Template DNA	max. 1 µg			
H ₂ O _{bidest}	ad 25 µL			

Table 2.3 Taq polymerase PCR reaction and program setup

10X Reaction Buffer	2.5 µL	Initial denaturation	94°C	2 min
10 mM dNTPs	0.25 µL	Denaturation	94°C	30 s
10 µM fwd Primer	1 µL	Annealing	50-72°C	10-30 s
10 µM rev Primer	1 µL	Elongation	72 °C	30s per kb
DNA Polymerase	0.5 µL	Final extension	72°C	5 min
Template DNA	max. 1 µg			
H ₂ O _{bidest}	ad 25 µL			

Table 2.4 List of Oligonucleotides

Name	Sequence (5' -> 3')
P13Flagfor	GAGCTCATGTAATAAACTAATCTGAAAATTTGTGA
P14Flagfor	GAGCTCATGTAAAAAATAAGTGATTGCCAAAGAGA
P13 1430 rev	GAATGGTTTGTCTGAATTTGATTCC
pPXV 5' fwd	TAAGATGAATGGAATATAATG
pPXV 3' rev	TTATTTAAGTGTGTTTCATTTA
M13 fwd	GTAAAACGACGGCCAG
M13 rev	CAGGAAACAGCTATGAC
GFP uni fwd	AGGAGAAGAACTTTTCACTGG
GFP uni rev	GAGTATTTTGTGATAATGGTCTGCTA

Oligonucleotides used for library preparation (section 2.7) are not listed. Information can be found in the kit manuals provided by the manufacturers.

2.2.2 Transformation of *E. coli*

Bacteria can be transformed with prepared plasmids by electroporation. Thereby, an electric impulse creates pores in the bacterial membrane, which then allow exogenous DNA to enter the cell. 50 μL of electrocompetent *E. coli* were thawed on ice, and 1 μL of plasmid (1-10 ng DNA) or ligation reaction (subsection 2.2.7) was gently added. After 5 min of incubation on ice, cells were transferred to an ice-cold electroporation cuvette, and an electric impulse of 2.0 kV was applied with an *E. coli* Pulser (Bio-Rad). Cells were recovered by adding 500 μL SOC-Medium and shaken at 200 rpm (1 h, 37 °C). 100 μL of cell suspension was spread out on selective LB plates followed by incubation at 37°C over-night. Colonies were checked for positive transformation by plasmid isolation, colony PCR and sanger sequencing.

SOC Medium

Tryptone	2% (w/v)
Yeast extract	0.5% (w/v)
Sodium chloride	10 mM
Potassium chloride	2.5 mM

The medium was autoclaved, and 20 mM MgCl_2 (sterile filtered) and 20 mM glucose (sterile filtered) were added.

2.2.3 Plasmid Isolation from *E. coli* by Alkaline Lysis

Single bacteria colonies were picked, and transformants were grown in 5 mL selective LB medium over-night at 37°C. 2 mL of bacteria cell suspension was collected (10,000 g, 1 min), and the pellet was resuspended in 340 μL Sol I. By adding 340 μL Sol II, the cells were lysed in SDS at a high pH for 5 to 10 minutes. For neutralization, Sol III was added, resulting in a pH shift to the acidic range. Thereby, proteins, genomic DNA, and cell debris precipitated while the plasmid renatured and remained in the supernatant. By centrifugation (13,000 g, 20 min), the supernatant containing the plasmid DNA was separated from the precipitate, transferred into a fresh reaction tube and the plasmid DNA was precipitated (30 min, -20°C) by adding 1 vol of isopropanol. After centrifugation (13,000 g, 10 min, 4°C) the pellet was washed twice with 70% ethanol (13,000 g, 5 min) and air-dried. The DNA was resuspended in 100 μL 10 mM Tris buffer (pH 8), and the concentration was determined with a microvolume UV/VIS photometer (NanoDrop, Thermo Fisher Scientific).

Sol I (Resuspension Buffer)

Tris	50 mM
EDTA	12 mM

Prior to adjusting the final volume of 100 mL with $\text{H}_2\text{O}_{dest}$, the pH was set to 8 with HCl and 5 mg of RNase A (500 $\mu\text{g}/\text{ml}$, Roche, # 11119915001) was added. The solution was stored at 4°C.

Sol II (Lysis Buffer)

NaOH	200 mM
SDS	1% (w/v)

Sol III (Neutralization Buffer)

Potassium acetate	3 M
-------------------	-----

The pH was adjusted to 7.0 with HCl and the solution was autoclaved at 121 °C for 20 min.

2.2.4 Plasmid Isolation in Large Scale (Midi-Prep)

To isolation of pure plasmid DNA (e. g. without bacterial toxins) in higher quantities, as needed for microinjection into the *Paramecium* macronucleus (2.3.6), the NucleoBond®extra Midi EF from MACHEREY-NAGEL was used following the manufacturers recommendations.

2.2.5 Agarose Gel Electrophoresis

The size of DNA fragments, such as e.g., from PCR, restriction digest of plasmids or DNA from chromatin preparations, was checked by agarose gel electrophoresis. During this procedure, nucleic acids are moved through an agarose matrix by applying an electrical field. The migration pattern is then compared to a mixture of DNA fragments of known length (DNA ladder) that is loaded in parallel. Depending on the expected fragment size, 0.8% to 3% (w/v) agarose gels were prepared in TAE buffer (1x). DNA was mixed with 6x loading dye purple (NEB) supplemented with GelRed (Merck). 80-130V was applied for ≈1 hour and gels were documented using a transilluminator (excitation at 312 nm) with the Gerix 1,000 documentation system (Biostep). Since GelRed is a fluorophore intercalating adjacent nucleotide base pairs, fragments of DNA became visible upon excitation with UV light.

Molecular Weight Marker

Low Molecular Weight DNA Ladder (NEB)
GeneRuler 1kb DNA-Ladder (Thermo Fisher Scientific)

2.2.6 Re-Isolation of DNA from Agarose Gels

To re-isolate DNA fragments of defined size, the MinElute Gel Extraction Kit (Qiagen) was used following the manufacturer's recommendations.

2.2.7 Ligation Procedure and Restriction Enzyme Analyses

PCR products (2.2.1) were re-isolated from agarose gels (2.2.6) and ligated over-night into the respective linearized plasmid using the T4 ligase (NEB). Plasmids were either treated with restriction enzymes (NEB) to validate their correct size, e.g., to check the insertion of the feeding fragment into the double T7 vector (2.3.5) or to linearize transgene carrying vectors prior to injection into the macronucleus (2.3.6).

Ligation Mix

Insert DNA	X µg
Vector DNA	X µg
T4 DNA Ligase Buffer (10x)	2 µL
T4 DNA Ligase	1 µL
H ₂ O _{dest}	ad 20 µL

Ligation was performed at 16°C over-night. Amounts of vector and insert DNA were calculated individually, taking size of DNA and molar ratios into account.

Standard Restriction Enzyme Digestion

DNA	1 µg
NEB buffer (10x)	5 µL
Enzyme	10-20 units per 1 µg DNA
H ₂ O _{dest}	ad 50 µL

Depending on the enzyme properties, restriction digestion was performed at 25°C or 37°C for 1 hour or over-night followed by heat inactivation at 65°C.

2.2.8 Sanger Sequencing

To validate the DNA sequence of the plasmids used for RNAi by feeding or injection into the *Paramecium* MAC, 750 ng of DNA mixed with specific primers were sent to MACROGEN (Amsterdam, Netherlands) for Sanger sequencing (EZ-seq service).

2.2.9 Preparation of Electrocompetent Bacteria

E. coli cells were grown on selective LB plates (+tetracycline) over-night and the next day, 10 mL LB medium was inoculated with one colony and incubated again over-night at 37°C. 1 L of LB medium was mixed with 10 mL of pre-culture and incubated at 37°C until the optical density of the bacterial suspension reached OD₅₉₅ = 0.8. The cells were chilled on ice, centrifuged (4.300 g, 10 min, 4°C) in 50 mL tubes, and the supernatant was discarded. Once all bacteria were harvested into one 50 mL tube, the pellet was washed to remove remaining salts four times with 10% (v/v) sterile glycerol by pipetting and centrifugation; finally, cells were resuspended in 3.5 mL 10% glycerol.

100 µL aliquots were flash-frozen in liquid nitrogen and stored at -80°C.

2.3 Handling of *Paramecium tetraurelia* and *Paramecium* Specific Methods**2.3.1 Growing of Clonal Cell Lines of Defined Age in Mass Cultures**

For all studies presented in this work, it was crucial to keep *Paramecium* cells in their vegetative state, carrying an intact MAC without any meiosis of MICs and MAC fragmentation, as this would be the beginning of autogamy, leading to massive genome rearrangements and synthesis of different kind of RNA species (chapter 1). To produce a *Paramecium* population of distinct age, an adopted protocol from Sonneborn, 1950, refined by Beisson et al., 2010, starting from a cell culture of mixed age was applied.

One single cell from a stock culture was isolated under a binocular microscope,

transferred in 300 μL WGP cultivation media in a depression of a three well depression slide and kept for 24 h at the desired temperature in a humidified chamber. Within 24 h, the cell divided giving rise to several clones, from which one single cell can be isolated and transferred to 300 μL fresh media again. The re-isolation of a single cell is repeated until the end of a week when the clonal cells should have undergone at least 20 divisions.

When the cells divide, they gain the capability to undergo autogamy. This was induced by the end of a week by transferring one cell to 500 μL fresh media and letting it rest for two days without adding new food media. With increasing cell density in the depression, starvation is induced, which leads to induction of autogamy in all cells in one depression. *Paramecium* cells are arrested in that stage of their life cycle until new food is supplied after two days. From this point on, cells end the sexual division and enter the vegetative state again. They will divide up to twenty times until they are able to start autogamy again. By monitoring the division rate, one can grow large cultures of cells omitting the risk of a beginning autogamy.

To grow mass cultures, cells were kept at a density of 50 to 500 cells per mL in large glass flask and were supplemented with fresh food every day. At least three times 100 μL of culture were monitored for cell density. To remove cell debris originating from paramecia and food bacteria, cultures were filtered over two layers of gauze. To completely exchange the media, cells were pelleted in pear-shaped flasks in an oil test centrifuge (2,000 rpm, min; Hettich Rotofix 46), washed in Volvic water (Danone Water, Germany), and pelleted again. The cell pellet can then be quickly transferred to fresh media.

2.3.2 Staining of Nuclei

To monitor cells for their vegetative state or to follow the successful induction of autogamy, staining of double-stranded DNA (dsDNA) with 4',6-diamidin-2-phenylindole (DAPI) was performed. Cells in a drop of 10 μL on a slide were mixed with 2 μL of 0.5 M EDTA, 1 μL of DAPI stock solution and incubated for five minutes in the dark. DAPI binds to AT-rich regions of dsDNA and shows strong fluorescence when excited with ultraviolet light, giving the chance to examine the shape of the MACs and MICs. Without covering, cells in the drop can be quickly examined under the fluorescent microscope at low magnification. The procedure can also be applied to an aliquot of fixed material or to a fraction of isolated MACs, just without adding EDTA.

DAPI Stock Solution

DAPI	1 μg
$\text{H}_2\text{O}_{dest}$	ad 1 mL

The solution was stored at 4 °C and always kept in the dark.

2.3.3 Isolating Serotype Pure Cell Lines

Paramecia express one distinct surface antigen at the outer ciliary membrane, which can be detected by antibodies in a serum. In a three well depression slide, 100 μL of cell culture (maximum 100 cells) were gently mixed with 1 μL of serum. Within the incubation time of approximately 30 min, antibodies bind to the antigen, leading to agglutination of neighboring cilia and the immobilization of the cells. Cells which

are effectively immobilized, sink to the bottom of the depression and are classified as having a specific serotype.

Polyclonal sera from immunized rabbits direct against surface antigen A, B, D, H were used in the presented studies (gift of James D. Forney, Purdue University, USA).

2.3.4 Trichocyst Discharge

Paramecia carry secretory granules under their cell surface, called trichocyst, which release their crystalline protein content into the environment upon induction of different irritations such as change in pH, mechanical stress or contact with predators. This reaction can be easily triggered by adding saturated picric acid (Morphisto) in a 1:1 ratio to a drop of cells. This method was used to monitor silencing efficiency upon transgene-induced silencing or feeding against the *ND169* reporter gene, which is, among other *ND* genes, involved in trichocyst discharge (Bonnemain et al., 1992). If silencing was successfully established, an impaired trichocyst discharge was observed.

2.3.5 RNAi by Feeding

To induce knock-down of a specific gene in *Paramecium*, feeding of double-stranded RNA (dsRNA) via bacteria is a well-established, rapid method, which was published by Galvani and Sperling, 2002 who made use of the RNAi by feeding protocol for *C. elegans* described by Timmons and Fire, 1998.

Since *paramecia* are bacterivores, they ingest *E. coli* producing homologous dsRNA against a target locus; the dsRNA is then escaping the food vacuole by an unknown mechanism and enters the pathway of RNA interference (RNAi).

The bacteria were transformed in advance with the L4440 vector, carrying two T7 promoters in inverted orientation flanking a sequence corresponding to the gene that should be targeted by RNAi (Figure 2.1). Upon induction of the T7 promoters, RNA is massively transcribed, reanneals to dsRNA and accumulates since the *E. coli* HT115(DE3) strain is RNase III deficient.

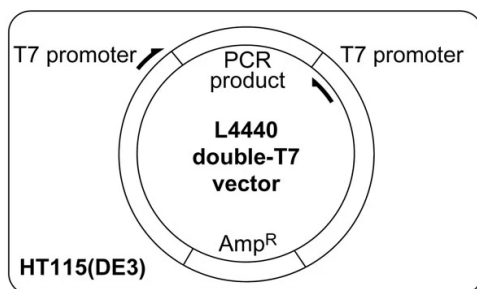


Figure 2.1 Map of the double T7 L4440 vector. A PCR product corresponding to a fragment of the gene that should be silenced by RNAi is cloned in between to inverted T7 promoters. The vector encodes for ampicillin resistance (Amp^R) while the *E. coli* HT115(DE3) strain carries a tetracycline resistance (Kamath et al., 2001).

To prepare feeding media, *E. coli* HT115DE3 were transformed with the respective L4440 vector by electroporation, selected for positive transformation on LB agar plates with ampicillin and tetracycline. As a pre-culture, 5 mL LB (+amp, +tet) was inoculated with one colony and incubated overnight (37°C, 200 rpm). The next day, the desired volume of LB media (+amp) (1/10 volume of the finally needed WGP media) was inoculated with 1:100 of the pre-culture and incubated (37°C, 200 rpm) until the bacteria suspension reached the OD₅₉₅ of 0.38-0.42. Since the T7 polymerase of *E. coli* HT115(DE3) is under control of a lac-repressor that can be blocked by IPTG, the production of

dsRNA by T7 polymerase was triggered by adding IPTG in a 1:400 ratio to the bacteria suspension. Induction of dsRNA production is carried out for additional 2.5 hours. Bacteria were pelletized (3300g, 10 min, 4°C) and resuspended, without any remaining LB media, in WGP media that was supplemented with IPTG (1:400 stock solution), ampicillin (1:1,000 stock solution) and β -sitosterol (1:5,000 stock solution). Paramecia were cultivated in the feeding media for up to four days at the desired temperature, and cultivation can be started in depression slides or flasks.

In parallel to feeding against a gene of interest, the *ND169* reporter gene was silenced as well. Additionally, WGP cultivation media bacterized with *K. pneumoniae* was fed to an aliquote of cells. Feeding phenotypes were monitored by division rate, serotype stability and trichocyst discharge, the latter being expected to be impaired upon feeding against the reporter gene *ND169*.

IPTG Stock Solution (400x)

Isopropylthiogalactoside	5% (w/v)
--------------------------	----------

Validation of dsRNA Synthesis

Prior to the resuspension of *E. coli* HT115 in WGP media, 2 mL of bacteria were pelletized (3,000 g, 10 min) and resuspended in 200 μ L 10 mM Tris-HCl buffer (pH 8.4). The cells were lysed for 10 min at 70°C in the water bath, followed by thoroughly mixing with 1 vol of phenol/chloroform/isoamylalcohol (Carl Roth). After centrifugation (13,000 g, 1 min), the aqueous phase was transferred into a fresh tube, and the RNA was precipitated by adding 1 vol of isopropanol and 0.3 M sodium acetate (pH 5.2). RNA was pelletized, washed twice (13,000 g, 10 min) with 80% ethanol and briefly air dried. The pellet was resuspended in 50 μ L Tris-HCl buffer (pH 8), and the size of the dsRNA was verified on a 1% agarose gel. The dsRNA migrates at the size of the corresponding feeding fragment (Table 2.5) plus the length of the multiple cloning site.

Table 2.5 Feeding fragments cloned into L4440 vector to target genes by dsRNA feeding.

Gene	Gene Accession Number (Paramecium DB)	Fragment Position in the CDS (coordinates)	Size Feeding Fragment (nt)
ND169	PTET.51.1.G0210080	scaffold51 21 from 137893 to 138302	410
PTIWI 13	PTET.51.1.G0480035	scaffold51 48 from 68340 to 69021	683
PTIWI 14	PTET.51.1.G1630015	scaffold51 163 from 26680 to 27402	723
DCR1	PTET.51.1.G0700179	scaffold51 70 from 310907 to 311875	968

2.3.6 Microinjection

Expression of modified genes such as the tagged fusion constructs can be achieved by the transformation of paramecia with DNA that is directly injected into the macronucleus. The ends of the DNA molecules are capped with *Paramecium* telomere sequences and maintained at high copy numbers by autonomous replication (Gilley et al., 1988; Bourgain and Katinka, 1991). By this, stable transformants can be kept while in the vegetative life cycle until MIC meiosis begins and the old MAC carrying the injected DNA molecule is destroyed.

The technique can be used to express GFP or FLAG tagged proteins or even for the silencing of endogenous homologous genes by injection of non-expressible truncated transgenes missing regulatory regions (e.g., truncated transgenes (section 1.1.4)) Ruiz et al., 1998.

100 µg of plasmid was linearized with 40 units restriction enzyme (AhdI; NEB) at 37°C over-night. Complete linearization was verified by loading 5 µL of the digest reaction on an 0.8% agarose gel.

Linearized DNA was extracted with 1 vol of alkaline phenol (vortex; centrifuge 13,000 g, 5 min) and precipitated from the supernatant over-night at -20°C by adding 2.5 vol of ethanol and 1/10 vol of 3 M sodium acetate (pH 9). The pellet was washed twice with 70% ethanol (centrifuge 13,000g, 5 min) and air-dried. To remove all dust particles which could clog the injection needle, the pellet was resuspended in 380 µL sterile water with 1/20 vol of sodium acetate (pH 9) and the solution was then filtered through an UltraFree 0.22 µm MC filter (Merck) (13,000g, 5 min). From that point on, all tubes and tips handling the DNA were rinsed with water in advance to avoid contamination by plastic or dust particles. The DNA was precipitated again, washed with 70% ethanol and air-dried. The pellet was dissolved in 5-10 µL sterile water just prior to the injection.

For injection, young cells of known age (2.3.1) were washed three times in Volvic water with 0.3% BSA and immobilized separately in a drop of 1 µL media on a slide, which was then covered with paraffin. The remaining surrounding media was removed with a micro pipet. Macronuclei of cells were injected using a micromanipulator (Eppendorf) and a microscope. Injected cells were recovered from the slide and washed in a depression with 300 µL WGP cultivation media (0.5x) and transferred to 300 µL WGP cultivation media (0.5x) once. Cells were kept at room temperature for one day and were then transferred to 31°C to grow large cultures or stored at 4 to 16°C to lower cell division rate and avoid any induction of autogamy by high cell densities.

Once the cells underwent several divisions, they were checked by PCR for successful integration of the transgene into the MAC.

2.4 Protein Specific Methods

2.4.1 Total Protein Isolation and Macronuclei Enrichment

To enrich for intact macronuclei, 500,000 *Paramecium* cells were filtered, washed, and starved in Volvic for 30 min. Subsequently, cells were washed in 100 mL 10 mM Tris-HCl buffer (pH 7), and finally collected in 2 mL. Cells were transferred into a pre-cooled potter homogenizer, mixed with 5 mL lysis buffer, and incubated on ice for 5 min. The cellular membrane was destroyed by 30-50 strokes in the homogenizer while the nuclear membranes remained intact. The cell lysate was transferred into a fresh 50 mL tube, topped with 40 mL wash buffer and centrifuged (2,700 g,

Table 2.6 List of injected transgenes (TG).

Name	Specifications	Plasmid Encoded Resistance	Origin
PTIWI 13 TG	pPXV derivate; Plasmid for injection; carries 3x Flag-Tag	Ampicillin	Rapahel Staudt
PTIWI 14 TG	pPXV derivate; Plasmid for injection; carries 3x Flag-Tag	Ampicillin	Rapahel
pTi (-/-) TG	pTI derivate; PLasmid for injection; truncated <i>ND169</i> and GFP under bidirectional promoter	Ampicillin	Simone

1 min, 4°C). The pellet was transferred into a fresh 50 mL tube and washed again twice. In-between, an aliquot of the nuclei pellet was examined under the fluorescent microscope for successful isolation by mixing 10 µL of nuclei with 1 µL of DAPI (1 µg/mL). After centrifugation (8,000 g, 10 min), the supernatant was removed, and the nuclei pellet was directly dissolved in 100 µL Laemmli sample buffer and boiled for 5 min in a water bath.

Total protein was isolated from at least 10,000 cells that were collected in 300 µL volume after washing and starvation in Volvic water. 100 µL of 4x Laemmli sample buffer were added, and the cells were lysed by boiling for 5 min in a water bath. Isolated proteins were aliquoted and stored at -20°C.

Lysis Buffer

Tris-HCl (pH 6.8)	10 mM
Sucrose	250 mM
MgCl ₂	10 mM

The detergent NP-40 was added freshly to a final concentration of 0.2% (v/v).

Wash Buffer

Tris-HCl (pH 7.4)	10 mM
Sucrose	250 mM
MgCl ₂	10 mM

Laemmli Sample Buffer (4x)

SDS	8% (w/v)
Tris-HCl (pH 6.8)	240 mM
Glycerol	40%
Bromphenol blue 0.01%	

Laemmli Sample Buffer (1x)

Laemmli sample buffer (4x)	25% (v/v)
β -mercaptoethanol	5% (v/v)

The solution was stored at RT in the dark for one week.

2.4.2 SDS-Gels for Western Blot

To analyze macronuclear or total protein isolates, sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS PAGE) was performed. By unfolding proteins with SDS and masking their electric charge, proteins can be separated in a polyacrylamide matrix based on their molecular weight by applying an electric field. The discontinuous gels were assembled following the recipes in Table 2.7. The separation gel was poured between two glass plates with defined space (1 mm) Mini-PROTEAN Tetra Cell Casting Module (Bio-Rad) and covered with isopropanol. Once the gel was polymerized, isopropyl alcohol was removed, the separation gel was poured on top, and a comb was inserted to create pockets for later sample application. Precast gels were wrapped in wet paper and stored at 4°C.

Table 2.7 SDS-gel composition

	Stacking Gel 5%	Separation Gel 8%	Separation Gel 16%
Acrylamide (30%, 29:1)	830 μ L	4.8 mL	9 mL
Stacking gel buffer	500 μ L	-	-
H ₂ Odest	3.6 mL	8.4 mL	4.2 mL
SDS (10%)	50 μ L	167 μ L	167 μ L
APS (10%)	10 μ L	100 μ L	100 μ L
TEMED	5 μ L	10 μ L	10 μ L

Precast gels were placed in a Mini-PROTEAN Tetra Vertical Electrophoresis Cell (Bio-Rad), covered with running buffer and equilibrated for 30 min at 70 V. In the meantime, protein samples were thawed and centrifuged (13,000 g, 10 min) to pelletize insoluble components. Up to 30 μ L were loaded on the gel while empty pockets were filled with 1x Laemmli sample buffer. Once the proteins synchronously entered the separating gel, 120 V were applied for up to 1.5 hours. To follow the separation of proteins and size estimation, 3 μ L Broad Range Color Prestained Protein Standard ladder (NEB) was loaded in parallel.

Running Buffer

Tris	25 mM
SDS	3.5 mM
Glycine	192 mM

The pH was adjusted to 8.27.

Stacking Gel Buffer

Tris	0.5 M
------	-------

The pH was adjusted to 6.8.

Separation Gel Buffer

Tris	1.5 M
------	-------

The pH was adjusted to 8.8.

2.4.3 Western Blot

Proteins separated by SDS PAGE were further analyzed by blotting them onto a membrane and labeling them immunologically with antibodies, a procedure called Western blotting. Gels were washed twice in blotting buffer and Whatman filter paper were equilibrated in blotting buffer aside from the nitrocellulose membrane. Three layers of Whatman filter paper were stacked in a transfer cell, followed by membrane, gel and three layers of Whatman filter paper. The transfer cell was wetted with blotting buffer, tightly sealed and 150 mA were applied for 1.5 to 2 hours. Afterward, the membrane was soaked in Ponceau S for 10 min to verify protein transfer, followed by subsequent washing in milk for 10 min three times. Thereby, free binding sites on the membrane were blocked entirely by protein from milk. The primary antibody targeting *Paramecium* proteins or the introduced FLAG-tag was hybridized over-night at 4°C with mild rotation followed by three subsequent washing steps in wash buffer, 10 min each, to remove unbound antibody. The secondary, peroxidase coupled antibody, was hybridized for 1 h at RT. The membrane was washed again three times and covered with WesternBright Sirius substrate (Advantsta) following the manufacturer's recommendations. The chemiluminescent signals were detected by a CCD-sensor and was capture with the Amersham Imager 600-Systems (GE Healthcare).

The HRP substrate was removed by subsequent washing steps, and membranes were stored for additional antibody decorations.

Blotting Buffer (10x)

Tris	250 mM
Glycine	192 mM
SDS	17 mM

The pH was adjusted to 8.27.

Blotting Buffer (1x)

Blotting buffer (10x)	10% (v/v)
Methanol	20% (v/v)

Ponceau Staining Solution

Ponceau S	1% (w/v)
Acetic acid	5% (v/v)

TBS (20x)

Tris	0.5 M
NaCl	3 M

The pH was adjusted to 7.5.

Wash Buffer

TBS (20x)	5% (v/v)
Tween-20	0.05% (v/v)

Blocking buffer was prepared by dissolving 5% (w/v) milk powder in wash buffer and passing the solution through filter paper.

2.4.4 Affinity Purification of Polyclonal Peptide Antibodies

Synthesis of peptides and immunization of rabbits was carried out by the group of Dr. Martin Jung, Medizinische Biochemie und Molekularbiologie, Universität des Saarlandes, Germany. Antibodies were affinity-purified from antisera collected at different time points after immunization using the SulfoLink coupling resin (Thermo Fisher Scientific).

10 mg of each peptide with a terminal cysteine used for immunization were dissolved in 2 mL coupling buffer. The coupling resin was prepared as recommended by the manufacturer and set up in a gravity-flow column. The resin was equilibrated by adding three times 2 mL coupling buffer. The column was closed, and the resin was incubated with 10 mg peptide in 3 mL coupling buffer for 15 min at RT with mild agitation, followed by incubation for 30 min without agitation. To block un-specific binding sites, the resin was washed twice with 4 mL wash buffer followed by incubation with 2 mL L-Cysteine solution for 30 min at RT with mild agitation followed by 30 min incubation without agitation. The resin was washed four times with 4 mL wash buffer and 0.05% NaN_3 and sealed with a frit on top.

To bind specific antibodies, the column was equilibrated with 20 mL of solution A at 4°C. 10 ml of the antiserum was diluted with 90 ml of solution A and applied to the equilibrated column using a peristaltic pump (Minipuls2, Gilson, Middleton, USA). The application speed was adjusted to ≈ 10 -15 ml/h. Once the flow-through was finished the next day, the column was washed successively with 20 mL solution A and solution C, each with a flow rate of 15-30 mL/h. The column to which the specific antibodies were bound was first washed with 20 mL of solution D leading to the elution of acid-sensitive antibodies from the resin. The eluates were collected as 1 mL aliquots by direct dropping into fresh tubes with 100 μL 1M Tris buffer (pH 8) to bring antibodies to a less harmful pH range as quickly as possible. The column was washed with 20 mL solution A supplemented with thimerosal. The column was stored at 4°C.

The resulting eluates were measured at the NanoDrop (Thermo Fisher Scientific), frozen at -20 °C and the individual fractions were tested for antibody specificity by Dot Blot and Western Blot analysis. In addition, the washing step (with solution C) can also be tested for immunoreactivity.

Coupling Buffer

Tris	50 mM
EDTA	5 mM

The pH was adjusted to 8.5.

Wash Buffer

NaCl	1 M
------	-----

L-Cysteine Solution

Cysteine	50 mM in coupling buffer
----------	--------------------------

Solutions for Coupling and Elution

Solution A	Tris-HCl	10mM	pH 7.5
Solution B	Tris-HCl	10 mM	pH 8.8
Solution C	Tris-HCl	10 mM	pH 8.8
	NaCl	0.5 M	
Solution D	Glycine	100 mM	pH 2.8 with HCl

Sodium Azide Solution

NaN ₃	0.05% (v/v)
------------------	-------------

2.4.5 Competition Assay and Dot Blot

To validate the specific reactivity of antibodies with the synthesized (Dr. Martin Jung, Homburg) or purchased peptides (Diagenode), rapid immunoblotting was performed, by dotting different amounts of peptides in 5 µL blocking buffer onto a nitrocellulose membrane. The membrane was blocked and decorated with primary and secondary antibodies as described (2.4.3).

Peptide competition assays were performed by blocking 2 µg of each antibody with a 10-fold excess of its corresponding peptide over-night at 4°C in milk with agitation. 1 pmol to 100 pmol of each peptide were blotted on a nitrocellulose membrane and decorated with blocked and unblocked antibodies as described (2.4.3).

Table 2.8 Peptides for competition assays

Name	Sequence	Origin
HsH3K9ac	NA, region of human histone H3	Diagenode, #C15410004
HsH3K27me3	NA, region of human histone H3	Diagenode, #C15410195
HsH3K4me3	NA, region of human histone H3	Diagenode, #C15410003
PtRBP1	SPHYTSHTNSPSPSYRSS-C	Dr. Martin Jung, UdS
PtPTIWI 13	C-DDAPPQARKNNKSPY	Dr. Martin Jung, UdS
PtPTIWI 14	C-QNWMQRLTAEIGDK	Dr. Martin Jung, UdS
PtH3K27me3	C-TKAARK(me3)TAPAVG	Dr. Martin Jung, UdS
PtH3K27me1	C-TKAARK(me)TAPAVG	Dr. Martin Jung, UdS
PtH3K27ac	C-TKAARK(ac)TAPAVG	Dr. Martin Jung, UdS

2.4.6 Immunostaining

Indirect immunofluorescence for localization of proteins in fixed cells was performed by using the protocol published by Frapporti et al., 2019. 10,000 *Paramecium* cells were washed and starved for 30 min in Volvic and collected in a final volume of 500 μ L. All following steps were carried out with gentle agitation. For permeabilization and mild fixation with 1% FA and 1.25% Triton, 500 μ L of permeabilization solution was added, and the cells were incubated for 30 min. For final fixation, 1 mL of the cells were transferred into a 15 mL tube and mixed with 7 mL of fixing solution to achieve cross-linking by 2.5% formaldehyde. After 10 min of incubation at gentle agitation, cells were washed twice for 5 min with blocking solution and collected by centrifugation (2,000g, 2 min). Fixed cells were stored in 1 mL blocking solution for up to one month at 4°C.

50 μ L of cells were transferred in a depression of a three well depression slide, the primary antibody was added to the 50 μ L at the desired final concentration and topped up with 300 μ L blocking solution. Cells were incubated with the antibody over-night at 4°C at gentle agitation. After washing by transferring cells with a micropipette into 300 μ L fresh blocking solution twice, 50 μ L of cells were incubated with the secondary antibody in a total of 300 μ L. After 1 hour of incubation in the dark and subsequent washing, cells were collected in 10 μ L and transferred to a thin glass slide. To preserve fluorescence, 1 μ L of mounting media (VECTASHIELD; Vector Laboratories) was added together with 1 μ L of DAPI (0.2 μ g/mL) to stain the nuclei. Cells were covered with a coverslip, sealed with nail polish, and stored at 4°C in the dark or were immediately examined under the fluorescence microscope (Axio Observer; Zeiss). Pictures were taken at a 400x or 630x magnification with and without Apotome2 (Zeiss). Exposure times and intensity of the LED were kept at the same values in-between the observation of the same set of immunostainings to assure comparability.

Table 2.9 Antibodies

Name	Host	Immunogen/Specification	Dilution	Origin
Polyclonal, ChIP-seq grade antibodies				
H3K9ac	rabbit	Sequence NA; ab against histone H3, acetylated lysine 9	1:2,000 WB, IF 2 µg ChIP	Diagenode, #C15410004
H3K27me3	rabbit	Sequence NA; ab against histone H3, trimethylated lysine 27	1:2,000 WB, IF 2 µg ChIP	Diagenode, #C15410195
H3K4me3	rabbit	Sequence NA; ab against histone H3, trimethylated lysine 4	1:2,000 WB, IF 2 µg ChIP	Diagenode, #C15410003
H3K9me3	rabbit	Sequence NA/ab against histone H3, trimethylated lysine 9	1:2,000 WB, IF 2 µg ChIP	Diagenode, #C15100146
IgG	rabbit	spectrum of IgG subclasses	2 µg ChIP	Diagenode, #C15100146
Custom antibodies				
RPB1	rabbit	SPHYTSHTNSPSPSYRSS-C; ab against <i>Paramecium</i> Polymerase II subunit Rpb1	1:250 WB, IF 10 µg ChIP	MJ
PTIWI13	rabbit	(C)-DDAPPQARKNNKSPY; ab against <i>Paramecium</i> Ptiwi13	1:250 WB, IF	Intavis peptides; MJ
PTIWI14	rabbit	(C)-QNWMQRLTAEIGDK; ab against <i>Paramecium</i> Ptiwi14	1:250 WB, IF	Intavis peptides; MJ
Others				
anti-FLAG, monoclonal	mouse	DYKDDDDK/ IgG1 subclass	1:500 WB, IF, 1 µg IP	Sigma, #F3165
anti-Tubulin, monoclonal	mouse	C-terminal α and β tubulins (glutamylated motif of α -tubulin)	1:1,000 WB, IF 1:100	Sigma, #T9822
Secondary antibodies				
anti-Rabbit IgG (H+L), polyclonal	goat	peroxidase-conjugated	1:5,000 WB	Jackson Immuno research, #111-035-045
anti-Mouse IgG (H+L), polyclonal	goat	peroxidase-conjugated	1:5,000 WB	Jackson Immuno research, #115-035-062
anti-Rabbit IgG (H+L), polyclonal	goat	F(ab') ₂ fragment; Alexa Fluor 594	1:2500 IF	Thermo Fisher Scientific, #A-21069
anti-Rabbit IgG (H+L), polyclonal	goat	F(ab') ₂ fragment; Alexa Fluor 568	1:2500 IF	Thermo Fisher Scientific, #A-11072

WB - Western Blot, IF - Immunostaining; MJ - Dr. Martin Jung, School of Medicine, Medical Biochemistry and Molecular Biology, Saarland University, Homburg, Germany.

PHEM Buffer (4x)

EDTA	40 mM
HEPES	100 mM
PIPES	240 mM
MgCl ₂	8 mM

pH was adjusted to 6.9, the solution was sterilized by filtration and stored at -20°C.

Permeabilization Solution (2% FA, 2.5% Triton)

PHEM buffer (4x)	25% (v/v)
Sucrose	4% (w/v)
Triton-X-100	2.5% (v/v)

Ingredients were mixed freshly for each experiment and 5 mL of 4% paraformaldehyde in PBS were added to achieve a final concentration of 2% FA in 10 mL.

Fixing Solution (2.8% FA, 0.8% Triton)

PHEM buffer (4x)	25% (v/v)
Sucrose	4% (w/v)
Triton-X-100	0.8% (v/v)

Ingredients were mixed freshly for each experiment and 7 mL of 4% paraformaldehyde in PBS were added to achieve a final concentration of 2.8% FA in 10 mL.

Blocking Solution

Bovine Serum Albumin (BSA)	2% (w/v)
Tween-20	0.05% (w/v)
TBS (20x)	5% (v/v)

2.4.7 Expression of Tagged Proteins and Immunoprecipitation

The pull down of fusion proteins carrying a known short peptide sequence (Tag) was performed with specific antibodies directed against the Tag epitope.

For the expression of recombinant proteins studied in the following chapters, pPXV vectors containing the open reading frame of a *Paramecium* gene and the coding sequence for three FLAG-Tag sequences at the recombinant proteins amino terminus (3x DYKDDDDK) were injected into the MAC of young, vegetative cells (2.3.1). Vectors were received as a gift from M. Valentine and J. Van Houten, Vermont, USA, and were successfully injected by Dr. Martin Simon into the vegetative MAC (2.3.6). Injected clones were screened by PCR on single injected cells after some divisions (2.2.1), and positive transgenic lines harboring the FLAG-fusion construct and/or the pTI^{-/-} transgene were grown for cell fixation for immunostaining (2.4.6, protein isolation (2.4.1) and immunoprecipitation.

Immunoprecipitation was performed by Dr. Martin Simon, applying the protocol adopted from Furrer et al., 2017. 500,000 cells were harvested in 2 mL lysis buffer and snap-frozen in liquid nitrogen. 1 mL of the cell lysate was cracked in a Douncer homogenizer and 1 mL was sonified in parallel until all visible MACs were destroyed. The lysate was centrifuged (15,000 g, 15 min, 4°C) and 50 µL of anti-FLAG M2 Magnetic Beads (#M8823, Sigma) were added to 1 mL supernatant and incubated overnight by gentle agitation at 4°C. The beads were washed with wash buffer five times

and finally re-suspended in 100 μ L. 10 μ L were mixed with 2.5 μ L Laemmli sample buffer (4x), boiled for 2 min in a water bath, and subsequently used for western blots. RNA was extracted from the residual 90 μ L with TRI reagent LS (#T3934, Sigma-Aldrich) according to the manufacturer's recommendation, followed by precipitation with isopropanol, glycogen and sodium acetate.

Lysis Buffer

Tris (pH 8.0)	50 mM
NaCl	150 mM
MgCl ₂	5 mM
DTT	1mM
Sodium deoxycholate	0.5% (w/v)
Triton-X-100	1% (v/v)
vanadyl ribonucleoside complex (Sigma)	2 mM
Glycerol	10% (v/v)

The lysis buffer was freshly supplemented with 1x Protease inhibitor complete tablet (Roche) without EDTA.

Wash Buffer

Tris (pH 8.0)	10 mM
NaCl	150 mM
MgCl ₂	1 mM
Glycerol	5% (v/v)

The wash buffer was freshly supplemented with 0.01% NP-40.

2.5 RNA Specific Methods

2.5.1 RNA Isolation

To isolate *Paramecium* RNA, 100,000 cells were spun down and washed twice with Volvic water (2,000 rpm, 2 min). Cells were starved at the respective cultivation temperature for 20 min in Volvic water, so they can finish cyclosis of food bacteria to reduce the amount of contamination by bacterial nucleic acids. Finally, the cells were pelleted (2,000 rpm, 2min) and lysed in 1 mL of TRIReagent for tissues or cells pellets (#T9424, Sigma-Aldrich) by vortexing strongly. The samples in Trizol can be stored at -20°C until further use. Once thawed, the lysate was incubated again at room temperature for 5 min to ensure the full dissociation of proteins from nucleic acids. 200 μ L chloroform was added, the sample was vortexed and after centrifugation (13,000 g, 5 min) the upper aqueous phase was transferred into a fresh RNase free tube without any carry over from the interphase containing proteins.

1 volume of ice-cold isopropanol was added and the RNA was precipitated overnight at -20°C. After centrifugation (13,000 g, 20 min, 4°C), the pellet was washed twice with 70% ethanol; after the final centrifugation step, the ethanol was completely removed and the RNA pellet was air-dried for about 5 min without over drying. The pellet was dissolved in 50 μ L RNase-free water, and RNA concentration

was measured using a NanoDrop (Thermo Fisher Scientific).

2.5.2 RNA Integrity Check

Total RNA isolates were run on denaturing agarose gels. By adding formaldehyde to the agarose, the RNA is denatured, and ribosomal RNAs become visible as discrete bands, a pattern that can be taken as a reference for RNA integrity. RNA that is highly degraded, which is detected as a staircase pattern of a numerous sizes for the ribosomal RNA, should not be used for subsequent library preparation.

2 µg RNA in 5 µL RNase/ DNase free water were mixed with 10 µL RNA loading dye. For full denaturation, the RNA was incubated at 65°C for 5 min and immediately transferred to an ice bath. RNA was loaded onto the agarose gel in 1x MOPS buffer, and 80V was applied for at least 1.5 h.

MOPS Buffer (10x)

MOPS	10 mM
NaCl	50 mM
EDTA	10 mM

The pH was adjusted to 7.0 with NaOH and the solution was autoclaved.

RNA Loading Dye (0.5x)

Formamide	50% (v/v)
Formaldehyde (37%)	16% (v/v)
MOPS buffer (10x)	10% (v/v)
Glycerol	17.5%(v/v)
Bromphenol blue (0.1% (w/v))	5% (v/v)
GelRed (10,000x)	1µL/mL

Denaturing Agarose Gel

Agarose	1.2 g
Formaldehyde (37%)	7.5 mL
MOPS buffer (1x)	ad 100 mL

Agarose was dissolved in MOPS buffer and the formaldehyde was added once the solution was cooled down.

For samples with low RNA yield, the concentration was measured with the Qubit 4 Fluorometer and Qubit RNA Assay Kit (#Q32852, Invitrogen), which allows accurate measurement for RNA sample concentrations between 250 pg/µL and 100 ng/µL. Integrity was checked with the Agilent RNA 6000 Pico Kit (#5067-1513, Agilent Technologies) run on a Bioanalyzer 2100 system (Agilent) using the Eukaryote total RNA Nano assay.

2.5.3 DNase I Treatment

To get rid of all DNA contaminating the RNA preparations, RNA was incubated with DNase I (Qiagen) for 20 min at RT. 50 μ L nuclease-free water and sodium acetate to a final concentration of 0.3 M was added and the RNA was purified using acid phenol (pH 4.5). RNA was precipitated with 2.5 Vol ethanol, washed twice with 70% ethanol and resuspended in the appropriate volume of nuclease-free water according to the subsequent protocol.

DNase I Digest Reaction

RNA	8-20 μ g
DNase I	2.5 μ L
RDD buffer	10 μ L
nuclease free water	ad 100 μ L

2.5.4 Gel Purification of Small RNAs

The fraction of small RNAs from total RNA extracts can be purified by size selection from denaturing polyacrylamide gels with 17.5% UREA and high-resolution for short fragments due to high acrylamide concentrations. Urea was dissolved in acrylamide and TBE buffer in an ultrasonic bath for at least 30 min and subsequently mixed with nuclease-free water, TEMED and APS. The gel mix was poured in between two glass plates (Bio-Rad Protean mini system), which were carefully cleaned with 0.1 M NaOH, 10% (w/v) SDS and isopropanol in advance. Once polymerized, the gels were loaded in a Protean cell and topped with 1x TBE buffer. Traces of urea were rinsed from the pockets with a pipet. 20 μ g of RNA were mixed with 2 vol siRNA loading dye, denatured (5 min, 90°C) and transferred on ice for 5 min. Samples were loaded next to a microRNA ladder (NEB, #N2102S) and 300V was applied for \approx 45 min. Gels were incubated in 50 mL SYBRGold solution for 10 min at mild agitation in the dark, rinsed twice with water and were examined on a blue light transilluminator (Biometra BLstar16, Analytik Jena).

sRNAs from 17-25nt length were cut in small gel pieces and transferred into a fresh 1.5 mL tube with 0.3 M NaCl. To elute the RNAs from the gel, the tube was agitated over-night at 4°C. Samples were passed through Costar Spin-X Centrifuge tube filters (Corning Life Sciences) to remove gel pieces and the RNA was precipitated with 2.5 vol ethanol, 0.3 M sodium acetate and 2 μ g glycogen (GlycoBlue, 15 mg/mL) over-night at -80°C. RNA was pelletized (13,000g, 30 min, 4°C) and washed twice with 80% ethanol, briefly air-dried, and dissolved in 3 to 5 μ L nuclease-free water. In some cases, RNA concentration was measured using the Qubit microRNA assay kit (#Q32880, Invitrogen), but in general, RNA was either directly stored at -80°C or used for subsequent protocols as periodate treatment (2.5.5) or library preparation (2.7.2).

Polyacrylamide Gel

Urea	8.4 g
Acrylamide (19:1, 30%)	8.75 mL
TBE (10x)	2 mL
Nuclease free water	ad 20 mL
TEMED	32 μ L
APS (25%)	17.5 μ L
Nuclease free water	ad 20 mL
TEMED	32 μ L
APS (25%)	17.5 μ L

siRNA Loading Dye (0.5x)

Formamide	950 μ L
Bromphenol blue (5%)	50 μ L
EDTA (200 mM)	5 μ L

Gel Staining Solution

SYBRGold	5 μ L in 50 mL TBE (1x)
----------	-----------------------------

2.5.5 Dissection of 3'-Modifications by Periodate Oxidation

Small RNAs can be modified at their 3' ends by a 2'-O-methyl group, a modification found on several RNA species in plants and animals. To check if *Paramecium* small RNAs also carry this modification, the RNAs were treated with periodate. Periodate cleaves the neighboring hydroxyl groups of the last sRNA nucleoside, and this produces a dialdehyde by the free hydroxyl groups, which are present in both 2' and 3' positions on the ribose. A subsequent β -elimination reaction removes the last nucleoside, and an RNA that is 1nt shorter is generated. Due to the elimination, the 3' ribose now carries a phosphate group which hinders 3' adapter ligation in the sRNA library preparation procedure and causes a loss of RNAs, that are not modified at their terminal ribose. RNAs carrying a 3' 2'-O-methylation, in contrast, cannot form dialdehydes and thus are protected from terminal elimination (Yu and Chen, 2010). 20 μ g of total RNA were pelletized and resuspended in 17.5 μ L Borax buffer I and 2.5 μ L 200 mM sodium periodate was added. The RNA was incubated for 10 min in the dark and subsequently mixed with 2 μ L glycerol and incubated for another 10 min to stop the reaction. The RNA was concentrated using a SpeedVac for \approx 30 min until only the RNA in 2 μ L glycerol remained, which was subsequently dissolved in 50 μ L Borax buffer II and incubated for 90 min at 45°C. Remaining salts were removed using a Sephadex G-25 column (GE) and the RNA was precipitated with sodium acetate and glycogen and dissolved in 5 μ L. RNA samples were separated on a denaturing polyacrylamide UREA gel (subsection 2.5.4) and size selected (17-25nt).

Borax Buffer I

Borax	4.375 mM
Boric acid	50 mM

The pH was adjusted to 8.6.

Borax Buffer II

Borax	33.75 mM
-------	----------

Periodate Solution

NaIO ₄	200 mM
-------------------	--------

2.6 Chromatin Specific Methods

The following methods describe crucial steps from the adapted NEXSON protocol (Nuclei EXtraction by SONication) published by Arrigoni et al., 2016, which aims to isolate nuclei from fixed cell material.

2.6.1 Fixation of Cells

2 – 3x10⁶ *Paramecium* cells from a dense, vegetative culture of known serotype were washed and starved in Volvic. Cells were pelleted in a 2 mL tube (3,000g, 1 min, RT). Media was removed, and cells were quickly resuspended in 1.5 mL fixing solution. Cells were fixed for 15 min at RT with gentle inversion. To quench formaldehyde, glycine was added to a final concentration of 125 mM and the tube was gently inverted. Cells were spun down (3.300g, 5 min, 4°C), washed in 2 mL ice-cold PBS (1x) followed by centrifugation (3.330 g, 3 min, 4°C) and washing in 2 mL ice-cold PBS (1x) with protease inhibitor cocktail (PIC; cOmplete EDTA-free, Roche). The sample was split in half, centrifuged (3.300 g, 5min, 4°C), and once the supernatant was removed entirely, the cells were frozen in liquid nitrogen and stored at -80°C.

Fixing Buffer (10x)

NaCl	100 mM
Tris-HCl (pH 8)	200 mM
EGTA	5 mM
EDTA	10 mM

Fixing Solution (1x)

Fixing buffer	10% (v/v)
Formaldehyde (16%)	1% (v/v)

The fixing solution was prepared freshly for each experiment.

Glycine	1.25 M
---------	--------

2.6.2 Isolation of Macronuclei from Fixed Material

Fixed, frozen cell pellets were thawed on ice and resuspended in Farnham lab buffer by gently pipetting. Aliquoted samples were split to $\approx 600,000$ cells per tube in a maximum volume of 500 μL buffer per tube. Aliquotes were transferred into pre-cooled Bioruptor tubes (Diagenode) and put in a special rotor. Cellular structures were broken by ultrasonication in a Bioruptor at 15 s on, 30 s off in 5 cycles at low intensity. 10 μL were mixed with 1 μL DAPI (1 $\mu\text{g}/\text{mL}$) and investigated for free MACs without remaining intact cells. Nuclei were pelletized (3,000 g, 5 min, 4°C), washed in 1 mL Farnham lab buffer and pelletized again (3,000 g, 5 min, 4°C). Isolated nuclei were either used for MNase digest (subsection 2.6.6) or shearing of chromatin in advance of chromatin immunoprecipitation (subsection 2.6.5).

Farnham Lab Buffer

PIPIES (pH 8.0)	5mM
KCL	85 mM
NP-40	0.5% (v/v)

The buffer was prepared freshly for each experiment and was supplemented with PIC.

2.6.3 Shearing of Chromatin

Nuclei pellets were thawed, resuspended in 500 μL shearing buffer, and transferred in fresh Bioruptor tubes. DNA was sheared by ultrasonication in a Bioruptor at 30 s on, 30s off in 5 cycles at high intensity. Sheared chromatin was transferred into fresh 1.5 mL tubes, cell debris were pelletized (16,000 g, 10 min, 4°C) and the supernatant containing the chromatin was transferred into fresh tubes. 50 μL were taken to validate the efficient shearing of the chromatin, remaining 100 μL aliquots were stored at -80°C.

Shearing Buffer

Tris-HCl (pH 8)	10 mM
SDS	0.1% (w/v)
EDTA	1 mM

The buffer was prepared freshly for each experiment and supplemented with PIC.

2.6.4 Quality Control for Chromatin Shearing

50 μL of sheared chromatin were adjusted to 200 μL with TE buffer, mixed with 2 μL Proteinase K (20 mg/mL, Merck,#03115879001) and 9 μL NaCl (5M) and incubated over-night at 65°C with constant shaking (500 rpm) to reverse all chromatin-crosslinks from formaldehyde. The DNA was purified by adding phenol chloroform, 0.3 M sodium acetate, 2 μL glycogen followed by centrifugation (13,000 g, 5min) and subsequently adding chloroform to the supernatant followed by centrifugation (13,000 g, 5min). The upper phase was transferred into a fresh tube and

treated with RNase A (2 μ L; 20 mg/mL, Roche) for 1 h at 52.5°C at 400 rpm and phenol chloroform extracted again. The DNA was precipitated with 1 Vol isopropanol, for at least 2 h at -20°C. Upon pelletizing and washing with 80% ethanol twice, the pellet was resuspended in 15 μ L TE buffer once the ethanol evaporated. DNA concentration was measured at the NanoDrop (Thermo Fisher Scientific) and 2 μ g were loaded onto a 1.5% agarose gel next to a low-range molecular ladder. Chromatin that was sheared to the size of 300 to 800bp was used for chromatin immunoprecipitations (2.6.5).

TE Buffer (10x)	
Tris-HCl (pH 8)	100 mM
EDTA (pH 8)	5 mM

2.6.5 Immunoprecipitation from Chromatin

8 μ g of adequately sheared chromatin was used for immunoprecipitation using the iDeal ChIP-seq kit for Histones (Diagenode, #C01010050). 2 μ g of antibodies against histone modifications and IgG or 10 μ g of custom RPB1 antibody were used. Additionally, 1 μ L of chromatin was put aside without mixing it with antibodies to serve as input. 20 μ L of DiaMag protein A-coated magnetic beads per IP were washed and reconstituted in ChIP buffer following the manufacturer's recommendations. Beads were mixed with antibodies and chromatin, and tubes were incubated over-night at 4°C on a rotating wheel at \approx 40 rpm. The next day, beads were washed and the chromatin was eluted from the beads in 400 μ L elution buffer at RT. 1 μ L input chromatin was mixed with elution buffer as well and treated in the same way as the IPed chromatin in the following steps. Samples were de-crosslinked, Proteinase K and RNase A treated and extracted. The DNA pellet was dissolved in 15 μ L TE buffer (1x, pH 8), and the concentration was measured using the Qubit4 Fluorometer 1x HS DNA Kit (Invitrogen).

2.6.6 MNase Treatment

Mononucleosomal DNA was isolated from fixed nuclei pellets using aliquots that correspond to the same fixed material used for ChIP (2.6.5). The following isolation of DNA covered by mononucleosomes was carried out as described in (Xiong et al., 2016). One aliquot of isolated nuclei was thawed on ice and resuspended in 1x MNase buffer and nuclei were counted in 10 μ L aliquots by staining with DAPI. Nuclei were split into portions of 20,000 per reaction. After centrifugation (3,000g, 5 min, 4°C) nuclei pellets were re-suspended in 500 μ L MNase reaction buffer. To each reaction, 10 or 128 gel units of MNase (NEB, #M0247S) were added and after incubation (10 min, 37°C, 450 rpm), the reaction was stopped by adding 1/10 Vol stop solution (5min, 450 rpm). MNase digested chromatin was de-crosslinked, Proteinase K and RNase A treated and extracted. As input, DNA without nucleosomes was treated with MNase to check for enzymes sequence bias and PCR bias. An aliquot of nuclei was treated with Proteinase K, extracted as described and 10 μ g DNA was treated with 0.1 units or 1.5 units MNase (5 min, 28°C) and extracted again. DNA was loaded onto a 3% agarose gel and mononucleosomal fractions (100-200bp) were re-isolated.

MNase Dilution/Storage Buffer

Tris-HCl (pH 7.5)	10 mM
NaCl	50 mM
EDTA	1 mM
Glycerol	50%

MNase was diluted to a concentration of 2 units/ μ L and stored at -20°C.

MNase Buffer (10x)

Tris-HCl (pH 8)	500 mM
CaCl ₂	50 mM

MNase Reaction Buffer (10x)

Tris-HCl (pH 8)	500 mM
CaCl ₂	50 mM
β - mercaptoethanol	10 mM
NP-40	1%

MNase Reaction Mix

MNase reaction buffer (10x)	50 μ L
BSA (100 μ g/mL)	5 μ L
MNase (2 units/ μ L)	0.5 μ L - 64 μ L
nuclease free water	ad 500 μ L

Stop Solution

EGTA	10 mM
EDTA	1 mM

2.7 Library Preparation and Sequencing

Different RNA species, either long RNA as mRNA or small RNA species can be converted into DNA using available kits, resulting in DNA products containing additional sequences necessary for sequencing by synthesis on an Illumina platform.

2.7.1 Transcriptome Library Preparation

For purification of all mature polyA-tailed transcripts for analyses of gene expression (transcriptomics), the NEBNext Ultra II Directional RNA Library Prep Kit for

Illumina (NEB, #7760S) with NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, #E7490) was used following the manufacturer's instructions. This approach uses oligo d(T) coupled paramagnetic beads which capture polyadenylated mRNAs. 1 µg DNase treated RNA in 25 µL was used as input material, volumes for all reactions were halved, and all adapters were diluted as recommended. The number of PCR cycles was adjusted to the starting material, ranging from 9-12 cycles. The PCR products were purified using AMPure XP beads (Beckman Coulter).

2.7.2 Small RNA Library Preparation

The fraction of small RNAs (17-25nt) purified from total RNA extracts using denaturing UREA PAGE (2.5.4) were ligated using the NEBNext® Small RNA Library Prep Set for Illumina (NEB,#E7330). Gel purified small RNAs were directly captured in 3 µL nuclease free water without additional quantification and used as starting material. All reaction volumes were halved, and the ligation of the 3' pre-adenylated adapter was carried out at 16°C for 18 hours to efficiently ligate small RNAs that carry a 3'2'-O-methyl group. PCR was performed with 10 cycles and PCR products were purified by 6% native PAGE followed by purification via Costar Spin-X centrifuge tube filters (Corning Life Sciences) and precipitation.

2.7.3 DNA Library Preparation

DNA libraries from material obtained by ChIP or MNase digestion (subsection 2.6.6) were prepared by using the NEBNext®Ultra™DNA Library Prep Kit for Illumina (NEB, E7370) with 10 ng input. All reaction volumes were halved, and the PCR was performed with 11 cycles using the KAPA Taq HotStart DNA polymerase (Kapa Biosystems, KK1512). The PCR products were purified using AMPure XP beads (Beckman Coulter).

2.7.4 Library Quantification and Quality Control

The concentration of purified PCR products was measured with the Qubit 4 Fluorometer 1x HS dsDNA Kit (Invitrogen). Size distribution of library fragment sizes was estimated using the Agilent High Sensitivity DNA Kit (Agilent Technologies, #5067-4626) run on a Bioanalyzer 2100 system (Agilent).

2.7.5 Sequencing

Libraries of the same chemistry were multiplexed according to the desired sequencing depth and sequenced on an Illumina HiSeq 2500 platform. Small RNA libraries were sequenced in Rapid Run Mode with 28bp single-end reads, while transcriptome libraries were sequenced with 100bp read length. ChIP and MNase DNA libraries were sequenced in High Output Run Mode with 100bp paired-end reads. Handling of the HiSeq 2500, clustering, and sequencing was carried out by Dr. Gilles Gasparoni at the Genetics/Epigenetics Department, Saarland University. Demultiplexing was carried out by Dr. Abdulrahman Salhab and Dr. Karl Nordstörn at the Genetics/Epigenetics Department.

2.8 Processing and Analyses of Sequencing Data

All reads analyzed in the presented studies were adapter and quality trimmed using Trim Galore (Krueger, 2015), that uses cutadapt (Martin, 2011). Trim Galore automatically detects public available adapter sequences and trims bases with a phred score below 20 by default, additionally sequences that become shorter than 20 nucleotides are removed. Output of the trimming process as well as the sequencing depth, expressed by the total read number, was evaluated using (*FastQC* 2015) or MultiQC report (Ewels et al., 2016).

2.8.1 small RNA Analyses

small RNA sequencing reads were trimmed with non-default parameters, allowing to precisely trim for sequences of 17-25nt length. small RNA reads were analyzed using the sRNA analysis offline tool RAPID (Read Alignment, Analysis, and Differential Pipeline, (Karunanithi, Simon, and Schulz, 2019)). Reads were aligned against the *Paramecium* reference genome (strain 51, version 2) (Arnaiz et al., 2012) using Bowtie2 (Langmead and Salzberg, 2012), and read statistics such as read length distribution and strandedness were analyzed by the *rapidStats* module that includes samtools (Li et al., 2009). The *rapidVis* module allows for visualization of the read statistics for all reads that aligned to the reference without mismatch. RAPID also allows the removal of contaminants supplied by the user, which are probably coming from food bacteria and other microbial contamination. Alignments were stored as bam files and read coverage along *Paramecium* scaffolds were visualized in the IGV browser (Robinson et al., 2011)

Simple read alignments against any reference in general can also be performed using the Geneious Prime 2020.1.2. software that uses Bowite2 and allows for coverage visualization.

sRNA Normalization Using RAPID

To allow fair comparisons between various datasets, the *rapidNorm* module performs normalization that accounts for the variability in the depth of sequencing between samples. Additionally, RAPID takes the *Paramecium* specific methods such as knock-down by feeding into account, a method which introduces huge amounts of small RNAs coming from the initially introduced dsRNA (primary siRNAs) which themselves trigger secondary siRNA synthesis. The KnockDown Corrected Scaling (KDCS) method calculates the normalized reads count by removing small RNAs that map to the feeding target region and finally allows comparison between samples. Thereby, the KDCS method scales the read counts of each siRNA mapping to an assigned region in a library to the library with the highest read counts (Karunanithi, Simon, and Schulz, 2019).

$$\left(\hat{R} = R \cdot \frac{M}{T - K} \right)$$

- \hat{R} - normalized read count
- R - read count for a region of interest that should be compared (e.g. NDgene region)
- T - total reads mapping to the genome

- K - sRNAs mapping to the feeding target region
- M - maximum overall values $(T_1 - K_1), \dots, (T_n - K_n)$ overall n samples

sRNA Sequence Logos and Nucleotide Count

small RNA were analyzed not only in terms of alignment statistics but also for their nucleotide composition. In order to visualize sequence conservation between sRNAs of the same length that originate from different loci, sequence logos were created using WebLogo3 (Crooks et al., 2004) on a local device, taking reads stored as fastq files as input. The output sequence logo consists of a stack of nucleotide symbols at each position, with the height of each stack representing the conservation at that position and error bars twice the height of the correction for small sample size. The overall nucleotide composition of small RNA populations was analyzed using a custom *Python* script (Christoph Kellner), and to analyze the nucleotide composition of genomic loci themselves, it was segmented in random 23nt bins custom *Python* script prior to composition counting.

Overlapping read pairs

Probabilities for overlapping reads from aligned sRNA reads were calculated using the small RNA signature analysis tool (Antoniewski, 2014), available on the Galaxy web platform (<https://mississippi.sorbonne-universite.fr>, key *small RNA signatures*). sRNAs of 17-25nt were mapped to each region of interest, allowing no mismatches or multimapper in Bowtie2, and bam alignment files were used as input to calculate overlaps of 1 to 25 nucleotides. Probabilities of overlaps of a distinct size were indicated by Z-scores, with high scores indicating higher probability.

2.8.2 Calculation of Gene Expression and Plasticity from mRNA Data

Expression values from polyA RNA data obtained from different cultivation conditions (Serotype A, B, D, H, as well as heat shock conditions (Cheaib et al., 2015)) (ENA PRJEB9464) were quantified using Salmon (v0.8.2) (Patro et al., 2017) with default parameters for all triplicates, and the mean of replicates was used in all analyzes. The relative abundance of transcripts is calculated as transcripts per million (TPM), allowing a fair comparison of reads between samples since read counts were normalized to sequencing depth and gene length. For indexing, the transcript annotation from the MAC genome of *P. tetraurelia* (version 2; strain 51 (Arnaiz et al., 2017)) was used. For identification of high plastic genes, the mean TPM for each gene over different conditions was calculated, and the absolute deviation from the mean for each gene was used to define plasticity. Genes with a large fluctuation were thus termed *plastic genes*.

2.8.3 ChIP-seq and MNase-seq Analyses

Paired-end reads obtained from ChIP-seq (2.6.5) and MNase-seq (2.6.6) experiments were trimmed and aligned against the *Paramecium* reference genome using either GEM mapper with default settings (Marco-Sola et al., 2012) or Bowtie2 in local mode, by setting the mismatch parameter to 1 (-N 1), allowing for one mismatch in the seed region (sub-sequence of a read used for the first alignment step). For ChIP-seq, duplicate reads were annotated by Picard tools (v1.115) (<http://broadinstitute>).

github.io/picard) while MNase-seq duplicate reads were removed by the subsequent DANPOS2 (Chen et al., 2013) operation.

Alignment files of replicates were analyzed for their overall read coverage correlation with *multiBamSummary*, *plotFingerprint* and *plotCorrelation* from the Deeptools package (Ramírez et al., 2016), whereby *plotFingerprint* for instance, aims to visualize the distance of input samples to specific pull-down experiments.

To identify read enrichment in distinct regions, peak calling was performed by using the *dpos* function of DANPOS2. Thereby, samples from different MNase digestion procedures were directly normalized against the MNase input samples ('naked DNA') at nucleotide resolution. Peak calling for ChIP-seq reads was performed using the *dpeak* function, including the normalization against the input reads. The position of peaks was pictured using DANPOS2 generated *wig* files in the IGV browser or the peak distribution was visualized using the *plot* function with default settings. Thus, DANPOS2 plots occupancy values, meaning the count of reads covering each base pair in the position, at genomic sites defined by the user, such as the transcription start site (TSS), the transcription termination site (TTS), the whole gene body and intergenic regions, the latter being defined as the region between annotated TTS and TTS.

Further, occupancy profiles at introns were created by creating a 20bp window centred on the first and last intron base of the 5'-exon-intron junction and the 3'-intron-exon junction. The nucleosome profile was plotted for 1,500bp around this window with x-axis centre representing the junctions.

By *plotProfile* and *plotHeatmap* of the Deeptools package (Ramírez et al., 2016) scaled enrichment plots were generated in addition.

Segmentation by ChromHMM

To learn how the distribution of histone marks and nucleosomes obtained by MNase- and ChIP-seq is probably linked to each other at different genomic sites, a chromatin state learning model was implemented by using ChromHMM (Hidden Markov Model) (Ernst and Kellis, 2012). Therefore, the *Paramecium* genome was binarized into 200bp bins, reflecting the expected nucleosome size, including some linker DNA, by *BinarizeBam*.

Subsequently, the binarized data was passed on to the *LearnModle* functionality to learn a chromatin state model with five different states. Thereby, ChromHMM discovers re-occurring combinatorial patterns of histone marks and nucleosomes. The states were further assigned to genomic sites of the *Paramecium* genome by using *intersect* from bedtools with *-f* 0.8 to 1, defining an overlap of a state with at least 80% to 100% with a defined region of interest.

Polymerase II Pausing Analyses

Pausing of the Pol II at the TSS prior to entering the effective elongation phase was studied by calculating a pausing index (PI). A region starting at 30bp upstream of the TSS to 300bp downstream of the TSS as *TSS region*, and a region starting at 300bp downstream of the TSS until the TTS as was quantified as *gene body*. The pausing index was calculated as a ratio of reads (in TPM) in the TSS region compared to reads in the gene body, by which genes with a pausing index lower than 1.5 were considered as not paused.

For comparative analyzes, the PI of external datasets from different species summarized in Table A.1 was examined, applying minimum read number thresholds in the respective regions. Corresponding mRNA quantification was done as in 2.8.2 using the respective genomic annotations mentioned in Table A.1.

Partial Correlations

A partial correlation of any two epigenetic marks of interest was calculated using the sparse partial correlation networks method after removing effects of other measured epigenetic marks in advance. The partial correlation approach thereby aims to build a network that represents global dependencies of epigenetic marks by extracting direct associations of histone modifications (Lasserre, Chung, and Vingron, 2013). For calculation, the signals for all epigenetic marks were normalized to gene body length and mRNA expression values (2.8.2) were included.

2.8.4 Phylogenetic Analyses and Protein Sequences Alignments

Studied proteins were analyzed for their phylogeny among protein (sub)clades as described in the respective chapter methods section. Parameters for ClustalX alignments are given in each section, besides further information on protein sequences.

2.9 Devices, Chemicals, Kits

This Thesis won't list the devices used for each of the listed methods, since the standard equipment can be found in every molecular biology department. Specific kits and devices can be found in each subsection of the Material and Methods chapter (chapter 2), including manufacturers information and catalogue numbers.

Company	City,Country
Agilent Technologies	Santa Clara, USA
Carl Roth	Karlsruhe, Germany
Carl Zeiss AG	Oberkochen, Germany
Hologic (incl. Diagenode)	Marlborough, USA
Illumina	San Diego, USA
MACHEREY-NAGEL	Düren, Germany
Merck	Darmstadt, Germany
New England Biolabs (NEB)	Ipswich, USA
Qiagen	Hilden, Germany
Thermo Fisher Scientific	Waltham, USA

2.10 Software, Packages, Web pages

Links were collected in Mai 2022 without a guarantee of continued maintenance of tools and web pages by the developer.

Tool	Link to Documentation
Bamtools	https://github.com/pezmaster31/bamtools/wiki
Bedtools	https://bedtools.readthedocs.io/en/latest/index.html
Biomart ParameciumDB	https://paramecium.i2bc.paris-saclay.fr/
Biopython	https://biopython.org/
Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
Tetrahymena DB	http://ciliate.org/
DANPOS2	https://sites.google.com/site/danposdoc/
Deeptools	https://deeptools.readthedocs.io/en/develop/
Galaxy	https://mississippi.sorbonne-universite.fr
Geneious Prime	https://www.geneious.com/prime/
IGV	https://software.broadinstitute.org/software/igv/
Matplotlib	https://matplotlib.org/
small RNA signatures	https://mississippi.sorbonne-universite.fr
ngLOC	http://ngloc.unmc.edu
Overleaf	https://de.overleaf.com
Python	https://www.python.org/
RAPID	https://rapid-doc.readthedocs.io/en/latest/index.html
Samtools	http://www.htslib.org/doc/samtools.html
SnapGene	https://www.snapgene.com/
Trim Galore	https://github.com/FelixKrueger/TrimGalore/
WebLogo	http://weblogo.threeplusone.com/

Chapter 3

The *Paramecium* Macronuclear Epigenome

Parts of this chapter were recently published at Genome Research (Drews et al., 2022, 09 March 2022, doi: 10.1101/gr.276126.121) as

Title

Broad domains of histone marks in the highly compact *Paramecium* macronuclear genome

Authors

Franziska Drews, Abdulrahman Salhab, Sivarajan Karunanithi, Miriam Cheaib, Martin Jung, Marcel H. Schulz, Martin Simon

3.1 Background

The results of the *Paramecium* MAC genome sequencing project published in 2006 and the improved annotation of transcription units in 2017 revealed striking peculiarities of the *Paramecium tetraurelia* MAC genome, which has a high coding density of 78% and tiny introns interspersing $\approx 40,000$ protein coding genes, which are separated by short intergenic regions. Those characteristics are quite divergent in comparison to *Tetrahymena thermophila* and even drastically different from the characteristics of metazoan genomes (Table 1.1), which becomes apparent by a fairly unfair comparison to the human genome, which has a mean intron size of 3kb and a coding density of 3.3%.

Furthermore, due to the high degree of polyploidy ($\approx 800n$), the MAC is full of DNA that must be separated in each cell division, which is realized by amitosis. Since centromeres are absent from the MAC, its division cannot be guided by the classical spindle apparatus, since kinetochores for microtubule attachment cannot be formed. Thus, MAC stretching results in uncontrolled separation of uncondensed chromosomes, and MAC chromosomes are randomly distributed to daughter nuclei. Although information on MAC and MIC genomes is accumulating, deeper insights into the chromatin organization of the DNA-crowded MAC are needed. The deposition of histone marks as guides for IES excision during formation of a new MAC in sexual development has been studied, but regulation of gene expression in *Paramecium*'s highly compact genome during vegetative growth has not been described. A recent study from vegetative MAC chromatin revealed a short NLR of ≈ 151 bp as one of the shortest repeat length in eukaryotes (Gnan et al., 2022). Studies from 1981 list unusual features such as interphase uncondensed chromatin and probable absence of heterochromatic regions, suggesting a different organization of chromatin. This was further supported by missing repressive marks such as 5-methylcytosine, well described for gene silencing in humans, and the absence of the H3K9me3 repressive histone mark which has never been detected in the vegetative MAC (Lhuillier-Akakpo et al., 2014; Ignarski et al., 2014; Samuel, Mackie, and Sommerville, 1981; Singh et al., 2018). Dynamic chromatin remodeling has been studied at several distinct loci in the vegetative MAC, such as at promoters of heat shock protein-encoding genes upon stress induction, but an integrated approach for description of the features of the MAC epigenome is missing.

Within the last 10 years, a couple of studies focused on the regulation of transcription events in the distinct nuclei occurring during *Paramecium* sexual development. It could be shown that *P. tetraurelia* possesses homologs of the Spt4 and Spt5 transcription factors, forming a complex responsible for RNA synthesis from the germline MIC, which are essential during sexual development (Owsian et al., 2022). Homologs of TFIIS, involved in Pol II transcriptional pausing in other eukaryotes, could be identified as well, with TFIIS4 being necessary for the synthesis of non-coding transcripts in the new developing MACs (Maliszewska-Olejniczak et al., 2015a). Interestingly, *Paramecium* possesses paralogs for these TFs that are differentially expressed during vegetative growth and in sexual development, probably assigning the paralogs to the different pathways resulting in specialization of the transcription machinery. Three TFIIS paralogs, two Spt4 paralogs and one Spt5 paralog are expressed in vegetative growth, the latter being indispensable for cell growth (Maliszewska-Olejniczak et al., 2015a; Gruchota et al., 2017a; Owsian et al., 2022). In addition to TFs, homologs of the FACT complex subunits SPT16 and SSRP1/Pob3 were identified (Vanssay et al., 2020). These histone chaperone complex subunits

were studied in developmental genome rearrangements, but data on their vegetative functions are missing. Again, paralogs of these genes show different expression levels, which is also the case for the six recently described *Paramecium* ISWI chromatin remodelers (Singh et al., 2022). Information on other components of the transcription machinery, such as Paf1, NELF or DSIF is missing.

The following study summarizes a first description of the vegetative *Paramecium* MAC epigenome including the streamlined experimental approach to collect nucleosome and histone modification data. Using results from pull-down experiments on the largest Pol II subunit Rpb1, a first description of Pol II processivity and gene expression regulation in a highly condensed genome on the whole genome level is presented. The study aims to answer, how *Paramecium* controls gene expression without canonical heterochromatin and short intergenic regions, which usually contribute to regulation of gene expression in higher eukaryotes.

3.2 Methods

Detailed protocols of the following methods can be found in the material and methods section (chapter 2).

Cell Culture (2.3.1)

Paramecium tetraurelia (strain 51) of serotype A were cultivated at 31°C, and dense cultures without any visible sign of autogamy induction (verified by DAPI staining) were fixed for MNase- and ChIP-seq and immunofluorescence staining.

MNase- and ChIP-seq (2.6.5,2.6.6)

MNase digest was performed with ten units (mild digest) and 128 units (heavy digest) of MNase (NEB) on biological replicates of fixed cells. Naked DNA digest was performed with 0.5 and 1.5 units in biological replicates. ChIP was performed using antibodies directed against H3K4me3, H3K27me3, H3K9ac, and Polymerase II subunit RPB1. Input libraries were generated from 1 µL of chromatin taken aside prior to pull downs.

Protein Sequences Alignment (2.8.4)

The following amino acid sequences of histone H3 subunits were used for phylogenetic analyses:

Homo sapiens: CAB02546, *Tetrahymena thermophila*: XP_001016594, *Paramecium tetraurelia*: PTET.51.1.P1080178, H3P1.

The following amino acid sequences of Rpb1 subunits were used for phylogenetic analyses:

Homo sapiens: P24928, *Schizosaccharomyces pombe*: NM001021568, *Saccharomyces cerevisiae*: YDL140C, *Tetrahymena thermophila*: 00538940, *Paramecium tetraurelia*: PTET.51.1.P1370127.

Additional Methods

- Antibody purification (2.4.4)
- Peptide competition assay, western blots and immunostaining (2.4.5, 2.4.3, 2.4.6)
- ClustalW alignment of RPB1 and histone H3 N-terminal amino acid sequences (2.8.4)
- Polymerase II pausing analyses (2.8.3)
- Partial correlation calculation and gene plasticity (2.8.3)

Data Deposition

All raw read data of this study has been deposited at European Nucleotide Archive (ENA), accession no. PRJEB46233.

3.3 Results

3.3.1 Histone H3 Modifications in the Vegetative *Paramecium* MAC

The organization of the vegetative MAC epigenome was analyzed first by having a global view on the appearance of histone modifications that are well described for their signals in cells undergoing sexual development (Lhuillier-Akakpo et al., 2014; Ignarski et al., 2014; Frapporti et al., 2019). Therefore, immunofluorescence stainings (subsection 2.4.6) using antibodies directed against the *Homo sapiens* histone H3 epitopes were performed with prior testing of antibody specificity.

Lhuillier-Akakpo et al., 2016 identified ten histone H3 proteins, five of them related to the canonical H3 of *H. sapiens* but still showing some divergence. Figure 3.1A shows the alignment of *Paramecium* H3P1 with the H3 sequences of *H. sapiens* and *T. thermophila*, which illustrates the insertion of an amino acid at the N-terminus and substitutions of amino acids.

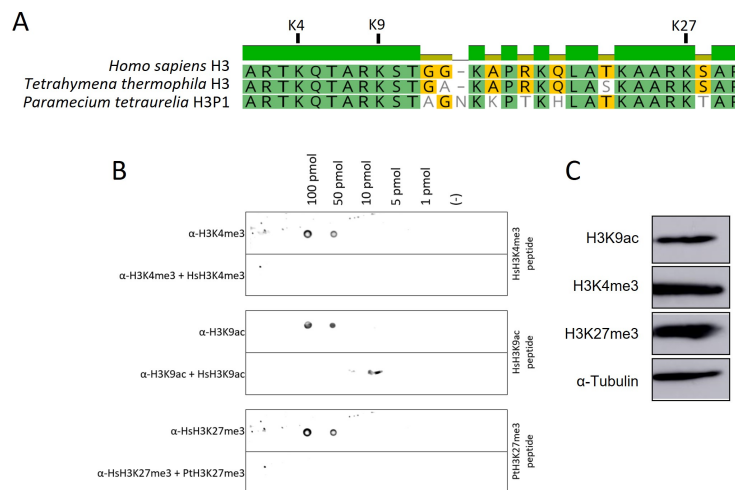


Figure 3.1 (A) Multiple sequence alignment (Global alignment, Blosum62) of histone H3 N-terminal sequences. **(B)** Peptides spotted on membranes (1-100 pmol) were decorated with respective antibodies (each top row) or antibodies that were blocked in advance with the corresponding peptide resulting in vanishing binding signals (each bottom row). **(C)** Detection of histone H3 modifications (H3, 15 kDa) in total protein lysates from *P. tetraurelia*; α -Tubulin served as a loading control (≈ 49 kDa).

To verify the specific binding of the antibodies used in the following study, peptide competition assays and dot blots were performed. Therefore, synthetic peptides covering histone H3 trimethylation at lysine 4 (H3K4me3), acetylation of lysine 9 (H3K9ac) and trimethylation of lysine 27 (H3K27me3) were bound on a membrane that was later decorated with respective antibodies. Besides the position of lysine 27, *Paramecium* shows an amino acid substitution compared to *H. sapiens* H3, which could impede binding of the human-specific antibody to the *Paramecium* H3. To test this, a *Paramecium*-specific peptide was synthesized (Dr. Martin Jung, Homburg, Germany) and spotted on the membrane. As shown in Figure 3.1B, all antibodies show specific binding to spotted peptides. In particular, the *Paramecium* specific peptide covering the K27 region can be detected by the publicly available antibody, as well as achieve blocking of the human H3 specific antibody. Furthermore, decoration of membranes carrying *Paramecium* total protein extracts revealed specific binding for the tested antibodies (Figure 3.1C).

Immunofluorescence stainings against H3K4me3 and H3K9ac revealed signals in the MAC for both modifications, with the H3K4me3 signal also appearing in the MICs (Figure 3.2). Detection of H3K27me3 exhibited low signals both in the MAC and MICs along with unspecific staining of the oral apparatus. This is in accordance with the results published by Ignarski et al., 2014, where the authors also describe low signals for H3K27 trimethylation.

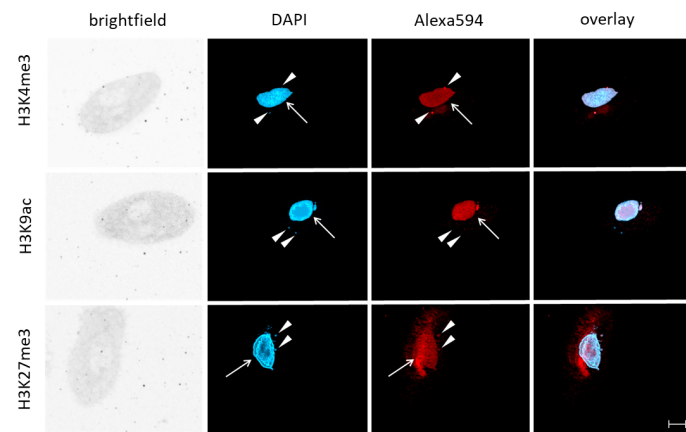


Figure 3.2 Fixed *Paramecium* cells were stained with DAPI (blue) while antibodies directed against the indicated histone modifications (H3K4me3, H3K9ac, H3K27me3) were labeled with a secondary Alexa594 conjugated antibody (red). Overlays of Z-stacks of magnified views are shown, with arrowheads pointing to the MICs and arrows indicating the position of the MAC. Brightfield and signal overlay are shown in the panels left and right, respectively. Scale bar 10 μm .

3.3.2 Nucleosome Patterns Unveiled by MNase-seq

Since histone modifications were detected by immunofluorescence staining and western blotting, the question was raised whether histone modifications could be associated with nucleosomes at different loci of the *Paramecium* MAC scaffolds. Therefore, the positioning of nucleosomes was analyzed by conducting a micrococcal nuclease (MNase) digest on fixed *Paramecium* material in parallel to Chromatin Immunoprecipitation (ChIP).

Arrigoni et al., 2016 published a protocol on nuclei isolation from fixed material (NEXSON), which was further adapted to the fixation of 2-3 million *Paramecium* cells monitored for their serotype and vegetative state of the MAC. Once fixed with formaldehyde, the chromatin state should not change during the following procedures, and nuclei could be isolated by mild ultrasonication. By this, MACs were enriched, and the tiny MICs were removed by washing, reducing contamination from MIC chromatin. The nuclei pellet was then split into halves, and one section was subsequently digested using MNase. Due to the enzymes characteristics, it was crucial to adjust the concentration of MNase to the number of nuclei in each sample to avoid over-digestion and loss of information on the nucleosome positioning.

Different MNase concentrations were tested, following recommendations kindly shared by Xiao Sean Cheng and Yifan Liu (University of Michigan, Department of Pathology) who recently adapted a MNase digest protocol on *Tetrahymena* nuclei but excluding cell fixation (Chen et al., 2016). The use of ten enzyme units resulted in a mild digest, meaning the appearance of a staircase pattern with intervals of approximately 150bp. The digestion using 128 units resulted in a higher degree of digestion, which is seen by the accumulation of DNA fragments the size of mononucleosomes ($\approx 150\text{bp}$) (Figure A.1). The mononucleosomal DNA was extracted from agarose gels and ligated with adapters for Next Generation Sequencing. The obtained MNase-seq

results should then give a high-resolution map of nucleosomes positioned along the *Paramecium* MAC scaffolds.

MNase-seq of two biological replicates for each digestion procedure (light/heavy) resulted in 22 to 78 million reads that were mapped to the *Paramecium* MAC genome. Read alignment files were loaded into the Integrative Genome Viewer (IGV) (Figure 3.3A), where regions with read accumulation became visible as broad peaks, as it is shown for several genes (top) and especially the transcription start site (TSS) and termination site (TTS) of one exemplary gene (bottom).

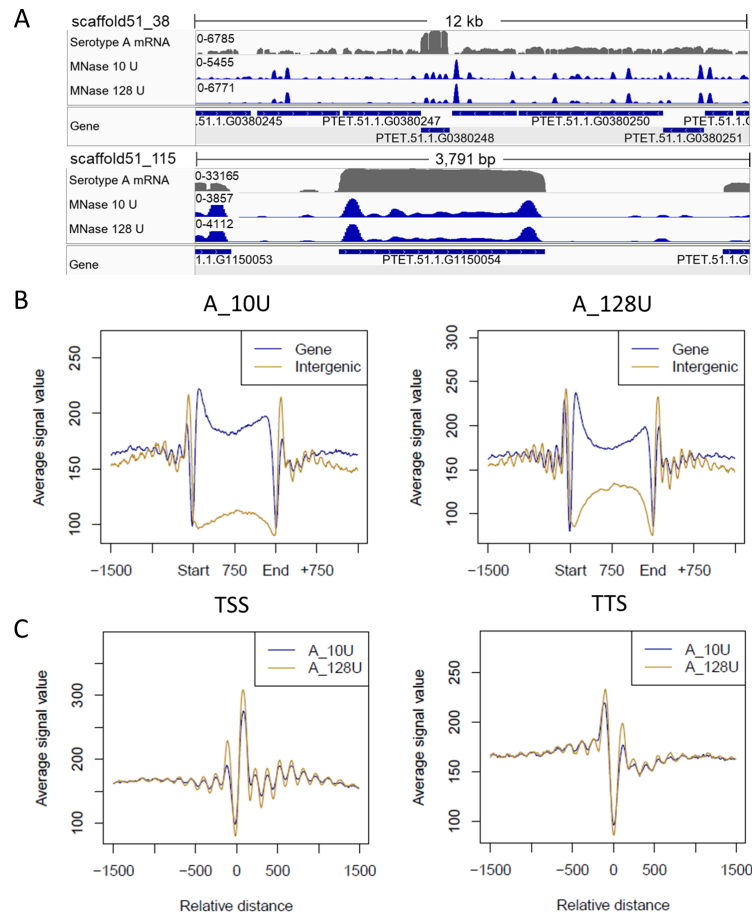


Figure 3.3 (A) Exemplary view of nucleosome distribution along MAC scaffolds of *Paramecium*, visualized using IGV browser. Top panel shows the peak distribution in a 12kb window while the lower panel shows the magnified view of one gene. The top row of each panel shows the coverage track from polyA mRNA-seq followed by the tracks for nucleosome occupancy obtained by light (10U) and heavy (128U) MNase digestion of *Paramecium* nuclei. (B) Metagenes plot of nucleosome distribution along the gene body and intergenic regions for MNase-seq data from mild and heavy digest. Genes and intergenic regions are stretched or shrunken to a length of 1,500bp, adding 1,000bp up- and downstream of the gene/intergenic region without scaling. (C) Profile plot of nucleosome distribution at the transcription start site (TSS) and the transcription termination site (TTS) in a 1,500bp window, respectively. Figure adapted from Drews et al., 2022.

For a more generalized understanding, alignment data was fed to the pipeline for *Dynamic analysis of nucleosome position and occupancy by sequencing* (DANPOS2). The pipeline defines the most preferred position of nucleosomes from paired-end sequencing data and calculates nucleosome occupancy as the count of reads that cover

each base pair in the genome (Chen et al., 2013). At the same time, input signals obtained from sequencing of naked DNA are taken into account and used for normalization. Naked DNA is obtained by performing a Proteinase K digest prior to MNase treatment, meaning that DNA free of any protective nucleosomes is digested by MNase, providing a background genome coverage resulting from MNase sequence cleavage preference. For visualization, the occupancy information for all genes was scaled into one window, resulting in a metagene plot, which aggregates coverage from multiple samples over regions such as genes to provide profiles of average coverage. Therefore, genes were scaled up or down to the size of 1,500bp and binned in 10bp bins; read coverage was calculated as occupancy for each bin. The plots in Figure 3.3B show the occupancy values for all genes from their transcription start site (TSS) to transcription termination site (TTS) with unscaled up- and downstream flanking regions, revealing a pronounced signal at the 5' and 3' end of gene bodies with a signal drop in the center of the gene body. The regions between the TTS and TSS of the neighboring genes or regions between two TSS/TTS (meaning the intergenic region) depending on the gene orientation, show an overall lower occupancy. The peaks after the TSS and prior to the TTS (Figure 3.3C) reflect the exemplary view shown in Figure 3.3A. Additionally, Figure 3.3C shows a strong nucleosome signal at the TSS, reminiscent of a +1 nucleosome in *Paramecium*, which has also been detected in other species such as humans and yeast. Following the +1, the second and third nucleosomes show a less intense occupancy signal which is not the case for occupancy profiles at the TSS for the other analyzed organisms. (Figure A.2). Remarkably, *Paramecium* shows a strong signal upstream of the TSS, a peak that resembles a -1 nucleosome. This signal seems to be stronger in the heavy digest method, but regardless of that, there is no substantial difference in the digestion methods (Figure 3.3B/C).

3.3.3 Coupling of Nucleosome Occupancy and Gene Expression

To answer whether nucleosome occupancy can be linked to the transcriptional status of a gene, either being high, low or not expressed (silent), all protein coding genes from three serotype A cultures were ranked by their mRNA expression values in five quantiles of approximately the same gene number ($\approx 8,000$) (Figure 3.4A) (Cheaib et al., 2015). For each gene in each quantile, the total occupancy was calculated for both MNase digestion procedures, which is the sum of all occupancy values in each bin along the gene body from TSS to TTS and the intergenic region. Figure 3.4A shows the trend for a higher total occupancy of genes with higher gene expression and lowest occupancy of intergenic regions. Since total occupancy could be biased towards the gene length, the individual peak height was analyzed in parallel (Figure 3.4B). This shows that high expressed genes have overall higher peak values - meaning a higher occupancy in specific regions - independent of gene length. From metagene plots it can be assumed that probably the +1 nucleosome is the main contributor to these high occupancy values.

Figure 3.4C shows profile plots at the TSS and TTS but for genes categorized by their expression (high, low, silent). In agreement with Figure 3.4A, the high expressed genes show a more pronounced +1 nucleosome occupancy, while the low expressed genes and the silent genes show an overall low occupancy which looks like a background noise signal.

Looking at *Paramecium*'s short introns, it becomes evident that those are flanked by well-positioned nucleosomes Figure 3.4D, independent of intron length (not shown).

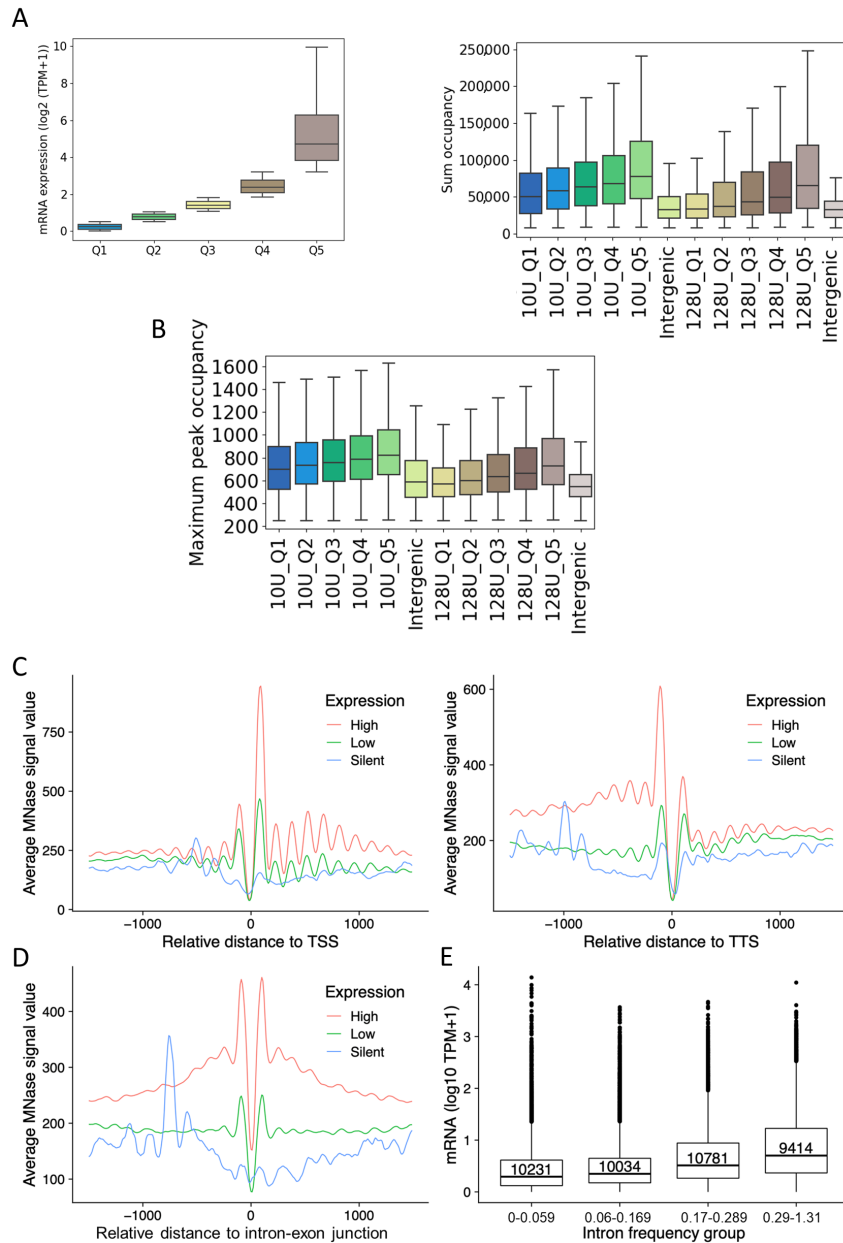


Figure 3.4 (A) Ranking of genes by their mRNA expression values from low to high (Q1-Q5) and total sum occupancy for genes in each expression quantile and intergenic regions. Occupancy values are shown for mild and heavy digest side by side. (B) Maximum peak occupancy along the gene body for all genes in each expression quantile and intergenic regions for 10U and 128U MNase digest. (C) Nucleosome profiles in relation to their distance (x-axis, zero) to TSS and TTS and intron-exon junction (D) is shown for gene categories based on their expression levels [19,090 high (TPM > 2); 20,001 low (TPM < 2); 1369 silent (TPM = 0)]. (E) Box plots showing the mRNA expression (y-axis; log₁₀ TPM+1) of genes with different intron frequency groups (number of introns per 100bp; x-axis). Figure adapted from Drews et al., 2022.

As MNase profiles suggest a general low occupancy of nucleosomes along gene bodies, intron-associated nucleosomes could be an exception to this. In fact, the correlation of the number of introns per 100bp (intron frequency) with gene expression levels (Figure 3.4E) shows increasing mRNA levels with increasing intron frequency, an effect that is independent of the gene length (Figure A.3). Thus, transcriptional

regulation might involve introns and their recruitment of nucleosomes to gene bodies.

3.3.4 *Paramecium's* Extraordinary -1 Nucleosome

The presence of the -1 nucleosome in *Paramecium* needs further investigation, as such a prominent peak has not been detected in the analog analyses of MNase data from *Tetrahymena*, *S. pombe*, or *D. melanogaster*, but they are apparent in humans (Figure A.2). In particular, the recent article by Gnan et al., 2022 does not describe these putative -1 nucleosomes in *Paramecium*. Differences in MNase profiles are not due to the bioinformatics pipelines because Figure A.4 still shows the absence of a putative -1 nucleosome peak when the DANPOS2 pipeline is applied on the data of Gnan et al., 2022. Therefore, the most likely contributor to the observed differences is the fixation of chromatin prior to MNase treatment: while Gnan et al., 2022 used fresh chromatin, the MNase profiles shown in Figure 3.4 were generated from formaldehyde-fixed material (see subsection 2.6.6), which likely results in a less harsh digest. Also, in *Tetrahymena*, light MNase digests indeed show a comparable weak -1 signal, which is lost upon heavy digest (Xiong et al., 2016), and it cannot be excluded that other MNase conditions applied to the analyses of yeast, flies, and human chromatin (Figure A.2) could produce alternative patterns.

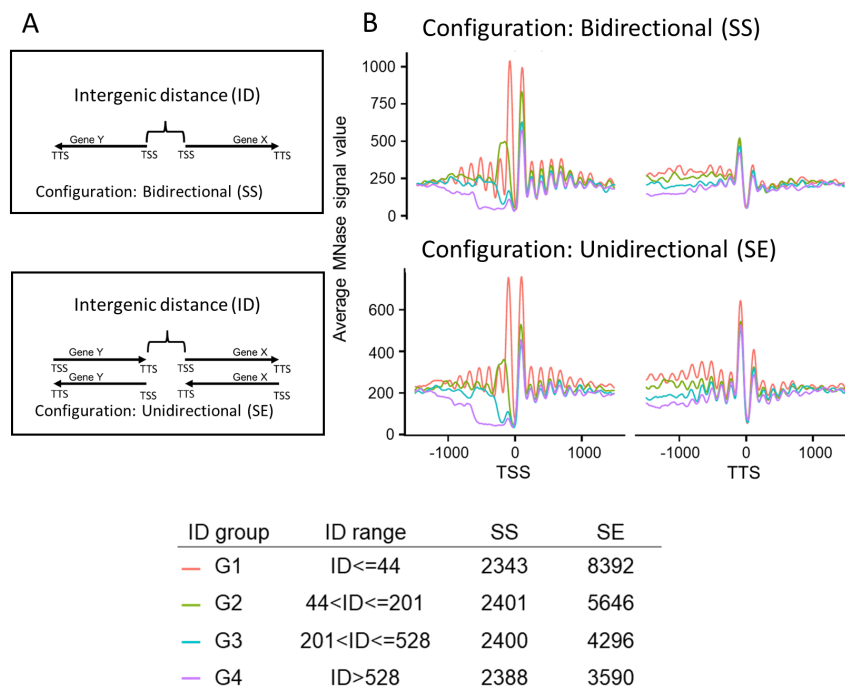


Figure 3.5 (A) Illustration of the gene orientation-based grouping plotted in the right panel. SS = bidirectional (Start-Start), SE = unidirectional (Start-End). (B) Nucleosome profiles in a 2kb window centered at the TSS (left) or the transcription termination site (TTS, right) for neighboring genes in SS and SE configuration. Genes were separated by the length of their intergenic distances. Color-coding can be found in the table below, which shows separation of genes by configuration and ID, ranked from short distances (G1) to long distances (G4). The last two columns indicate the numbers of genes in each configuration and ID group.

The questions of whether there is any rationale beyond the positioning of -1 nucleosomes and if this could be due to *Paramecium's* short intergenic regions were raised.

Therefore, *Paramecium* genes were dissected based on two parameters: their separating intergenic distance and orientation of neighboring genes.

Orientation analysis included bidirectional promoter genes, where the two start sites of both genes are adjacent (Start-Start, SS) and unidirectional genes, where one start site is paired with the end of the other gene (Start-End, SE, Figure 3.5A). Amongst those two categories, genes were classified into four groups based on their intergenic distance, with Group 1 having really short intergenic distances and Group 4 covering the longest intergenic distances (longer than 528bp). The number of genes in each category is given in Figure 3.5 (Table).

Figure 3.5B shows nucleosome positioning of genes amongst those categories at the TSS and the TTS. Most apparent, putative -1 nucleosomes are much more pronounced in genes with short 5'-intergenic regions below 44bp (Group 1, green line), and this is true for the SE and SS orientation.

TTSs also show well-positioned nucleosomes at the 3' end of ORFs, which are more pronounced in the SE configuration and regardless of intergenic distance. The absence of -1 nucleosomes in genes with longer intergenic regions (Group 2 to Group 4) indicates that these putative -1 nucleosomes are either +1 nucleosomes or TTS nucleosomes of upstream genes, but not the labile -1 nucleosome with rapid turn-over as described in budding yeast (Dion et al., 2007). However, the nucleosomes observed for neighboring genes could have a function in regulating gene expression. The correlation of neighboring gene expression shows a high degree of coregulation - but regardless of the bi-/unidirectional configuration (Figure A.5A/B). Although some genes have the same bidirectional promoter and/or short intergenic distances, there is still a level of uncoupled gene expression independent of the neighboring gene. However, genes with bidirectional promoters tend to have a longer intergenic distance (Figure A.5C), suggesting that selection pressure acts on these regions to separate bidirectional genes from each other.

3.3.5 Combinatorial Patterns of Histone Marks

The organization of the *Paramecium* chromatin landscape was analyzed not only in terms of nucleosome occupancy and positioning but also regarding distribution of histone modifications. In addition to immunofluorescence staining, western blots (Figure 3.1) and MNase-seq (Figure 3.3, chromatin immunoprecipitation (ChIP)) experiments were performed on the same fixed chromatin material as used for MNase-seq analysis. The study presented in this chapter overall covers four replicates of ChIP-seq experiments using anti H3K4me3, H3K9ac, and H3K27me3 antibodies and two replicates of MNase-seq experiments (Figure 3.6A). ChIP-seq signals were again visualized using the IGV browser after aligning the reads to the *Paramecium* MAC genome, where the signals were quite broad and did not reveal sharp peaks as expected from, for example, the results of the metazoan H3K4me3 ChIP-seq results (Kundaje et al., 2015). The common procedure to analyze ChIP-seq signals and functionally annotate them is to identify narrow regions of enrichment that pass a given threshold, termed peak calling. Due to the broad signals for all ChIP-seq signals, a different approach was applied, using the ChromHMM (Chromatin state discovery and characterization) software (Ernst and Kellis, 2012).

The software learns and characterizes the combinatorial pattern of histone marks amongst different samples and defines the chromatin states that appear along the epigenome of an organism. It can be used to discover de novo the major re-occurring combinatorial and spatial patterns of marks, and based on a multivariate Hidden Markov Model, the epigenome of an organism can be systematically annotated.

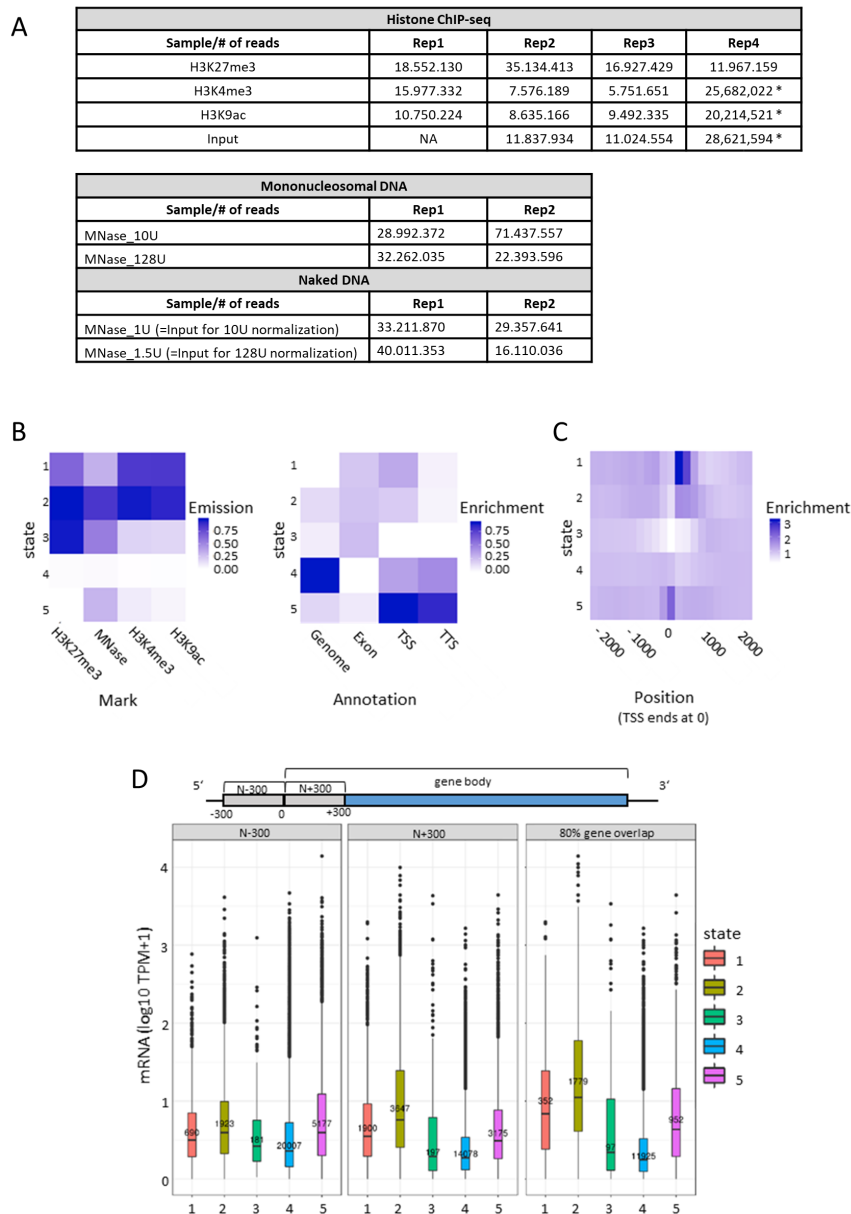


Figure 3.6 (A) Read statistics of ChIP-seq and MNase-seq samples analyzed in the presented study. * Samples were down-sampled to 10 million reads. (B) Chromatin state assignments are shown as a heatmap of emission parameters from a five-state ChromHMM model (left). Each row corresponds to a ChromHMM state, and each column represents a different epigenetic mark. The darker the color of an epigenetic mark for a state, the higher the probability of observing that epigenetic mark in that state. Heatmap showing the overlap fold enrichment of each ChromHMM state (row) in different genomic annotations (columns, right). (C) Fold enrichment of each state in 200bp bins within a 2kb window around the transcription start site (TSS). (D) Box plots showing the mRNA expression (y-axis; \log_{10} TPM+1) of genes whose loci overlap at least by 80% with a respective state (right). Additionally, genes were separated by their assigned state in 300bp upstream of the TSS (N-300), the first 300bp of the gene body (N+300), and mRNA expression values of these genes are plotted (left, middle). Sketch on top of the plots visualizes the arrangement of the three analyzed regions. Figure adapted from Drews et al., 2022.

For ChromHMM application, the genome was binned into 200bp segments and patterns of combinatorial, spatial re-occurring histone marks and nucleosome signals

were determined. The software parameters, which allowed for five different states, were defined as optimum since increasing the number of possible combinatorial patterns resulted in similar states. The software output is summarized as a heatmap, which shows the contribution of each histone mark signal to each chromatin state, with high emission values corresponding to high signal intensities. Those states can quantitatively be assigned to specific traits of the genomes such as exons, TTS, TSS and the whole genome (Figure 3.6B). The most apparent finding is the pattern of **state 4**, which is defined as being free of almost any signal, and that state is attributed to the highest percentage of the genome. This goes along with the metagene plots (Figure 3.3) for MNase signals, showing the gene body and intergenic regions having low nucleosome signals and consequently, fewer sites for introducing introducing histone marks, except for the pronounced +1 nucleosome in high expressed genes. The MNase and histone mark patterns covered in **states 1-3** and **state 5** show a quite dynamic combinatorial pattern of all the marks analyzed. The distribution of states varies at different genomic sites, with **state 1** and **2**, being enriched in the active chromatin marks H3K4me3 and H3K9ac, accumulating at the TSS, while **state 5**, mostly covering MNase signal, is enriched at the TSS and TTS. These observation again go along with the profiles in metagene plots. In contrast, **state 3**, which is enriched for H3K27me3, is depleted in these regions. For higher resolution, the distribution of states is plotted in Figure 3.6C, which shows the distribution up-and downstream of the TSS. The functionality of the marks in gene expression was analyzed by dissecting genes into sub-regions: the region 300bp upstream of the TSS, 300bp downstream of the TSS, and the whole gene body from TSS to TTS. For each of these regions, the states for each gene were assigned. Thereby, a gene has to overlap by at least 80% with a state, which consequently leads to the exclusion of many genes from the analysis ($\approx 15,000$ out of 40,000 coding genes are shown). For each gene with a particular state in a given region, the mRNA level was plotted (Figure 3.6D).

The analysis visualizes the open chromatin state along most parts of the genome as the highest number of genes is found in **state 4**, and strikingly, this state is assigned with the lowest gene expression. This leads to the conclusion that gene silencing in the vegetative MAC is associated with genomic regions that consist mainly of accessible nucleosome-free DNA. On the contrary, the highest gene expression values can be determined for genes associated with **state 2**. This state groups all epigenetic marks, even the H3K27 trimethylation, which argues against the repressive function of this mark in *Paramecium*.

For completeness, genes assigned to **state 3**, show low expression values, but their number is relatively low, which does not allow for a conclusion on the repressive function of H3K27me3. The activating role of H3K4me3 and H3K9ac can be deduced from **state 1**, which has an enrichment for these marks. Genes covered by state 1 have a higher gene expression than genes covered by state 3, 4, 5.

The patterns of state association to genes do not change drastically concerning -300/+300bp of the TSS, although differences in gene expression levels amongst different genes at -300bp seem to be less robust. Probably, the upstream region does not contribute to gene regulation as strong as +300bp, which is consistent with the assumption that active histone modifications at the +1 nucleosomes are associated with gene transcription. Therefore, the +1 nucleosome seems to be a strong contributor to gene regulation.

3.3.6 Gene Expression Regulation by Polymerase II Occupancy

The +1 nucleosome was extensively studied in yeast and fly because it is a specific site that regulates the pausing, release, and elongation of RNA polymerase II (Pol II) by recruiting chromatin remodeling factors and Pol II CTD phosphorylation. Consequently, the pronounced *Paramecium* +1 nucleosome could serve the same functions. Elucidation of Pol II footprints along expressed and silent genes was approached by ChIP-seq using a specific antibody as it was done for the three histone modifications. Commercially available antibodies target the heptad repeat of the Rpb1 CTD, which is conserved between multiple species and accumulates up to 50 times in the human CTD, offering multiple antibody binding sites.

However, the *Paramecium* Pol II diverges from metazoan and even unicellular Pol II, because it does not have the heptad serine-rich repeats as visualized by an alignment of the C-terminal sequences (Figure 3.7A), making it essential to produce an own antibody. The designed antibody, targeting a serine rich stretch (N-SPHYTSHTNSPSPSYRSS-C), was affinity-purified from rabbit serum and validated for its specificity by immunofluorescence staining and western blots. Immunofluorescence shows an enrichment in the MAC, as the MAC is responsible for gene expression throughout vegetative growth (Figure 3.7B). Immunofluorescence stainings were also performed in autogamous cells (Dr. Jacek Nowak, personal communication) and signals could be detected in the old MAC in early stages of autogamy. In these MACs, ncRNA transcription occurs, probably by Pol II, prior to the scnRNA selection mechanism (see Figure 1.8).

Furthermore, western blots using protein from enriched MAC nuclei fractions (verified by histone H3 detection) showed enrichment of a band corresponding to the size of the *Paramecium* Rpb1 subunit (Figure 3.7C). ChIP-seq signals were again plotted for all genes along the gene body, with genes being separated by their expression in nine quantiles. Quantile nine, covering the highest expressed genes, also shows the highest Pol II occupancy along the gene body. All genes show a drop of Pol II signal directly at the TSS and TTS, and genes seem to be evenly covered with Pol II along the ORF (Figure 3.7D).

3.3.7 Pausing Regulation with a Highly Divergent Polymerase II CTD

Pol II transition from initiation of transcription to elongation involves the phosphorylation of sites in the heptad repeat and the recruitment of elongation complexes. Additionally, release from transient series of Pol II stalling events (pausing) followed by resumption of transcription is also regulated by orchestrated CTD modification. Since *Paramecium*'s CTD is so divergent compared to other species (Figure 3.7), the question was raised, how transcriptional regulation and release from pausing is regulated.

If Pol II pauses at a given site can be estimated by the pausing index (PI), which is simply calculated by dividing the number of Pol II ChIP-seq reads at the TSS by the number of reads along the gene body. If the PI is larger than 1.5 (having more reads at the TSS than in the downstream gene body), a gene is categorized as paused; otherwise, genes are termed not-paused (Figure 3.8A).

The PI was calculated not only for *Paramecium* but also for *Tetrahymena*, yeast, and humans (Table A.1), revealing some striking differences. In *Paramecium*, $\approx 65\%$ of protein-coding genes are not paused, which is thrice the amount of paused genes; a phenomenon that cannot be seen in other species (Figure 3.8B). The comparison

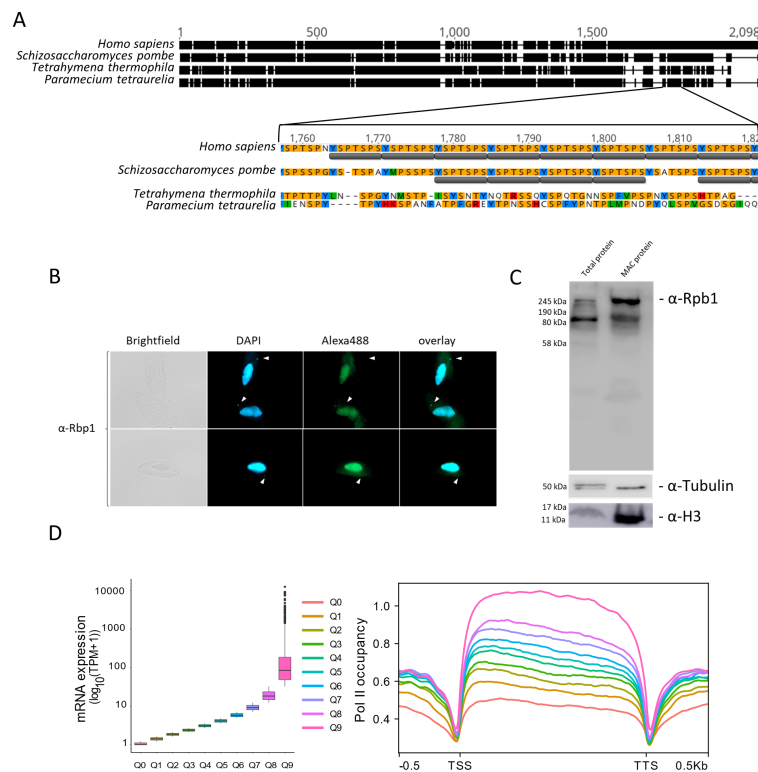


Figure 3.7 (A) Multiple sequence alignment of the RNA polymerase II Rpb1 subunit in different organisms. The C-terminal end of RPB1 is zoomed in to show the difference in conserved regions of some ciliates to other organisms. Grey bars underline the Pol II heptad repeat. **(B)** Localization of Polymerase II by immunofluorescence staining and western blots. Primary anti-Rpb1 antibody was labeled with Alexa488-conjugated secondary antibody (green), and nuclei were stained with DAPI (blue). Arrowheads point at micronuclei. Representative overlays of Z-Stacks are shown. Scale bar is 10 μ m. **(C)** Protein lysate from whole cells (total protein) and protein from fractions of enriched macronuclei (MAC protein) were blotted; membrane was decorated with antibodies against Rpb1 (200 kDa), α -Tubulin (49 kDa), and histone H3 (15 kDa) as loading control. Figure adapted from Drews et al., 2022.

indeed is not fair, since threshold for minimum gene length and minimum reads compares PI of $\approx 6,000$ human genes against $\approx 26,000$ *Paramecium* genes.

Having a look at the Pol II occupancy profiles for paused and not-paused genes (Figure 3.8C), *Paramecium* Pol II seems to be distributed differently along the ORF: Pol II is evenly distributed amongst genes of both categories, having an overall higher occupancy amongst not-paused genes. While in *Tetrahymena* and humans the occupancy for not-paused genes is increasing toward the 3' end of the ORF, in *Paramecium*, not-paused genes show a slightly decreasing pattern towards the 3' end. Although gene lengths are significantly different amongst species, analysis for genes of the same length revealed the same pattern (not shown).

For paused genes, *Tetrahymena* and yeast show a peak at the TSS with a strong drop in occupancy towards the 3' end, a pattern that cannot be seen in *Paramecium*. It is tempting to speculate that the +1 nucleosome in the three other species has a functional role in regulating pausing. In comparison, the +1 nucleosome in *Paramecium* seems to be less involved in this particular process.

Paramecium's Pol II occupancy shows a clear drop at the 5' and 3' non-coding regions, suggesting less regulatory function coming from intergenic regions and regulation of gene expression happens inside the ORF. The correlation of mature polyA mRNA levels with Pol II occupancy shows lower mRNA levels of paused genes in *T. thermophila* and *S. pombe*, while in *Paramecium*, expression levels are almost equal amongst paused and not-paused genes (Figure 3.8D). Only in *H. sapiens* paused genes show higher expression values.

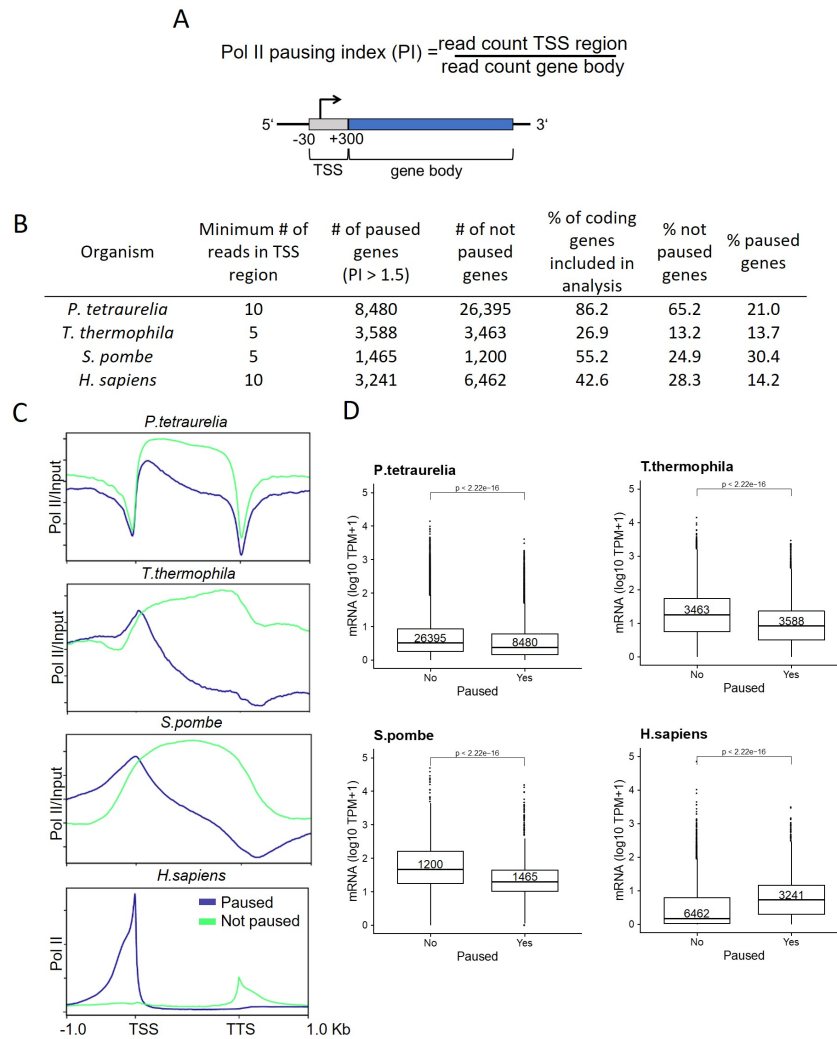


Figure 3.8 (A) Graphical representation of the regions included in Polymerase II pausing index (PI) calculation. Paused genes have a $PI \geq 1.5$. (B) Table summarizes numbers of paused/not paused genes for selected organisms and how many genes are included in the pausing index analysis. (C) Same as the Pol II enrichment profiles in Figure 3.7D, but genes are split based on the status of Pol II pausing. (D) Box plots of gene expression for paused/ not paused genes. Figure adapted from Drews et al., 2022.

3.3.8 How do Epigenetic Marks Orchestrate Gene Expression?

In which manner all epigenetic marks analyzed in this study, are distributed along the *Paramecium* genome is shown by the heatmaps in Figure 3.9A. Genes were separated by their expression values as previously and ranked from high to silent. For each category, genes were ordered from long to short, as indicated on the most left

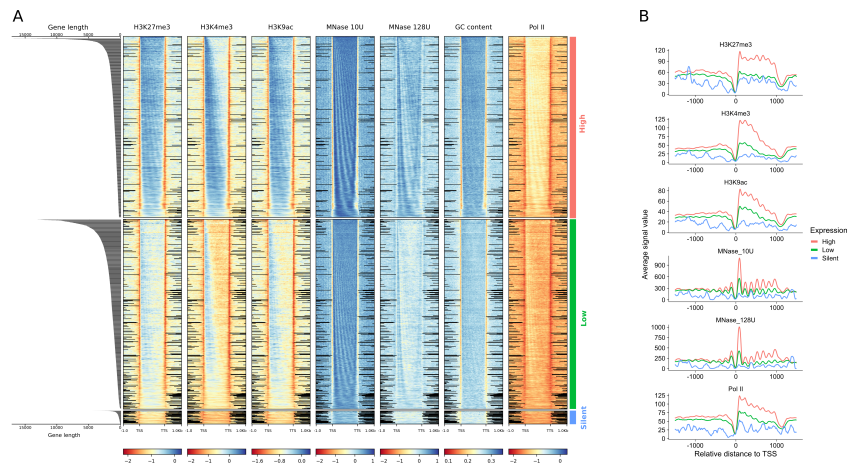


Figure 3.9 (A) Distribution of epigenetic marks along genes with different expression levels. Heatmaps show the input-normalized enrichment values for epigenetic marks. Genes (rows) are split into three categories based on gene expression: High (TPM>2), low ($0 < \text{TPM} < 2$) and silent (TPM=0), and genes are sorted by decreasing order of gene length, which is visualized by the length distribution graph on the left. Distance shown on the x-axis is scaled, i.e. all genes (TSS-TTS) are either stretched or shrunk to a length of 1,500bp, adding 1,000bp up- and downstream of the gene without scaling. Missing data in intergenic regions is indicated by black lines. **(B)** Distribution of epigenetic marks for a subset of $\approx 4,000$ genes with discrete length of $\approx 1.2\text{kb}$. Plots show the signal at the TSS and the TTS and in the upstream intergenic regions of genes belonging to the similar expression categories as in **B**. The plots avoid any kind of scaling, allowing for fair comparison of signals. Figure adapted from Drews et al., 2022.

side of the heatmaps. In addition to the nucleosome profile and histone marks, Pol II distribution was plotted as well as the GC content of each gene, which is crucial to analyze since DNA sequence composition itself influences nucleosome organization (Xiong et al., 2016).

The active marks H3K4me3 and H3K9ac show, along with the nucleosome profiles (MNase 10U, MNase 128U) and Pol II occupancy, an accumulation at the 5' region with decreasing signal intensity along the ORF, which is apparent for high expressed genes and becomes even more obvious when having a look on genes of the same length (Figure 3.9B). Nevertheless, signals are still high toward the 3' end of the ORF and correlate with gene expression. Although the repressive histone mark H3K27me3 shows a less pronounced enrichment at the 5' region of the ORF compared to other marks, the signal values still correlate with gene expression.

In contrast, silent genes have the faintest signals for all epigenetic marks and thus, in *Paramecium* unoccupied DNA seems to be the main regulator of transcriptional inactivation. Additionally, all epigenetic marks are low at the 5' and 3' flanking non-coding regions, showing a clear drop which again contributes to the hypothesis that intergenic regions hardly contribute to the regulation of gene expression.

Nucleosomes seem to be strongly phased in all gene categories, a pattern that becomes apparent for shorter genes. Nucleosome occupancy is correlated with higher gene expression, as shown in Figure 3.3. The phasing pattern resembling nucleosome positions is seen in the Pol II heatmap and for almost all epigenetic marks. The histone marks phasing pattern follows not only the nucleosomes but also the GC content, which oscillates in position and quantity. Therefore, the GC content acts as a *cis*-regulator for nucleosome positioning and gene expression in *Paramecium* as well.

Pol II phasing, following the pattern of nucleosomes, probably indicates that there exists an association of Pol II and nucleosomes along the ORF, and the high occupancy of both leads to higher gene expression. It seems like Pol II is pausing at each nucleosome, probably indicating a mechanism of inefficient elongation. In fact, shorter genes show higher levels of mRNA than longer genes, probably being transcribed more efficiently (Figure A.6).

Heatmap signals indicate an existing interplay of histone marks, *cis*-factor regulated nucleosome phasing, Pol II pausing with gene length and gene expression. Therefore, the correlation of all epigenetic marks was calculated with each other, including mRNA data (Figure A.7). All epigenetic marks are positively correlated (Pearson's correlation > 0.6) with each other and with mRNA (Pearson's correlation > 0.30), which is especially noteworthy in terms of H3K27 trimethylation, which is classically associated with transcriptional silencing.

3.3.9 Correlation of Epigenetic Features

The positive correlation of H3K27me3 with mRNA levels and other histone marks and its presence throughout the ORF raises the question of the role of H3K27me3 in the vegetative *Paramecium* MAC. To answer this and gain more insight into the regulation of gene expression by epigenetic features, mRNA data from varying cultivation conditions was gathered to ask for the contribution of each factor to differential gene expression. The mRNA data sets comprise transcriptomic data from different environmental states, such as heat shock, cultivation at 4°C and the information from four different *Paramecium* serotypes (A,B,D,H) (Cheaib et al., 2015).

Genes that show large expression variants during vegetative growth among different growth conditions (termed *high plastic genes*) (see subsection 2.8.2 for details) appear to be dynamically regulated and were separated from housekeeping genes, which in contrast have a robust expression among different conditions. The genes were classified into four groups of plasticity (G1-G4), with G4 covering the genes with the largest variation. For genes in these four plasticity groups, chromatin states based on ChromHMM segmentation are visualized in (Figure 3.10A/B). The states show gradual differences along the groups, with the most apparent increase in ChromHMM **state 4** (blue) and the decrease in state 2 (gray), the latter covering all epigenetic marks together, while state 4 shows almost no signal for all marks (see Figure 3.6). This pattern suggests that epigenetic marks are used for gene regulation in reaction to environmental changes and are not

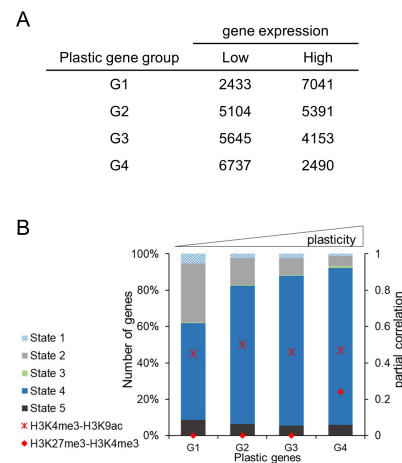


Figure 3.10 (A) The number of high and low expressed genes in each plastic gene group. (B) Distribution of chromatin states among plastic gene groups. Only genes with a ChromHMM state overlapping of at least 80% were included (see Figure 3.6). Partial correlation values for H3K4me3-H3K27me3 (circle) and H3K4me3-H3K9ac (asterisk) are shown in red for each group. Figure adapted from Drews et al., 2022.

used for gene regulation in reaction to environmental changes and are not

only used to control gene expression, similar to an *on-off* switcher. To add a higher resolution to the understanding of combinatorial patterns of epigenetic marks, as it was described for bistable domains harboring H3K4me3 and H3K27me3 signals, partial correlation between different histone marks for the plastic gene groups were calculated (see Methods section 2.8.3). Figure 3.10B shows an increase in partial correlation of H3K4me3/ H3K27me3 only for most plastic genes, suggesting that the interplay between histone marks varies in the four considered groups. Contradictory, the correlation of both active marks, H3K4me3 with H3K9ac, does not change regardless of gene plasticity.

3.4 Discussion

The genome of *Paramecium tetraurelia* has been extensively studied in terms of guided genome rearrangement processes during sexual development involving the biogenesis of different sRNA species and the precise elimination of distinct genomic sequences by histone modifications (see section 1.2.2). However, studies of the organization of the vegetative MAC epigenome and observations on MAC gene expression regulation during vegetative growth are lacking on the molecular level.

How the DNA is organized in the nucleus of this unicellular organism is not as trivial as it seems: the macronuclear DNA is extremely gene rich and intergenic regions are the shortest amongst various species (see Table 1.1). Additionally, due to the high degree of polyploidy, the MAC is full of DNA that needs to be precisely accessed to regulate gene expression. Although unicellular, excluding the need for tissue-specific gene regulation, *Paramecium* still reacts to external stimuli, which needs to be realized on the gene expression level. Because positioning of nucleosomes and placement of histone modifications are key players of the epigenetic toolkit for gene expression regulation, their functional role in the vegetative *Paramecium* MAC was studied and expanded by the study of Polymerase II processivity.

The most striking observation is the correlation of higher nucleosome occupancy in actively transcribed genes. In contrast, genes that are silent in expression are devoid of nucleosomes, and those regions, low in signals for nucleosomes and histone modifications, comprise most parts of the genome. Furthermore, the short intergenic regions are nucleosome-free as well. This leads to the idea that DNA being less covered by nucleosomes, being naked so to speak, is the predominant chromatin conformation, while an attraction of nucleosomes to a specific site seems to be crucial for induction of gene expression.

The observed particularities in the organization of the *Paramecium* nucleosomes are quite the opposite to the organization of nucleosomes in other organisms: a long-held dogma describes naked DNA that is accessible for the transcription machinery, and thus open chromatin is transcribed, while gene inactivation is regulated by DNA wrapping around nucleosomes, making gene regulatory regions less accessible. For *Paramecium*, this model seems to be inapplicable since nucleosome patterns are almost the opposite. Because chromatin condensation is carried out by recruiting the linker histone H1 (Nalabothula et al., 2014), the question raised is whether this is simply not possible due to a missing H1 ohnolog that could not be identified in *Paramecium*. If this is a cause or consequence of the short nucleosome repeat length with linker DNA of just a few bases, is only one of the open questions (Gnan et al.,

2022), although the missing H1 ohnolog could simply be so divergent that it cannot be detected.

Not only the placement of nucleosomes and their association to linker histone H1 are contributors to the regulation of transcriptional silencing, but the introduction of the H3K27 trimethylation is a well-described mark for guidance of low to no transcription. By analyzing this mark and its combinatorial pattern with other marks from ChromHMM segmentation analysis, it becomes obvious that this histone modification is not exclusively correlated to low gene expression in the *Paramecium* epigenome. Basically, the repressive function of H3K27me3 could be blocked by the introduction of a phosphate group on neighboring serine as it was described for H3K9me3 blocking by H3S10 phosphorylation (Fischle et al., 2005). Since this serine is not found in any of the *Paramecium* H3 proteins, this event is unlikely (Lhuillier-Akakpo et al., 2016). In several studies of *Paramecium* developmental genome rearrangements, it was reported that H3K27me3 accumulates mainly in new developing MACs after induction of sexual development (Vanssay et al., 2020; Frapporti et al., 2019), and H3K27me3 was transiently found in the MIC during the first meiotic division and in the fragments of the maternal MAC (Lhuillier-Akakpo et al., 2014). Since cells were monitored extensively prior to fixation for nuclei isolation, contamination from the MAC fragments or Anlagen could be ruled out. Additionally, the experimental approach specifically enriches for MACs, limiting MIC contamination. However, if some MICs contaminate nuclei isolates, they should not contribute to the general outcome of detectable histone mark patterns, since the ≈ 800 MAC chromosome copies would overshadow faint signals from diploid MIC contamination. In fact, this was evaluated in a previously performed ChIP experiment by Cheaib and Simon, 2013 (prior to NEXSON protocol adaption), and DNA from MICs could not be detected from H3/H3K9ac pull-downs, arguing against contaminations from MIC chromatin.

If H3K27me3 is not exclusively associated with repression, what is the regulatory role of H3K27me3 in gene expression then? The partial correlation revealed that genes with dynamics in their expression amongst different conditions show higher correlations for the active mark H3K4me3 and H3K27me3 than stably expressed housekeeping genes. This observation is striking, as it resembles the condition of a bistable domain (Sneppen and Ringrose, 2019), being associated with poised genes, which are always on the edge of being transcribed or silent, simply by the removal of one of the respective marks. As fascinating as it seems, the MAC polyploidy still needs to be kept in mind since one cannot dissect which of the ≈ 800 copies of the MAC chromosomes locus is associated with the respective mark. If *Paramecium*, for instance, would use gene dosage to regulate gene expression level, one would expect different ratios of marks: some copies silent, some copies active. Answers could be given from re-ChIP experiments (Kinkley et al., 2016), which would consist of two sequential pull-downs from one chromatin input: so H3K27me3 could only be pulled down if H3K4me3 was present on the same molecule and pulled down in the first IP.

It cannot be distinguished right now that both modifications appear on the same molecule, as this would need to be seen in mass spectrometry analysis from long peptides covering both K4 and K27. At least in *Tetrahymena*, the co-occurrence of H3K4me3 and H3K27me3 has not been detected (Taverna et al., 2007) (personal communication). Additionally, although cells are vegetative and do not undergo genome rearrangements, a culture is not synchronized in terms of cell cycle, so data is always gathered from a mixture of cell stages and even from varying MACs, since each amitosis results in imbalances of MAC chromosome copy numbers between

daughter cells.

In a previous study, increased H3K27me3 levels in association with decreased levels of H3K4me3 at an endogenous reporter gene have been shown to go along with siRNA mediated silencing (Götz et al., 2016). This supports the idea of H3K4/K27me3 ratio controlling gene expression levels and could be a more ancient mechanism rather than an invention of multicellular organisms, using those ratios, especially during developmental programs to keep epigenetic plasticity (Kumar et al., 2021).

From the data presented, it is tempting to speculate that the epigenetic repertoire already has the capacity to manifest vegetative gene expression by poisoning genes during development, although *Paramecium* is unicellular. *Paramecium* has been shown to inherit gene expression patterns, for example, where the transcription of a surface antigen follows the expression pattern of the cytoplasmic parental cell. However, the mechanism of this transmission is not yet understood, and it needs to be studied if dynamic histone modification ratios are involved (Baranasic et al., 2014; Simon and Plattner, 2014).

ChIP-seq Reveals Broad Domains Instead of Narrow Peaks

The key regulator, in addition to positioning nucleosomes in transcription-prone genes, appears to be the introduction of H3K4me3 and, to a lesser extent, H3K9ac. Both marks are well described for being associated with positive transcriptional regulation. This being said, the distribution of marks seems to be, again, divergent in *Paramecium*: they did not reveal sharp peaks but broad domains. These domains can be interpreted as domains ensuring an ongoing transcription elongation as it was claimed for tumor-suppressor genes (Chen et al., 2015) but also for genes shaping cellular identity. Those broad domains are associated with H3K4me3 and pausing Pol II, thereby controlling the robustness of transcription.

A Divergent Mechanism of Transcriptional Elongation

Pol II pausing is/allows for fine-tuning of transcription and the pausing pattern in correlation to broad histone marks domains was examined, having in mind, that *Paramecium* Pol II does not exhibit the serine rich heptamer repeats being the site for highly orchestrated phosphorylation patterns regulating different phases of transcription. Nevertheless, phosphorylation could be introduced as an activating mark on the CTD, which is still serine-rich although not organized in repeats. The differentiation of varying phosphorylation states of Pol II along gene bodies would hit the nail on the head and reveal mechanistic of transcriptional regulation, but the polyclonal serum against a peptide, including unphosphorylated serines, could miss CTD variants being phosphorylated. Still, Pol II in *Paramecium* can be detected in the center of ORFs and at the 3' regions, regions where Pol II in mammals is usually phosphorylated at the CTD. One can speculate that the antibody still detects different conformations of Pol II CTD.

From the pausing index and phasing pattern results, with *Paramecium* Pol II kind of stalling at each nucleosome along the ORF, it can be suggested that pausing in *Paramecium* occurs differentially in comparison to other species. Since shorter genes show a tendency for higher expression, regulation of transcriptional elongation in longer genes is probably not that sufficient, resulting in stalling of Pol II along the whole ORF. Since Pol II does not have to transcribe huge intergenic regions as it does in humans, the question arises if Pol II really needs an advanced elongation regulation. In yeast and humans, nucleosomes placed at intron boundaries contribute to

co-transcriptional and alternative splicing (Patrick et al., 2015), the latter only detected at low rates in *Paramecium* (Jaillon et al., 2008). Furthermore, highly efficient splicing appears to be regulated by GC content and not by nucleosome positioning at intron boundaries (Gnan et al., 2022). Another possible role for introns can be hypothesized from MNase-seq data: Since genes with higher intron frequency show higher gene expression levels, introns seem to recruit more nucleosomes into ORFs and thus contribute to increased transcription levels.

How Pol II is released from stalling at nucleosomes is well understood for yeast and humans, involving the release of NELF and assembly of the Mediator. This system seems to be divergent in *Tetrahymena* and probably *Paramecium*, as both miss homologs of NELF and the critical regulator of Pol II processivity in transcription and elongation, the Mediator complex. Further, a set of transcription-associated proteins were shown to be highly divergent in *Tetrahymena* (Garg et al., 2019; Tian, Mochizuki, and Loidl, 2019). However, the core component (Med31) that is highly conserved among eukaryotes was identified in *Tetrahymena* and allows for efficient transcription (Garg et al., 2019). Furthermore, in *Paramecium*, not all components of the Paf complex could be identified, which is involved in regulating elongation, 3'-end processing, and histone modification deposition (Jaehning, 2010). Especially, the subunit Paf1, involved in serine phosphorylation of the CTD of Pol II, is missing, which fits the missing serine repeats of the CTD. Preliminary silencing experiments on the subunits of the Paf complex revealed reduced division rates upon slicing, leading to the assumption that although divergent, the Paf complex contributes to cell viability, probably by transcriptional regulation.

Due to the lack of canonical elongation systems coupled with a lack of conserved serine residues, transcriptional elongation in *Paramecium* seems to be regulated differently. As discussed above, broad H3K4me3 domains, along with increased occupancy of Pol II in gene bodies, could be an alternative control of transcription by buffer domains. It seems tempting to speculate that this form of Pol II buffering represents an alternative or perhaps an ancient form of elongation control.

Cis-factors in Nucleosome Positioning and Methodical Limitations

The interpretation of nucleosome positioning profiles obtained from MNase-seq must always be handled with care since the enzyme's preference for AT-rich regions and its ability to digest up to the nucleotide size can result in over-digestion and loss of nucleosome information. This was indeed extensively reviewed from fly and yeast data (Chereji, Bryson, and Henikoff, 2019). Although the nucleosome profiles obtained in this study (Figure 3.3, Figure A.4) are somewhat different from the profiles published by Gnan et al., 2022, especially in terms of -1 nucleosomes upstream of the TSS, this must not raise conflicts in biological interpretation. Fixation of chromatin and mild digestions preserve some nucleosome patterns lost by harsh digestion. Since both, nucleosome data and ChIP-seq data, revealed the same patterns (Figure 3.9) of signal distribution in genic and intergenic regions, the conclusions drawn from ChromHMMM seem to be fair; arguing against the detection of nucleosome profiles simply due to MNase cleavage preferences.

In addition to experimental approaches and enzyme sequence preferences, *cis*-factors such as the DNA sequence itself also contribute to nucleosome organization along chromosomes, with GC content highly favoring nucleosome positioning (Tillo and Hughes, 2009). In *Paramecium*, the overall genome GC content is quite low ($\approx 28\%$) but it has been shown that GC content in gene bodies is higher, which could probably contribute to the regulation of gene expression (Meyer and Liu, 2014). There

is an ongoing discussion about the DNA sequence preferences of nucleosomes, and also MNase-seq can generate a signature of higher occupancy at GC-rich regions, on naked as well as on DNA covered by nucleosomes (Chung et al., 2011). To overcome the limitations of this bias, unoccupied DNA was treated in parallel during all experiments and was used for normalization and background reduction of MNase cleavage preferences.

To add another layer of complexity, nucleosome profiles are not just affected by GC content and MNase-bias or technical issues such as PCR bias but also by *trans*-acting factors such as nucleosome remodelers (Xiong et al., 2016). Indeed, discovered nucleosome profiles by Xiong et al., 2016 followed GC oscillations, but also shifts in nucleosomal peaks upon transcriptional activation of some genes, probably due to remodeling complexes, were seen. Thereby, the authors compared profiles between the silent MIC showing no transcription and no activity of remodelers and compared the information of MNase-seq from MICs to those of transcriptional active MACs. A similar study in *Paramecium* would shed more light on the regulation of the highly polyploid MAC in the presence of an active transcription machinery, but isolation of *Paramecium* MIC chromatin is the pie in the sky, at least today.

Not only the bare DNA sequence acts as a *cis*-regulator but also introduced nucleotide modifications have regulatory functions, such as the well-described cytosine methylation (5mC), important for developmental regulation and gene silencing (e.g. Bird, 2002). 5mC can recruit binding proteins comparable to the histone code readers and reduce DNA bendability, although little is known about the effect on nucleosome positioning (Ngo et al., 2016). However, 5mC could not be detected in *P. tetraurelia* and *Tetrahymena pyriformis* (Singh et al., 2018; Hattman et al., 1978). In contrast, N_6 -methyladenine(6mA) was identified in *T. thermophila*, being found preferentially in the linker DNA of well-positioned nucleosomes of Pol II transcribed genes (Wang et al., 2017; Luo et al., 2018) in the MAC. In *Paramecium* MACs, 6mA is enriched between well-positioned nucleosomes that are positively correlated with gene expression (Hardy et al., 2021), which fits the described observations (Figure 3.4): the more nucleosomes, the more 6mA, the more transcription.

3.4.1 Outlook

The presented study revealed unusual features of an unusual genome with open, potentially accessible chromatin that is transcriptionally silent while the occupied gene regions are expressed. Thereby, gene expression is regulated by a divergent Polymerase II complex that seems to crawl along gene bodies, probably due to its association with divergent elongation complexes. The presented study is the first description of *Paramecium tetraurelia*'s MAC epigenome, providing a fundament to dig deeper into the analysis of gene expression regulation in a genome with unusual features.

To build on this, the distribution of histone marks canonically associated with elongation such as H3K36me3 and other DNA binding proteins should be analyzed by ChIP experiments to elucidate the regulation of Pol II. Six homologs of the TFIIS transcription factors necessary for the release of Pol II from backtracking events have been identified based on sequence homology, but the functional role of the vegetatively expressed paralogs have not been studied yet. By pull-down experiments either by the Pol II specific antibody or from transgenic cell lines expressing a tagged Rpb1 subunit, interaction partners of the transcription machinery such as TFIIS and others will be characterized. Technically, the footprint of TFs on DNA can also be gathered by MNase digests, but because of their relatively labile binding to

the DNA in comparison to nucleosome wrapping, their footprint is lost upon harsh digests and upon size selection prior to MNase-seq, thus, their footprints are not included in the presented study.

Furthermore, transcriptional elongation can be studied in terms of the accumulation of not successfully polyadenylated nascent transcripts in the absence of subunits of the Paf1 complex described to be involved in promoting elongation. Pull-downs of Pol II-associated RNAs could shed more light on processivity and efficiency of elongation and, when extended to developmental programs, on how Pol II is positioned on genes upon sexual induction. These experiments should answer (i) how Pol II is released from stalling without canonical elongation complexes and (ii) how Pol II is regulated in development when lncRNA has to be transcribed for scanRNAs.

Since antibody specificity to different phosphorylation states at the CTD cannot be clearly assigned, mass spectrometry analysis of peptides of the Pol II CTD could shed light on the orchestration of phosphorylation. It has been shown that not all heptad repeats need to become phosphorylated in a specific pattern and *Paramecium*'s CTD is therefore still capable of building a platform for transcription factors (Suh et al., 2016).

If the chromatin is mostly open and unprotected, how is spurious transcription prevented? One could think of histone modifications that went missing in the presented ChIP-seq studies, resulting in unoccupied chromatin states from ChromHMM segmentation analysis, but since MNase-seq data revealed chromatin mainly to be naked, this seems unlikely. From extensive phylogenetic studies, homologs of well described chromatin-binding proteins such as heterochromatin-binding protein 1 or HMG-1, the latter being involved in DNA-bending, should be identified, and their role in protecting the DNA from spurious transcription will be studied. Thereby, knowledge from nonhistone proteins being involved in developmental chromosome processing and chromatin binding in *Tetrahymena* (Yao et al., 2007; Kataoka and Mochizuki, 2015) will be extended to *Paramecium*'s vegetative MAC landscape.

It has been shown that heat shock responses result in massive transcriptome alterations which also involves dynamic chromatin remodeling. Still, a global view on chromatin alterations upon stress induction is missing and could now be identified by the established workflow of nuclei isolation, MNase- and ChIP-seq, nucleosome positioning and detection of combinatorial patterns. These regulations probably also involve chromatin remodelers like ISWI and chromatin binding proteins to be identified from mass spectrometry.

The functionality of histone modifications in development and their guidance by a PCR2 complex on probably nascent transcripts has been described recently in *Paramecium* (Miró-Pina et al., 2022) as well as in *Tetrahymena* (Xu et al., 2021), proposing a mechanism comparable to CTGS in yeast but with the drastic result of DNA elimination. In *Paramecium*, it has been shown that TGS can appear not only in development but also at an endogenous locus upon transgene-induced silencing, a mechanism comparable to paramutation, involving core proteins of the RNAi machinery. Two vegetatively expressed proteins of this machinery, *Paramecium* Piwi proteins, seem to be involved, and results from transgene-induced silencing experiments and Ptiwi characterizations will be presented in the following chapter.

Chapter 4

Paramecium tetraurelia Piwi Proteins Silence on the Chromatin Level

Parts of this chapter were recently published in RNA Biology (Drews et al., 2021, 18 October 2021, DOI: 10.1080/15476286.2021.1991114) as

Title

Two Piwis with Ago-like functions silence somatic genes at the chromatin level

Authors

Franziska Drews, Sivarajan Karunanithi, Ulrike Götz, Simone Marker, Raphael deWijn, Marcello Pirritano, Angela M. Rodrigues-Viana, Martin Jung, Gilles Gasparoni, Marcel H. Schulz, Martin Simon

4.1 Background

Due to three successive WGDs, *Paramecium tetraurelia* possesses 15 Piwi (Ptiwi 1-15) proteins, with five of them being extensively studied in terms of their specific roles in *Paramecium* sexual development. Ptiwis responsible for scanRNA and ies-RNA biogenesis show high expression in developmental stages, such as Ptiwi 01/09 and Ptiwi 10/11, while others exhibit a more basal expression pattern throughout the vegetative life cycle. Two of these vegetatively expressed Ptiwis, Ptiwi 13 and Ptiwi 14, were recently described for their participation in the transgene-induced silencing pathway, while Ptiwi 13 additionally contributes to siRNA biogenesis from exogenously introduced double-stranded RNA (feeding-pathway) (Götz et al., 2016; Bouhouche et al., 2011).

Silencing of endogenous genes by truncated transgenes (TG) in *Paramecium* has been described for the first time by Ruiz et al., 1998 and follow-up studies revealed some of the essential molecular mechanisms involved. By injection of DNA corresponding to an endogenous gene into the *Paramecium* MAC but missing up-stream and down-stream regulatory regions resulted in a silencing phenotype of the endogenous gene in *trans* based on sequence homology, and this effect seems to arise in a dosis-dependent manner. Based on these observations, the authors postulated that an aberrant mRNA from the injected transgene could be the initial trigger for induction of silencing and in 2001 they discovered that especially transgenes lacking the 3' UTR trigger silencing (Galvani and Sperling, 2001). This silencing mechanism involves sRNAs and can result in the establishment of stably repressed chromatin or post-transcription RNA degradation (Ruiz et al., 1998), as it was described for cosuppression in fungi, plants, and animals.

TG-induced silencing involves several proteins of the *Paramecium* RNAi-machinery, some of them being exclusively linked to the TG pathway, while others also participate in both TG-induced silencing and processing of dsRNA from exogenous templates. In addition to Ptiwi 13/14, two of the four *Paramecium* Rdrp proteins, Rdr2 and Rdr3, have been shown to be involved in the biogenesis of primary and secondary siRNAs, as well as Dcr1. Usually, the initial trigger for RNAi is a dsRNA, which is probably generated by bidirectional transcription of the transgene locus. However, in the presence of a truncated transgene in *Paramecium*, unspliced aberrant transcripts accumulate, corresponding to the transgene. Almost no antisense transcripts can be detected, which is likely due to a rapid turn-over and processing of the antisense strands into small RNAs. Importantly, sRNAs also cover genomic regions that are not part of the TG but map to the endogenous locus; enabling transitivity, although to a really low extent. No RNAi component seems to be exclusively essential for the accumulation of such 2° siRNAs, but absence of one of each component affects the accumulation of 1° siRNAs. These primaries are generated upon Rdr2 activity on the sense, spliced template synthesizing long antisense RNAs.

Another class of long, unspliced transcripts is probably processed in an Rdr3-dependent manner. This polymerase, which has a highly divergent catalytic domain, was shown to be involved in transcriptional gene silencing (TGS) of surface antigen genes, and, moreover, silencing of Rdr3 resulted in altered expression of several genes, especially in down-regulation of chromatin remodeling genes and up-regulation of genes for transcriptional activity. Hence, Rdr3 could link gene regulation to heterochromatin formation at the transgene and endogenous locus as well as regulate loci apart from TG-induced silencing (Marker et al., 2010; Götz et al., 2016).

Despite extensive studies revealing the function of different RNAi components in TG-induced silencing, it remains unclear what the initial trigger for silencing is and why two Ptiwi proteins are involved. Ptiwi 13 and 14 have been shown to behave differently in TG-induced silencing at altered temperatures, which could be related to the efficiency of silencing (Pirritano et al., 2018). Still, the underlying molecular mechanism is not yet fully understood. Since changes on the chromatin level upon TG-induced silencing were observed in *Paramecium*, the question arises if Ptiwi-bound sRNAs act like piRNAs, which aim to preserve genome integrity by establishing repressive heterochromatin at transposon loci.

During *Paramecium* development, Ptiwi 01/09/10/11 perform such functions, by loading Dicer products and guiding heterochromatin formation on transposon remnants that must be removed during sexual development (e.g., Furrer et al., 2017; Miró-Pina et al., 2022). These observations indicate that *Paramecium* Ptiwis act like Agos but with Piwi-like functions. Traditionally, Piwi proteins select ssRNA from long ssRNA templates and process them into functional sRNAs by trimming with further amplification. Agos, on the contrary, select one ssRNA strand from sRNA duplexes, which are produced by Dicer/Dicer-like proteins. While Agos form the RISC with mRNAs and siRNAs, Piwi proteins historically load 5' U piRNAs in the germline generated independently from Dicer, which further become size trimmed followed by an amplification mechanism, either by the ping-pong cycle or involvement of an RNA-dependent RNA polymerase. It has been proposed that Piwis specialized for transcriptional gene silencing (TGS) in the nucleus, possibly through the targeting of histone modifications, while Agos specialized for post-transcriptional gene silencing (PTGS) in the cytoplasm. Probably all piRNAs in metazoans become 3' 2'-O-methylated after PIWI loading, and this 3' modification likely protects sRNAs from uridylation and degradation. On the contrary, all plant sRNAs, si- and miRNAs, are methylated at the 3' end of both strands of the dsRNA duplex, suspecting that this modification is involved in different pathways (Yang et al., 2006; Li et al., 2005).

The functional segregation of Piwis from Agos is not just black and white; especially the piRNA pathway seems to be highly divergent and not exclusively associated with the silencing of transposons to protect the germline, which was supposed to be their ancestral function, but Piwis also seem to be involved in silencing of somatic genes.

Since *Paramecium*, as all other ciliates studied so far, does not harbor any Agos, it serves as an excellent model to study the wide variety of the Piwi protein machinery. In the following study presented, Ptiwi 13 and 14 bound sRNAs were examined in the context of TG-induced silencing, and several questions were addressed: Why are two Ptiwi proteins involved in transgene-induced silencing, and what are the characteristics of Ptiwi-bound sRNA? Do they show any sequence preferences as was shown for Piwis in *Drosophila*? Could both Ptiwi 13 and 14 perform divergent functions such as Aubergine/Ago3 and PIWI in piRNA biogenesis and amplification? Since vegetatively expressed Piwi proteins in Tetrahymena (Twi) show loading of sRNAs from highly divergent loci such as pseudogenes (Kurth and Mochizuki, 2009; Couvillion et al., 2009), the following study aimed to answer if Ptiwi 13 and Ptiwi 14 load sRNAs apart from the TG-mechanism and can take part in the regulation of endogenous genes.

4.2 Methods

Detailed protocols for the following methods can be found in the material and methods section (chapter 2).

Cell Culture

Paramecium tetraurelia (strain 51 and d4-2) of serotype A were cultivated at 31°C. Potential autogamy induction was excluded via DAPI staining and dense cultures were collected for immunofluorescence staining and IPs. Transgenic cell lines were grown from 14 ° C to 26 ° C until proof of stable transgene expression.

Immunoprecipitation (IP)

Vegetative *Paramecium* cells were simultaneously injected with a linearized plasmid carrying the pTI-/- transgene and plasmid encoding a Ptiwi 13-FLAG or Ptiwi 14-FLAG fusion construct (2.4.7). Cells injected with only the pTI-/- transgene served as a control.

Phylogenetic Analyses

Alignments of Argonaute proteins were performed using Muscle with default parameters (Edgar, 2004). The set of *Paramecium* Ptiwi protein sequences (Bouhouche et al., 2011) was expanded by the curated amino acid sequence of the putative pseudogene Ptiwi 04, which was annotated using its paralog Ptiwi 05 amino acid sequence as a template. Neighbor-Joining method with 1000 bootstrap replicates (Saitou and Nei, 1987; Felsenstein, 1985) was used to infer the evolutionary history of Argonaute proteins. Evolutionary distances were computed using the Poisson correction method with distances measured by number of amino acid substitutions per site (Zuckerandl and Pauling, 1965). Ambiguous positions were removed by pairwise deletion. The final dataset comprised 1.703 positions. MEGA X (Kumar et al., 2018) was used to perform evolutionary analyses.

The following amino acid sequences of Argonaute proteins were used for phylogenetic analyses:

Tetrahymena thermophila: Twi1-Twi11 (Tetrahymena DB), *Paramecium tetraurelia*: Ptiwi 1-15 (Paramecium DB), *Homo sapiens* Ago1: HGNC:3262, *Homo sapiens* Piwil1: HGNC:9007, *Caenorhabditis elegans* Prg1: D2030.6, *Drosophila melanogaster* Aubergine: FBgn0000146, *Drosophila melanogaster* Piwi: FBgn0004872, *Schizosaccharomyces pombe*: SPCC736.11, *Chlamydomonas reinhardtii* Ago1-3: <https://www.plantgdb.org/>

Additional Methods

- Antibody purification (2.4.4)
- Peptide competition assay, western blots and immunostaining (2.4.5, 2.4.3, 2.4.6)
- Immunoprecipitation (IP) sRNA enrichment and periodate treatment (2.4.7, 2.5.4, 2.5.5)
- RNA isolation (2.5.1)

-
- sRNA library preparation and sequencing (2.7.2)
 - sRNA analyses, sRNA signatures, nucleotide content calculation and weblogs (2.8.1, 2.8)

Data Deposition

All raw read data of this study has been deposited at European Nucleotide Archive (ENA), accession no. PRJEB38766.

4.3 Results

4.3.1 Ptiwi Phylogeny and Localization

The evolutionary relationship of the 15 *Paramecium* Ptiwi proteins to other proteins of the Argonaute clade is shown in Figure 4.1A. The phylogenetic tree shows the separation of Argonaute proteins into two subclades: the Ago subclade based on *Homo sapiens* Ago1 and the Piwi subclade based on *Drosophila melanogaster* Piwi. The tree reveals a clustering of *Paramecium* Ptiwi proteins with the metazoan Piwis (green), which are separated from Agos (yellow). Piwi proteins of another ciliate, *Tetrahymena thermophila*, show the same clustering, which fits the description of *Tetrahymena thermophila* having only Piwis but no Agos, just like *Paramecium* (Seto, Kingston, and Lau, 2007).

The two Ptiwis involved in transgene-induced silencing, Ptiwi 13 and 14, do not show a close relationship and therefore do not appear to be a result of one of the whole genome duplication. Ptiwi 14 has one ohnolog, Ptiwi 08, which is expressed during development. Both Ptiwi 13 and 14 were analyzed for their catalytic domains. The catalytic DEDH tetrad of the PIWI domain responsible for slicer activity on the targeted mRNA was identified (Figure B.1), as well as key residues of the PAZ and MID domain (Nakanishi et al., 2013; Bouhouche et al., 2011). Unlike, for example, Ptiwi 01/09, which show an up-regulation during the developmental process that involves sRNA biogenesis and shuttling, Ptiwi 13 and Ptiwi 14 do not show such alterations in gene expression. Although hidden by extreme up-regulation of other genes in Figure 4.1B, both Ptiwis show vegetative expression and these levels are not altered in development.

Among species, Piwi proteins have been described for their ability to shuttle between the nucleus and the cytoplasm depending on the specialized function, and Ptiwis that act in genome rearrangements are likely to shuttle between different nuclei. Antibodies specifically directed against each of the Ptiwis were produced and applied in immunofluorescence staining, specificity was verified by competition assays and western blots (Figure B.3A,B,D). Subsequently, *Paramecium* cells were injected with fusion constructs carrying Ptiwi 13-FLAG or Ptiwi 14-FLAG transgenes to clarify the location, and then stable transgenic lines were forwarded to immunoprecipitation (IP).

Ptiwi 13 showed a cytosolic localization for each approach, using the custom antibody specific to Ptiwi or the FLAG antibody. Several structures could be observed in the cytoplasm, probably due to artifacts of the fixation and binding of proteins to the ER. Ptiwi 13 could also be detected in the MAC by the specific antibody, although to a lesser extent (20% of cells; Figure B.3C), which was also seen in western blots using protein from MAC enriched fractions (Figure B.3D).

Ptiwi 14 localized in the MAC but still emits a faint cytoplasmic signal, which is fostered by a specific Ptiwi 14 signal in the MAC fraction in western blots (Figure B.3). Since injection of FLAG-fusion constructs induces the over-expression of the respective Ptiwi protein (being under the control of the *Paramecium* endogenous calmodulin-promoter directing high gene expression), signals must be compared between custom, specific antibody signals applied to uninjected cell lines and the anti-FLAG signals from overexpression cell lines. Indeed, there were slight differences in localization patterns, but in total, overexpression of both proteins did not interfere with cell division rates (not shown) and localization signals (compared to Figure B.3C). Thus, FLAG-fusion transgenes were used for IP experiments described in the following sections.

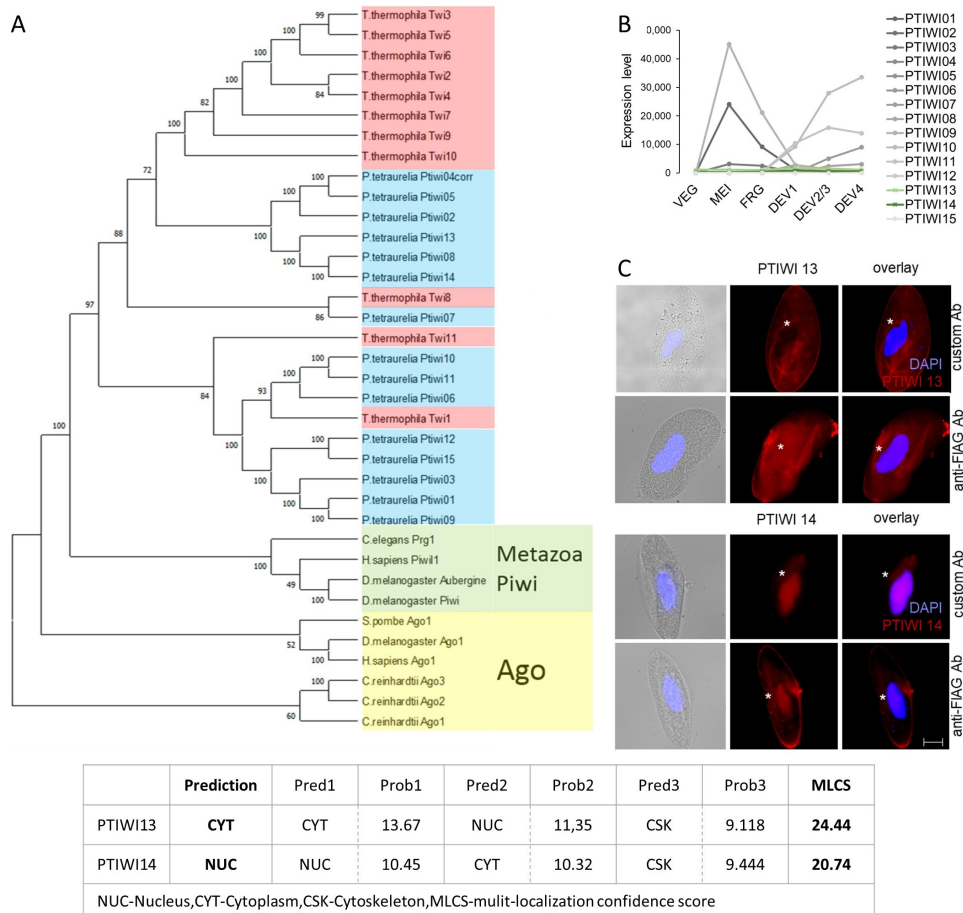


Figure 4.1 (A) Phylogenetic tree of *Paramecium tetraurelia* Ptiwi proteins (blue) in relation to Piwis (Twi) of *Tetrahymena thermophila* (red) and metazoan Piwis (green) and Agos (yellow). Support values are given at nodes. **(B)** Expression values for all 15 Ptiwis at different time points in *Paramecium* development. Data is taken from an autogamy timecourse performed by (Arnaiz et al., 2017; Arnaiz, Meyer, and Sperling, 2020). VEG-vegetative, MEI-micronuclear meiosis, FRG-fragmented MAC, DEV1-4-Development of Anlagen. **(C)** Localisation of Ptiwi proteins in vegetative *Paramecium* cells injected with Ptiwi 13-FLAG (top) or Ptiwi 14-FLAG (bottom). Cells were analyzed by immunofluorescence staining using custom anti-Ptiwi 13 or anti-Ptiwi 14 antibodies labeled with secondary Alexa594-conjugated antibody (red). Cells were additionally stained with anti-FLAG antibody. Other panels show DAPI (in blue), brightfield, and overlay of DAPI and Alexa594 signal with white asterisk indicating MAC position. Representative overlays of Z-stacks are shown. Scale bar is 10 μm and exposure is 2 s. **(Table)** Results of the localization prediction using the ngLOC method. Figure adapted from Drews et al., 2021.

Localization was also predicted *in silico* from amino acid sequences using the ngLOC method (King et al., 2012) that predicted a cytosolic localization for Ptiwi 13 and a nuclear localization for Ptiwi 14 in addition to a multi-localization confidence score for both Ptiwis, with the probability for Ptiwi 13 shuttling between compartments being slightly higher. From these results, Ptiwis appear to have different subcellular localization preferences, although showing some appearance in the other respective compartment, Ptiwi 14 in the MAC and Ptiwi 13 in the cytosol.

4.3.2 sRNA Loading Preferences

Ptiwi-FLAG-transgenes were each injected into transgenic cells harboring the pTI^{-/-} transgene. This transgene contains a truncated version of the endogenous *ND169* gene, which consequently causes the silencing of the endogenous locus by an RNAi-dependent mechanism (Marker et al., 2010). Furthermore, a GFP-marker is introduced under the control of the same bidirectional promoter as the truncated *ND169* gene, allowing for the rapid screening of transgenic cell lines by the fluorescence signal.

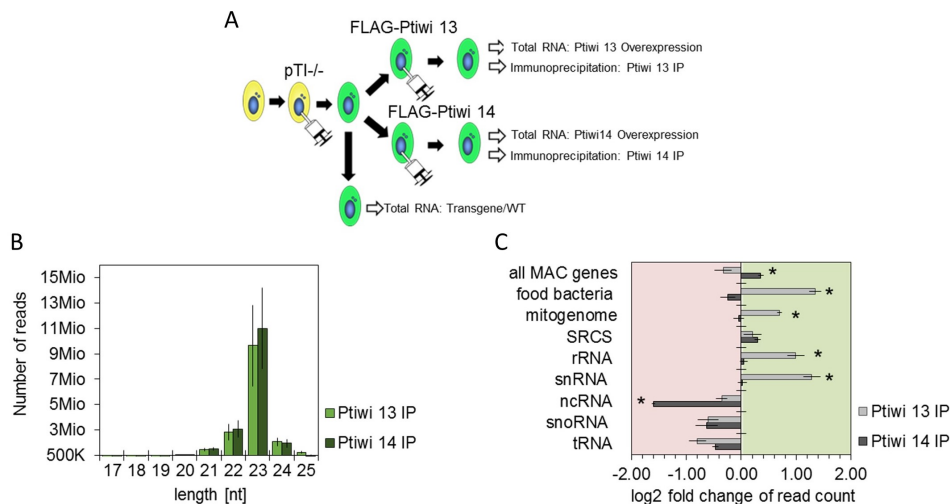


Figure 4.2 (A) Experimental setup. A single cell was injected with the pTI^{-/-} transgene. Once a stable line was established (green), cells were injected with Ptiwi 13/14-FLAG fusion constructs. RNA was isolated from only pTI^{-/-} injected lines (Transgene/WT), overexpression lines prior to IP, and after specific pull-downs. **(B)** Read length distribution of total reads from Ptiwi IPs after adapter and quality trimming. Three IP samples each were analyzed. **(C)** Relative enrichment of RNA reads from Ptiwi IPs as in **B** mapping to different categories of templates. The enrichment was calculated in reference to each individual Ptiwi overexpressing line. * p-value < 0.005. Figure adapted from Drews et al., 2021.

Figure 4.2A shows the experimental approach of the presented study. To verify specific pull-downs of FLAG-fusion Ptiwi proteins, lysates, and pull-downs were analyzed by western blots, decorating the membranes with anti-FLAG antibody. For all IP approaches, specific bands were detected (Figure B.3E). The IP procedure allows the isolation of Ptiwi-bound small RNAs which were prepared for small RNA sequencing in triplicates.

Knowing that piRNA and also siRNA in plants carry a special modification at their 3' end (3' 2'-O-methylation), particular importance was placed on the detection of this modification: aliquots of small RNA isolates from IPs were treated with periodate, followed by a β -elimination resulting in a 3' phosphate on unmodified sRNA, hindering the inclusion in the sRNA library preparation protocol for subsequent sequencing. Otherwise, if sRNAs carry a 3' 2'-O-methyl group, they are resistant to periodate treatment and can enter the sequencing procedure.

To gain insight into the overall composition of Ptiwi-bound sRNAs, their length distribution was analyzed after read preprocessing, including quality and adapter trimming. Figure 4.2B shows that sRNAs, pulled down in both Ptiwi IPs, have a predominant length of 23nt (Figure 4.2B), which is the known siRNA length in *Paramecium*

(Götz et al., 2016; Karunanithi et al., 2019; Karunanithi et al., 2020). To find differences in loading preferences, Ptiwi IP reads were mapped against several different templates mimicking the availability of a mixed RNA template pool (Figure 4.2C). The number of mapped reads for each IP was quantified in relation to the abundance of mapping reads obtained from Ptiwi overexpression lines without enrichment by IP. Those sRNAs represent the overall appearance of sRNAs in Ptiwi transgene lines, whereas Ptiwi IPs should show enrichment or loss of sRNA classes, which is visualized by fold changes in Figure 4.2C.

Obviously, Ptiwi 13 enriches a broad spectrum of RNAs from different origins, such as rRNA and snRNA and, more specifically, Ptiwi 13 enriches for sRNAs from exogenous templates such as mitochondrial RNAs and RNA from food bacteria.

Ptiwi 14, on the contrary, shows an enrichment for sRNAs from MAC protein-coding genes and a set of sRNAs produced from distinct loci in *Paramecium* (small RNA cluster, SRC (Karunanithi et al., 2019)). sRNAs from other RNA templates are quite underrepresented in Ptiwi 14 IPs.

4.3.3 Transgene-Induced Silencing: Loading of 1° and 2° siRNAs

After describing the overall binding preference of endo- and exogenous templates, it was now essential to study the role of each Ptiwi in transgene-induced silencing, first described by Bouhouche et al., 2011.

The structure of the transgene described above is shown in Figure 4.3 below the endogenous *ND169* locus, which becomes silenced at the chromatin level upon injection of the truncated *ND169* transgene. Endogenous *ND169* is not involved in any RNAi mechanism but serves as a reporter gene due to its function in the final step of trichocyst discharge: silencing hinders trichocyst extrusion, which can be examined by stimulating cells with acid, and non-discharge serves as a quick control for successful induction of silencing (Froissard et al., n.d.).

The truncated transgene was modified since its first description by Marker et al., 2010, and now contains two deletions: one at the 5' coding region (ND-1) and one at the 3' coding region including the downstream UTR (ND-2). The truncated gene still contains five introns (Götz et al., 2016). The design of the transgene, having deleted regions, allowed the detection of not only primary (1°) siRNAs, but also secondary (2°) siRNAs. In a wildtype cell line, the *ND169* locus does not produce any sRNAs, so sRNAs mapping to deleted regions must appear in transgene-induced silencing manner. This would be the production of 2° siRNAs by transitivity of 1° siRNAs attacking the endogenous *ND169* locus, probably by targeting a nascent transcript. Therefore, sRNAs mapping to the ND-1 and ND-2 loci are termed 2° siRNAs in the following.

Conversely, regions existing in the endogenous gene and the transgene, called ND-gene, can be assigned with 1° and 2° siRNAs that cannot be dissected from each other. However, due to the overall low abundance of 2° siRNAs (ten times less than 1° (Götz et al., 2016)), the NDgene region shows predominately 1° siRNA accumulation.

Having a look at the *ND169* locus simply by aligning reads to all loci, it becomes evident that both Ptiwis load siRNAs originating from all loci, so 1° and 2° siRNAs (Figure 4.3C). Figure 4.3D shows only reads mapping to the the ND-1 region as representative for 2° siRNA producing loci, since ND-2 mapping reads were low abundant. Those sRNAs are predominant in length, showing a distinct size peak of 23nt for both 1° and 2° siRNAs. When again comparing the loading preferences among

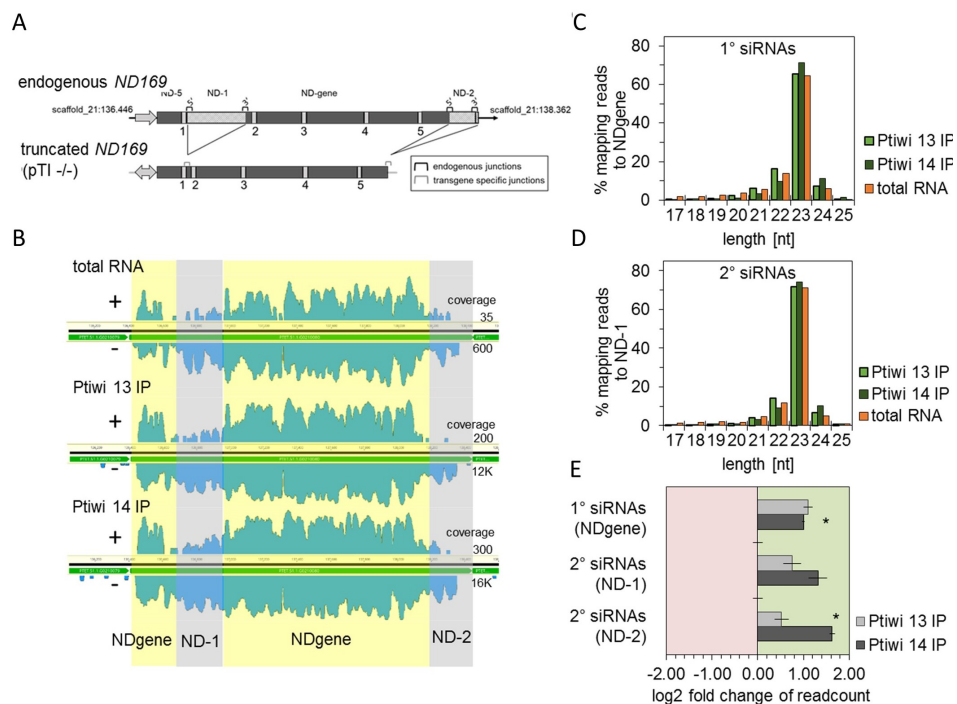


Figure 4.3 (A) Schematic of the truncated transgene is shown below the endogenous *ND169* locus. Introns are numbered and brackets symbolize specific junctions. Regions are not part of the transgene (ND-1 and ND-2) are shaded. (B) Patterns of siRNAs mapped to the endogenous *ND169* locus are shown as coverage tracks with prior separation of siRNAs by their direction (sense/antisense). Regions accounting for 1° siRNAs (NDgene, yellow) and 2° siRNAs (ND-1 and ND-2, grey) were highlighted. Coverage track shows data for one replicate each in log scale with numbers on the right indicating untransformed sense and antisense coverage. (C) Read length distribution as in Figure 4.2B but for 1° siRNAs and 2° siRNAs from Ptiwi IPs and total RNA from pTI-/- injected cells mapping to the NDgene locus and (D) the ND-1 locus (E) Relative enrichment of RNA reads from Ptiwi IPs mapping to different regions of the transgene. Enrichment was calculated as in Figure 4.2C. * p-value <0.005. Figure taken from Drews et al., 2021.

both Ptiwi IPs, taking into account the overall abundance of sRNA from overexpression, it becomes obvious that Ptiwi 14 loads more 2° siRNAs (Figure 4.3E). Since these arise from the endogenous locus, it is tempting to speculate that these 2° siRNAs arise from a nascent transcript in the MAC and are subsequently loaded by Ptiwi 14.

sRNA Loading Preferences I: Strandedness

The specific loading preference for 23nt sRNAs becomes even more obvious when one looks at the ratio of all available sRNAs that could be loaded (total RNA) and the actual loaded ones (Figure 4.4A). For several RNAs, such as those from food bacteria, mitochondria, or rRNA, Ptiwis specifically select the 23nt RNAs from a pool of sRNA, which seem to be produced by different mechanisms. For example, mitochondrial mapping sRNAs comprise $\approx 30\%$ 23nt sRNAs in the total RNA fraction, while in Ptiwi 13 and 14 IPs, almost 70% of the loaded mitochondrial mapping RNAs are 23nt long, arguing for a size selection mechanism. Transgene-associated RNAs, in contrast, appear to be of precise length, even in the total RNA sample, suggesting

that these RNAs were already generated in a more precise manner prior to Ptiwi loading.

Not only length of the sRNA but also directionality (sense/ antisense) and the 3' end modification can be an indicator for mechanism of siRNA biogenesis and preferences of each Ptiwi protein. The antisense ratio of transgene-associated siRNAs is greater than 0.96; therefore, Ptiwis load predominantly antisense siRNAs (Figure 4.4B) which was already suggested from coverage plots (Figure 4.3). Additionally, those RNAs

must carry a 3' modification as it could be concluded from periodate treatment: the antisense ratio was not altered upon periodate treatment, indicating that antisense siRNAs are resistant and thus carry a 3' 2'-O-methylation (Figure 4.4C). To be more precise, antisense ratio even raised upon periodate treatment, indicating that few sense RNAs are not modified, arguing against the occurrence of dsRNA methylation at both strands as it is described for plants.

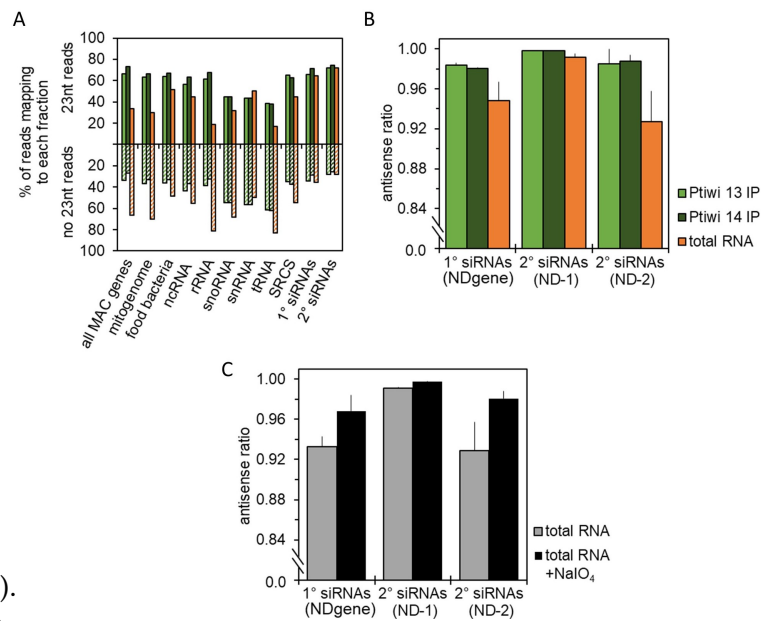


Figure 4.4 (A) Number of 23nt reads mapping to each indicated genomic feature and transgene regions were related to the total number of reads of other sizes. Proportions were calculated for the mean of Ptiwi IPs and RNA from pTI^{-/-} injected cells serving as control (total RNA). (B) Average antisense ratio of reads from Ptiwi IPs triplicates. (C) Average antisense ratio of small RNAs from duplicate pTI^{-/-} transgene samples (total RNA, untreated) and the small RNAs from the same samples treated with sodium periodate (+NaIO₄). Figure taken from Drews et al., 2021.

sRNA Loading Preferences II: Nucleotide Composition

It has been well described that Argonaute proteins themselves have sequence preferences for sRNA loading, e.g. for Piwi in *Drosophila* or Ago1 in *Araidopsis*. Moreover, it has been observed for *Paramecium* that Dicer enzymes have sequence cleavage preference, introducing sRNA signatures prior to Ptiwi-loading (Sandoval et al., 2014).

Based on this knowledge, sequence logos for 23nt antisense transgene-associated siRNAs (being the most abundant ones) were analyzed, which revealed a 5' uridine (5' U) preference for 1° and 2° siRNAs in the total RNA fraction (Figure 4.5A) as well as for the Ptiwi-bound sRNAs (Figure 4.5B and Figure B.4). Comparison of logos between both Ptiwis reveals a 5' U preference that is remarkably seen for Ptiwi 14 IPed RNAs.

Sequence logos themselves could give an idea of the Ptiwi-bound siRNAs biogenesis mechanism (Figure 4.5C). If Dicer would be involved, a 5' U preference should result in an A preference at position 21 in the opposite strand. This is not seen in the logos

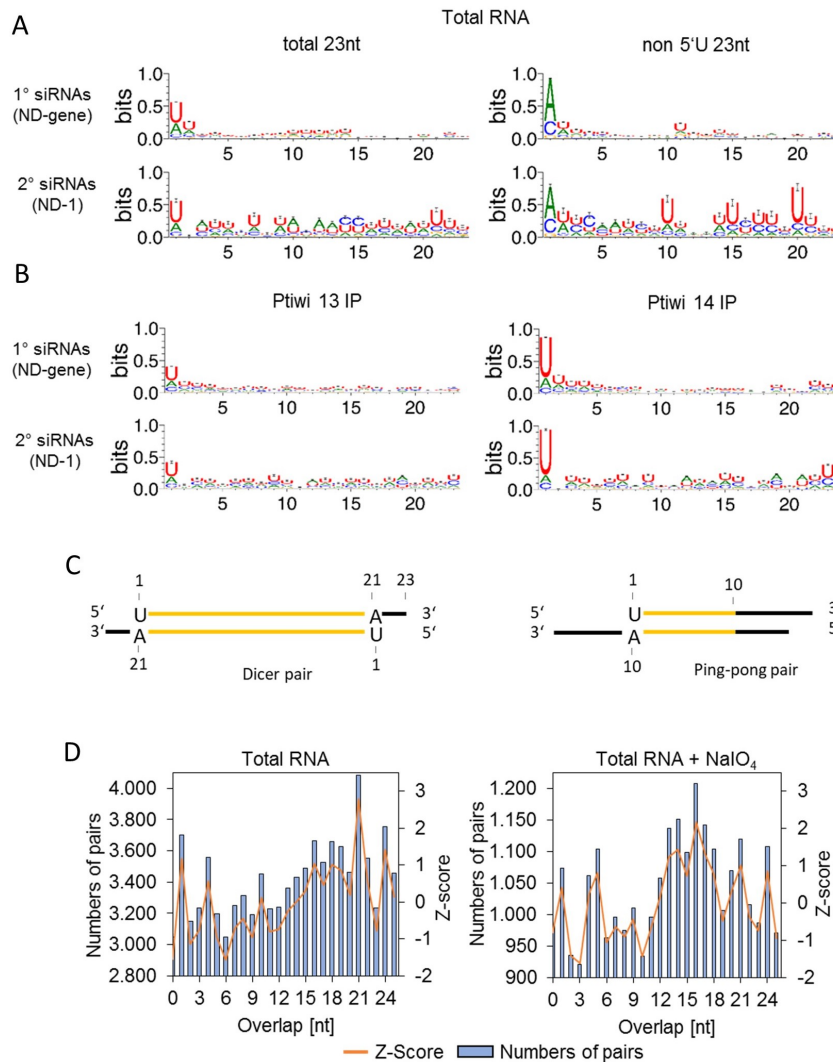


Figure 4.5 (A) Sequence logos of 23nt antisense reads from pTI^{-/-} injected cell lines (total RNA) mapping to transgene regions. Logos were created for either all sequences or the ones without 5'-U. (B) Sequence logos of 23nt antisense reads from Ptiwi IPs. (C) Cartoon of sRNA logos and overlap signatures to be expected from each of the different sRNA biogenesis mechanisms. (D) Overlaps of small RNAs from 17 to 25nt of untreated, total RNA and the same small RNAs treated with sodium periodate (+NaIO₄) were calculated. Y-axis shows numbers of predicted pairs of the respective size. Postive Z-scores (orange) indicate high probability of overlapping-pair formation.

of Ptiwi-bound siRNAs. The same holds true for a ping-pong signature: for the amplification mechanism, piRNAs show an A preference at position 10 in non-5'-U reads, which cannot be detected as well.

Since sequence logos did not reveal the biogenesis mechanism of siRNAs, reads were analyzed for their overlapping pattern, expecting 10nt or 21nt overlaps as displayed in (Figure 4.5C). Overlapping pairs prediction (Figure 4.5D) displays a slight peak of reads that overlap by 21nt. This resembles 23nt reads overlapping with 3' 2nt overhangs, which is well described for Dicer generated siRNAs (Ma, Ye, and Patel, 2004) (Figure 4.5C). The overlap is not as pronounced as in siRNA duplex analysis from *Drosophila* RNA (Antoniewski, 2014) since Ptiwis load only one strand (Figure 4.4) while the passenger strand is degraded, resulting in a number of reduced of overlapping pairs.

Strand-specific selection is again documented in Figure 4.5 D (right), showing that for periodate-treated siRNAs, overlapping pairs of 21nt could not be significantly predicted, leading to the conclusion that mainly antisense siRNAs are methylated, and the siRNA mate is lost. Furthermore, those siRNAs also show a 5' U preference, suggesting that methylation and sequence preference cooccur on the same molecule (Figure B.5).

sRNA Loading Preferences III: U-Content

The sequence logos revealed not only a 5' U preference but an overall high uridine content of bound siRNAs (Figure 4.5). To examine whether this preference is due to the template sequence itself or selective loading of siRNAs, all siRNAs mapping to the transgene and endogenous *ND169* locus were analyzed. Although most stretches of the *ND169* locus show an antisense preference, the proximal region of the promoter (ND-5) shows divergent patterns (Figure 4.6A): the ratio of sense to antisense siRNAs is almost 50:50. In particular, this region offers a higher uridine content in the sense strand (Figure 4.6B) which is also seen in Ptiwi-bound sRNAs: when comparing the U-content of *in silico* diced ND regions to the U-content of total RNA and Ptiwi-bound RNAs mapping each loci, it becomes apparent that the sense bias of the promoter proximal ND-5 region correlates with the enrichment of U-rich sRNAs, mainly introduced by Ptiwi 14 selective loading (Figure 4.6C).

For all other regions, the U-content of the more abundant antisense 1° and 2° siRNAs is higher than that of sense siRNAs and especially, the U-content of Ptiwi IPed RNAs is higher than for *in silico* diced sRNAs. Thus, Ptiwi 13 and Ptiwi 14 strand selection seems to include a general preference for U-rich sRNAs.

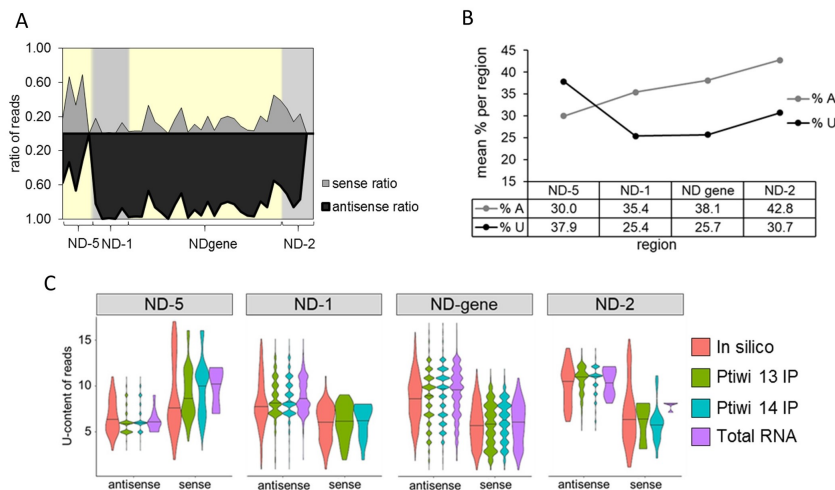


Figure 4.6 (A) Simplistic coverage track (50nt windows) of total sense/antisense reads from pTI-/- injected cells mapping to the *ND169* locus. (B) Adenine and uridine percentage of the sense RNA transcript from each region. (C) U-content of reads from Ptiwi IPs with violin density representing number of reads. *In silico* data is generated by counting U-content of all possibly generated 23mers of the DNA sequence. Figure taken from Drews et al., 2021.

4.3.4 Ptiwis Load sRNAs from Endogenous sRNA Producing Clusters

The data presented in the previous section indicate that Ptiwis are involved in the loading of siRNAs associated with transgenes with specific preferences, revealing

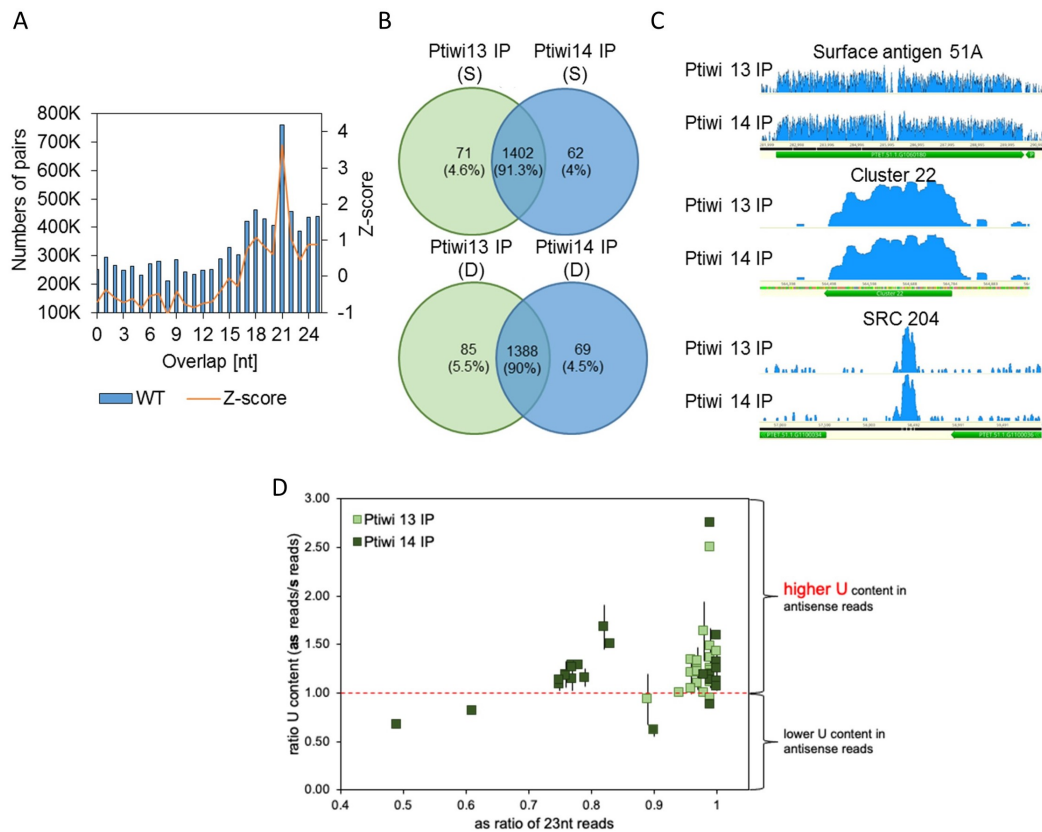


Figure 4.7 (A) Overlaps of endogenous 17-25nt sRNAs isolated from wildtype RNA, calculated as in Figure 4.5D. (B) Venn diagrams of sRNAs mapping to SRCs (small RNA clusters) found in Ptiwi IPs. Numbers indicated amount of SRCs detected and the proportion of each fraction of the total found SRCs. Sonified (S) and dounced (D) correspond to sample preparation technique. (C) Coverage tracks of Ptiwi-bound sRNAs mapping to endogenous loci (surface antigen 51A, cluster22 and SRC 204; data not normalized). (D) 23nt reads from Ptiwi IPs were mapped to SRCs and U-content of each read was counted. Data was filtered for the top 23 SRCs with the highest antisense ratio and a minimum of 5,000 mapping reads. Mean of median U-content of antisense reads in ratio to U-content of sense reads is plotted on y-axis while x-axis shows the antisense ratio of read count. Deviation in ratio of U-content for three IP replicates is represented by error bars.

some divergence between both examined Ptiwis. Since Ptiwi 13 and 14 are expressed during *Paramecium* vegetative growth irrespective of a transgene background, it was worth examining their role in the *Paramecium* endogenous siRNA biogenesis pathway. siRNAs originating from distinct loci recently described in the *Paramecium* genome (Karunanithi et al., 2019) show a read length preference of 23nt and a predominant overlap of 21nt (Figure 4.7A), which is comparable to the investigated transgene-associated sRNAs, again indicating a Dicer-dependent biogenesis pathway.

For the 2,602 SCRs, most can also be identified in IPs of Ptiwi 13 and 14 (Figure 4.7B,C); strikingly, the majority of them (> 90%) are loaded in both Ptiwis, independently of the sample preparation procedure. Ptiwis do not only load sRNAs from described clusters, but also from surface antigen genes that produce sRNAs in a manner not yet fully understood and, from a large cluster on scaffold 22 (cluster22), which lies between two convergent genes and produces a high amount of sRNAs, which were

also identified in both IPs (Figure 4.7C). In terms of SRC derived sRNAs, the Ptiwi-loaded ones which have high antisense ratios, also show a higher U-content in antisense reads. This is seen by computing the U-content of the antisense reads and correlating it with the overall antisense ratio of at least 5,000 reads from specific SRCs (Figure 4.7D). This again leads to the hypothesis of loading preferences of Ptiwis for uridine-rich sRNAs, as it was already shown for transgene-associated siRNAs.

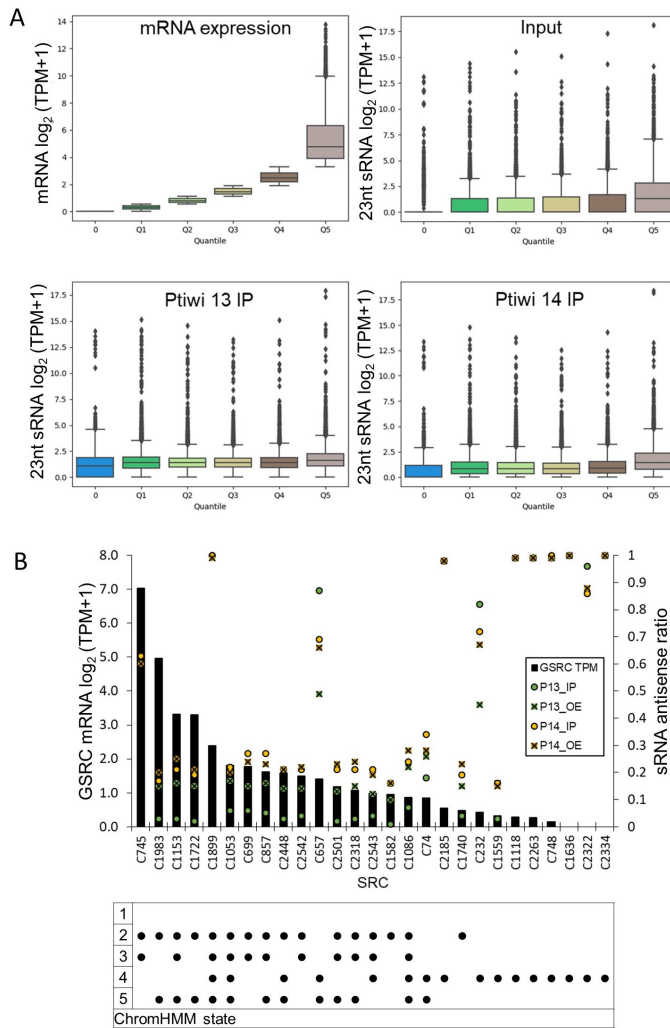


Figure 4.8 (A) Protein-coding genes were classified by their expression from low to high (0,Q1-Q5) and accumulation of 23nt sRNAs was calculated for all genes in each expression group. sRNAs from wildtype and both Ptiwi IPs were analyzed amongst triplicates. **(B)** Genes associated with SRCs (GSRCs) were analyzed according to their mRNA expression level (black bars) and the antisense ratio of small RNAs mapping to the corresponding SRC located in the respective gene (squares/triangles) Only those GSRCs were analyzed, which could be assigned to 100% to one SRC. SRCs had to raise at least 2,000 100% mapping reads of 23nt length and which were shared among all four samples. Below chromatin states allocated to each SRC are indicated (chapter 3, Figure 3.6)

sRNAs originating from endogenous loci in genes are loaded by both Ptiwis which seems to be irrespective of the gene expression level. Genes in all quantiles classified from low to high expression produce small RNAs (Figure 4.8) (Input), and highly expressed genes are associated with more sRNAs. However, sRNAs from all expression groups are loaded into both Ptiwis, even those from silent genes seem to be enriched. A clear function of these sRNAs cannot be defined, since the precise read alignment procedure only allows analyses of sRNAs that act on the respective gene/mRNA (*in cis*), while it is known that sRNAs are certain to also act on divergent RNA templates (*in trans*) - templates which could not be identified in this study.

A rather simple approach to identify the role of Ptiwi-bound sRNAs is summarized in Figure 4.8B: Genes with one SRC (GSRCs) were ranked by their expression and the antisense ratio of sRNAs mapping to each SRC were plotted aside. The accumulation of antisense, Ptiwi-bound

siRNAs in GSRCs with low expression (most right side of the plot) could be an indicator for a classical siRNA-mediated silencing mechanism. Nevertheless, this does not seem to be the exclusive function of antisense siRNAs, since also high expressed GSRCs are associated with high amounts of antisense siRNAs.

Furthermore, the chromatin states at each SRC were analyzed, gathering data on vegetative wildtype MAC chromatin described in chapter 3. Although it cannot be withdrawn that Ptiwi overexpression due to the experimental IP approach led to changes on the chromatin level, the analyzes could give a preliminary insight in sRNA appearance in clusters associated with chromatin states at the endogenous level apart from the transgene-induced silencing model. SRCs in silent genes (right side of the plot) are associated with Ptiwi-bound antisense siRNAs and the chromatin state 4, which is almost devoid of any epigenetic mark and associated with low gene expression along the MAC epigenome. This observation has to be considered as provisional, since only a small, highly filtered set of GSRCs was investigated. Additionally, SRCs have to be analyzed in the chromatin state of the whole GSRC: do these sites specifically recruit marks in GSRCs or is the whole GSRC already covered with activating histone marks? Based on MNase data it can be suggested, that gene bodies show low epigenetic signals and SRCs could be clusters for attraction of nucleosomes and histone modifications. For example, the SRC in the highest expressed GSRC (C745, most left side of Figure 4.8B) lies in the middle of a gene, arguing against the SRC association to state 2 and 3 simply due to its location at the TSS and association with the +1 nucleosome.

4.4 Discussion

The aim of the presented study was to answer why two of the 15 *Paramecium* Ptiwi proteins are involved in transgene-induced silencing, ending with the unexpected finding that both Ptiwis also load sRNA from *Paramecium* endogenous sRNA producing clusters apart from transgene-induced silencing. Thus, transgene-induced silencing and endogenous sRNA accumulation seem to share some genetic requirements, making the transgene system a suitable model to study endogenous sRNA accumulation. Moreover, transgene-mediated silencing by which additional copies of a gene cause silencing of all endogenous homologous loci is not yet fully understood in many species, including *Paramecium*. The study of two involved proteins was supposed to shed light on the molecular mechanism.

Biochemical Properties of Ptiwi-bound sRNAs

From the classical separation of Ptiwis from Agos, by selecting and processing ssRNAs from long templates by trimming rather than selecting strands from small RNA duplexes, and knowing that *Paramecium* does not harbor Agos but Ptiwis, a preference for sRNA loading similar to that of Piwi proteins was expected. Moreover, both investigated Ptiwis should be able to generate their own sRNA products since both have the catalytic tetrad necessary for RNA cleavage.

However, 1° transgene-associated siRNAs have been shown to be mainly dependent on previous Dicer cuts (Götz et al., 2016), ruling out a ping-pong-like sRNA biogenesis pathway involving both Ptiwis. This was fostered by an overlap of 21nt in bulk sequencing of transgene-associated siRNAs, probably resulting from Dicer cuts, and a missing internal A-preference in Ptiwi-IPed siRNAs, which could have been expected from classical PIWI-bound sRNAs.

Hence, although the ping-pong cycle in *Drosophila* involves different PIWIs, which would argue for the involvement of two Ptiwi proteins, this does not hold true for the examined TG-mechanism in *Paramecium*. In contrast, Ptiwis load strands from Dicer products rather than contribute to a ping-pong like sRNA amplification mechanism, thereby exposing a non-canonical function of Ptiwis.

The functionality of sorting primary and secondary RNAs into different Ptiwis is ruled out by the presented data as well since both Ptiwis load 1° and 2° siRNAs. From these results, it seems like both Ptiwis act somewhat redundantly.

The argument of redundancy could be emphasized by several observations on Ptiwi-bound siRNAs: Both Ptiwis load sRNA from endo- and exogenous templates as well as 1° and 2° transgene-associated siRNAs, have sequence preferences for 23nt anti-sense siRNAs and show a slight 5' U preference. But paying attention to differences in those categories, it can be assumed that each Ptiwi has specific preferences.

Especially, localization studies by western blots and immunofluorescence staining attribute both Ptiwis to shuttling between cytoplasm and nucleus while Ptiwi 14 being more prone to enter the MAC. Shuttling for Argonaute proteins is not uncommon and is described, for example, for plant AGO4, which loads siRNAs in the cytoplasm and shuttles to the nucleus. Although localization signals from immunofluorescence stainings are not strong for both Ptiwis, the shuttling and localization is additionally fostered by in-silico predictions. It has been shown in 2016 by Götz et al., that transgene-induced silencing involves changes on the chromatin level at the endogenous locus and silencing-signals must be transmitted to the nucleus, probably by shuttling of Ptiwi-siRNA complexes between cell compartments. However, direct evidence for Ptiwi 13/14 associated guidance of sRNAs and induction of chromatin changes, as could be shown from cross-linking studies in yeast and plants, where Ago interacts directly with nascent transcripts in the nucleus (e.g. (Wierzbicki et al., 2009)), is still missing in *Paramecium*.

Apart from differences in localization prediction, Ptiwi 14 furthermore shows stronger 5' U preferences for loaded siRNAs and loads more 2° siRNAs than Ptiwi 13, which in turn preferentially binds more sRNAs from exogenous templates. Given that Ptiwi 13 is involved in the feeding-pathway in *Paramecium*, its preference for loading exogenous RNAs fits quite well. In fact, loading preferences and selectivity for sRNAs from Dicer products have already been described for Ptiwis being involved in genome rearrangements, selecting scanRNAs or iesRNAs from Dcl2/3 and Dcl5 products. Here, different Ptiwis are involved (Ptiwi 01/09 and Ptiwi 10/11) as ies- and scanRNAs are generated in a different manner. Sequence preferences such as 5' UNG signature for iesRNAs are generated by Dicers rather than by the individual Ptiwis strand selection.

Transgene-Induced Silencing and Endogenous siRNA Accumulation Show Similar Characteristics

In comparison to developmental sRNAs, 23nt transgene-associated RNAs carry a 3' 2'-O-methylation, which has not been studied for *Paramecium*'s scan-/iesRNAs yet. This modification is found in various sRNA species, differing among pathways, probably protecting sRNAs from exonucleases and regulating the sRNA binding-affinity to different Ago/Piwi proteins (Farazi, Juraneck, and Tuschl, 2008): In flies, single-stranded piRNAs as well as siRNAs are modified at their 3' end and so are

miRNA and siRNA duplexes in plants, while miRNAs in flies are not modified (Horwich et al., 2007). The modification is introduced by the methyltransferase Hen1, first identified in *Arabidopsis* (Yu et al., 2005) and Hen1 homologs were also identified in *Paramecium* (Marker et al., 2010) and *Tetrahymena* (Kurth and Mochizuki, 2009). The introduction of the 3' methyl group seems to contribute to selection of strands by Ptiwi 13/14 in TG-induced silencing pathway. This strand selection is enforced by a 5' U preference and an overall higher U-content. Interestingly, Hen1 has been localized in the *Tetrahymena* MAC during sexual development, although a vegetative MAC localization was not precluded. Therefore, Ptiwi 14 shuttling to the MAC and interaction with Hen1 could be a feasible model for functional RISC assembly in *Paramecium* (Kurth and Mochizuki, 2009).

Apart from the transgene-induced silencing pathway, the presented study revealed that the two vegetatively expressed Ptiwis are also involved in the loading of sRNA from *Paramecium* endogenous clusters. The role of these sRNAs is not fully understood, since, on the global level, those siRNAs show no preferences or a link to transcriptional repression. However, Ptiwis again selects antisense sRNAs from clusters with higher U-content in their antisense DNA strand. Elucidating how Ptiwis act on the chromatin level in vegetative growth is one of the upcoming challenges. From GSRC analysis, silent loci with higher antisense siRNAs showed absence of epigenetic marks. This analysis is reduced to an extreme level at the moment as only a small subsets of GSRCs has been analyzed but one could think of antisense siRNAs counteracting gene expression by hindering nucleosome assembly at gene bodies. How exactly this siRNA trigger could be translated to the chromatin level cannot be envisioned from the collected data, but aberrant transcripts from pausing Pol II (chapter 3, Figure 3.7) can be prime candidates for siRNA attacks and the recruitment of chromatin-modifying enzymes. Thereby they would hinder nucleosome recruitment, which would be mechanistically different to CTGS in yeast.

In *Tetrahymena*, 27-30nt developmental scnRNAs carry a 3' 2'-O-methylation which is introduced by the methyltransferase Hen1 on ssRNA after Twi1-scanRNA loading in the parental MAC (Kurth and Mochizuki, 2009). Endogenous 23-24nt sRNAs from distinct clusters seem to be unmodified (Kurth and Mochizuki, 2009; Lee and Collins, 2006; Couvillion et al., 2009), which also holds true for sRNAs associated with the *Paramecium* vegetative cluster, as previously shown by siRNAs originating from cluster 22 (Marker et al., 2010). Interestingly, constitutively expressed 23-24nt siRNAs in *Tetrahymena* mapped to a small number of genomic loci show an overall higher U-content and are Dicer-dependent (Couvillion et al., 2009). The function of these siRNAs has not yet been identified, and further studies on this sRNA class are missing, but at least for a subset of 23-24nt sRNAs, it has been shown, that they are indeed 3' methylated and loaded into Twi8, which localizes to the vegetative MAC. Thus, these sRNA-Twi complexes could contribute to TGS (Couvillion et al., 2009). More studies are needed to examine whether (i) the *Paramecium* Hen1 homolog is responsible for methylation of Ptiwi 13/14 bound siRNAs, if (ii) this occurs in the nucleus in interaction with Ptiwis, and if (iii) also SRCS are affected. It seems like Ptiwi 13/14 act on the endogenous loci and transgene and in the same manner, with slight 5'U preferences not as strong as in scanRNAs and no mutual-exclusively loading of 1° and 2° siRNAs.

Strand selection in other species is controlled by not only 3' modifications and 5' nucleotide preferences, guiding the sorting of sRNAs into each Argonaute pathway, but also by thermodynamics of binding and binding quality to the targeted mRNA.

Todesco et al., 2010 showed that miRNA binding and regulation are impaired by introducing mRNA mimics that impair perfect binding of the miRNA to mRNA and consequently alter mRNA abundance. Number of hydrogen bonds at 5' ends influences which strand is loaded into RISC (Schwarz et al., 2003). Targeting of sense mRNAs would then indeed bias for loading of antisense siRNAs, as it is seen in Ptiwi IPs, and also infer with loading of sense siRNAs from the ND-5 region. However, the higher uridine content in the sense strand should shift the preference towards sense siRNA loading, resulting in a sense/antisense siRNA population from this region.

4.4.1 Outlook

Whole genome duplications in *Paramecium* resulted in a plethora of Ptiwi paralogs that show subfunctionalization in different pathways during a *Paramecium*'s life cycle. Ptiwis participate in processing of exogenous RNA, developmental genome processing by shuttling different small RNA species, and - as shown by the presented study - contribute to transgene-induced silencing additionally to loading of endogenous RNAs. Several mechanistic aspects, asking for complex biochemical approaches, need to be enlightened to understand the Ptiwi operation and coordinated loading fully.

In terms of TG-induced silencing, the objective is to illuminate the interaction of Ptiwi 13 and 14 with RNAi machinery proteins, specifically methyltransferase Hen1. Ptiwi pull-down experiments with subsequent mass spectrometry analyses should list interaction partners, expecting Ptiwi 14 to be associated with nuclear shuttling proteins. By periodate treatment of Ptiwi-IPed RNAs, the biochemical modification of 3' methylation could be shown, which could further extend to pull-downs in Hen1 knockdown background. The experiment would need to be performed vice versa to verify interaction of partners.

Preferences for sRNA uridine composition could be verified by injection of transgenes with altered nucleotide sequences, e.g. higher or lower U-content. Seeing whether U-rich sRNAs will be more preferentially loaded and lead to efficient silencing is of great future interest. The question of which sRNA properties prone them to be loaded into Agos or Piwis is of future interest amongst various species, and studies in *Paramecium* could contribute to understanding Argonaute protein loading preferences.

Apart from TG-induced silencing, a characterization of endogenous small RNA producing clusters and conditional changes in abundance and Ptiwi loading is of high interest. Although shared among serotypes, some clusters show a serotype-specific expression, allowing for differentiation by sRNA accumulation rather than by mRNA expression. These clusters were identified using short stack that de-novo predicts miRNAs by searching for RNA-folding and stem-loop structures. No miRNA cluster have been identified for the *Paramecium* genome. Nevertheless, it has been postulated in yeast that transcripts show backfolding to give rise to dsRNA hairpins which can be processed by Dicer, resulting in accumulation of endogenous sRNAs, thereby regulating stress responses and transcription termination. Such biosynthesis mechanism for endogenously arising sRNAs has not been investigated for *Paramecium* loci until this day (Woolcock et al., 2012; Castel et al., 2014).

RNAs from the $\approx 2,600$ cluster show a predominant length of 23nt with a slight antisense preference but also accumulate sense siRNAs. $\approx 1,300$ SRCS lie in protein-coding genes (GSRCs), probably assigning endogenous sRNAs to an extent to gene regulation. The sRNAs are predominately produced from mRNAs, cause no intron

mapping sRNAs could be detected, ruling out a de-novo synthesis on a genomic template. Thereby, a paradox can be observed: genes with high gene expression show an accumulation of sRNAs, raising doubts on a silencing function of the small RNAs on GSRCs. These observations were further promoted by Ptiwi-bound sRNAs that are linked to high and low expressed genes (Figure 4.8). sRNAs from clusters are dependent of both Rdr1 and Rdr2, and especially Rdr1 knockouts leads to massive transcriptome changes, even for other RNAi components such as Ptiwi 12, 13 and 14; thus RNAi components are linked to endogenous gene regulation. Furthermore, phased clusters are Rdrp dependent and show a higher gene expression upon Rdrp slicing.

The function of these endogenous sRNAs remains to be uncovered, but few pathways have been shown to involve endogenous sRNAs: siRNAs control the number of DNA copies in DNA replication in *Oxytricha* (Ciliates, Spirotrichea (Khurana et al., 2018)) and, as already mentioned, RNAi components shape transcriptional responses to stress and further control transcription in parallel to replication by sRNAs in yeast (Woolcock et al., 2012; Castel et al., 2014).

Such, if loss of endogenous (antisense) siRNAs would result in changes in chromatin conformation upon transcriptome alterations or induce the latter ones consequently demands for MNase- or ChIP-seq experiments from Ptiwi silencings. This will be investigated in follow-up studies and give a broader insight in the function of endogenous sRNAs.

Chapter 5

General Discussion and Future Perspective

Controlled regulation of gene expression is the main contributor to an organism's ability to react to external stimuli, fluctuating environmental conditions, and optimal energy management. However, maintaining a robust and precise gene expression, uncoupled from external variability, is equally essential. Both processes, gene plasticity and expression stability, are tightly regulated by core proteins of the transcriptional machinery as well as epigenetic modulators, which control accessibility to genes by reversible changes in the conformation of the chromatin. The genetic toolbox of organisms to regulate gene expression differs among kingdoms, with the unicellular ancestor of metazoans already having a rich repertoire of genes that are required for cell signaling and transcriptional regulation. It is believed that the cooption of ancestral genes into new functions was an essential mechanism in the evolution of multicellularity and tissue differentiation. The latter processes involve complex gene regulatory networks, receptor-ligand evolution, and epigenetic memory (Sebé-Pedrós, Degnan, and Ruiz-Trillo, 2017).

Albeit multiple unicellular model organisms have been extensively studied in terms of their gene regulatory machinery and serve as good models for a broad understanding of biological dogmas (e.g. the core set of the Pol II complex is highly conserved from yeast to human (Spåhr et al., 2009)), differences in genome architecture and catalytic residues must result in divergent ways of gene expression regulation among species. Unicellular organisms such as ciliates show regulated gene expression profiles in reaction to external stimuli. They harbor factors of guided gene expression and a proportion of the repertoire of metazoans transcription factors in addition to epigenetic inheritance. Therefore, ciliates provide excellent models to study the first glances of the complexity of early multicellular cell types and the regulatory principles that orchestrate them, including transcription factors and chromatin dynamics (Sebé-Pedrós, Degnan, and Ruiz-Trillo, 2017; Drews, Boenigk, and Simon, 2022). In particular, ciliates provide excellent models for studying epigenetic landscapes, as they harbor a transcriptionally active nucleus and a silent germinal nucleus in an isogenic background.

Paramecium tetraurelia is the favorite ciliated model organism of many researchers around the world, focusing on understanding gene expression patterns in the *Paramecium* life cycle and epigenetic inheritance involving genome rearrangements and RNAi components. Since the *Paramecium* MAC genome was fully sequenced in 2006 and gene annotations are consistently updated along with protein predictions, multiple transcription factors and chromatin regulators have recently been identified. Although there are accumulating data on developmental programs, it is poorly understood how differential gene expression is realized at molecular levels in vegetative growth. Hence, the presented studies aimed to extend the understanding of gene regulation in the vegetative nucleus, shedding light on transcriptional elongation, *Paramecium*'s histone code, and vegetative RNAi components shaping the epigenetic landscape.

5.1 Interaction of Epigenetic Key Players in a Crowded Nucleus

The *Paramecium* MAC responsible for vegetative gene expression was analyzed in terms of regulation of gene expression by epigenetic key players and cross-talk events between each other. Having analyzed the genome characteristics of *Paramecium* MAC in addition to the genome characteristics of other (multicellular) organisms, it became clear that the gene organization and the high degree of coding density

is significantly different from even closely related (compared to mammals) ciliates such as *Tetrahymena* (Oligohymenophorea) (Table 1.1). Thus, the hardware for gene expression machinery to work on is strongly different and should consequently call for a divergent expression machinery.

In mammals, gene expression is tightly controlled by distal regulatory regions. These regions, called enhancers, are bound by TFs and modified by histone modifications; they often are located up to kilobases away from the gene they regulate and are in intergenic or intronic regions, so they can control transcription by chromatin looping (Shlyueva, Stampfel, and Stark, 2014). The human genome consists of $\approx 50\%$ intergenic regions, and in the intragenic regions, the proportion of exons is low ($\approx 12\%$). Therefore, the human genome provides space for enhancers and other regulatory regions, which have initially been thought to be junk DNA (Kenny et al., 2020). In *Paramecium*, the highly condensed genome with small intergenic regions ($\approx 30\%$ of the genome) and tiny introns ($\approx 7\%$), a mechanism of distal gene regulation by enhancers cannot be envisioned (Table 1.1). In the yeast *S. cerevisiae*, having at least shorter intergenic regions than *H. sapiens* and genome features more comparable to *P. tetraurelia*, enhancer-like elements upstream of genes have been identified (Johnston, 1987).

Nevertheless, paramecia can be grown under several different conditions, and such gene regulation without long promoters or distal enhancers/silencers is feasible. Proximal regulatory regions upstream of genes themselves provide platforms for binding an assembly of the pre-initiation complex and establishment of activating histone marks, further promoting transcriptional activation. In *P. tetraurelia*, a promoter region 270bp upstream of the TSS has been described to be involved in mutual exclusive gene expression of surface antigen genes, thus controlling differential gene expression (Martin et al., 1994). Again, in humans, promoters can be as long as $\approx 10\text{kb}$, as investigated by TFIID binding (Kim et al., 2005). Furthermore, a region downstream of TSS also control differential surface antigen gene expression in *P. tetraurelia*, but the mechanism is poorly understood (Leeck and Forney, 1996). Recruiting the + 1 nucleosome to downstream regions might be a prime candidate for regulation, as these nucleosomes have been shown to be mainly linked to high gene expression (Figure 3.4).

In *Paramecium*, short intergenic regions that separate two genes still allow gene regulation and, especially, the uncoupling of genes in close proximity. One would assume that genes separated just by short intergenic/promoter regions cannot be regulated independently, resembling an operon-like structure in prokaryotes. This structure is not conserved in *Paramecium*, and even short promoters allow differential gene expression (Figure A.5). It is of high future interest how assembly of the transcription pre-initiation complex is realized in the light of space and time: short bi-directional promoters (SS, Figure 3.5) still need to allow for PIC assembly and blocking of spurious transcription. Interestingly, non-coding promoter antisense RNAs (PAS) have been shown to be regulated by chromatin remodeling, movement of the -1 nucleosome, and transcripts themselves contribute to Pol II promoter proximal pausing release and transcript degradation (Yang, 2022). In the light of closely neighboring nucleosomes, and a divergent Pol II pausing pattern, identification of long antisense RNAs in bidirectional promoters in *Paramecium* total RNA-seq is of high future interest.

Among ciliates, *Paramecium tetraurelia* is not the species with the compactest genome: *Halteria grandienella* and *Oxytricha trifallax* have the most compact known genomes (Zheng et al., 2021). In particular, these ciliates have MAC genomes organized

in nanochromosomes, meaning short chromosomes the size of a single gene. Gene regulation in these ciliates is accompanied by chromosome copy number regulation, where high copy numbers correlate with higher gene expression. It is tempting to speculate that those raised copy numbers are used to produce more gene products, a mechanism that is controlled by small RNAs. It is not clear to this day, what the reason for this high coding density is, but one could speculate that these species tend to have short, compact transcripts for protein-coding genes to optimize their metabolism. Apart from sRNA participation, at least *O. trifallax* has been shown to use well-positioned nucleosomes and DNA modifications to regulate gene expression, where N_6 -methyladenine(6mA) and nucleosome recruitment are linked to high gene expression (Beh et al., 2019), which is indeed partially comparable to *Paramecium*'s gene expression regulation by chromatin factors.

In *Paramecium*, many genes are linked on one chromosome, such copy number determination is likely not involved in gene dosage regulation since it would result in regulation of many genes at the same time. However, transgenes injected into the *Paramecium* MAC are maintained at their copy number level over vegetative fission (Garnier et al., 2004; Götz et al., 2016). Hence, there probably exists a not yet uncovered mechanism in *Paramecium* to control copy numbers. These listed examples simply indicate, that ciliates invented different pathways to control differential gene expression and thereby involve nucleosome positioning machinery.

As mentioned, the compact *Paramecium* genome demands transcriptional flexibility and regulation without space for long-distance interactions and enhancer-like modes of gene activation. Thus, experimental approaches were emphasized on chromatin modifications and the characterization of RNAi components, since the epigenetic regulators were determined to be involved in developmental genome processing and likely also shape vegetative expression.

5.1.1 High Coding Hardware Is Unprotected

MNase- and ChIP-seq studies on vegetative MAC chromatin revealed that the nucleosomal organization of *Paramecium*'s highly condensed genome does not follow well established textbook knowledge. The default state of the chromatin is open - more precisely speaking, most parts of the genome are not protected by nucleosomes. Indeed, expression is regulated by attracting nucleosomes to genes in conjunction with histone modifications (Figure 3.4).

The open conformation of a genome, exposing bare nucleotides rather than wrapping them in higher order chromatin structures, first of all seems contradictory. UV radiation and other toxic traits cause severe damage such as DNA double-strand breaks that cannot always be healed by the nuclear machinery and delay or impede the cell cycle. This is of special relevance in cancer cells, cause most tumor types show defects in DNA damage responses (Luijsterburg and Attikum, 2011). Chromatin protects from radiation and chemical agents (Takata et al., 2013) but also activates chromatin remodeling and deposition of histone modifications, which are crucial for the repair of DNA lesions. Interestingly, SWI ATPase remodelers are inactivated in cancer cells, so they may function as tumor-suppressors in healthy tissues. DNA lesions further promote the accumulation of KU complexes for non-homologous end joining and DNA repair (Luijsterburg and Attikum, 2011). *Paramecium* possesses components of the NHEJ-machinery and these components have

been shown to be involved in development, when IES become excised from developing MAC chromosomes, which involves KU70/80 heterodimers that anchor PiggyMAC transposase to ensure efficient coupling of DNA excision and DSB repair. Paralogs of KU proteins are also vegetatively expressed and are likely to perform DNA repair function, but little is known about how they act in the highly polyploid MAC (Abello et al., 2020). It is tempting to speculate, that high polyploidy allows tolerance of DNA damage in some sites of the ≈ 800 chromosome copies. In *Tetrahymena*, proteins of the DNA repair machinery and transposon-like proteins act together with chromatin binders (Shieh and Chalker, 2013), which illustrates that ciliates can serve as promising candidates to study DNA damage response in a clinical context.

However, *Paramecium* does not need to rely on the highly accessible, exposed MAC chromatin as genetic material, since the transcriptionally silent MIC serves as a genetic backup. This nucleus is located mostly in a protective pocket close to the MAC and possesses centromer-specific histone variants. Apart from this, there is no further knowledge about MIC chromatin. Repressive histone marks such as H3K27me3 and H3K9me3 have been described as absent from MIC by some research groups (Lhuillier-Akakpo et al., 2014), but the studies are based on immunofluorescence stainings and are highly dependent on antibody specificity. Repressive marks appear to accumulate in sexual induction and MIC meiosis (Lhuillier-Akakpo et al., 2014; Lhuillier-Akakpo et al., 2016) such that at least, *Paramecium* has the genetic repertoire to protect its genome in the MIC by canonical heterochromatin. This is crucial, since the MIC is full of transposable elements and their remnants which must be repressed to avoid their insertion in distant sites and even genes. By immunofluorescence stainings, at least in the presented study (chapter 3), both repressive marks could be identified (data not shown).

Despite that, compared to studies on *Tetrahymena* MIC chromatin with nucleosome phasing patterns but lower occupancy compared to MAC chromatin (Xiong et al., 2016), the chromatin organization in the *Paramecium* MIC remains elusive. Since Guérin et al., 2017 recently published a protocol on *Paramecium* MIC enrichment, it is expected that accumulating data on the organization of MIC chromosomes will be published soon. Nevertheless, chromatin-specific methods will need to be scaled to low input amounts in the future to gain insight into nucleosome organization and histone modifications in the MIC. Therefore, it is of high future interest to follow dynamic chromatin patterns upon MIC transcriptional activation at the beginning of sexual development. Mechanisms of transposon and elimination of IES are well understood as dependent on chromatin remodeling, but how expression of relevant key players takes place is poorly understood. Additionally, since the old MAC fragments still show gene expression, the interplay of gene regulation in old and new MACs must be tightly controlled. Elucidating this network is of great interest and would benefit from a low input ChIP-/ATAC-seq approach from MIC/Anlagen material.

From the study presented, MAC chromatin is the opposite to the textbook of DNA being mainly covered by protective chromatin. Allowing for beneficial mutations or deletions can be one mechanism to enhance offspring fitness. As a consequence of high chromosome copy numbers and random segregation by amitosis of the MAC, beneficial mutations can be assorted into daughter cells, thus allowing for selection of a cell population with accumulating advantageous traits. This phenomenon is most frequently observed in *Tetrahymena*: here, the polyploidy of $\approx 45n$ allows for

quick phenotypic assortment, where different chromosomal variants become separated by amitosis and give rise to daughter cells that can be homozygous for one loci upon several divisions. Artificially induced deletions in some of the MAC chromosomes have recently been shown to segregate into phenotypically wildtype and mutant lines with increasing amitotic divisions: thus, phenotypic assortments may also occur on chromosome variants in *Paramecium* (Nekrasova et al., 2019; Drews, Boenigk, and Simon, 2022).

Why *Paramecium* opens up its MAC chromatin to this extreme level is not understood to this day. Since the study covers the first description of the *Paramecium* nucleosome landscape by MNase- and ChIP-seq, methodical limitations may have contributed to the loss of information on the chromatin organization. In some Dinoflagellates, eukaryotes belonging to the Alveolata clade like *Paramecium*, DNA is not wrapped around nucleosomes but rather lysine-rich proteins, although the organisms have the genetic repertoire to express histone variants. Furthermore, these organisms have a large amount of DNA in their nucleus and show mitosis by attachment of an extra nuclear spindle apparatus, comparable to amitosis in Heterotrichia. The ratio of protein to DNA in Dinoflagellata is not 1:1 as in other organisms that wrap DNA around nucleosomes but more imbalanced to the side of higher DNA content (Talbert and Henikoff, 2012). Further, the Dinoflagellata genome size is strongly correlated with nuclear volume and cell size, suggesting that selection on cell size could influence genome size, which seems to be an evolutionary selection pressure driving *Paramecium*'s high copy number as well. In *Paramecium*, the footprints of non-histone proteins have not been investigated, and additional information from other ciliates such as *Tetrahymena* is missing. Data is accumulating on histone binders and chromatin remodelers, but if these ciliates additionally use nonhistone proteins to pack the high amounts of DNA in their MACs or guide DNA bending by e.g. HMG-1 as mammals do (Landsman and Bustin, 1993), is not known. The MNase- and ChIP-seq protocol was applied with subsequent enrichment of DNA fragments corresponding to the size of mononucleosomal DNA. Such, smaller fragments protected by additional DNA binders likely got counter-selected by the protocol. Combinatorial approaches in using ATCA- or DNase-seq to detect open chromatin in context with nucleosome positions obtained by MNase digest could add more resolution to the understanding of DNA storage in the MAC. In conclusion of protocol adaption to investigate chromatin landscape, parameters such as fixation time, temperature, fixative concentration have to be carefully considered and adjusted to the biological aspect of TFs dynamics and nucleosome binding strength. Even among ciliates, patterns of nucleosome positioning are quite diverse, which can be a consequence of different methodical approaches but probably also reflects the varying epigenetic repertoires of the investigated species.

Apart from DNA damages, another aspect of open and accessible chromatin must be considered, which is spurious transcription. In A/T rich genomes, such as the *Paramecium* genome, A/T stretches can infer nucleosome positioning and thus can be a determinant for transcription initiation. Consequently, the genes' GC content is higher compared to intergenic regions in *Paramecium*. The suppression of spurious transcription is regulated by chromatin remodelers such as ISWI, recruited by H3K36me3 in yeast (Wade and Grainger, 2018; Smolle et al., 2012). In mouse embryonic stem cells, CpG methylation in gene bodies prevents spurious transcription. This DNA modification is not found in *Paramecium*, however, 6mA exists which is located especially in the 5' region of gene bodies as it was shown for *Tetrahymena*

as well. Nonetheless, how *Paramecium* prevents spurious transcription on the chromatin level in gene bodies is not known. One could think of a Pol II being less possessive, thus nascent spurious transcripts can be detected early by the exosome. At the same time, the cells need to make sure that enough Pol II complex is available to express house keeping genes which would counterselect for multiple Pol II complexes inefficiently stalling along chromosomes. Understanding how spurious transcription is prevented is of high future and clinical interest, since activation of cryptic promoters is linked to carcinoma development. In this context, loss of DNA methyltransferase 3 activity results in loss of gene body methylation, increasing Pol II occupancy, transcripts of cryptic RNA and consequently altered proteomes. Vice versa, in healthy tissues, Pol II recruits the introduction of H3K36me3 and CpG methylation preventing spurious transcription (Neri et al., 2017).

5.1.2 Recruitment of Epigenetic Regulators to Genes Drive Expression

The presented studies show that nucleosome occupancy in genes is positively correlated to gene expression. The patterns of nucleosome positioning and occupancy that influence gene expression in different organisms have already been discussed. But the general dogma that nucleosome recruitment and higher occupancy are linked to reduced transcription does not hold true in *Paramecium*.

The shown data reflect the current situation in the vegetative *Paramecium* MAC under standard cultivation conditions. It is of high future interest to disturb the steady-state level of transcription by altering environmental conditions, to truly understand how the epigenetic landscape changes to induce transcriptome alterations. Cheaib et al., 2015 established a protocol for cold, heat shock, and starvation conditions during *Paramecium* cultivation and described massive transcriptome alterations involving differential expression of chromatin-modifying enzymes like histone deacetylase isoforms and nuclear assembly proteins. By MNase- and CHIP-seq approaches, it will be identified, how nucleosome occupancy on genes that become up-or down-regulated changes and if these changes appear rapidly and are maintained over longer time periods or even when the initial trigger is removed. Preliminary MNase data from cultures of serotype H, cultivated at 14°C, indicate that the overall occupancy pattern among all genes is not globally altered. This experiment fostered the results from MNase approaches with patterns of -1 and +1 nucleosomes and higher occupancy in high expressed genes, thus confirming robustness of the technical approach. Additionally, when sorting for differentially expressed genes, their nucleosome patterns were not drastically altered, indicating that adaptation to cold stress does not involve massive chromatin remodeling as a consequence. Nevertheless, studies on histone modifications are missing, so that conclusion on PTMs on histones cannot be drawn. In *Oxytricha*, in vivo fluctuations of 6mA in between nucleosomes did not alter the nucleosome profile and density, which is why the authors concluded, that nucleosome patterns must be regulated by *cis* factors which in consequence also contribute to nucleosome positioning in *Paramecium* (Beh et al., 2019).

Implication of chromatin conformation changes upon stress induction has been investigated in a case study in *Tetrahymena*, exploring nucleosome patterns upon starvation (Sheng et al., 2021). Thereby, the authors describe not only a more pronounced phasing and occupancy pattern in genes being down-regulated upon starvation and but also detect an altered accumulating of 6mA in linker DNA. Both epigenetic regulators - nucleosome positioning and DNA methylation - therefor

contribute to transcriptome alterations. Thereby, the level of gene up- or down-regulation positively correlates with levels of 6mA, thus 6mA contributes to gene regulation upon starvation. Nevertheless, the signals are not black and white, since down-regulated genes still show 6mA in linker DNA. Nucleosome positioning is globally increased upon starvation, whereby reduced replication rates need to be taken into account, resulting in less nucleosome perturbing.

The patterns of 6mA distribution and nucleosome positioning are divergent to *Paramecium* and *Oxytricha*, the latter showing fuzzy nucleosomes in down-regulated genes. Still, in *Paramecium* and *Oxytricha*, 6mA is also located in linker DNA and shapes nucleosome positioning (Beh et al., 2019; Hardy et al., 2021). Although patterns of these epigenetic modulators are divergent among ciliates, it can be concluded that they contribute to fine tuning in gene expression - even in unicellular organisms (Sheng et al., 2021). Interestingly, the functions of 6mA are rather divergent even in eukaryotes, as 6mA can be linked to gene activation and epigenetic silencing. Thus, even more distantly related ciliates such as *Blepharisma* (Heterotrichea) interestingly use 5mC and 6mA while *Stentor* (Heterotrichea) uses only cytosine methylation to shape gene expression (Wang et al., 2017).

The proportion of *cis*-acting DNA modifications varies amongst ciliate species and little is known about the function of this modification (Wang et al., 2017). Apart from DNA modifications, the sequence composition itself contributes to nucleosome positioning. In *Tetrahymena*, GC content contributes to positioning of nucleosomes - being highly positioned in the TSS - forming an array of downstream positioned nucleosomes and GC oscillations within *Tetrahymena* gene bodies contributing to spaced nucleosomes (Beh et al., 2019). Equally to *Paramecium* MNase data, the authors describe labile nucleosomes in intergenic regions in *Tetrahymena*. Future research will show, how distinct nucleosome positioning mechanisms operate in the context of numerous other regulatory codes within the genome, including the maintenance of transcription factor binding site, translational efficiency, mRNA splicing and higher secondary structures or chromatin intermingling of ≈ 800 chromosome copies.

The nucleosome itself is the basic building block of chromatin and mainly contributes to gene expression regulation (Kornberg and Lorch, 2020). Apart from being a steric obstacle, the nucleosome shows modifications on histone tails orchestrating gene expression. The presented study introduced a combinatorial approach to study gene expression regulation for the first time in *Paramecium*. Approaches to understand dynamic combinatorial patterns of epigenetic marks have been described in mammals, especially integrating information of (partially) methylated domains, histone marks and RNA-seq data. The studies aid to differentiate among cell-type specific epigenomes and transcriptional states (Salhab et al., 2018). Further studies on well described histone modifications being implicated in transcriptional regulation such as H3K27ac and H3K36me3 as well as H3K14ac will indeed shed light on the epigenomic landscape of *Paramecium*, taking divergences in amino acid sequences into account. Just to list one example: amino acids surrounding the analogous H3K79me epitope, thought to be involved in maintaining an open chromatin conformation (Talbert and Henikoff, 2021b), are not conserved in *Paramecium*, meaning that the occurrence of this mark cannot be elucidated by using custom antibodies in ChIP-seq experiments. To be precise: the presented study did not focus on any modifications on histone variants or the incorporation of those as it was investigated in *Stylonychia* development (Postberg et al., 2018). Additionally, modification at Histone H4 that are described to be crucial for chromatin folding and dosage compensation in males (Talbert and Henikoff, 2021b) and have not been investigated.

However, data of the combinatorial pattern (Figure 3.6) allow for preliminary insights and future research needs to enlighten *Paramecium*'s histone code on a more profound level.

5.1.3 Processivity of Polymerase II: Beyond the CTD

The listed key players of gene expression regulation investigated in the presented studies, like nucleosome positioning, histone modifications and small RNA processing, contribute to RNA Polymerase II recruitment and transcription initiation, elongation and termination. Thereby, Pol II processivity and transition from pausing to elongation is regulated by modification of amino acids in the heptad repeat of the C-terminal domain. How this kind of phosphorylation pattern is established in *Paramecium*'s highly divergent CTD domain is not understood to this date. Schüller et al., 2016 and others argue, that Pol II CTD phosphorylation does not occur on all heptad repeats in the CTD and still, Pol II is active. Effectors like the CDK9-kinase in mammals, involved in CTD phosphorylation, were not investigated on the homology level in *Paramecium* to this date. Such, how and when these kinases act on the highly divergent heptad repeats is of high future interest. By coordinated digest of Pol II CTD and mass spectrometry analysis, orchestrated heptad specific phospho-sites could be investigated.

Comparing yeast and human CTD phosphorylation, the former show divergent patterns, meaning, less Ser2 phosphorylation. This observation is probably based on the lower amount of introns and the generally shorter genes, resulting in decreased elongating of Pol II. It is therefore promising, to take *Paramecium* into account for future studies, since this organism has such highly condensed genome and no heptad repeat structure. In this context, the emerging field of *nuclear condensates* will benefit from *Paramecium* as model organism: it is thought that Pol II, associated factors, and the DNA to be transcribed find each other in nuclear spheres (condensates) and researches aim to understand, which nuclear properties contribute to the spatial compartmentalization. Elucidating, how transcription domains form in a MAC full of ≈ 800 chromosome copies in concordance with a divergent CTD will tell more about the evolutionary conservation of transcriptional systems (Bhat, Honson, and Guttman, 2021).

It has been shown, that polycomb silencing by the PRC1/PRC2 complex leads to recruitment of an RNA endonuclease when chromatin is not fully closed by repressive histone marks. The endonuclease cleaves nascent RNA and releases Pol II which entered elongation thus RNA degradation seems to have also crucial functions in heterochromatin mediated gene silencing (Zhou et al., 2022a). If such mechanisms are present in *Paramecium* has not been analyzed yet but seems to be an interesting molecular mechanism to investigate in future studies.

5.1.4 Small RNAs Shape Gene Expression

Apart from chromatin examinations, the presented studies also cover sRNAs as epigenetic key regulators in the highly condensed *Paramecium* MAC genome (chapter 4). RNAs have been described in mammalian systems to recruit chromatin remodeling complexes, while lncRNAs recently were termed as "tentacles to recruit and induce chromatin changes" in *cis* and guide either up- or down-regulation of gene expression (Neve et al., 2021). Besides from lncRNAs, small RNAs can contribute to chromatin remodeling which is well described for CTGS in yeast. In

Paramecium likewise to other ciliated species, small RNA-guided chromatin remodeling is described in developmental processes with accumulating data (e.g. (Miró-Pina et al., 2022; Singh et al., 2022)). Nevertheless, little is known about vegetative RNAi apart from processing of dsRNA from bacteria and silencing of homologous genes.

Vegetative regulation of gene expression and especially fine tuning of gene expression is regulated by miRNAs in mammals. Thereby, one miRNA can regulate hundreds of target genes in a complex regulatory network. Since small RNAs emerging from small RNA producing clusters identified in *Paramecium* were not analyzed in terms of off-targets effects, conclusions on regulation of multiple genes in *trans* cannot be drawn (Karunanithi et al., 2019). Studies by Hu et al., 2021 and others on single-cell miRNA profiles show fluctuating miRNA levels among different cells in one tissue. Thus, *Paramecium* can contribute to our understanding of small-RNA mediated transcription regulation in single cells and under different conditions. Nevertheless, single-cell transcriptomics and description of small RNA populations from single cells was not successful to this day.

Several examples exist, that uncovered gene regulation mechanisms by small RNAs apart from the miRNA pathway. In *D. melanogaster*, endogenous siRNAs regulate expression of metabolic stress response genes. These are generated in an Dicer depended manner and loss of Dicer function leads to reduced life span and hypersensitivity to oxidative stress (Lim et al., 2011). However, the authors cannot conclude how disrupted sRNA biogenesis results in gene regulation, but speculate that endo-siRNAs coordinate the expression of multiple target genes that act together to regulate energy homeostasis. Therefore, studies on the *Paramecium* vegetative RNAi system will help to understand sRNA mediated regulation of homeostasis. Moreover, in *C. elegans* sRNAs were shown to regulate genes of the phosphorus metabolism. These siRNAs are dependent on an Rdrp since the silencing of Rdrp resulted in alerted gene expression of phosphorous metabolism (Asikainen et al., 2008).

The first description of *Paramecium's* vegetative small RNAs did not allow for conclusions on regulatory functions. These siRNAs have no sense/antisense preferences and their accumulation does not result in silencing of homologous genes. Thus it is questionable whether these siRNAs are important for fine-tuning gene expression, or perhaps a matter of maintaining fitness in an changing environment. In agreement with Okamura and Lai, 2008 the question rises, if sRNA accumulation is a regulatory mechanism that generates species-specific characters during evolution?

In mouse oocytes, pseudogenes have been shown to regulate corresponding genes and this regulation involves siRNAs that are generated from gene-pseudogene pair transcripts. Genes with abundant pseudogene-derived siRNAs show an increase in expression upon loss of Dicer activity. Pseudogenes have long been considered to be non-functional artifacts of transposition pathways that act on protein-coding mRNAs. In some cases, regulatory roles have been postulated for pseudogenes, largely through antisense mechanisms. Tam et al., 2008 postulate a role for a subset of mammalian pseudogenes in the production of functional siRNAs. The vegetative SCRS of *Paramecium* are distributed among different genomic categories, such as genes, intergenic regions, and 117 are located in pseudogenes (Karunanithi et al., 2019). However, it could not be identified whether sRNA-mediated regulation of pseudogenes occurs. It would be of high interest to investigate pseudogene regulation and regulation of ohnologous genes resulting from whole genome duplication. Nevertheless, studying these s challenging since parameters of siRNA targets and off-targets effects must be considered.

Table 5.1 Table summarized features of sRNAs bound by Ptiwi proteins (left) and features of sRNAs assigned to different loci/pathways (right). Asterisk indicates reduced information. Data is collected from (Karunanithi et al., 2019; Götz et al., 2016; Marker et al., 2014)

Ptiwi13/14-bound sRNA		active SAg associated	SRC associated	GSRC associated	TG associated	
directionality	antisense	directionality	sense	sense and antisense	sense and antisense	antisense
size preference	23nt	size preference	23nt	23nt	23nt, phased	23nt
signature	5'U, overall U-content	signature	5' A/U	overall U-content	overall U-content	5'U
3' modification	yes	3' modification	nA	nA	nA	yes
Dicer-signature	21nt overlap	Dicer-signature	nA	21nt	21nt	21nt
		Dependency on RNAi component*	RDR1/3; Ptiwi13/14 loading	RDR1/2; Ptiwi13/14 loading	RDR1/2; Ptiwi13/14 loading	RDR2/3; Ptiwi13/14 loading
		Chromatin status	coverage of all SAgS with marks; silent SAgS accumulate more (silent) states	Epigenetic marks in SRCs with high sense siRNA accumulation and high GSCRs TPM		remodelling upon TG induced-silencing at endogenous loci*

sRNAs have also been shown to control splicing in mammals. In a 2012 study, the authors identified endogenous sRNAs that accumulate at the 3' end of introns. Ago2 loaded with sRNA recruits H3K9me3 and HP1, leading to reduced Pol II elongation and thus promotes the incorporation of variant exons (Ameyar-Zazoua et al., 2012). Although alternative splicing in *Paramecium* does not occur (Jaillon et al., 2008), splicing efficiency is still controlled by factors of the epigenetic toolbox. Positioned nucleosomes at exon boundaries could thus contribute to higher gene expression (Figure 3.4E), and if this guided by endogenous sRNAs will be studied in the future.

Figure 5.1 schematically summarizes vegetative MAC chromatin and different players in epigenetic regulation that were characterized in the presented studies. MAC chromosomes with short intergenic regions and high coding density (blue) show genes with divergent expression levels, whereas genes under the same bidirectional promoter are not shown. Genes of high expression are characterized by a well-positioned, pronounced +1 nucleosome and nucleosome-devoid gene bodies. Thereby, Pol II is stalling at every nucleosome. The divergent machinery of the Polymerase II elongation complex is not shown, but will be studied in future.

mRNAs from expressed genes can serve as templates for Rdr1 and Rdpr2 as it was shown by studies from Karunanithi et al., 2019. The dsRNA is then processed by a Dicer or Dicer-like protein into small dsRNA, of which preferably the antisense strands are loaded by Ptiwi 13 and Ptiwi 14. Shuttling of these vegetative sRNAs is likely and fostered by the current data summarized in chapter 4. Nucleus shuttling 23nt sRNAs show accumulation in distinct clusters and are imagined to recruit nucleosomes to gene bodies or favor nucleosome remodeling. This is highly speculative and needs to be investigated by Ptiwi IP mass spectrometry or MNase nucleosome profiling upon Ptiwi silencing. The endogenous loci, either regions attracted by small RNAs or the TSS and TTS show distribution of histone marks that appear in a combinatorial, plastic pattern, while plasticity is indicated by dashed modifications. The + 1 nucleosome, indicate by the most pronounced peak next to the gene start, is associated with H3K4me3 and H3K9ac while the repressive H3K27me3 is

depleted. Still, this mark is not exclusively found in silent genes. In gene bodies, nucleosomes at SRC sites are modified by all marks and future research has to show, if these sRNAs (Table 5.1, SRC/GSRC associated) contribute to distribution of histone modifications (Figure 4.8B).

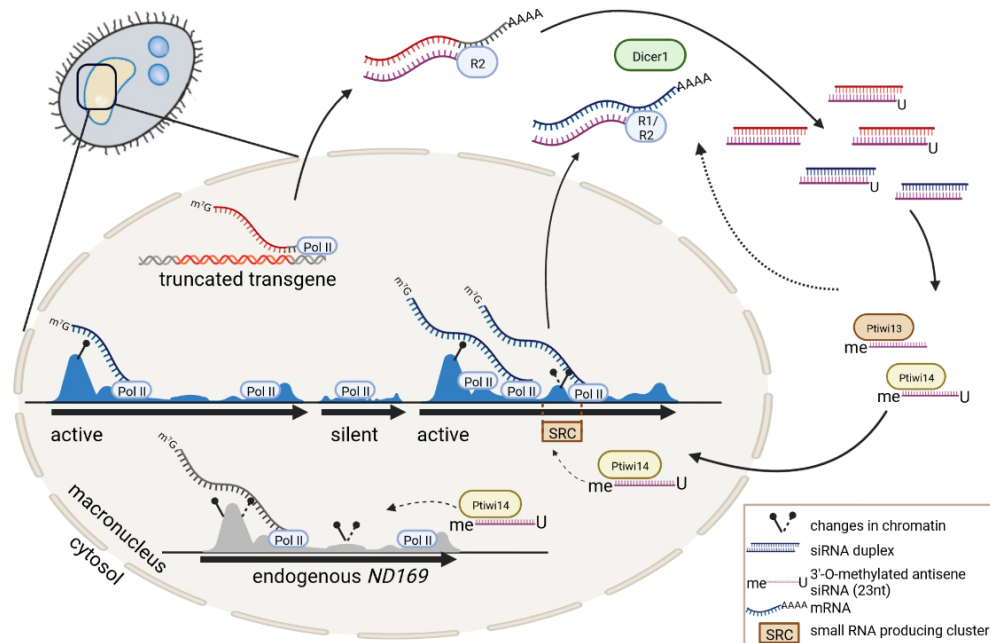


Figure 5.1 Model on *Paramecium* endogenous and exogenous siRNA pathways including chromatin conformation changes. The figure is centered on the vegetative MAC (beige), excluding MIC chromatin and developmental DNA rearrangements. See text for details. Created with BioRender.com.

Götz et al., 2016 showed dynamic chromatin remodeling at endogenous locus of *ND169* upon transgene-induced silencing. Apart from endogenous clusters producing sRNAs and/or attract nucleosomes, the model (Figure 5.1) includes the de-novo telomere capped transgene chromosome (orange). The chromatin conformation of this locus has not been described yet and how injected transgenes are packed by proteins is not known. The transgene is transcribed by Polymerase II producing aberrant transcripts as seen by long RNAs covering the initially injected vector sequences. These transcripts probably serve as templates for Rdr2/3 and the generated dsRNA is further processed by Dicers/Dicer-likes into dsRNA duplexes. Again, Ptiwis contributing to vegetative RNAi also load transgene associated sRNAs (Table 5.1, TG associated). These sRNAs are likely to shuttle to the nucleus and attack the endogenous locus (grey), probably by interacting with a nascent transcript. Chromatin remodeling occurs at the endogenous locus, thus resembling the CTGS model in yeast.

5.1.5 Future Perspective: Surface Antigen Expression Regulation by Epigenetic Marks

Surface antigen expression seems to depend on chromatin marks, since active and silent genes are covered with nucleosomes possessing different ratios of histone marks between active and silent genes. Preliminary experiments, silencing of Rdr3, resulted in massive changes on the chromatin landscape, precisely on the level of histone modifications rather than on the nucleosome occupancy or +1 nucleosome

positioning, as well as in the loss of mutual exclusive SAg expression: upon silencing of *Rdr3*, the mRNAs of all SAgS were detected and the cells were expressing multiple SAgS at the same time on their surface. First results from histone enrichment and western blots (Tobias Beckröge) indicate a global loss of H3K27me3. In a wildtype situation, active SAgS accumulate sense siRNAs of unknown function (Table 5.1, active SAg associated), but upon *Rdr3* silencing, all SAgS accumulate antisense siRNAs in high amounts. From these results, it seems likely that a link between small RNAs and chromatin remodeling in gene expression, comparable to TGS in yeast, exists. Among unicellulars, surface antigen regulation is realized differentially: *Plasmodium falciparum* silences SAgS by heterochromatin formation and on the transcriptional level whereas *Giardia lamblia* transcribes all VSP genes and post-transcriptionally cleaves mRNA of the non-expressed genes using the RNAi pathway while expression of SAgS is additionally regulated by histone marks (Kulakova et al., 2006). In *Trypanosoma brucei*, RNAi-guided genome rearrangements steer surface antigen variability and expression in addition to regulation by chromatin changes. *Paramecium* SAg expression seems to be regulated on the post-transcriptional level (Simon, Marker, and Schmidt, 2006b) but the role of sRNAs and nucleosome coverage in association with histone modifications remains elusive until this day. Studies from mutants impaired in surface antigen A expression ((Matsuda and Forney, 2005) and Tobias Beckröge) revealed accumulation of sense siRNAs corresponding to the mutated surface antigen gene besides the accumulation of other SAg associated siRNAs. This implicates a functional association of siRNAs to SAg expression. How antigen shifts are regulated by these siRNAs and chromatin alterations will be studied in future, following antigen shifts of serotype pure cultures subjected to varying temperatures.

The pipelines established by the presented studies, covering MNase-, ChIP-, total, and small RNA-seq with subsequent bioinformatic approaches like nucleosome profiling will contribute to the idea of *Paramecium*'s vegetative gene expression machinery focusing on the dynamic regulation of particular gene groups.

Bibliography

- Abello, Arthur et al. (2020). “Functional diversification of *Paramecium* Ku80 paralogs safeguards genome integrity during precise programmed DNA elimination”. In: *PLoS genetics* 16.4, e1008723.
- Adams, Mark D et al. (2000). “The genome sequence of *Drosophila melanogaster*”. In: *Science* 287.5461, pp. 2185–2195.
- Allis, C David and Thomas Jenuwein (2016). “The molecular hallmarks of epigenetic control”. In: *Nature Reviews Genetics* 17.8, pp. 487–500.
- Almouzni, Genevieve and Howard Cedar (2016). “Maintenance of epigenetic information”. In: *Cold Spring Harbor Perspectives in Biology* 8.5, a019372.
- Ameyar-Zazoua, Maya et al. (2012). “Argonaute proteins couple chromatin silencing to alternative splicing”. In: *Nature structural & molecular biology* 19.10, pp. 998–1004.
- Antoniewski, Christophe (2014). “Computing siRNA and piRNA overlap signatures”. In: *Methods in molecular biology (Clifton, N.J.)* 1173, 135–146.
- Arnaiz, Olivier, Eric Meyer, and Linda Sperling (2020). “ParameciumDB 2019: integrating genomic data across the genus for functional and evolutionary biology”. In: *Nucleic acids research* 48.D1, pp. D599–D605.
- Arnaiz, Olivier et al. (2012). “The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences”. In: *PLoS Genet* 8.10, e1002984.
- Arnaiz, Olivier et al. (2017). “Improved methods and resources for paramecium genomics: transcription units, gene annotation and gene expression”. In: *BMC genomics* 18.1, p. 483.
- Aronica, Lucia et al. (2008). “Study of an RNA helicase implicates small RNA–noncoding RNA interactions in programmed DNA elimination in *Tetrahymena*”. In: *Genes & development* 22.16, pp. 2228–2241.
- Arrigoni, Laura et al. (2016). “Standardizing chromatin research: a simple and universal method for ChIP-seq”. In: *Nucleic acids research* 44.7, e67–e67.
- Asikainen, Suvi et al. (2008). “Functional characterization of endogenous siRNA target genes in *Caenorhabditis elegans*”. In: *BMC genomics* 9.1, pp. 1–10.
- Aury, Jean-Marc et al. (2006). “Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*”. In: *Nature* 444.7116, pp. 171–178.
- Baldi, Sandro, Philipp Korber, and Peter B Becker (2020). “Beads on a string—nucleosome array arrangements and folding of the chromatin fiber”. In: *Nature structural & molecular biology* 27.2, pp. 109–118.
- Baranasic, Damir et al. (2014). “Genomic characterization of variable surface antigens reveals a telomere position effect as a prerequisite for RNA interference-mediated silencing in *Paramecium tetraurelia*”. In: *mBio* 5.6, e01328–14.
- Barnett, Audrey and E Steers (1984). “Antibody-induced membrane fusion in *Paramecium*”. In: *Journal of cell science* 65.1, pp. 153–162.
- Bastiaanssen, Carolien and Chirlmin Joo (2021). “Small RNA-directed DNA elimination: the molecular mechanism and its potential for genome editing”. In: *RNA biology* 18.11, pp. 1540–1545.

- Beale, GH (1952). "Antigen variation in *Paramecium aurelia*, variety 1". In: *Genetics* 37.1, p. 62.
- Beh, Leslie Y et al. (2019). "Identification of a DNA N6-adenine methyltransferase complex and its impact on chromatin organization". In: *Cell* 177.7, pp. 1781–1796.
- Beisson, Janine et al. (2010). "Maintaining clonal *Paramecium tetraurelia* cell lines of controlled age through daily reisolation". In: *Cold Spring Harbor Protocols* 2010.1, pdb-prot5361.
- Bétermier, Mireille and Sandra Duharcourt (2014). "Programmed rearrangement in ciliates: *Paramecium*". In: *Microbiology spectrum* 2.6, pp. 2–6.
- Bhat, Prashant, Drew Honson, and Mitchell Guttman (2021). "Nuclear compartmentalization as a mechanism of quantitative control of gene expression". In: *Nature Reviews Molecular Cell Biology* 22.10, pp. 653–670.
- Bhaumik, Sukesh R, Edwin Smith, and Ali Shilatifard (2007). "Covalent modifications of histones during development and disease pathogenesis". In: *Nature structural & molecular biology* 14.11, pp. 1008–1016.
- Bird, Adrian (2002). "DNA methylation patterns and epigenetic memory". In: *Genes & development* 16.1, pp. 6–21.
- Bischerour, Julien et al. (2018). "Six domesticated PiggyBac transposases together carry out programmed DNA elimination in *Paramecium*". In: *Elife* 7, e37927.
- Bonnemain, Hugues et al. (1992). "Interactions between genes involved in exocytotic membrane fusion in *paramecium*." In: *Genetics* 130.3, pp. 461–470.
- Bouhouche, Khaled et al. (2011). "Functional specialization of Piwi proteins in *Paramecium tetraurelia* from post-transcriptional gene silencing to genome remodelling". In: *Nucleic acids research* 39.10, pp. 4249–4264.
- Bourgain, Florence M and Michael D Katinka (1991). "Telomeres inhibit end to end fusion and enhance maintenance of linear DNA molecules injected into the *Paramecium primaurelia* macronucleus". In: *Nucleic acids research* 19.7, pp. 1541–1547.
- Boyle, Alan P et al. (2008). "High-resolution mapping and characterization of open chromatin across the genome". In: *Cell* 132.2, pp. 311–322.
- Brahma, Sandipan and Steven Henikoff (2020). "Epigenome regulation by dynamic nucleosome unwrapping". In: *Trends in biochemical sciences* 45.1, pp. 13–26.
- Buenrostro, Jason D et al. (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature methods* 10.12, pp. 1213–1218.
- Bühler, Marc, André Verdel, and Danesh Moazed (2006). "Tethering RITS to a nascent transcript initiates RNAi-and heterochromatin-dependent gene silencing". In: *Cell* 125.5, pp. 873–886.
- Burkhart, Kirk B et al. (2011). "A Pre-mRNA-associating factor links endogenous siRNAs to chromatin regulation". In: *PLoS genetics* 7.8, e1002249.
- Carone, Benjamin R et al. (2010). "Paternaly induced transgenerational environmental reprogramming of metabolic gene expression in mammals". In: *Cell* 143.7, pp. 1084–1096.
- Carradec, Quentin et al. (2015). "Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate *Paramecium tetraurelia*". In: *Nucleic acids research* 43.3, pp. 1818–1833.
- Castel, Stephane E et al. (2014). "Dicer promotes transcription termination at sites of replication stress to maintain genome stability". In: *Cell* 159.3, pp. 572–583.
- Catania, Francesco et al. (2009). "Genetic diversity in the *Paramecium aurelia* species complex". In: *Molecular biology and evolution* 26.2, pp. 421–431.

- Cedar, Howard and Yehudit Bergman (2009). "Linking DNA methylation and histone modification: patterns and paradigms". In: *Nature Reviews Genetics* 10.5, pp. 295–304.
- Cerutti, Heriberto and J Armando Casas-Mollano (2006). "On the origin and functions of RNA-mediated silencing: from protists to man". In: *Current genetics* 50.2, pp. 81–99.
- Chalker, Douglas L and Meng-Chao Yao (2011). "DNA elimination in ciliates: transposon domestication and genome surveillance". In: *Annual review of genetics* 45, pp. 227–246.
- Chapman, Rob D et al. (2008). "Molecular evolution of the RNA polymerase II CTD". In: *Trends in Genetics* 24.6, pp. 289–296.
- Cheaib, Miriam and Martin Simon (2013). "Dynamic chromatin remodelling of ciliate macronuclear DNA as determined by an optimized chromatin immunoprecipitation (ChIP) method for *Paramecium tetraurelia*". In: *Applied microbiology and biotechnology* 97.6, pp. 2661–2670.
- Cheaib, Miriam et al. (2015). "Epigenetic regulation of serotype expression antagonizes transcriptome dynamics in *Paramecium tetraurelia*". In: *DNA Research* 22.4, pp. 293–305.
- Chen, Kaifu et al. (2013). "DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing". In: *Genome research* 23.2, pp. 341–351.
- Chen, Kaifu et al. (2015). "Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes". In: *Nature genetics* 47.10, pp. 1149–1157.
- Chen, Xiao et al. (2014). "The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development". In: *Cell* 158.5, pp. 1187–1198.
- Chen, Xiao et al. (2016). "Enzymatic and chemical mapping of nucleosome distribution in purified micro- and macronuclei of the ciliated model organism, *Tetrahymena thermophila*". In: *Science China Life Sciences* 59.9, pp. 909–919.
- Chen, Ying-Jiun C, Evangelia Koutelou, and Sharon YR Dent (2022). "Now open: evolving insights to the roles of lysine acetylation in chromatin organization and function". In: *Molecular cell*.
- Cheng, Chao-Yin et al. (2020). "The evolution of germ-soma nuclear differentiation in eukaryotic unicells". In: *Current Biology* 30.10, R502–R510.
- Chereji, Răzvan V, Terri D Bryson, and Steven Henikoff (2019). "Quantitative MNase-seq accurately maps nucleosome occupancy levels". In: *Genome biology* 20.1, pp. 1–18.
- Chung, Ho-Ryun et al. (Dec. 2011). "The Effect of Micrococcal Nuclease Digestion on Nucleosome Positioning Data". In: *PLOS ONE* 5.12, pp. 1–8.
- Couvillion, Mary T et al. (2009). "Sequence, biogenesis, and function of diverse small RNA classes bound to the Piwi family proteins of *Tetrahymena thermophila*". In: *Genes & development* 23.17, pp. 2016–2032.
- Cramer, Patrick et al. (2000). "Architecture of RNA polymerase II and implications for the transcription mechanism". In: *Science* 288.5466, pp. 640–649.
- Crooks, Gavin E et al. (2004). "WebLogo: a sequence logo generator". In: *Genome research* 14.6, pp. 1188–1190.
- Cuatrecasas, Pedro, Sara Fuchs, and Christian B. Anfinsen (1967). "Catalytic Properties and Specificity of the Extracellular Nuclease of *Staphylococcus aureus*". In: *Journal of Biological Chemistry* 242.7, pp. 1541–1547.
- Czech, Benjamin et al. (2018). "piRNA-guided genome defense: from biogenesis to silencing". In: *Annual review of genetics* 52, pp. 131–157.

- Dion, Michael F et al. (2007). "Dynamics of replication-independent histone turnover in budding yeast". In: *Science* 315.5817, pp. 1405–1408.
- Draizen, Eli J et al. (2016). "HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants". In: *Database* 2016.
- Drews, Franziska, Jens Boenigk, and Martin Simon (2022). "Paramecium epigenetics in development and proliferation". In: *Journal of Eukaryotic Microbiology*, e12914.
- Drews, Franziska et al. (2021). "Two Piwis with Ago-like functions silence somatic genes at the chromatin level". In: *RNA biology* 18.sup2, pp. 757–769.
- Drews, Franziska et al. (2022). "Broad domains of histone marks in the highly compact Paramecium macronuclear genome". In: *Genome Research* 32.4, pp. 710–725.
- Drotos, Katherine HI et al. (2022). "Throwing away DNA: programmed downsizing in somatic nuclei". In: *Trends in Genetics*.
- Duret, Laurent et al. (2008). "Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia: a somatic view of the germline". In: *Genome research* 18.4, pp. 585–596.
- Edgar, Robert C (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic acids research* 32.5, pp. 1792–1797.
- Epstein, LLOYD M and JAMES D Forney (1984). "Mendelian and non-Mendelian mutations affecting surface antigen expression in Paramecium tetraurelia". In: *Molecular and cellular biology* 4.8, pp. 1583–1590.
- Ernst, Jason and Manolis Kellis (2012). "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature methods* 9.3, pp. 215–216.
- Ewels, Philip et al. (2016). "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19, pp. 3047–3048.
- Farazi, Thalia A, Stefan A Juranek, and Thomas Tuschl (2008). "The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members". In: *FastQC* (2015). URL: <https://qubeshub.org/resources/fastqc>.
- Felsenstein, Joseph (1985). "Confidence limits on phylogenies: an approach using the bootstrap". In: *evolution* 39.4, pp. 783–791.
- Fire, Andrew et al. (1998). "Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*". In: *nature* 391.6669, pp. 806–811.
- Fischle, Wolfgang et al. (2005). "Regulation of HP1–chromatin binding by histone H3 methylation and phosphorylation". In: *Nature* 438.7071, pp. 1116–1122.
- Forney, JD et al. (1983). "Structure and expression of genes for surface proteins in Paramecium". In: *Molecular and Cellular Biology* 3.3, pp. 466–474.
- Frapporti, Andrea et al. (2019). "The Polycomb protein Ezh1 mediates H3K9 and H3K27 methylation to repress transposable elements in Paramecium". In: *Nature communications* 10.1, pp. 1–15.
- Froissard, M. et al. (n.d.). "Novel secretory vesicle proteins essential for membrane fusion display extracellular-matrix domains". In: *Traffic* 5.7 (), pp. 493–502.
- Furrer, Dominique I et al. (2017). "Two sets of piwi proteins are involved in distinct sRNA pathways leading to elimination of Germline-Specific DNA". In: *Cell reports* 20.2, pp. 505–520.
- Galvani, Angélique and Linda Sperling (2001). "Transgene-mediated post-transcriptional gene silencing is inhibited by 3 non-coding sequences in Paramecium". In: *Nucleic acids research* 29.21, pp. 4387–4394.
- (2002). "RNA interference by feeding in Paramecium." In: *Trends in genetics: TIG* 18.1, pp. 11–12.

- Garg, Jyoti et al. (2019). "The med31 conserved component of the divergent mediator complex in *Tetrahymena thermophila* participates in developmental regulation". In: *Current Biology* 29.14, pp. 2371–2379.
- Garnier, Olivier et al. (2004). "RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*". In: *Molecular and cellular biology* 24.17, pp. 7370–7379.
- Gilley, D et al. (1988). "Autonomous replication and addition of telomere-like sequences to DNA microinjected into *Paramecium tetraurelia* macronuclei". In: *Molecular and Cellular Biology* 8.11, pp. 4765–4772.
- Gnan, Stefano et al. (2022). "GC content but not nucleosome positioning directly contributes to intron-splicing efficiency in *Paramecium*". In: *Genome Research*, gr-276125.
- Götz, Ulrike et al. (2016). "Two sets of RNAi components are required for heterochromatin formation in trans triggered by truncated transgenes". In: *Nucleic Acids Research* 44.12, pp. 5908–5923.
- Grattepanche, Jean-David et al. (2018). "Microbial diversity in the eukaryotic SAR clade: Illuminating the darkness between morphology and molecular data". In: *BioEssays* 40.4, p. 1700198.
- Griffiths-Jones, Sam et al. (2007). "miRBase: tools for microRNA genomics". In: *Nucleic acids research* 36.suppl_1, pp. D154–D158.
- Grimaud, Charlotte et al. (2006). "RNAi components are required for nuclear clustering of Polycomb group response elements". In: *Cell* 124.5, pp. 957–971.
- Gruchota, J. et al. (2017a). "A meiosis-specific Spt5 homolog involved in non-coding transcription". In: *Nucleic Acids Research* 45.8, pp. 4722–4732.
- Gruchota, Julita et al. (2017b). "A meiosis-specific Spt5 homolog involved in non-coding transcription". In: *Nucleic acids research* 45.8, pp. 4722–4732.
- Guérin, Frédéric et al. (2017). "Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements". In: *BMC genomics* 18.1, pp. 1–17.
- Guo, Min et al. (2008). "Core structure of the yeast spt4-spt5 complex: a conserved module for regulation of transcription elongation". In: *Structure* 16.11, pp. 1649–1658.
- Gutbrod, Michael J and Robert A Martienssen (2020). "Conserved chromosomal functions of RNA interference". In: *Nature Reviews Genetics* 21.5, pp. 311–331.
- Hangauer, Matthew J, Ian W Vaughn, and Michael T McManus (2013). "Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs". In: *PLoS genetics* 9.6, e1003569.
- Hansen, Jeffrey C et al. (2018). "The 10-nm chromatin fiber and its relationship to interphase chromosome organization". In: *Biochemical Society Transactions* 46.1, pp. 67–76.
- Hardy, Alexis et al. (Jan. 2021). "DNAModAnnot: a R toolbox for DNA modification filtering and annotation". In: *Bioinformatics*. btab032.
- Harlen, Kevin M and L Stirling Churchman (2017). "The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain". In: *Nature reviews Molecular cell biology* 18.4, pp. 263–273.
- Hattman, STANLEY et al. (1978). "Comparative study of DNA methylation in three unicellular eucaryotes". In: *Journal of bacteriology* 135.3, pp. 1156–1157.
- Henikoff, Steven and Ali Shilatifard (2011). "Histone modification: cause or cog?" In: *Trends in Genetics* 27.10, pp. 389–396.
- Hoehener, Cristina, Iris Hug, and Mariusz Nowacki (2018). "Dicer-like enzymes with sequence cleavage preferences". In: *Cell* 173.1, pp. 234–247.

- Hogg, Simon J et al. (2020). "Targeting the epigenetic regulation of antitumour immunity". In: *Nature reviews Drug discovery* 19.11, pp. 776–800.
- Holoch, Daniel and Danesh Moazed (2015). "RNA-mediated epigenetic regulation of gene expression". In: *Nature Reviews Genetics* 16.2, pp. 71–84.
- Horwich, Michael D et al. (2007). "The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC". In: *Current biology* 17.14, pp. 1265–1272.
- Hu, Tao et al. (2021). "Single-cell transcriptomes reveal characteristics of micror-nas in gene expression noise reduction". In: *Genomics, Proteomics & Bioinformatics* 19.3, pp. 394–407.
- Huang, He et al. (2015). "Quantitative proteomic analysis of histone modifications". In: *Chemical reviews* 115.6, pp. 2376–2418.
- Ignarski, Michael et al. (2014). "Paramecium tetraurelia chromatin assembly factor-1-like protein PtCAF-1 is involved in RNA-mediated control of DNA elimination". In: *Nucleic acids research* 42.19, pp. 11952–11964.
- Iyer, Lakshminarayan M et al. (2008). "Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes". In: *International journal for parasitology* 38.1, pp. 1–31.
- Jaehning, Judith A (2010). "The Paf1 complex: platform or player in RNA polymerase II transcription?" In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1799.5-6, pp. 379–388.
- Jaillon, Olivier et al. (2008). "Translational control of intron splicing in eukaryotes". In: *Nature* 451.7176, pp. 359–362.
- Janssens, Derek H et al. (2022). "CUT&Tag2for1: a modified method for simultaneous profiling of the accessible and silenced regulome in single cells". In: *Genome biology* 23.1, pp. 1–19.
- Jenuwein, Thomas and C David Allis (2001). "Translating the histone code". In: *Science Signaling* 293.5532, p. 1074.
- Johnston, Mark (1987). "A model fungal gene regulatory mechanism: the GAL genes of *Saccharomyces cerevisiae*". In: *Microbiological reviews* 51.4, pp. 458–476.
- Kamath, R S et al. (2001). "Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *Caenorhabditis elegans*." In: *Genome biology* 2.1, RESEARCH0002.
- Kapusta, Aurélie et al. (2011). "Highly precise and developmentally programmed genome assembly in *Paramecium* requires ligase IV-dependent end joining". In: *PLoS genetics* 7.4, e1002049.
- Karunanithi, Sivarajan, Martin Simon, and Marcel H. Schulz (Apr. 2019). "Automated analysis of small RNA datasets with RAPID". In: *PeerJ* 7, e6710.
- Karunanithi, Sivarajan et al. (2019). "Exogenous RNAi mechanisms contribute to transcriptome adaptation by phased siRNA clusters in *Paramecium*". In: *Nucleic acids research* 47.15, pp. 8036–8049.
- Karunanithi, Sivarajan et al. (2020). "Feeding exogenous dsRNA interferes with endogenous sRNA accumulation in *Paramecium*". In: *DNA Research*.
- Kataoka, Kensuke and Kazufumi Mochizuki (2015). "Phosphorylation of an HP1-like protein regulates heterochromatin body assembly for DNA elimination". In: *Developmental cell* 35.6, pp. 775–788.
- Katz, Laura A (2001). "Evolution of nuclear dualism in ciliates: a reanalysis in light of recent molecular data." In: *International Journal of Systematic and Evolutionary Microbiology* 51.4, pp. 1587–1592.

- Katzmarski, Natalie et al. (2022). "Reply to: 'Lack of evidence for intergenerational inheritance of immune resistance to infections'". In: *Nature Immunology* 23.2, pp. 208–209.
- Kelly, Theresa K et al. (2012). "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules". In: *Genome research* 22.12, pp. 2497–2506.
- Kenny, Nathan J et al. (2020). "Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*". In: *Nature communications* 11.1, pp. 1–11.
- Ketting, René F (2011). "The many faces of RNAi". In: *Developmental cell* 20.2, pp. 148–161.
- Khurana, Jaspreet S et al. (2018). "Small RNA-mediated regulation of DNA dosage in the ciliate *Oxytricha*". In: *RNA* 24.1, pp. 18–29.
- Kim, Tae Hoon et al. (2005). "A high-resolution map of active promoters in the human genome". In: *Nature* 436.7052, pp. 876–880.
- King, Brian R et al. (2012). "ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes". In: *BMC research notes* 5.1, p. 351.
- Kinkley, Sarah et al. (2016). "reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4+ memory T cells". In: *Nature communications* 7.1, pp. 1–13.
- Klar, Amar JS (1998). "Propagating epigenetic states through meiosis: where Mendel's gene is more than a DNA moiety". In: *Trends in genetics* 14.8, pp. 299–301.
- Klattenhoff, Carla et al. (2009). "The *Drosophila* HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters". In: *Cell* 138.6, pp. 1137–1149.
- Kleeman, Elizabeth A, Carolina Gubert, and Anthony J Hannan (2022). "Transgenerational epigenetic impacts of parental infection on offspring health and disease susceptibility". In: *Trends in Genetics*.
- Kleinendorst, Rozemarijn WD et al. (2021). "Genome-wide quantification of transcription factor binding at single-DNA-molecule resolution using methyl-transferase footprinting". In: *Nature protocols* 16.12, pp. 5673–5706.
- Kornberg, Roger D and Yahli Lorch (2020). "Primary role of the nucleosome". In: *Molecular Cell* 79.3, pp. 371–375.
- Krueger, Felix (2015). "Trim galore". In: *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* 516, p. 517.
- Kulakova, Liudmila et al. (2006). "Epigenetic mechanisms are involved in the control of *Giardia lamblia* antigenic variation". In: *Molecular microbiology* 61.6, pp. 1533–1542.
- Kumar, Dharendra et al. (2021). "Decoding the function of bivalent chromatin in development and cancer". In: *Genome research* 31.12, pp. 2170–2184.
- Kumar, Sudhir et al. (2018). "MEGA X: molecular evolutionary genetics analysis across computing platforms". In: *Molecular biology and evolution* 35.6, pp. 1547–1549.
- Kundaje, Anshul et al. (2015). "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539, pp. 317–330.
- Kurth, Henriette M and Kazufumi Mochizuki (2009). "2-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in *Tetrahymena*". In: *Rna* 15.4, pp. 675–685.
- Lander, ES et al. (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921.

- Landsman, David and Michael Bustin (1993). "A signature for the HMG-1 box DNA-binding proteins". In: *Bioessays* 15.8, pp. 539–546.
- Langmead, Ben and Steven L Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4, p. 357.
- Lasserre, Julia, Ho-Ryun Chung, and Martin Vingron (Sept. 2013). "Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks". In: *PLOS Computational Biology* 9.9, pp. 1–12.
- Lee, Suzanne R and Kathleen Collins (2006). "Two classes of endogenous small RNAs in *Tetrahymena thermophila*". In: *Genes & development* 20.1, pp. 28–33.
- Lee, Suzanne R et al. (2021). "Disruption of a 23–24 nucleotide small RNA pathway elevates DNA damage responses in *Tetrahymena thermophila*". In: *Molecular biology of the cell* 32.15, pp. 1335–1346.
- Leeck, Charles L and James D Forney (1996). "The 5' coding region of *Paramecium* surface antigen genes controls mutually exclusive transcription." In: *Proceedings of the National Academy of Sciences* 93.7, pp. 2838–2843.
- Lepere, Gersende et al. (2008). "Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*". In: *Nucleic Acids Research* 37.3, pp. 903–915.
- (2009). "Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*". In: *Nucleic acids research* 37.3, pp. 903–915.
- Lhuillier-Akakpo, Maoussi et al. (2014). "Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements". In: *PLoS Genet* 10.9, e1004665.
- Lhuillier-Akakpo, Maoussi et al. (2016). "DNA deletion as a mechanism for developmentally programmed centromere loss". In: *Nucleic acids research* 44.4, pp. 1553–1565.
- Li, Bing, Michael Carey, and Jerry L Workman (2007). "The role of chromatin during transcription". In: *Cell* 128.4, pp. 707–719.
- Li, Heng et al. (2009). "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–2079.
- Li, Junjie et al. (2005). "Methylation protects miRNAs and siRNAs from a 3-end uridylation activity in *Arabidopsis*". In: *Current biology* 15.16, pp. 1501–1507.
- Lim, Do-Hwan et al. (2011). "The endogenous siRNA pathway in *Drosophila* impacts stress resistance and lifespan by regulating metabolic homeostasis". In: *FEBS letters* 585.19, pp. 3079–3085.
- Long, Hong-An et al. (2013). "Accumulation of spontaneous mutations in the ciliate *Tetrahymena thermophila*". In: *Genetics* 195.2, pp. 527–540.
- Lorch, Yahli et al. (2011). "Selective removal of promoter nucleosomes by the RSC chromatin-remodeling complex". In: *Nature structural & molecular biology* 18.8, pp. 881–885.
- Luger, Karolin et al. (1997). "Crystal structure of the nucleosome core particle at 2.8 Å resolution". In: *Nature* 389.6648, pp. 251–260.
- Luijsterburg, Martijn S and Haico van Attikum (2011). "Chromatin and the DNA damage response: the cancer connection". In: *Molecular oncology* 5.4, pp. 349–367.
- Luo, Guan-Zheng et al. (2018). "N 6-methyldeoxyadenosine directs nucleosome positioning in *Tetrahymena* DNA". In: *Genome biology* 19.1, pp. 1–12.
- Ma, Jin-Biao, Keqiong Ye, and Dinshaw J Patel (2004). "Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain". In: *nature* 429.6989, pp. 318–322.
- Maliszewska-Olejniczak, K. et al. (2015a). "TFIIS-Dependent Non-coding Transcription Regulates Developmental Genome Rearrangements". In: *PLoS Genet* 11.7, e1005383.

- Maliszewska-Olejniczak, Kamila et al. (2015b). "TFIIS-dependent non-coding transcription regulates developmental genome rearrangements". In: *PLoS genetics* 11.7, e1005383.
- Mannironi, Cecilia, William M Bonner, and Christopher L Hatch (1989). "H2A. X. a histone isoprotein with a conserved C-terminal sequence, is encoded by a novel mRNA with both DNA replication type and polyA 3 processing signals". In: *Nucleic acids research* 17.22, pp. 9113–9126.
- Mao, Hui et al. (2015). "The Nrde pathway mediates small-RNA-directed histone H3 lysine 27 trimethylation in *Caenorhabditis elegans*". In: *Current Biology* 25.18, pp. 2398–2403.
- Marco-Sola, Santiago et al. (2012). "The GEM mapper: fast, accurate and versatile alignment by filtration". In: *Nature methods* 9.12, p. 1185.
- Marker, Simone et al. (2010). "Distinct RNA-dependent RNA polymerases are required for RNAi triggered by double-stranded RNA versus truncated transgenes in *Paramecium tetraurelia*". In: *Nucleic acids research* 38.12, pp. 4092–4107.
- Marker, Simone et al. (2014). "A forward genetic screen reveals essential and non-essential RNAi factors in *Paramecium tetraurelia*". In: *Nucleic Acids Research* 42.11, pp. 7268–7280.
- Marmignon, Antoine et al. (2014). "Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate *Paramecium tetraurelia*". In: *PLoS genetics* 10.8, e1004552.
- Martienssen, Robert and Danesh Moazed (2015). "RNAi and heterochromatin assembly". In: *Cold Spring Harbor perspectives in biology* 7.8, a019323.
- Martin, Linda D et al. (1994). "DNA sequence requirements for the regulation of immobilization antigen A expression in *Paramecium tetraurelia*". In: *Developmental genetics* 15.5, pp. 443–451.
- Martin, Marcel (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet. journal* 17.1, pp. 10–12.
- Matsuda, Atsushi and James D Forney (2005). "Analysis of *Paramecium tetraurelia* A-51 surface antigen gene mutants reveals positive-feedback mechanisms for maintenance of expression and temperature-induced activation". In: *Eukaryotic Cell* 4.10, pp. 1613–1619.
- Matzke, Marjori A and Rebecca A Mosher (2014). "RNA-directed DNA methylation: an epigenetic pathway of increasing complexity". In: *Nature Reviews Genetics* 15.6, pp. 394–408.
- McClintock, Barbara (1950). "The origin and behavior of mutable loci in maize". In: *Proceedings of the National Academy of Sciences* 36.6, pp. 344–355.
- Meyer, Clifford A. and X. Shirley Liu (2014). "Identifying and mitigating bias in next-generation sequencing methods for chromatin biology". In: *Nature Reviews Genetics* 15.11, pp. 709–721.
- Meyer, E, F Caron, and B Guiard (1984). "Blocking of in vitro translation of *Paramecium* messenger RNAs is due to messenger RNA primary structure". In: *Biochimie* 66.5, pp. 403–412.
- Meyer, Eric (1992). "Induction of specific macronuclear developmental mutations by microinjection of a cloned telomeric gene in *Paramecium primaurelia*." In: *Genes & development* 6.2, pp. 211–222.
- Meyer, ERIC, FRANCOIS Caron, and A Baroin (1985). "Macronuclear structure of the G surface antigen gene of *Paramecium primaurelia* and direct expression of its repeated epitopes in *Escherichia coli*". In: *Molecular and cellular biology* 5.9, pp. 2414–2422.

- Minnoye, Liesbeth et al. (2021). "Chromatin accessibility profiling methods". In: *Nature Reviews Methods Primers* 1.1, pp. 1–24.
- Miró-Pina, Caridad et al. (2022). "Paramecium Polycomb repressive complex 2 physically interacts with the small RNA-binding PIWI protein to repress transposable elements". In: *Developmental Cell* 57.8, pp. 1037–1052.
- Nagai, Shigeki et al. (2017). "Chromatin potentiates transcription". In: *Proceedings of the National Academy of Sciences* 114.7, pp. 1536–1541.
- Nakanishi, Kotaro et al. (2013). "Eukaryote-specific insertion elements control human ARGONAUTE slicer activity". In: *Cell reports* 3.6, pp. 1893–1900.
- Nalabothula, Narasimharao et al. (2014). "The chromatin architectural proteins HMGD1 and H1 bind reciprocally and have opposite effects on chromatin structure and gene regulation". In: *BMC genomics* 15.1, pp. 1–14.
- Nekrasova, Irina et al. (2019). "Loss of a Fragile Chromosome Region leads to the Screw Phenotype in Paramecium tetraurelia". In: *Genes* 10.7, p. 513.
- Neri, Francesco et al. (2017). "Intragenic DNA methylation prevents spurious transcription initiation". In: *Nature* 543.7643, pp. 72–77.
- Neve, Bernadette et al. (2021). "Long non-coding RNAs: the tentacles of chromatin remodeler complexes". In: *Cellular and Molecular Life Sciences* 78.4, pp. 1139–1161.
- Ngo, Thuy et al. (2016). "Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability". In: *Nature communications* 7.1, pp. 1–9.
- Nielsen, Erik, Yun You, and James Forney (1991). "Cysteine residue periodicity is a conserved structural feature of variable surface proteins from Paramecium tetraurelia". In: *Journal of molecular biology* 222.4, pp. 835–841.
- Nonet, Michael, Doug Sweetser, and Richard A Young (1987). "Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II". In: *Cell* 50.6, pp. 909–915.
- Noto, Tomoko et al. (2015). "Small-RNA-mediated genome-wide trans-recognition network in Tetrahymena DNA elimination". In: *Molecular cell* 59.2, pp. 229–242.
- Nowacki, Mariusz, Włodzimierz Zagorski-Ostoja, and Eric Meyer (2005). "Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia". In: *Current Biology* 15.18, pp. 1616–1628.
- Nowacki, Mariusz et al. (2008). "RNA-mediated epigenetic programming of a genome-rearrangement pathway". In: *Nature* 451.7175, pp. 153–158.
- Oberbeckmann, Elisa et al. (2019). "Absolute nucleosome occupancy map for the Saccharomyces cerevisiae genome". In: *Genome research* 29.12, pp. 1996–2009.
- Okamura, Katsutomo and Eric C Lai (2008). "Endogenous small interfering RNAs in animals". In: *Nature reviews Molecular cell biology* 9.9, pp. 673–678.
- Orias, Eduardo, Deepankar Pratap Singh, and Eric Meyer (2017). "Genetics and epigenetics of mating type determination in Paramecium and Tetrahymena". In: *Annual Review of Microbiology* 71, pp. 133–156.
- Owsian, Dawid et al. (2022). "The transient Spt4-Spt5 complex as an upstream regulator of non-coding RNAs during development". In: *Nucleic acids research* 50.5, pp. 2603–2620.
- Patrick, Kristin L et al. (2015). "Genetic interaction mapping reveals a role for the SWI/SNF nucleosome remodeler in spliceosome activation in fission yeast". In: *PLoS genetics* 11.3, e1005074.
- Patro, Rob et al. (Apr. 2017). "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4, pp. 417–419.
- Pepenella, Sharon, Kevin J Murphy, and Jeffrey J Hayes (2014). "Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure". In: *Chromosoma* 123.1, pp. 3–13.

- Pirritano, Marcello et al. (2018). "Environmental Temperature Controls Accumulation of Transacting siRNAs Involved in Heterochromatin Formation". In: *Genes* 9.2.
- Postberg, Jan et al. (2018). "27nt-RNAs guide histone variant deposition via 'RNA-induced DNA replication interference' and thus transmit parental genome partitioning in *Stylonychia*". In: *Epigenetics & chromatin* 11.1, pp. 1–22.
- Preer, John R (1976). "Quantitative predictions of random segregation models of the ciliate macronucleus". In: *Genetics Research* 27.2, pp. 227–238.
- Preer, John R, Louise B Preer, and Bertina M Rudman (1981). "mRNAs for the immobilization antigens of *Paramecium*". In: *Proceedings of the National Academy of Sciences* 78.11, pp. 6776–6778.
- Prescott, David M (1994). "The DNA of ciliated protozoa". In: *Microbiological reviews* 58.2, pp. 233–267.
- Price, David H (2018). "Transient pausing by RNA polymerase II". In: *Proceedings of the National Academy of Sciences* 115.19, pp. 4810–4812.
- Ragunathan, Kaushik, Gloria Jih, and Danesh Moazed (2015). "Epigenetic inheritance uncoupled from sequence-specific recruitment". In: *Science* 348.6230, p. 1258699.
- Ramírez, Fidel et al. (2016). "deepTools2: a next generation web server for deep-sequencing data analysis". In: *Nucleic acids research* 44.W1, W160–W165.
- Reik, Wolf, Wendy Dean, and Jorn Walter (2001). "Epigenetic reprogramming in mammalian development". In: *Science* 293.5532, pp. 1089–1093.
- Reisner, AH, Janet Rowe, and RW Sleight (1969). "Tertiary structure of the soluble surface proteins of *Paramecium*". In: *Biochemistry* 8.11, pp. 4637–4644.
- Robinson, James T et al. (2011). "Integrative genomics viewer". In: *Nature biotechnology* 29.1, pp. 24–26.
- Rössle, Privatdozent Dr Robert (1905). "Spezifische Sera gegen Infusorien." In: *Archiv für Hygiene und Bakteriologie* 54, p. 1.
- Ruiz, Françoise et al. (1998). "Homology-dependent gene silencing in *Paramecium*". In: *Molecular biology of the cell* 9.4, pp. 931–943.
- Saitou, Naruya and Masatoshi Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Molecular biology and evolution* 4.4, pp. 406–425.
- Salhab, Abdulrahman et al. (2018). "A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains". In: *Genome biology* 19.1, pp. 1–13.
- Samuel, Charlotte, John Mackie, and John Sommerville (1981). "Macronuclear chromatin organization in *Paramecium primaurelia*". In: *Chromosoma* 83.4, pp. 481–492.
- Sandoval, Pamela Y et al. (2014). "Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting". In: *Developmental Cell* 28.2, pp. 174–188.
- Sawka-Gądek, Natalia et al. (2021). "Evolutionary plasticity of mating-type determination mechanisms in *Paramecium aurelia* sibling species". In: *Genome biology and evolution* 13.2, evaa258.
- Schmidt, Helmut J (1988). "Immobilization antigens". In: *Paramecium*. Springer, pp. 155–166.
- Schooneveld, Eleni van et al. (2015). "Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management". In: *Breast cancer research* 17.1, pp. 1–15.
- Schüller, Roland et al. (2016). "Heptad-specific phosphorylation of RNA polymerase II CTD". In: *Molecular cell* 61.2, pp. 305–314.

- Schwarz, Dianne S et al. (2003). "Asymmetry in the assembly of the RNAi enzyme complex". In: *Cell* 115.2, pp. 199–208.
- Scott, Jill M et al. (1994). "Non-Mendelian inheritance of macronuclear mutations is gene specific in *Paramecium tetraurelia*". In: *Molecular and cellular biology* 14.4, pp. 2479–2484.
- Sebé-Pedrós, Arnau, Bernard M Degnan, and Iñaki Ruiz-Trillo (2017). "The origin of Metazoa: a unicellular perspective". In: *Nature Reviews Genetics* 18.8, pp. 498–512.
- Segal, Eran and Jonathan Widom (2009). "Poly (dA: dT) tracts: major determinants of nucleosome organization". In: *Current opinion in structural biology* 19.1, pp. 65–71.
- Sellis, Diamantis et al. (2021). "Massive colonization of protein-coding exons by selfish genetic elements in *Paramecium* germline genomes". In: *Plos Biology* 19.7, e3001309.
- Seto, Anita G, Robert E Kingston, and Nelson C Lau (2007). "The coming of age for Piwi proteins". In: *Molecular cell* 26.5, pp. 603–609.
- Sharma, Upasna et al. (2018). "Small RNAs are trafficked from the epididymis to developing mammalian sperm". In: *Developmental cell* 46.4, pp. 481–494.
- Sheng, Yalan et al. (2020). "The completed macronuclear genome of a model ciliate *Tetrahymena thermophila* and its application in genome scrambling and copy number analyses". In: *Science China Life Sciences* 63.10, pp. 1534–1542.
- Sheng, Yalan et al. (2021). "Case Study of the Response of N6-Methyladenine DNA Modification to Environmental Stressors in the Unicellular Eukaryote *Tetrahymena thermophila*". In: *Mosphere* 6.3, e01208–20.
- Shieh, Annie Wan Yi and Douglas L Chalker (2013). "LIA5 is required for nuclear reorganization and programmed DNA rearrangements occurring during *tetrahymena* macronuclear differentiation". In: *PLoS One* 8.9, e75337.
- Shlyueva, Daria, Gerald Stampfel, and Alexander Stark (2014). "Transcriptional enhancers: from properties to genome-wide predictions". In: *Nature Reviews Genetics* 15.4, pp. 272–286.
- Sijen, Titia et al. (2007). "Secondary siRNAs result from unprimed RNA synthesis and form a distinct class". In: *Science* 315.5809, pp. 244–247.
- Simon, Martin and Helmut Plattner (2014). "Unicellular eukaryotes as models in cell and molecular biology: critical appraisal of their past and future value". In: *International review of cell and molecular biology* 309, pp. 141–198.
- Simon, Martin C, Simone Marker, and Helmut J Schmidt (2006a). "Inefficient serotype knock down leads to stable coexistence of different surface antigens on the outer membrane in *Paramecium tetraurelia*". In: *European journal of protistology* 42.1, pp. 49–53.
- (2006b). "Posttranscriptional control is a strong factor enabling exclusive expression of surface antigens in *Paramecium tetraurelia*". In: *Gene Expression, The Journal of Liver Research* 13.3, pp. 167–178.
- Simon, Martin C and Helmut J Schmidt (2005). "Variety of serotypes of *Paramecium primaurelia*: single epitopes are responsible for immunological differentiation". In: *Journal of Eukaryotic Microbiology* 52.4, pp. 319–327.
- (2007). "Antigenic Variation in Ciliates: Antigen Structure, Function, Expression 1". In: *Journal of Eukaryotic Microbiology* 54.1, pp. 1–7.
- Singh, Aditi et al. (2018). "Determination of the presence of 5-methylcytosine in *Paramecium tetraurelia*". In: *PloS one* 13.10, e0206667.
- Singh, Aditi et al. (2022). "RNA-mediated nucleosome depletion is required for elimination of transposon-derived DNA." In: *bioRxiv*.

- Singh, Deepankar Pratap et al. (2014). "Genome-defence small RNAs exapted for epigenetic mating-type inheritance". In: *Nature* 509.7501, pp. 447–452.
- Smolle, Michaela et al. (2012). "Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange". In: *Nature structural & molecular biology* 19.9, pp. 884–892.
- Sneppen, Kim and Leonie Ringrose (2019). "Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent". In: *Nature communications* 10.1, pp. 1–18.
- Sommerville, John (1969). "Serotype expression in *Paramecium*". In: *Advances in microbial Physiology*. Vol. 4. Elsevier, pp. 131–178.
- Song, Ji-Joon et al. (2004). "Crystal structure of Argonaute and its implications for RISC slicer activity". In: *science* 305.5689, pp. 1434–1437.
- Sonneborn, TM (1954). "The relation of autogamy to senescence and rejuvenescence in *Paramecium aurelia*". In: *The Journal of Protozoology* 1.1, pp. 38–53.
- (1975). "The *Paramecium aurelia* complex of fourteen sibling species". In: *Transactions of the American Microscopical Society*, pp. 155–178.
- Sonneborn, TM et al. (1950). "The cytoplasm in heredity." In: *Heredity* 4, pp. 11–36.
- Sonneborn, Tracy Morton (1937). "Sex, sex inheritance and sex determination in *Paramecium aurelia*". In: *Proceedings of the National Academy of Sciences of the United States of America* 23.7, p. 378.
- (1950). "Methods in the general biology and genetics of *Paramecium aurelia*". In: *Journal of Experimental Zoology* 113.1, pp. 87–147.
- Spåhr, Henrik et al. (2009). "Schizosaccharomyces pombe RNA polymerase II at 3.6 Å resolution". In: *Proceedings of the National Academy of Sciences* 106.23, pp. 9185–9190.
- Sparmann, Anke and Maarten van Lohuizen (2006). "Polycomb silencers control cell fate, development and cancer". In: *Nature Reviews Cancer* 6.11, pp. 846–856.
- Stein, Chad B et al. (2019). "Decoding the 5' nucleotide bias of PIWI-interacting RNAs". In: *Nature communications* 10.1, p. 828.
- Steurer, Barbara et al. (2018). "Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II". In: *Proceedings of the National Academy of Sciences* 115.19, E4368–E4376.
- Struhl, Kevin and Eran Segal (2013). "Determinants of nucleosome positioning". In: *Nature structural & molecular biology* 20.3, pp. 267–273.
- Suh, Hyunsuk et al. (2016). "Direct analysis of phosphorylation sites on the Rpb1 C-terminal domain of RNA polymerase II". In: *Molecular cell* 61.2, pp. 297–304.
- Svoboda, Petr (2020). "Key mechanistic principles and considerations concerning RNA interference". In: *Frontiers in Plant Science* 11, p. 1237.
- Swart, Estienne C et al. (2014). "Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion". In: *Nucleic acids research* 42.14, pp. 8970–8983.
- Szerlong, Heather J and Jeffrey C Hansen (2011). "Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure". In: *Biochemistry and Cell Biology* 89.1, pp. 24–34.
- Takata, Hideaki et al. (2013). "Chromatin compaction protects genomic DNA from radiation damage". In: *PloS one* 8.10, e75622.
- Talbert, Paul B and Steven Henikoff (2012). "Chromatin: packaging without nucleosomes". In: *Current Biology* 22.24, R1040–R1043.
- (2017). "Histone variants on the move: substrates for chromatin dynamics". In: *Nature reviews Molecular cell biology* 18.2, pp. 115–126.
- (2021a). "Histone variants at a glance". In: *Journal of cell science* 134.6, jcs244749.

- Talbert, Paul B and Steven Henikoff (2021b). "The Yin and Yang of Histone Marks in Transcription". In: *Annual review of genomics and human genetics* 22, pp. 147–170.
- Talbert, Paul B, Michael P Meers, and Steven Henikoff (2019). "Old cogs, new tricks: the evolution of gene expression in a chromatin context". In: *Nature Reviews Genetics* 20.5, pp. 283–297.
- Tam, Oliver H et al. (2008). "Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes". In: *Nature* 453.7194, pp. 534–538.
- Taverna, Sean D et al. (2007). "Long-distance combinatorial linkage between methylation and acetylation on histone H3 N termini". In: *Proceedings of the National Academy of Sciences* 104.7, pp. 2086–2091.
- Tian, Miao, Kazufumi Mochizuki, and Josef Loidl (2019). "Non-coding RNA transcription in Tetrahymena meiotic nuclei requires dedicated mediator complex-associated proteins". In: *Current Biology* 29.14, pp. 2359–2370.
- Tillo, Desiree and Timothy R Hughes (2009). "G+ C content dominates intrinsic nucleosome occupancy". In: *BMC bioinformatics* 10.1, pp. 1–13.
- Timmons, Lisa and Andrew Fire (1998). "Specific interference by ingested dsRNA". In: *Nature* 395.6705, pp. 854–854.
- Todesco, Marco et al. (2010). "A collection of target mimics for comprehensive analysis of microRNA function in Arabidopsis thaliana". In: *PLoS genetics* 6.7.
- Tremethick, David J (2007). "Higher-order structures of chromatin: the elusive 30 nm fiber". In: *Cell* 128.4, pp. 651–654.
- Trojer, Patrick and Danny Reinberg (2007). "Facultative heterochromatin: is there a distinctive molecular signature?" In: *Molecular cell* 28.1, pp. 1–13.
- Vanssay, Augustin de et al. (2020). "The Paramecium histone chaperone Spt16-1 is required for Pgm endonuclease function in programmed genome rearrangements". In: *PLoS genetics* 16.7, e1008949.
- Vishnoi, Anchal and Sweta Rani (2017). "MiRNA biogenesis and regulation of diseases: an overview". In: *MicroRNA Profiling*, pp. 1–10.
- Waddington, Conrad H (1942). "The epigenotype". In: *Endeavour* 1, pp. 18–20.
- Wade, Joseph T and David C Grainger (2018). "Spurious transcription and its impact on cell function". In: *Transcription* 9.3, pp. 182–189.
- Wang, Gang G, C David Allis, and Ping Chi (2007). "Chromatin remodeling and cancer, Part I: Covalent histone modifications". In: *Trends in molecular medicine* 13.9, pp. 363–372.
- Wang, Yuanyuan et al. (2017). "N6-adenine DNA methylation is associated with the linker DNA of H2A. Z-containing well-positioned nucleosomes in Pol II-transcribed genes in Tetrahymena". In: *Nucleic acids research* 45.20, pp. 11594–11606.
- Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews genetics* 10.1, pp. 57–63.
- Weintraub, Harold and Mark Groudine (1976). "Chromosomal Subunits in Active Genes Have an Altered Conformation: Globin genes are digested by deoxyribonuclease I in red blood cell nuclei but not in fibroblast nuclei." In: *Science* 193.4256, pp. 848–856.
- Wierzbicki, Andrzej T et al. (2009). "RNA polymerase V transcription guides ARGONAUTE4 to chromatin". In: *Nature genetics* 41.5, pp. 630–634.
- Wood, V et al. (2002). "The genome sequence of Schizosaccharomyces pombe". In: *Nature* 415.6874, pp. 871–880.
- Woolcock, Katrina J et al. (2011). "Dicer associates with chromatin to repress genome activity in Schizosaccharomyces pombe". In: *Nature structural & molecular biology* 18.1, pp. 94–99.

- Woolcock, Katrina J et al. (2012). "RNAi keeps Atf1-bound stress response genes in check at nuclear pores". In: *Genes & Development* 26.7, pp. 683–692.
- Wu, Steven J et al. (2021). "Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression". In: *Nature biotechnology* 39.7, pp. 819–824.
- Xiong, Jie et al. (2016). "Dissecting relative contributions of cis-and trans-determinants to nucleosome distribution by comparing Tetrahymena macronuclear and micronuclear chromatin". In: *Nucleic acids research* 44.21, pp. 10091–10105.
- Xu, Jing et al. (2021). "A Polycomb repressive complex is required for RNAi-mediated heterochromatin formation and dynamic distribution of nuclear bodies". In: *Nucleic acids research* 49.10, pp. 5407–5425.
- Yang, Fan (2022). "Promoter antisense RNAs: beyond transcription by-products of active promoters". In: *RNA biology* 19.1, pp. 533–540.
- Yang, Zhiyong et al. (2006). "HEN1 recognizes 21–24 nt small RNA duplexes and deposits a methyl group onto the 2' OH of the 3' terminal nucleotide". In: *Nucleic acids research* 34.2, pp. 667–675.
- Yao, Meng-Chao et al. (2007). "Identification of novel chromatin-associated proteins involved in programmed genome rearrangements in Tetrahymena". In: *Journal of cell science* 120.12, pp. 1978–1989.
- Yu, Bin and Xuemei Chen (2010). "Analysis of miRNA modifications". In: *Plant MicroRNAs*. Springer, pp. 137–148.
- Yu, Bin et al. (2005). "Methylation as a crucial step in plant microRNA biogenesis". In: *Science* 307.5711, pp. 932–935.
- Zhang, Ke et al. (2008). "Roles of the Clr4 methyltransferase complex in nucleation, spreading and maintenance of heterochromatin". In: *Nature structural & molecular biology* 15.4, pp. 381–388.
- Zhao, Xiaolu et al. (2019). "RNAi-dependent Polycomb repression controls transposable elements in Tetrahymena". In: *Genes & development* 33.5-6, pp. 348–364.
- Zheng, Weibo et al. (2021). "The compact macronuclear genome of the ciliate *Halteria grandinella*: A transcriptome-like genome with 23,000 nanochromosomes". In: *Mbio* 12.1, e01964–20.
- Zhou, Haining et al. (2022a). "Rixosomal RNA degradation contributes to silencing of Polycomb target genes". In: *Nature* 604.7904, pp. 167–174.
- Zhou, Yuanyuan et al. (2022b). "Absolute quantification of chromosome copy numbers in the polyploid macronucleus of *Tetrahymena thermophila* at the single-cell level". In: *Journal of Eukaryotic Microbiology*, e12907.
- Zuckerkindl, Emile and Linus Pauling (1965). "Evolutionary divergence and convergence in proteins". In: *Evolving genes and proteins*. Elsevier, pp. 97–166.

Appendix A

Supplementary Material Chapter 3

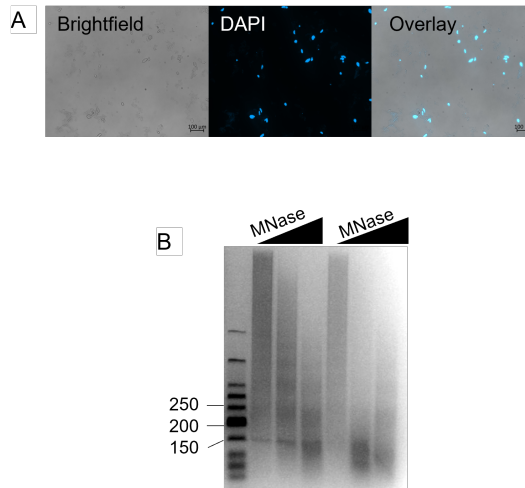


Figure A.1 (A) *Paramecium* MACs were isolated by ultrasonication from fixed material and mixed with DAPI for verification of MAC integrity. (B) Nucleosomal ladder of 2 μ g DNA from MNase digested chromatin in two exemplary replicates. Samples were treated with increasing units of enzyme and were loaded on a 3% agarose gel in ascending order (1U,10U and 128U).

Table A.1 Datasets used for nucleosome profile analyses (left) and pausing analyses (right)

Organism	Sequence type	GEO accession number	SRR accession number	Organism	Sequence type	GEO accession number
<i>Schizosaccharomyces pombe</i>	naked DNA	GSE140920	SRR10528270	<i>Tetrahymena thermophila</i>	Pol II	GSE77583
<i>Schizosaccharomyces pombe</i>	MNase	GSE52170	SRR1821723	<i>Tetrahymena thermophila</i>	Input DNA	GSE77583
<i>Schizosaccharomyces pombe</i>	MNase	GSE141676	SRR10611800	<i>Tetrahymena thermophila</i>	mRNA	GSE130336
<i>Drosophila melanogaster</i>	naked DNA	GSE69177	SRR2038260, SRR2038261	<i>Schizosaccharomyces pombe</i>	Pol II	GSE115636
<i>Drosophila melanogaster</i>	Mnase - High 1N	GSE69177	SRR2038276, SRR2038277	<i>Schizosaccharomyces pombe</i>	Input DNA	GSE115636
<i>Homo sapiens</i>	naked DNA	GSE100401	SRR5749438	<i>Schizosaccharomyces pombe</i>	mRNA	GSE115636
<i>Homo sapiens</i>	MNase - 1000U	GSE100401	SRR5749432, SRR5749433	<i>Homo sapiens</i>	Pol II	GSE98368
<i>Tetrahymena thermophila</i>	MNase heavy digest	GSM2055773	SRX1590945	<i>Homo sapiens</i>	mRNA	GSE98368
<i>Tetrahymena thermophila</i>	MNase light digest	GSM2055775	SRX1590947			
<i>Tetrahymena thermophila</i>	naked DNA	GSE64061	SRR2041674			
<i>Tetrahymena thermophila</i>	MNase light digest	GSE64061	SRR2041661			

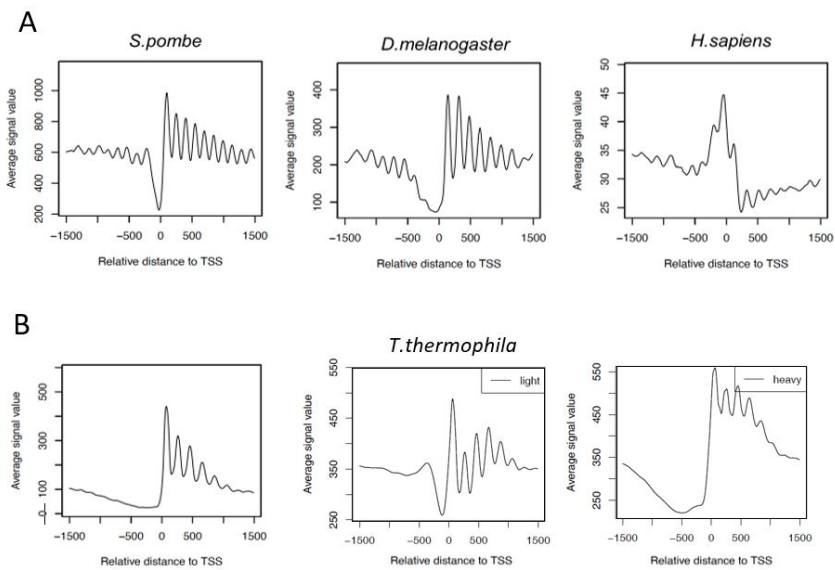


Figure A.2 (A) Plot of nucleosome distribution at the TSS (1500bp up- and downstream) plotted with DANPOS2 for *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*. TSS were annotated by EST analysis (Wood et al., 2002; Adams et al., 2000; Lander et al., 2001). **(B)** Same plot as in **A**, but for *Tetrahymena thermophila* MNase-seq analyses from varying fixation/digestion protocols. TSS annotation is predicted from CAP-seq data (Arnaiz et al., 2017). Left: mild fixation+light digest (SRR2041661); middle: no fixation+light digest (Rep1, GSM2055775); right: no fixation+heavy digest (Rep1, GSM2055773).

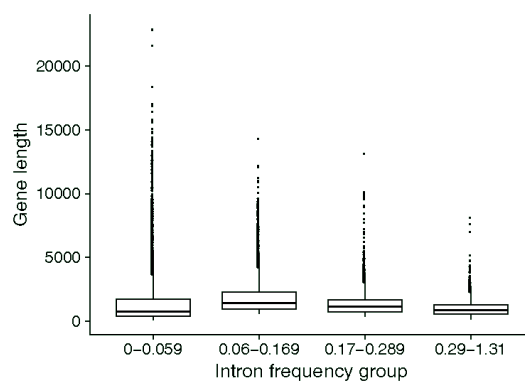


Figure A.3 Distribution of gene length across different intron frequency groups.

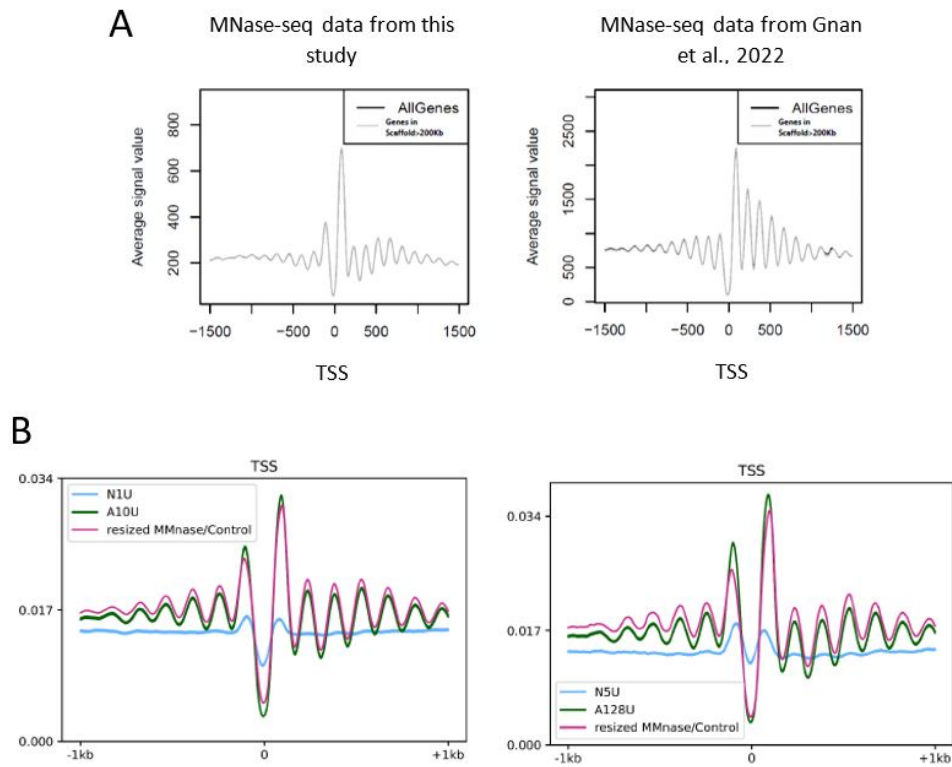


Figure A.4 (A) Nucleosome distribution at the TSS obtained by plotting data from this thesis (left) and the data from (Gnan et al., 2022) (right), both analyzing nucleosome profiles of *Paramecium* chromatin. Plots were created for all genes and genes on scaffolds ≥ 200 kb using the DANPOS2 pipeline. **(B)** Nucleosome profile plot at the TSS kindly provided by Gnan et al., 2022 using their nucleosome analysis pipeline and MNase-seq data from this thesis (A10U).

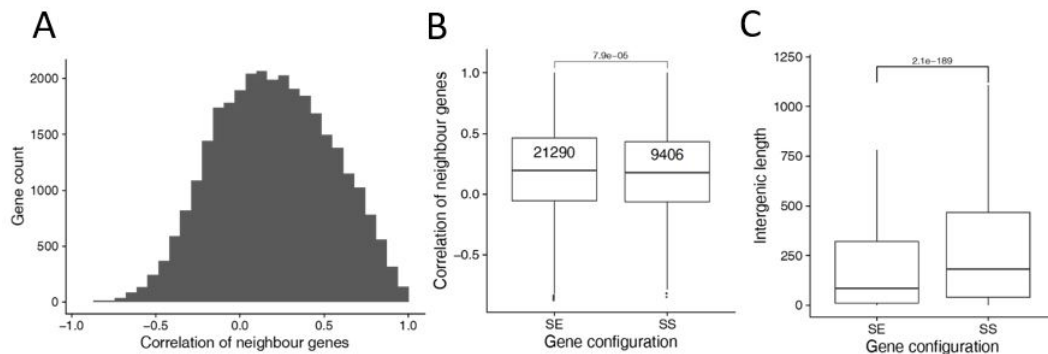


Figure A.5 (A) Pearson's correlation coefficient of neighboring genes' expression from different serotypes/cultivation conditions (Cheaib et al., 2015). **(B)** Pearson's correlation coefficient of neighboring genes expression for genes with different configurations (bi-/unidirectional). **(C)** Length of intergenic regions for genes with the same configurations as in B. P-values are based on a two-tailed Wilcoxon test.

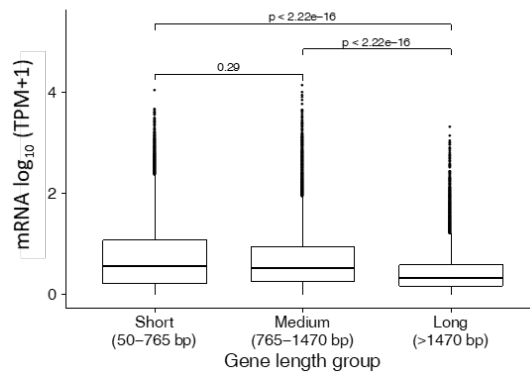


Figure A.6 Expression of genes ($\log_{10}(\text{TPM}+1)$), separated by length.

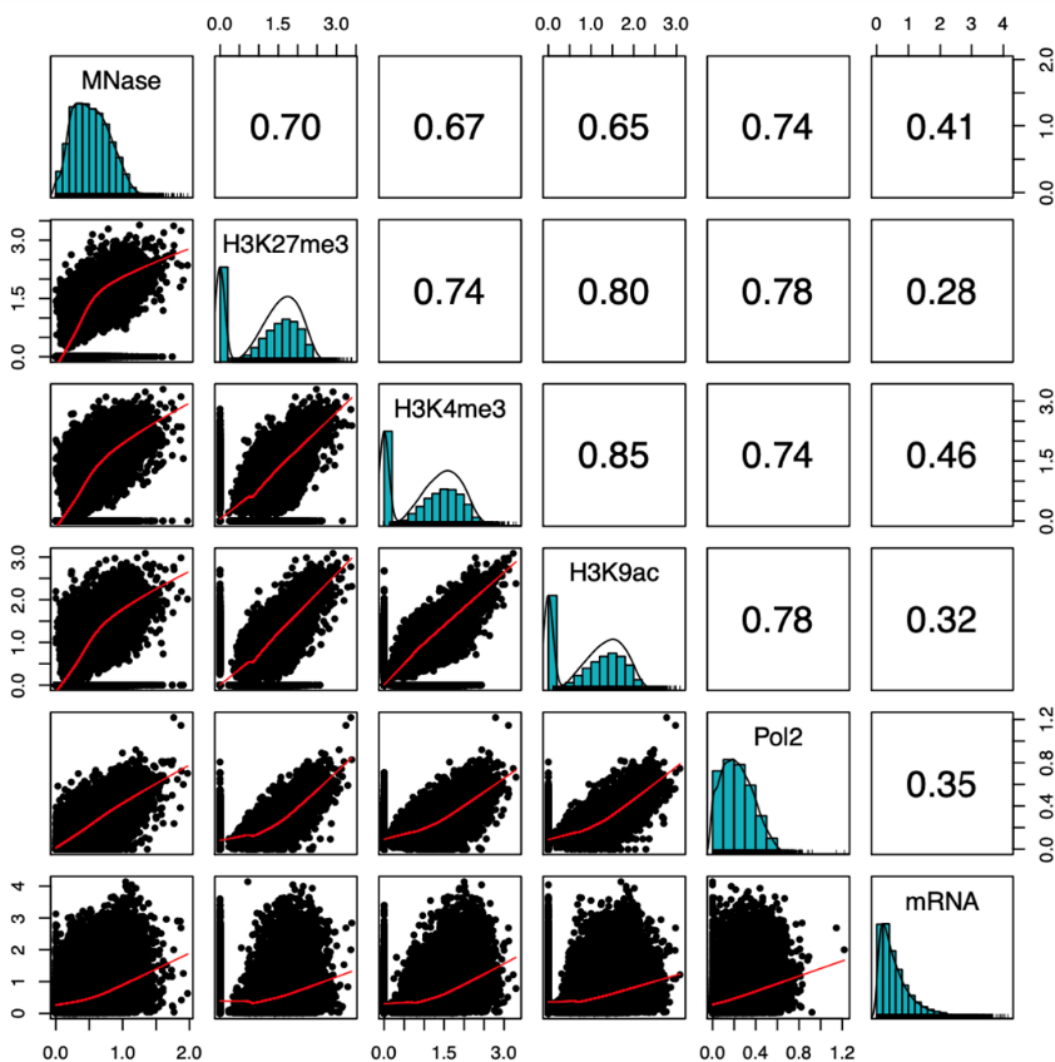


Figure A.7 Distribution plot of each epigenetic mark and mRNA (TPM) are shown along the diagonal. For the respective variables mentioned along the x- and y-axis of each box, the Pearson's correlation coefficients (above the diagonal) are shown. The y-axis of scatter plots belongs to the variable mentioned along the horizontal line of that plot. All values are \log_{10} transformed with a pseudo count of 1.

Appendix B

Supplementary Material Chapter 4

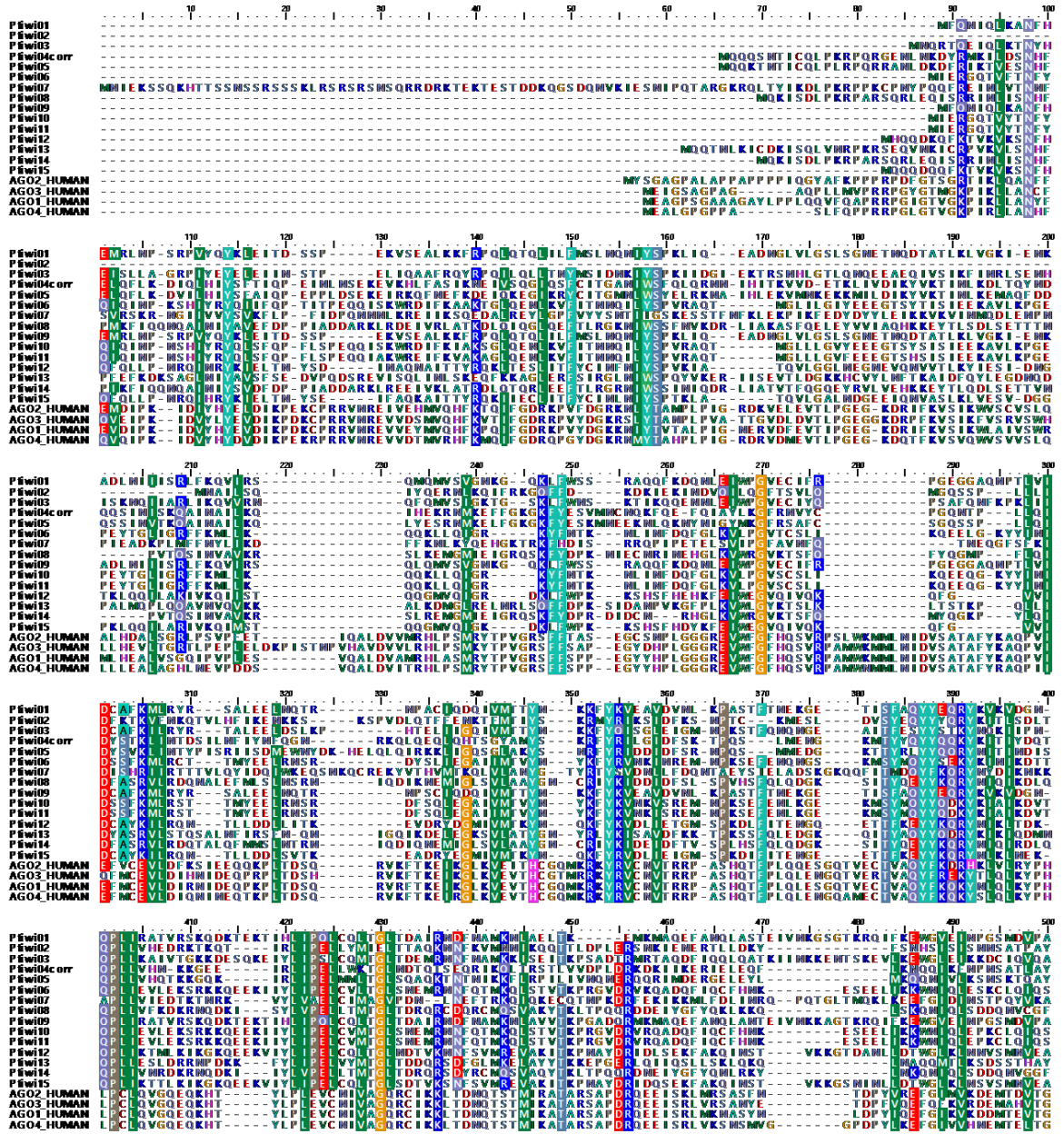


Figure B.1

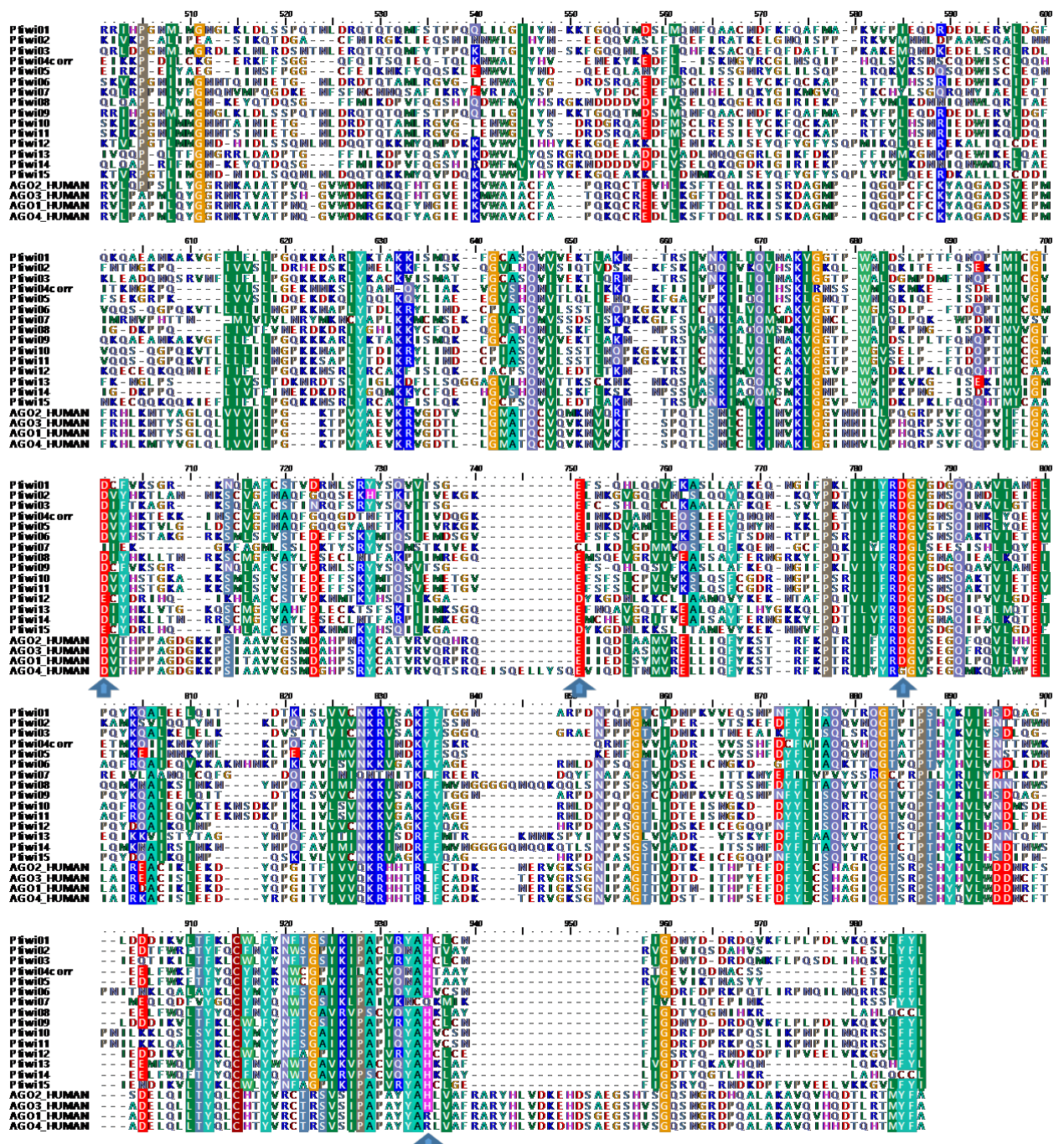


Figure B.2 Extension of Figure B.1. ClustalX alignment of *Paramecium tetraurelia*'s 15 Ptiwi amino acid sequences. Blue arrows (bottom) indicate conserved catalytic residues. Human Ago1-4 are included as references.

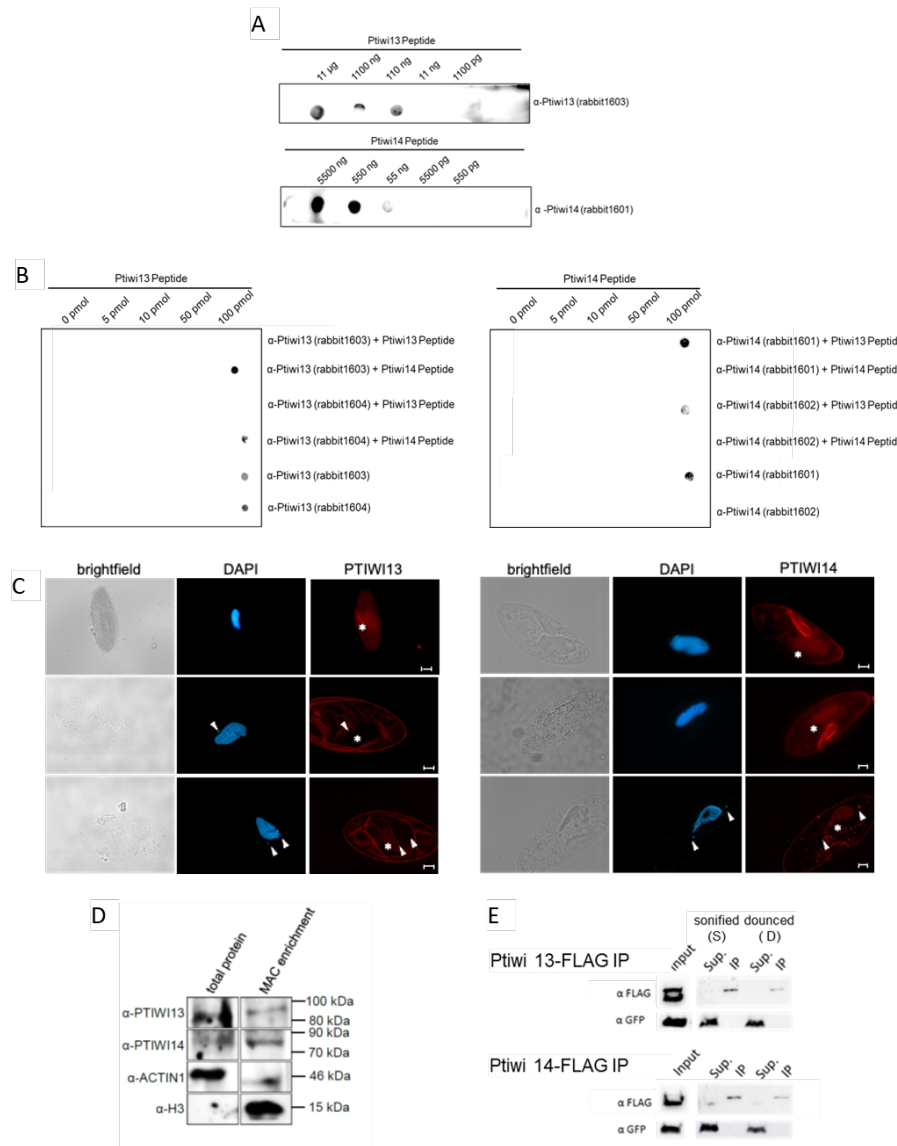


Figure B.3 (A) Dot blot assay using custom antibodies against *P. tetraurelia* Ptiwi 13 and Ptiwi 14. Indicated peptides were spotted in different amounts. Hybridization with the respective antibodies showed reactivity with the corresponding peptide. (B) Peptide competition assay. 0-100 pmol of each peptide were spotted. The antibodies were pre-incubated with the indicated peptide in 100x excess and the mix was used to decorate the membrane. Left: Top row shows blocking of α -Ptiwi 13 antibody with Ptiwi 13 peptide which results in loss of antibody binding. Blocking of the α -Ptiwi 13 antibody with the Ptiwi 14 peptide does not have an effect on the binding affinity to the spotted Ptiwi 13 peptide (mock, 2nd row). Competition assay was performed for antibodies from serum of two immunized rabbits (rabbit1603/1604). The last two rows show antibody binding without pre - incubation. Same patterns were observed for the Ptiwi 14 antibody (right). (C) Localization of Ptiwi proteins in vegetative *Paramecium* cells. Cells were analyzed by immunofluorescence staining using custom antibodies (red) as in Figure 4.1. (D) Protein from whole cell lysate (total protein) and protein from MAC enrichments were decorated with custom antibodies against α -Ptiwi 13 and α -Ptiwi 14. α -Actin and histone antibody α -H3 serve as loading controls and verification of MAC enrichment. Estimated molecular weights (from ParameciumDB): Ptiwi 13 91.9 kDa, Ptiwi 14 91.4 kDa, Actin1 41.7 kDa and histone H3 15.8 kDa. (E) Control Western blots for the IPs using α -FLAG antibody for Ptiwi detection and α -GFP (Sup.-Supernatant). Two different setups of the IPs used sonication (S) and douncing (D) for cell lysis, the latter remains MAC structure but permeabilized.

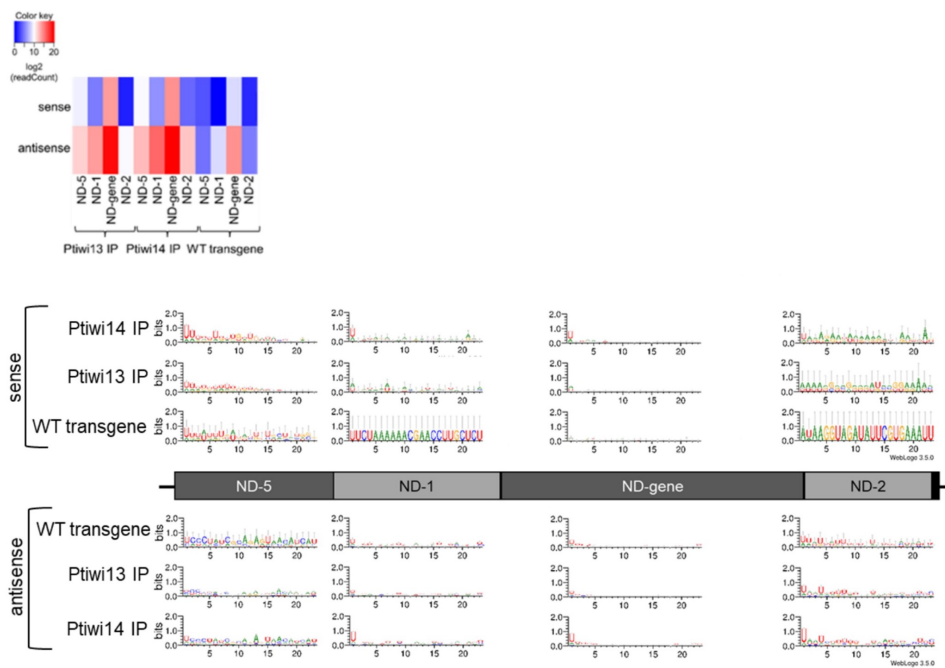


Figure B.4 Sequence logos of 23nt siRNAs in Ptiwi IPs. Sequences were separated by their direction (sense, top; antisense, bottom). Heat maps as in Figure B.5

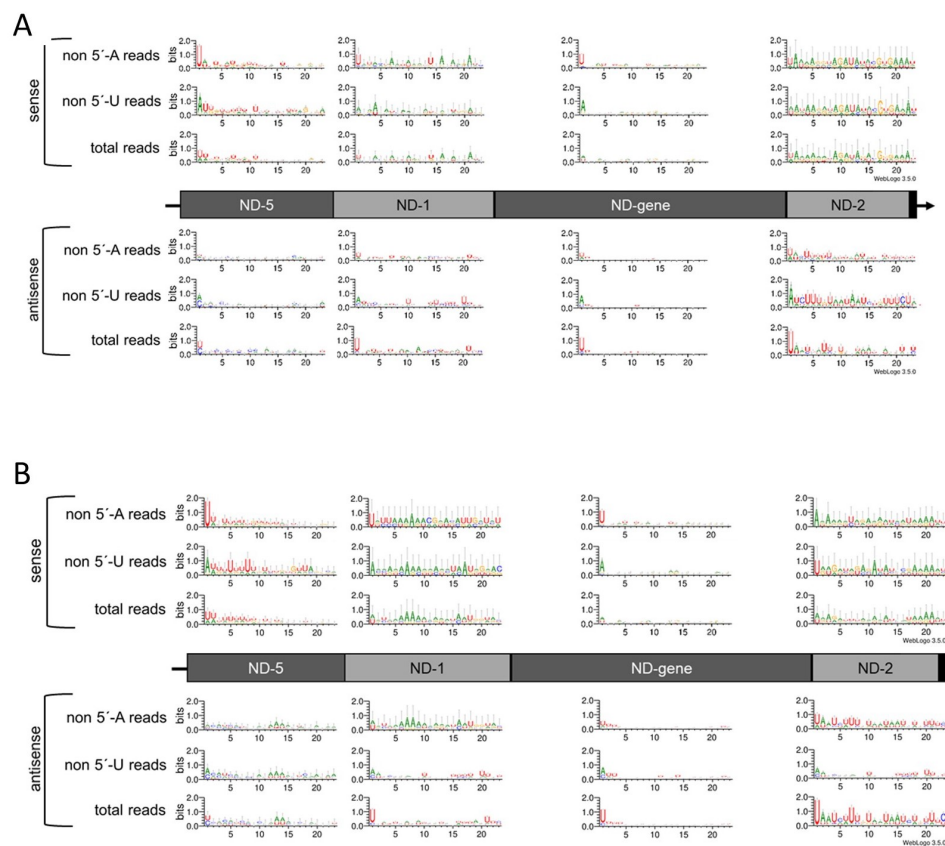


Figure B.5 (A) Logos of 23nt sRNAs from ICL KD and **(B)** ICL KD periodate treated RNA. Heat maps show log₂ of the number of reads from each sample used for logo production.

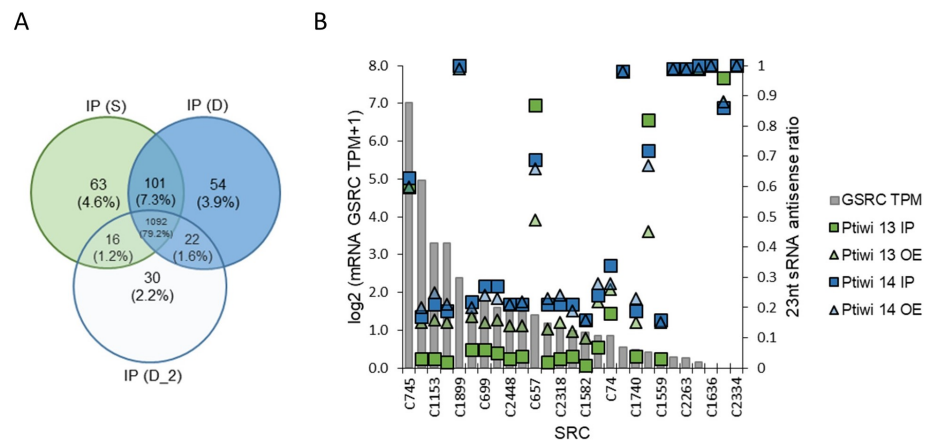


Figure B.6 (A) Analysis of cluster-associated sRNAs in Ptiwi IPs. Sample preparation is described in Figure B.3. Venn diagram indicates the number of SRCs appearing with ≤ 1 TPM in IPs. Below the percentage of covered SRCs is given. **(B)** Genes associated with SRCs (GSRCS) were analyzed according to their mRNA expression level (grey bar) and the antisense ratio of small RNAs mapping to the SRC located in the respective gene (squares for the Ptiwi IPs and in triangles for the overexpression lines). Only those GSRCS were analyzed which can be assigned to 100% to one SRC. SRCs had to be shared among all four samples and were filtered for having at least 2,000 and 100% mapping reads of 23nt length.

Appendix C

Manuscripts

Research

Broad domains of histone marks in the highly compact *Paramecium* macronuclear genome

Franziska Drews,^{1,2} Abdulrahman Salhab,^{3,7} Sivarajan Karunanithi,^{4,5,7,8}
Miriam Cheaib,² Martin Jung,⁶ Marcel H. Schulz,^{4,5} and Martin Simon^{1,2}

¹Molecular Cell Biology and Microbiology, Faculty for Mathematics and Natural Sciences, University of Wuppertal, 42219 Wuppertal, Germany; ²Molecular Cell Dynamics, Centre for Human and Molecular Biology, Saarland University, 66123 Saarbrücken, Germany; ³Genetics/Epigenetics, Centre for Human and Molecular Biology, Saarland University, 66123 Saarbrücken, Germany; ⁴Cluster of Excellence, Multimodal Computing and Interaction, Saarland University and Department for Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany; ⁵Institute for Cardiovascular Regeneration, Goethe-University Hospital, 60590 Frankfurt, Germany; ⁶School of Medicine, Medical Biochemistry and Molecular Biology, Saarland University, 66421 Homburg, Germany

The unicellular ciliate *Paramecium* contains a large vegetative macronucleus with several unusual characteristics, including an extremely high coding density and high polyploidy. As macronuclear chromatin is devoid of heterochromatin, our study characterizes the functional epigenomic organization necessary for gene regulation and proper Pol II activity. Histone marks (H3K4me3, H3K9ac, H3K27me3) reveal no narrow peaks but broad domains along gene bodies, whereas intergenic regions are devoid of nucleosomes. Our data implicate H3K4me3 levels inside ORFs to be the main factor associated with gene expression, and H3K27me3 appears in association with H3K4me3 in plastic genes. Silent and lowly expressed genes show low nucleosome occupancy, suggesting that gene inactivation does not involve increased nucleosome occupancy and chromatin condensation. Because of a high occupancy of Pol II along highly expressed ORFs, transcriptional elongation appears to be quite different from that of other species. This is supported by missing heptameric repeats in the C-terminal domain of Pol II and a divergent elongation system. Our data imply that unoccupied DNA is the default state, whereas gene activation requires nucleosome recruitment together with broad domains of H3K4me3. In summary, gene activation and silencing in *Paramecium* run counter to the current understanding of chromatin biology.

[Supplemental material is available for this article.]

The degree of epigenetic differentiation and the organization of eukaryotic genomes are usually adapted to the complexity of an organism: Chromatin serves as an additional layer of information, either for manifestation of gene expression patterns, for the cyclic condensation of chromosomes, or for microtubule-assisted separation of DNA in mitotic divisions. Chromatin further influences the proper processing of functional mRNAs as histone modifications influence Pol II dynamics and its interaction with RNA modifying components, such as the capping enzyme or the spliceosome.

Paramecium tetraurelia is a unicellular organism belonging to the SAR clade (including Stramenophiles, alveolates, and Rhizaria), which is as distant to plants, fungi, and animals. *Paramecium* is a ciliate, a phylum of Alveolata, and shows an unusual nuclear feature: Although unicellular, these cells already differentiate between germline and soma by germline micronuclei (MICs) and somatic macronuclei (MACs). Both differ in structural and functional aspects. MICs are small (1–2 μm) and transcriptionally inactive during vegetative growth, because the large (~30-μm) MACs transcribe all necessary genes to allow for cell proliferation (Bétermier and Duharcourt 2014). During sexual reproduction, haploid meiotic nuclei are reciprocally exchanged and fuse to a zy-

gote nucleus: This creates new MICs and MACs, whereas the new developing MAC (anlagen) already transcribes some genes involved in development (Furrer et al. 2017; Rzeszutek et al. 2020).

The genomic structures between MICs and MACs are quite different. MICs contain thousands of short transposon remnants (internal eliminated sequences [IESs]), which become deleted by a RNAi-related mechanism during macronuclear development (Allen and Nowacki 2020). The MAC differs from the MICs by the absence of IESs and transposons (Guérin et al. 2017). In addition, MAC chromosomes are tiny in size, usually <1 Mb, because MIC chromosomes are fragmented into many (about 200) different MAC chromosomes. These are amplified then to about 800 copies each, resulting in a massive polyploidy. The separation of that many DNA molecules, approximately 200 MAC chromosomes × 800n, is realized by amitotic divisions of the MAC: Replicated DNA becomes distributed to daughter nuclei without chromosome condensation and a typical mitotic spindle. The latter would be useless as the absence of centromeres (Lhuillier-Akakpo et al. 2016) and, consequently, kinetochores would not allow for attachment of microtubules.

In 2006, the *Paramecium* macronuclear genome project revealed two highly unexpected findings: (1) an exceptionally

⁷These authors contributed equally to this work.

⁸Present address: Institute of Molecular Biology (IMB), 55128 Mainz, Germany

Corresponding author: masimon@uni-wuppertal.de

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.276126.121>.

© 2022 Drews et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

high number of genes (about 40,000), most of them resulting from three successive whole-genome duplications, and (2) an exceptionally high coding density of 78%. The latter is owing to tiny introns, predominantly 25 bp in length, and small intergenic regions (352 bp on average) (Aury et al. 2006).

Chromatin during the amitotic M phase remains uncondensed, suggesting that the MAC does not harbor the full genetic requirements to create highly condensed chromatin. In addition, interphase chromatin was reported to show several unusual features compared with other species based on chromatin spread preparations, for instance, the finding of several unusual filament types and the appearance of a low level of polyteny between individual transcription nodes (Samuel et al. 1981). Classical heterochromatin is believed to be absent from the MAC, although a deeper biochemical insight in the MAC chromatin organization is still missing. The same holds true for the presence of classical repressive histone marks in the vegetative MAC, raising the question of how gene repression is regulated. Another epigenetic mark, 5-methylcytosine, is known to be involved in the negative regulation of gene expression in many eukaryotes. However, 5-methylcytosine is reportedly absent in MAC DNA (Singh et al. 2018).

Hence, the contribution of dynamic MAC chromatin modifications to the regulation of gene expression remains poorly understood in ciliates. We know from other organisms that chromatin marks have functions in RNA processing and active elongation of transcription. Current studies of mammalian chromatin report functions for well-positioned nucleosomes in the context of Pol II phosphorylation and interaction with RNA modifying enzymes. This raises the question of how such a regulation is realized in ciliates, specifically in *Paramecium*.

+1 nucleosome positioning, for instance, was indicated to correlate with Pol II pausing and increased recruitment of negative elongation factor (NELF) (Jimeno-González and Reyes 2016). Whereas initiation of transcription is accompanied by phosphorylation of serine 5, P-TEFb was shown to mediate the conversion of the Pol II complex from its initiation to the processive elongation form, which includes phosphorylation of serine 2 (Egloff and Murphy 2008; Buratowski 2009). Promoter proximal pausing is known to be controlled by the negative regulators NELF and DSIF, whereas the C-terminal domain (CTD) of Pol II interacts with the capping components for 5'-capping of the nascent mRNA. Similarly, polyadenylation and splicing are controlled by both the CTD of Pol II and correctly positioned nucleosomes (Böhm and Östlund Farrants 2011). Especially for the latter aspect, alternative splicing has been implicated to be regulated by alternative CTD phosphorylation regulated by the SWI/SNF chromatin remodeling complex (Batsché et al. 2006), rich heptad repeat. Although we do not know much about these mechanisms in ciliates, we suspect them to differ from the above-described CTD regulation and interaction with additional components in metazoans. This suspicion arises from the missing Pol II heptameric repeats in *Paramecium*, which likely also affect the interacting complexes owing to a coevolutionary effect. One of those complexes involved in transcription coactivation and elongation, the Mediator complex, for instance, significantly differs from *Tetrahymena* to other species (Zhao and Liu 2019). As a consequence, we currently do not understand the role of the ciliate epigenome architecture concerning Pol II activity in terms of initiation, elongation, pausing, and interaction with complexes. In this work, we aim to understand the epigenomic organization of the polyploid vegetative MAC of *P. tetraurelia*. These cells contain two diploid and transcriptionally silent micronuclei, which

divide by classical mitosis during cellular fission, whereas the MAC divides amitotically: Stretching and outlining results in uncontrolled separation of uncondensed chromosomes (Fig. 1A). The interpretation of any MAC epigenome data requires a look for the genomic structure of the chromosomes. During their processing from MIC chromosomes after sexual recombination, heterochromatic regions such as telomeres, centromeres, satellites, and transposons become eliminated in addition to about 45,000 transposon remnants called IES elements (Fig. 1B). Fragments undergo de novo telomere addition, resulting in small acentromeric chromosomes with a size of <1 Mb. These chromosomes exist at varying lengths owing to imprecise eliminations of repeated sequences (Duret et al. 2008). Compared with other species, even the related ciliate *Tetrahymena*, the *Paramecium* MAC genome shows an extremely high coding density of ~80%, with small intergenic regions and tiny introns of 25 nt on average (Aury et al. 2006).

Results

Unusual properties of the macronuclear genome

The mechanisms of DNA elimination described above during development of the *Paramecium* MAC result in a highly compact genome with striking differences in comparison to *Schizosaccharomyces pombe* and individual metazoans (Fig. 2A,B).

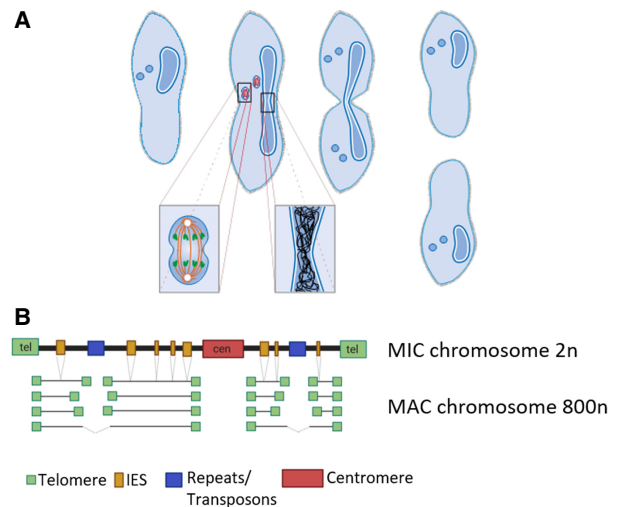


Figure 1. *Paramecium* vegetative cell divisions and chromosomal structure of MIC and MAC. (A) *Paramecium tetraurelia* showing two generative MICs and one vegetative MAC. Cell division involves mitotic separation of condensed MIC chromosomes and amitotic separation of uncondensed MAC chromosomes. While MICs and MAC divide, the nuclear envelope remains at both nuclei. (Figure courtesy of Jens Boenigk and Martin Simon.) (B) Chromosomes of the diploid MIC are large and contain centromeres and telomeres similar to canonical eukaryotic chromosomes. In addition, they consist of about 45,000 internal eliminated sequence (IES) elements and repeats (transposons, minisatellites). During macronuclear development after sexual reproduction (not shown here), telomeres, centromeres, repeats, and IESs become eliminated by different mechanisms. Although IESs are precisely excised, elimination of repeats and, presumably, centromeres is imprecise, resulting in fragmentation into heterogenous macronuclear chromosomes (with rare fusion of fragments). All macronuclear fragments show de novo telomere addition and amplification to 800n (created with BioRender [https://biorender.com]).

Drews et al.

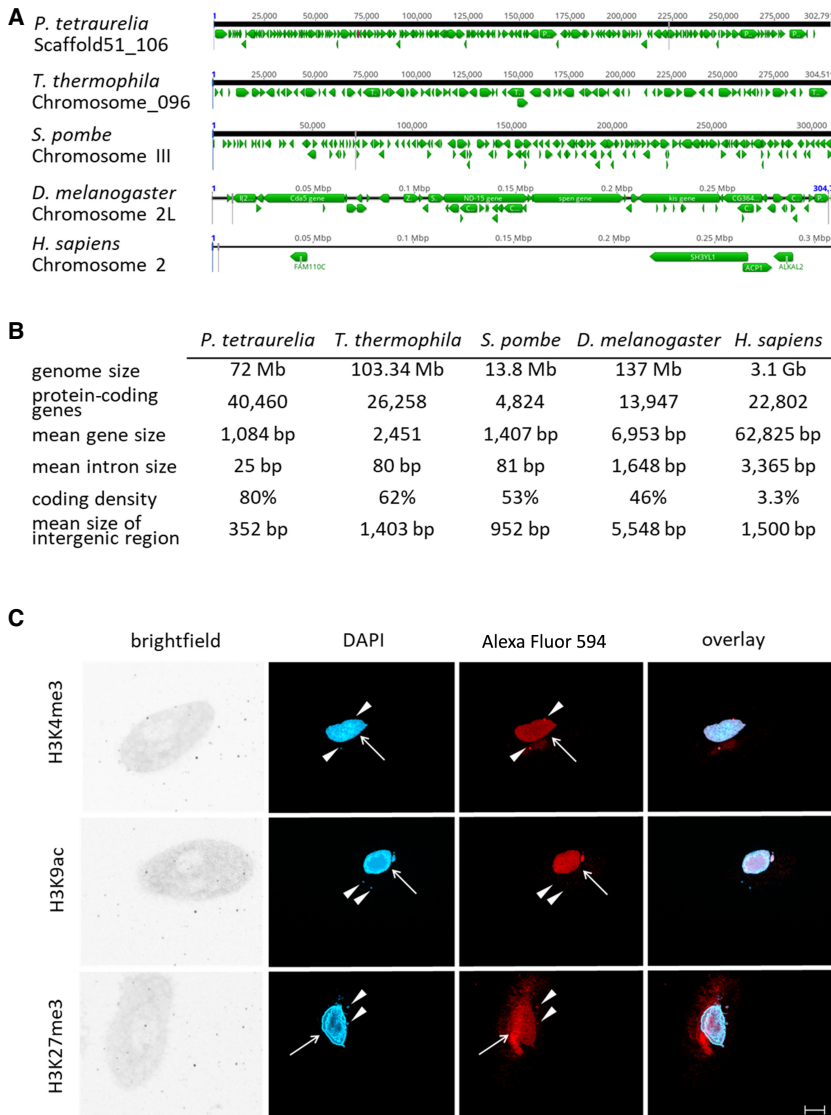


Figure 2. Features of the *Paramecium* genome in comparison to other organisms. (A) Comparisons of distribution of genes (green arrows) along the chromosomes of selected organisms to highlight the variation in coding density (*P. tetraurelia*, *Tetrahymena thermophila*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, *Homo sapiens*). A window of 300 kb is shown for each chromosome in a genome browser. (B) Summary of genomic features of the same organisms named in A. For details on collected data, see Methods. (C) Detection of histone modifications in vegetative *Paramecium* nuclei by immunofluorescence staining. DNA in the nuclei is stained with DAPI (blue), and antibodies directed against the three indicated modifications (H3K4me3, H3K9ac, H3K27me3) were labeled with a secondary Alexa Fluor 594 conjugated antibody (red). Arrowheads point at micronuclei; arrows indicate position of the macronucleus. Other panels show brightfield and overlay of signals. Representative overlays of Z-stacks of magnified views are shown. Scale bar, 10 μ m.

To quantify global epigenome organization in *Paramecium*, we first investigated the distribution of histone H3 modifications in the vegetative MAC, because histone modifications are major contributors to chromatin architecture. Immunofluorescence analysis with histone H3-specific antibodies shows H3K4me3 and H3K27me3 occurring in both MICs and the MAC, whereas H3K9ac is present in the MAC only (Fig. 2C). The MAC H3K27me3 signal is usually weak in immunofluorescence, similar to earlier reports (Ignarski et al. 2014), and shows slight unspecific

staining of extranuclear structures as the oral apparatus. To test the specificity of the antibodies for their respective target, competition assays using dot-blot were performed and are shown in Supplemental Figure S1.

Low nucleosome occupancy in intergenic regions and silent genes

To characterize nucleosome positioning, mononucleosomal DNA was isolated after digestion of MAC chromatin with micrococcal nuclease (MNase). Reads were mapped to the genome and normalized against a digest of naked DNA, resulting in discrete peaks for both setups using 10 or 128 U MNase (Fig. 3A), corresponding to light and heavy digestion. As the figure suggests that intergenic regions show low nucleosome occupancy, we separately analyzed coding genes and intergenic regions, the latter being defined as the region in between the transcription start site (TSS)/transcription termination site (TTS) of the gene of interest and the TSS/TTS (depends on the orientation) of the upstream gene. Figure 3B shows that genes show increased nucleosome occupancy in the 5'- and 3'-coding regions associated with drops in occupancy in flanking noncoding regions. The latter indeed show general low occupancy (Fig. 3C). For further quantification, we dissected genes by their expression levels (Fig. 3D) and calculated the associated nucleosome occupancy. Figure 3E shows the MNase signals quantified in intergenic regions and quantiles of genes. Intergenic regions show the lowest nucleosome occupancy. Please note that these values are not normalized for the individual gene length of groups, given in Supplemental Figure S2A. In support of these analyses, Supplemental Figure S2B shows the occupancy only of the most prominent nucleosome (+1) in these gene groups. Genes show increasing nucleosome occupancy with increasing gene expression levels. This is an unexpected result, as unoccupied DNA is believed to be highly accessible for Pol II and therefore usually defined as active chromatin. Our results here suggest that this is the opposite in the *Paramecium* MAC.

Prominent +1 nucleosomes mark actively transcribed genes

We aim to analyze the nucleosome positioning and occupancy in genes more in detail. Genomic analysis of MNase data revealed well-positioned +1 and -1 nucleosomes at the TSS (Fig. 4A). Especially the presence of -1 nucleosomes differs from analog analyses

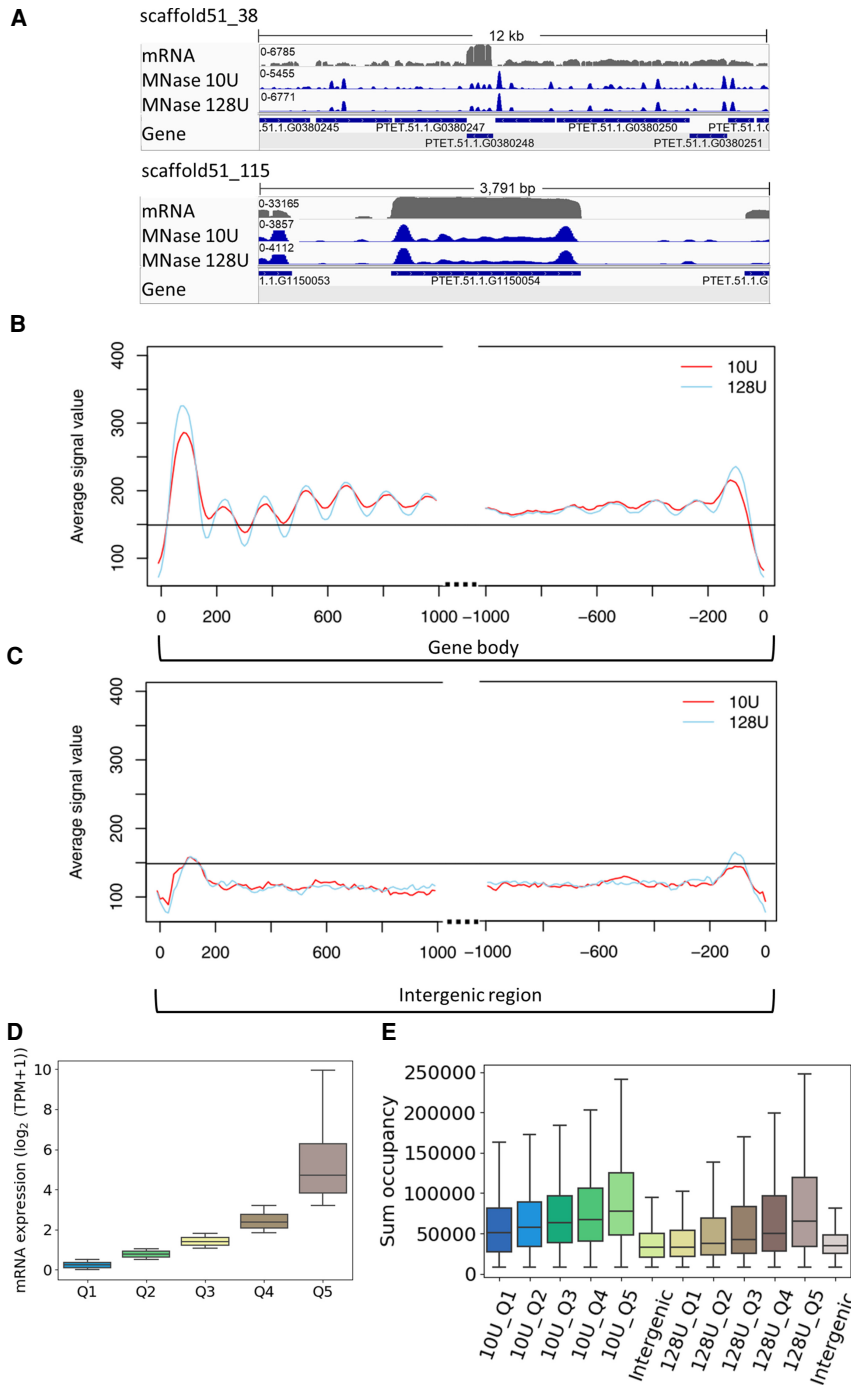


Figure 3. MNase-seq results reveal well-positioned +1 nucleosomes. (A) Exemplary view of nucleosome distribution along the MAC scaffolds of *Paramecium*. Top panel shows the peak distribution in a 12-kb window, and the bottom panel shows the magnified view on one gene. For both panels, the top row shows the coverage track from poly(A) mRNA-seq followed by the tracks for nucleosome occupancy obtained by light (10 U) and heavy (128 U) MNase digestion of *Paramecium* nuclei. Coverage tracks were visualized using the Integrative Genomics Viewer (IGV) browser (Robinson et al. 2011). (B) Profile plot of nucleosome distribution at the transcription start site (TSS; left) and transcription termination site (TTS; right) for genes >1 kb and digestion conditions as in A. The plot organization resembles the nucleosome profile along the gene body/intergenic region with dotted lines indicating excluded regions in the center of both plots. (C) Same plot as in B, but for intergenic regions >1 kb. Horizontal line is drawn to aid comparison between B and C. (D) Ranking of genes by their mRNA expression values from low to high (Q1–Q5) and (E) total sum occupancy for the genes in each expression quantile and the intergenic regions. Occupancy values are shown for mild and heavy digest side by side.

of MNase data from *Tetrahymena*, *S. pombe*, *Drosophila melanogaster*, but they are apparent in humans (Supplemental Fig. S3). As such, their presence in *Paramecium* is surprising and requires additional analysis. In addition, the comparison to other species shows that downstream nucleosomes (downstream from +1) in *Paramecium* are apparently much less pronounced; already, the +2 nucleosome signal is roughly background, which is in contrast to *Tetrahymena*, *S. pombe*, and *Drosophila* showing slightly decreasing peak values inside the gene bodies (Supplemental Fig. S3). The recent paper of Gnan et al. (2022) did not identify these putative –1 nucleosomes in *Paramecium*. This difference is not owing to the bioinformatics pipelines, because Supplemental Figure S3 shows still the absence of putative –1 nucleosomes when our MNase pipeline is applied on the data of Gnan et al. (2022). We therefore conclude that the difference is owing to the MNase conditions. We used formaldehyde-fixed material in contrast to fresh chromatin. It seems suitable that our MNase digests are weaker compared with the relatively harsh conditions on native chromatin. Lighter MNase digests can obtain signals of nucleosomes, which are otherwise hidden: For example, in *Tetrahymena*, light MNase digests indeed show a weak –1 signal, which was similar to our data Xiong et al. (2016). We added the MNase profiles of the latter data of *Tetrahymena*, analyzed with our MNase pipeline to Supplemental Figure S3. As a result, one indeed needs to take the MNase conditions into account. We cannot exclude that other MNase conditions applied to the analyses of yeast, flies, and human chromatin (Supplemental Fig. S3) could produce alternative patterns. In the following, we aimed to see whether the positioning of –1 nucleosomes could be owing to short intergenic regions. We therefore dissected the *Paramecium* genes owing to two parameters: intergenic distance and orientation of genes. We considered bidirectional promoter genes, in which the two start sites of both genes are adjacent (start–start [SS]), or unidirectional genes, in which one start site is paired with the end of the other gene (start–end [SE]) (Supplemental Fig. S4A). These two categories were additionally classified into four groups based on their intergenic distance. The number of genes in each category is given in Figure 4B. Figure 4C shows nucleosome positioning of

Drews et al.

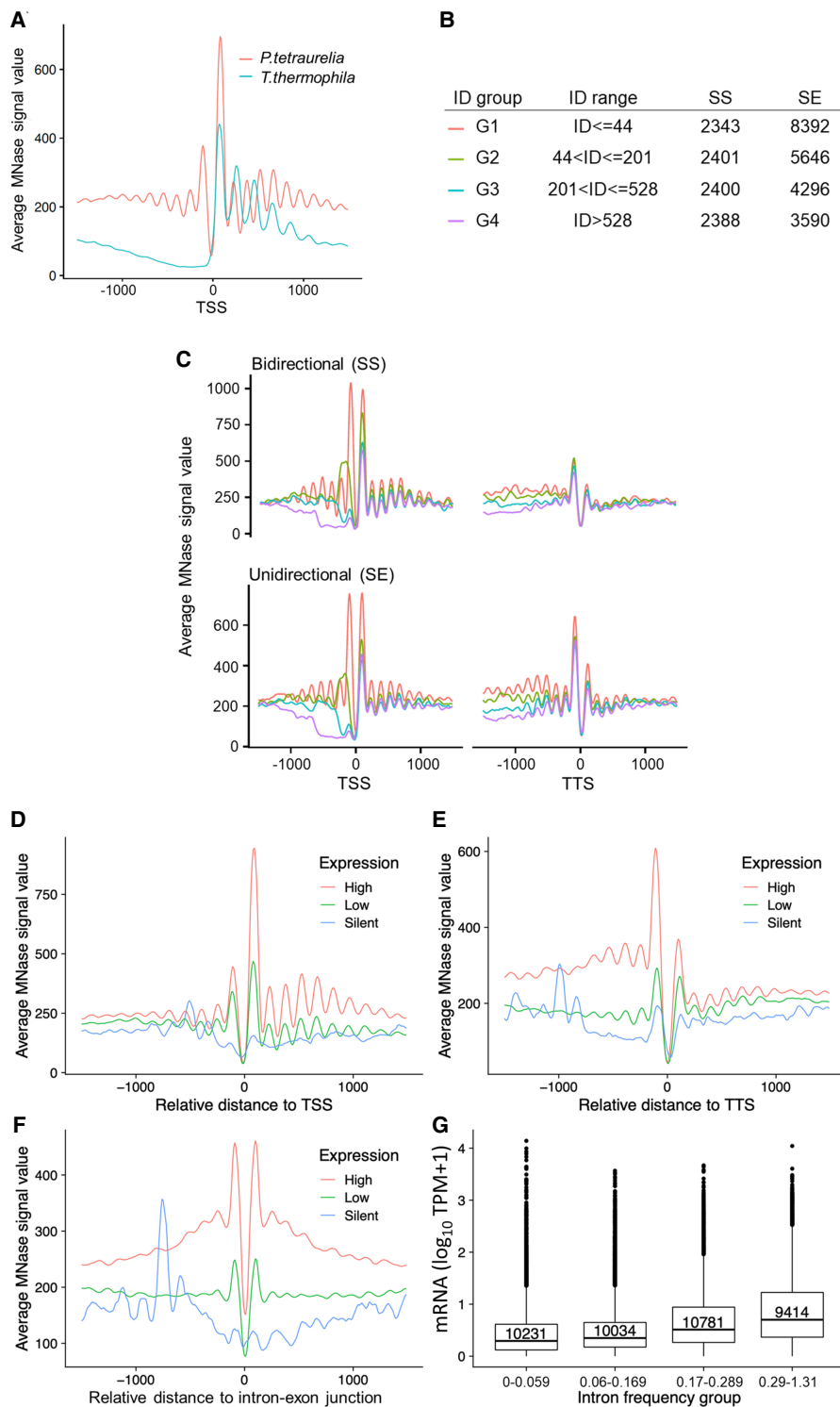


Figure 4. Positioning of nucleosomes in relation to gene expression. (A) Profile plot for nucleosome distribution relative to the transcription start site (TSS) for all analyzed *Paramecium* genes. Signal for 1000 bp upstream of and downstream from the TSS is shown. For comparison, MNase-seq data from *T. thermophila* were plotted in the same manner. (B) Dissection of neighboring *Paramecium* genes based on their configuration and intergenic distance (ID). Table shows separation of genes by configuration and ID, ranked from short distances (G1) to long distances (G4). The last two columns indicate numbers of genes in each configuration and ID group. (C) Nucleosome profiles in a 2-kb window centered at the TSS (left) or the TTS (right) for neighboring genes in SS and SE configuration are shown. Genes were additionally separated by the length of intergenic distances; see color-coding in B. The nucleosome profiles in relation to their distance (x-axis) to TSS (D), TTS (E), and intron-exon junction (F) are shown for gene categories based on their expression levels. (G) Box plots showing the mRNA expression (y-axis; $\log_{10} \text{TPM} + 1$) of genes with different intron frequency groups (number of introns per 100 bp; x-axis). A Kruskal-Wallis test showed that the expression distribution between all pairs of intron frequency groups is significantly different ($P < 2.2 \times 10^{-16}$).

these categories at the TSS and the TTS. Most apparent, putative -1 nucleosomes are much more pronounced in genes with short 5'-intergenic regions >142 bp, and this is true for the SE and the SS configuration. In addition, TTSs also show well-positioned nucleosomes at the ultimate 3'-end of ORFs, which are more pronounced in the SE configuration. The absence of -1 nucleosomes in genes with a longer intergenic region let us conclude that these putative -1 nucleosomes are either $+1$ or TTS nucleosomes of upstream genes, but no true -1 nucleosomes. They could, however, have a function in regulation of both genes, being "coincidental" -1 nucleosomes in point of view of our analysis.

We consequently asked for a potential coregulation of genes at bidirectional promoters. Correlation analysis of neighboring genes suggests a high degree of coregulation of all neighbor genes regardless of the configuration (Supplemental Fig. S4A,B). However, Supplemental Figure S4C shows that we cannot identify a higher degree of coregulation in genes under the same bidirectional promoter, suggesting that even short intergenic distances are sufficient to control regulation of gene expression independently of the neighbor gene. However, our data indicate that genes with bidirectional promoters tend to have a longer intergenic distance (Supplemental Fig. S4D), suggesting that selection pressure acts on these regions to separate bidirectional genes from each other. Gene length itself seems not to have a strong effect on TSS and TTS nucleosome positioning (Supplemental Fig. S5).

We sought to investigate whether nucleosome positioning is changed with differences in gene expression levels (Fig. 4D,E). At both ends of a gene, TSS and TTS, well-positioned nucleosomes can be found in highly expressed genes only. In contrast, these regions and also gene bodies of silent genes appear to be almost devoid of well-positioned nucleosomes.

We can detect well-positioned di-nucleosomes around introns (Fig. 4F). As mentioned, the 25-nt introns are among the shortest reported in eukaryotes (Russell et al. 1994). Intron splicing appears to result from efficient intron definition, rather than exon definition as in multicellular species, although only 3 nt define the 5'- and 3'-splice sites (Jaillon et al. 2008). Our data do not reveal any associations of intron nucleosomes with intron length (Supplemental Fig. S6A). As our MNase data suggest a general low occupancy of nucleosomes in gene bodies, intron-associated di-nucleosomes could be an exception to this. We correlated the intron frequency (number of introns per 100 bp) with gene expression levels (Fig. 4G) and found increasing mRNA levels with increasing intron frequency, an effect independent of the gene length (Supplemental Fig. S6B). Thus, introns in *Paramecium* may be involved in transcriptional regulation by recruitment of nucleosomes to gene bodies.

Broad histone mark domains in gene bodies

To extend the chromatin analysis to histone modifications, chromatin immunoprecipitation followed by sequencing (ChIP-seq) was performed from vegetative cells. We used the NEXSON procedure (Arrigoni et al. 2016) involving isolation of intact MACs without MICs. Another advantage of this procedure was that we were able to use the very same MAC preparations for both MNase- and ChIP-seq. We used antibodies for the activation-associated marks H3K9ac and H3K4me3, as well as an antibody for the repressive mark H3K27me3. It is necessary to note here that H3 variants have been described in *P. tetraurelia* (Lhuillier-Akakpo et al. 2016): Divergent and putative development-associated H3 variants cannot be detected with the antibodies used here; it is not likely

that these antibodies can dissect the five H3 variants expressed during vegetative growth, which means that ChIP should detect all of these variants, as well as the putative H3.3. The observed ChIP-seq signatures of these three marks showed rather broad signals, which were not comparable to sharp peaks of metazoan ChIP-seq signals. Thus, we refrained from a peak-calling approach and used ChromHMM (Ernst and Kellis 2012) to segment the entire MAC genome into 200-bp bins, representing approximately the resolution of a nucleosome including a spacer region, for de novo determination of reoccurring combinatorial and spatial signal patterns. We found that five different chromatin states (CSs) could be observed (trying to increase the number of states resulted in highly similar states, and we continued all further analyses with five states). Heatmaps in Figure 5A show the contribution of the individual signals to each CS and, on the right, the quantitative assignment of each CS to different regions of the genome. We abbreviate all five CSs as CS1 to CS5.

One major finding of the segmentation is represented in CS4. ChromHMM defines this state as being almost free of any signal; this state is moreover attributed to the highest percentage of the genome (Fig. 5A, right). This may support our previous assumption that a high amount of MAC DNA is free of nucleosomes and therefore also of transcription-altering histone marks. In contrast, MNase and histone mark signals can be found in CS1–CS3 and CS5. Their ChromHMM signature shows dynamic combinations between the three investigated histone marks, and the occurrence of these states also varies in different genomic areas. Focusing on histone marks around the TSS, CS1 and CS2, both enriched in H3K9ac and H3K4me3, show strong accumulation at the $+1$ nucleosome (Fig. 5B). All other CSs show depletion at $+1$, especially CS3, which suggests that especially H3K27me3 is depleted at these gene loci.

To go deeper into the role of the individual marks and states in association with gene expression, we dissected genes into categories overlapping with a CS (1) for $>80\%$ of the entire gene body, (2) with first 300 bp of the ORF, or (3) with 300 bp of the noncoding upstream region. We consequently correlated this with the gene expression level of these genes (Fig. 5C). Genes with high levels of H3K9ac and H3K4me3 (CS1) are highly expressed. Focusing to the role of H3K27me3, its high abundance in CS2, associated genes showing the highest expression level, is an argument against a repressive function of this histone mark. Only few genes (91) can be attributed to CS3, the only state in which the H3K27me3 signal dominates over H3K4me3 and H3K9ac; although the genes appear to be quite lowly expressed, the small number of genes does not allow for a conclusion about a possible repressive function of H3K27me3.

Genes associated with CS5 show low levels of H3K4me3 and H3K9ac with the absence of H3K27me3, and these genes show an intermediate gene expression level. CS4 shows the lowest gene expression level and, in agreement with the quantitative analysis, the highest number of genes. We conclude that gene silencing in the MAC is associated with genomic loci that consist predominantly of free and accessible DNA. Comparing the 80% gene overlap category to the upstream and the 5'-coding region, our analysis indicates that the upstream region contributes less to gene regulation. Mainly the 5'-CDS and the ORF appear to be involved in gene regulation, which fits to our conclusions from MNase data. We can therefore conclude that gene transcription is mainly associated with high levels of H3K9ac and H3K4me3 at the $+1$ nucleosome. We do not see direct evidence for a repressive function of H3K27me3. These results now raise several questions, especially

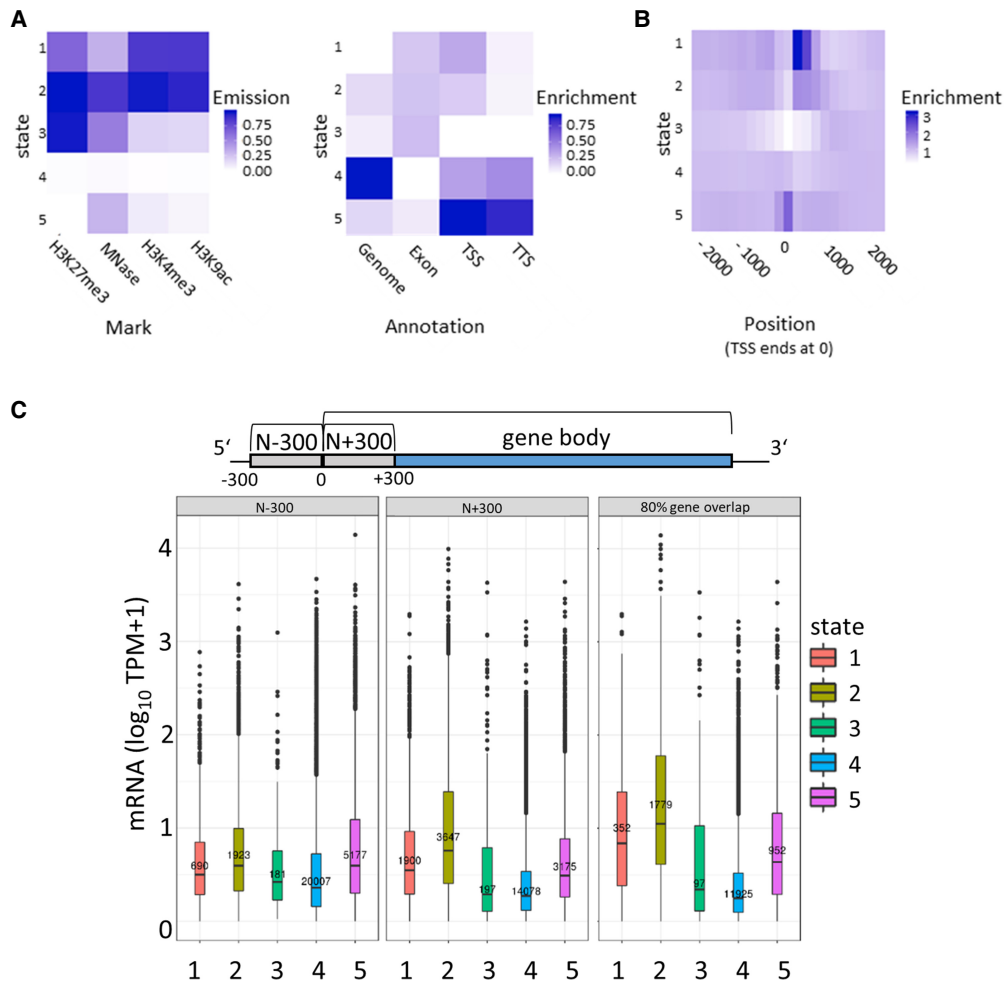


Figure 5. Segmentation analysis using ChromHMM. (A) The chromatin state (CS) assignments are shown as a heatmap of emission parameters from a five-state ChromHMM model (*left*). Each row corresponds to a ChromHMM state, and each column represents a different epigenetic mark. The darker the color of an epigenetic mark for a state, the higher the probability of observing that epigenetic mark in that state. Heatmap showing the overlap fold enrichment of each ChromHMM state (row) in different genomic annotations (columns; *right*). Enrichment values are obtained from the overlap enrichment functionality of ChromHMM with a column-specific color scale. (B) The fold enrichment of each state in 200-bp bins within a 2-kb window around the TSS is shown. Enrichment values are obtained from the neighborhood enrichment functionality of ChromHMM with a uniform color scale. (C) Box plots show mRNA expression (y -axis; $10 \text{ TPM} + 1$) of genes whose loci overlap at least 80% with a respective state (*right*). Additionally, genes were separated by their assigned state in 300 bp upstream of the TSS (N - 300) and the first 300 bp of the gene body (N + 300), and mRNA expression values of these genes are plotted (*left, middle*). Sketch on *top* of the plots visualizes the arrangement of the three analyzed regions.

about the role of the prominent +1 nucleosome in transcriptional activation: Could this be a place for RNA Polymerase II pausing in order to regulate gene expression?

Pol II occupancy correlates with gene expression levels

To characterize Pol II occupancy and activity, it is important to note that *Paramecium* Pol II diverges from conserved metazoan and most unicellular Pol II. In *Paramecium*, as well in *Tetrahymena*, the consensus serine-rich repeats are missing, but the CTD shows overall a high percentage of serines (Fig. 6A). As commercial Pol II antibodies target the heptamers in the CTD, we had to produce an antibody of our own against the *P. tetraurelia* CTD of RPB1. After affinity purification and specificity checks by IF and western blots of cellular fractions (Supplemental Fig. S7), ChIP was performed as described. Figure 6B shows high Pol II occupancy of genes showing

high expression and vice versa. Here, the analysis of all genes of the genome results in a quite equal distribution of Pol II along the ORF.

We consequently asked whether Pol II pausing at the +1 nucleosome can be observed, and we calculated a pausing index (PI) by dividing the Pol II coverage of the TSS by the coverage of the gene body (Fig. 6C). Dissecting paused and nonpaused genes by a threshold of PI larger than 1.5, we compared Pol II occupancy of *Paramecium* to other species. Figure 6D shows that *Paramecium* is the only species with similar occupancy of paused and nonpaused genes. The overall distribution of *Paramecium* Pol II is highly different to other species. In humans, *S. pombe*, and *Tetrahymena*, non-paused genes show increasing coverage along the ORF (for detailed heatmaps, see Supplemental Fig. S8A). This is different in *Paramecium*, in which non-paused genes show in general higher occupancy and less decrease along the ORF. Considering the huge differences in gene length distribution for the different species, we

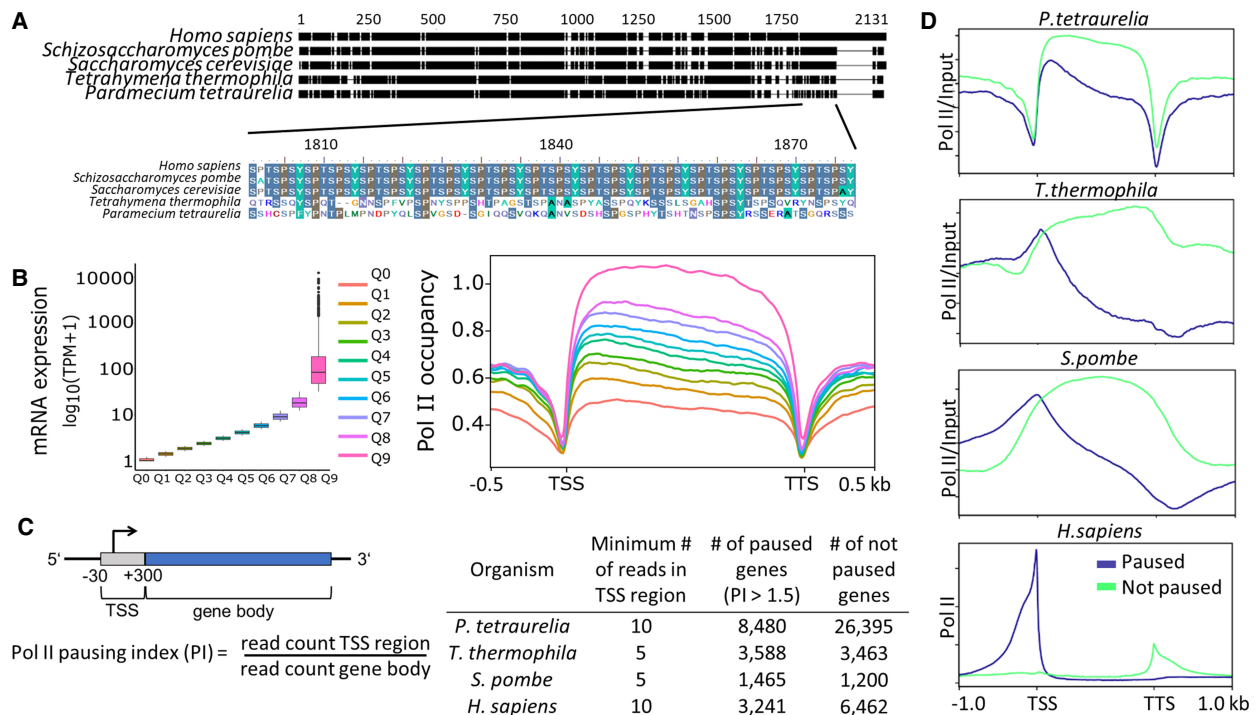


Figure 6. Analysis of RNA polymerase II pausing. (A) Multiple sequence alignment of the RNA polymerase II enzyme's RPB1 subunit in different organisms is shown. The C-terminal end of RPB1 is zoomed in to show the difference in conserved regions of some ciliates to other organisms. For details, see Supplemental Methods. (B, left) Box plots of gene expression (y-axis; \log_{10} TPM) split in 10 quantiles are shown; higher quantiles mean higher expression. (Right) Pol II enrichment (y-axis) profiles in respective quantiles are shown. Distance shown on the x-axis is scaled; that is, all genes (TSS–TTS) are either stretched or shrunk to a length of 1500 bp. A 500-bp window upstream of and downstream from the gene loci is included. Enrichment profiles were plotted using deepTools2. (C) A graphical representation of the regions included in polymerase pausing index (PI) calculation is shown. We categorized a gene as paused if the $PI \geq 1.5$. The table summarizes numbers of paused/not paused genes for selected organisms (Supplemental Table S1 contains details on Pol II data sets). (D) Same as the Pol II enrichment profiles in B, but genes are split based on the status of Pol II pausing.

additionally analyzed subsets of genes with approximately the same length (Supplemental Fig. S8B) and still observed the similar Pol II distribution as shown in Figure 6D. The pattern of *Paramecium* appears different to other species, suggesting that regulated pausing at the +1 nucleosome occurs only rarely. This is to some extent also true for *Tetrahymena* and yeast with the difference that paused genes here show a clearer peak at the TSS along with a strong decrease along the ORF. Such patterns cannot be identified in *Paramecium*. *Paramecium* in contrast shows a clear drop in Pol II occupancy before the TSS and at the TTS: This seems in agreement with our hypothesis that regulation of gene expression occurs mainly inside ORFs. We further analyzed whether pausing is associated with reduced full-length mRNA production. Supplemental Figure S8C shows that we see a significantly lower expression of paused genes in *Tetrahymena* and *S. pombe*; only in humans do paused genes show higher mRNA levels. Thus, Pol II pausing may indeed be a mechanism of gene regulation, but used in a different manner. Especially in *Paramecium*, the mRNA levels between paused and nonpaused genes show the smallest differences, although significant, suggesting that pausing is more involved in fine-tuning transcription rather than on/off switching.

H3K4me3 is the most important predictor of gene expression

Integrating all the data generated, we started by characterizing their distribution over all genes categorized by two factors, namely, gene expression and gene length. Figure 7A shows the input nor-

malized profiles of different epigenetic marks and GC content based on the gene expression groups. Genes in heatmaps are sorted by gene length. MNase, Pol II, H3K4me3, and H3K9ac show accumulation in the 5'-CDS in expressed genes with decreasing intensity along the ORF. However, most signals are still high and correlate to gene expression level in the 3'-CDS. The 5'-accumulation is not that pronounced in H3K27me3, which shows more equal distribution along the ORF. Hence, we further investigated how the epigenetic marks are distributed along the gene structure, based on their length. MNase signals show a strongly phased pattern in all categories of gene expression, which becomes apparent when genes are sorted by length. Supplemental Figure S10A shows a strong positive correlation of exon length and nucleosome counts in exons. Similarly, nucleosome occupancy is positively correlated with gene expression (Fig. 7A). Similar to the strongly phased signals of MNase, we observe that Pol II signals are also phased and show positive association with gene expression.

All epigenetic marks are consistently low at 5'- and 3'-non-coding regions, showing a clear gap in all analyses and thus fostering the assumption that intergenic regions hardly contribute to gene regulation. All silent genes have very faint signals of all epigenetic marks, supporting our conclusion that lowly occupied nearly naked DNA is a hallmark of gene inactivation in *Paramecium*.

The visualization in the heatmaps in Figure 7A reveals a phasing pattern for almost all marks, as genes are ordered by gene length in each expression group. This means that nucleosomes are indeed well positioned in all ORFs and along the entire length, but with

Drews et al.

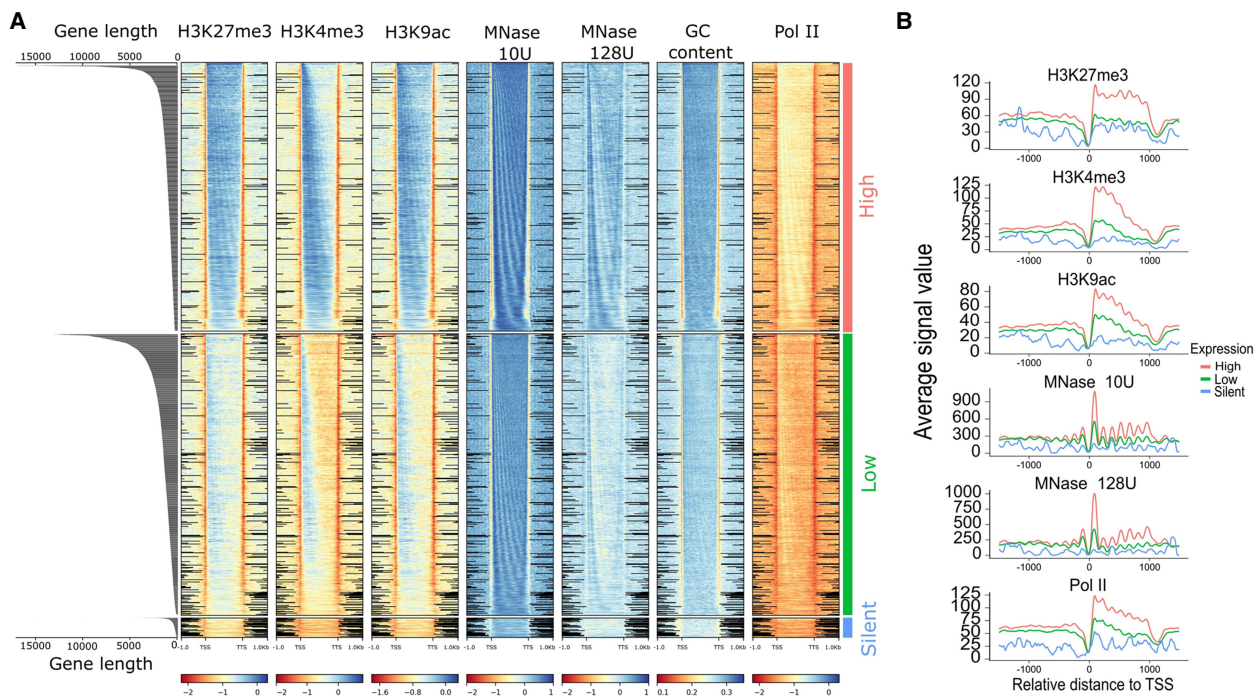


Figure 7. Distribution of epigenetic marks in different transcriptomic groups. Heatmaps show the input normalized enrichment values for different epigenetic marks. Genes (rows) are split into three categories based on gene expression—high (TPM > 2), low (0 < TPM < 2), and silent (TPM = 0)—and are sorted by decreasing order of gene length in each, which is visualized by the length distribution graph on the left. Distance shown on the x-axis is scaled; that is, all genes (TSS–TTS) are either stretched or shrunken to a length of 1500 bp, adding 1000 bp upstream of and downstream from the gene. Heatmaps were plotted using deepTools2; black lines in intergenic regions reflect missing data at this position. (B) Distribution of epigenetic marks for a subset of 4000 genes with discrete length of ~1.2 kb. Plots show the signal in the upstream intergenic region, the TSS and the TTS of genes belonging to the similar expression categories as in B.

varying intensity, owing to differences in gene expression. As one will have assumed then that the histone marks need to follow the nucleosome pattern, this follows also the GC content oscillations in position and quantity. As such, this *cis*-factor likely contributes nucleosome positioning and, consequently, gene expression. We investigated the effects of gene length and mRNA levels and observed that shorter genes show higher mRNA levels (Supplemental Fig. S9), and as such, gene length itself appears to be a factor limiting transcriptional efficiency. We observe the phasing pattern also for Pol II occupancy. This would suggest that Pol II shows association with nucleosomes along the entire ORF, and the higher Pol II occupancy in highly expressed genes does not indicate that this association is a mechanism of transcriptional inhibition. In agreement with the conclusion from the PI analyses, this Pol II nucleosome association appears to be a mark of highly expressed genes, although one could get the impression that Pol II stops at every single nucleosome, which could also be an argument for inefficient elongation. Figure 7B shows the signals of the epigenetic marks in a subset of genes with similar gene length (~1200 kb), thus avoiding the projection of small and large genes. As we observed some intriguing patterns of histone marks, especially of H3K27me3, which is abundant in highly expressed genes, we checked the correlation of all epigenetic marks with each other with mRNA (Supplemental Fig. S10A). We observed that all epigenetic marks are positively correlated (Pearson's correlation > 0.6) with each other, and with mRNA (Pearson's correlation > 0.30). We wondered about the individual contribution of gene characteristics and epigenomic marks to gene expression. Thus, we construct-

ed a machine learning classifier to predict genes as highly or lowly expressed using epigenetic features and genic features (see Methods). Our model is based on a random forest algorithm, which accurately predicts gene expression with an average area under the precision-recall curve (PR-AUC) of 0.74 and 0.76 for genic or epigenetic features, respectively. The model combining all information performed best (PR-AUC of 0.82) (Fig. 8A). These differences were statistically significant (Supplemental Fig. S10B). The experiments in Figure 8A were performed using histone marks in the complete gene body. When quantification is restricted to the proximal TSS region (TSS + 300 bp), performance decreased (Supplemental Fig. S10C), supporting a role of those marks throughout the gene body.

Further, we interrogated the best-performing model on the importance of each feature in obtaining the classification (Fig. 8B). According to the feature importance values calculated on our best-performing model, H3K4me3, intron frequency, and gene length are the top three features required to classify gene expression. Intergenic length and H3K27me3 are among the least important features for our model. The presence of H3K27me3 in the whole gene body, with its high correlation to other histone marks and highly expressed genes, does raise the question of the role of H3K27me3 in MAC nucleosomes of *Paramecium*.

H3K4me3 and H3K27me3 co-occur at plastic genes

We consequently asked for the contribution of individual features to gene regulation. We used RNA-seq data from environmental states that include four different serotypes at different

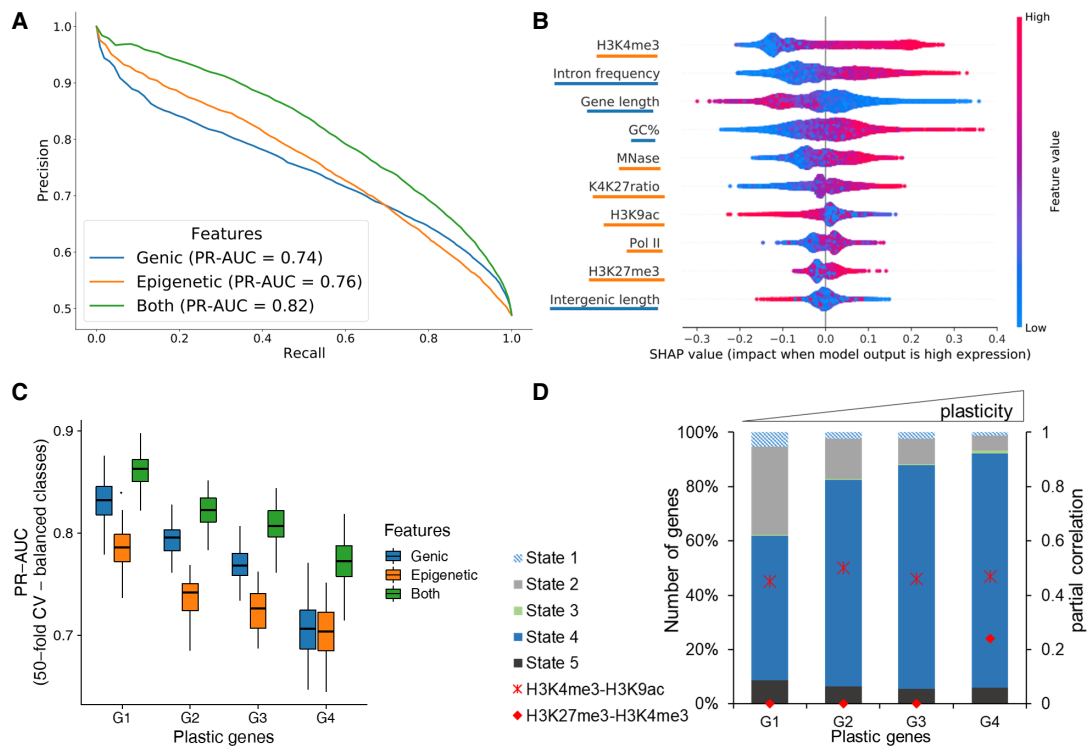


Figure 8. Prediction of gene expression by epigenetic marks and genomic features. (A) Results of classifying low and high gene groups using different data (genic: related to gene structure; epigenetic: using abundance of histone marks and MNase; both: genic and epigenetic). PrecisionRecall curve with average values from a 40-fold cross-validation with random forests indicating features by different colors. (B) Analysis of feature importance using both genic and epigenetic features (underlining color indicates type on y-axis; see legend for A). Features are listed in decreasing order of classification importance from top to bottom. The importance (SHAP value; x-axis) of a feature for each gene illustrates its contribution to classification as high or low, with positive and negative SHAP values, respectively. The color gradient depicts the feature value in scale from low to high, for example, the length of a gene (third row). For example, long genes strongly contribute to the prediction of lowly expressed genes. (C) Genes were separated into four groups by their plasticity, which is defined by a large variation in gene expression among different conditions. The box plot shows the distribution of classifier performance values for genes with different plasticity (50-fold CV-based PR-AUC) for the same three feature sets as A. The number of genes in each plastic gene group was randomly subsampled to have equal number of genes in high and low expressed category. (D) Distribution of CSs among plastic gene groups. We only included genes with a ChromHMM state overlap of at least 80% (see Fig. 5). Additionally, partial correlation values for H3K4me3-H3K9ac (cross) and H3K4me3-H3K27me3 (circle) are red for each group.

temperatures, starvation, heat shock, and cultivation at 4°C (Cheaib et al. 2015). Using those data, we dissected genes showing large expression variations (high plasticity) during the vegetative growth in different environments to identify dynamically regulated genes from housekeeping genes (see Methods) (Supplemental Fig. S11). We defined four classes of plasticity (G1–G4), in which G4 genes showed the largest variation. We again used the random forest algorithm to analyze whether genic/epigenetic factors contribute to the accuracy of gene expression prediction for each gene plasticity group. The performance of expression prediction decreased for genes with higher plasticity (Fig. 8C). Thus, plasticity of gene expression seems to be accompanied with additional and unknown features contributing to gene regulation.

To get further insights, we checked the CSs based on our ChromHMM segmentation of the four categories of plastic genes (Fig. 8D). These show gradual differences, with most apparent increase that of CS4 and decrease that of CS2. This suggests that epigenetic marks are used not only for control of gene expression but moreover for gene regulation. We studied the differences of histone marks of these categories in more detail and calculated the partial correlation between different modifications (see

Methods). Figure 8D shows an increase in partial correlation of H3K4me3/H3K27me3 for the most plastic genes only, suggesting that the interplay between histone marks varies in the four considered groups.

Discussion

Genomic and epigenomic paradoxes

At first glance, the genomic structure of the *Paramecium* MAC seems paradoxical. Although *Paramecium* is extremely gene-rich, with approximately 40,000 genes (Aury et al. 2006), the size limitations of intergenic regions and introns provide only restricted capacity for differential gene regulation. This is different compared with genomic/epigenetic features in metazoans, because unicellular organisms do not need to differentiate into distinct tissues with all the known epigenetic manifestations to guarantee for cell type-specific gene expression patterns. However, the *Paramecium* epigenome still needs to manage dynamic regulation of gene expression and proper transcription of mRNA. We know that histone marks do not just control condensation and

transcriptional on/off switches but interact with capping enzymes, splicing factors, and elongation factors to guarantee mature mRNA synthesis (Jimeno-González and Reyes 2016).

Thus, we aimed to answer the question in which manner the MAC epigenome signature is associated with transcriptional regulation in this ciliate. Its nucleosomes appear to hold some secrets as recent results show that the nucleosome repeat length is only ~151 bp, which means that the linker DNA between nucleosomes is only few base pairs long (Gnan et al. 2022). Our data show that nucleosome occupancy appears to be associated in general with active transcription, because segmentation of MNase and ChIP data shows a large number of genes where our setup detects only low or no signals (CS4 in Fig. 5). Correlation of this CS with gene expression indicates that low nucleosome occupancy, regardless of the histone marks, is associated with silent or lowly expressed genes. One could therefore interpret naked or lowly occupied DNA as a default state, which needs to be occupied with nucleosomes first to become transcribed into mRNA. As such, the epigenome of *Paramecium* appears paradoxical as well, as gene inactivation becomes realized by low nucleosome occupancy, and this is contrary to the classical models.

Textbooks describe gene inactivation by a hierarchical chromatin folding from open 10-nm fibers to condensed and higher occupied 30-nm filaments. Active transcription accompanied by open, accessible chromatin in mammals was highly supported in the last years by many studies of DNA accessibility using ATAC, NOME, DNase-seq, or methods free of enzymatic steps like sedimentation velocity centrifugation (Klemm et al. 2019; Nordström et al. 2019; Ishihara et al. 2021). Our data do not support this model for *Paramecium* MAC chromatin, suggesting a different chromatin-associated mechanism of gene inactivation. This raises many more questions about how, in particular, spurious and aberrant transcription of Pol II in open regions is inhibited or whether this could be tolerated to some extent. In most species, condensation of chromatin is accompanied with linker histone H1 recruitment and studies on *Drosophila* chromatin show H1 occurring exclusively at closed heterochromatic loci (Nalabothula et al. 2014). We are not able to identify a MAC histone H1 variant in *Paramecium*, supporting the idea of condensation-free gene inactivation. To be precise, we have to distinguish MAC and MIC linker histones in ciliates. *Tetrahymena* has distinct MAC- and MIC-specific H1 histones, where the MAC version (Hho1) is non-essential (Schulman et al. 1987). *HHO1* knockouts show an overall decondensation of MAC chromatin (Huang et al. 1999). Indeed, a Hho1 homolog is not present in *Paramecium*, or it may be more divergent to identify. However, the recent findings of Gnan et al. (2022) showed an extremely short linker DNA length between *Paramecium* nucleosomes compared with other species, and the investigators speculate that this could correlate with the absence of a canonical H1 ortholog.

Bistable H3K4/K27me3 as a mark of poised genes?

Another question we followed is whether the H3K27me3 could be involved in gene inactivation. Our ChIP data do not suggest that H3K27me3 is associated exclusively with silent or lowly expressed genes. When we asked for the function of this modification in the vegetative MAC, its role is unlikely the condensation of chromatin, and the segmentation shows H3K27me3 co-occurring in varying ratios with the H3K9ac and H3K4me3. Our data suggest that genes with high regulation dynamics show an increasing correlation for H3K27me3 and H3K4me3. This is one of the best-studied bivalent

domains for poised chromatin, where chromatin is placed into a waiting state for future activation, and this was described to occur particularly in embryonic stem cells (Pan et al. 2007; Zhao et al. 2007). There is an ongoing debate whether poised chromatin is bistable or bivalent, the latter representing a background population of fragments with active and silent marks, whereas bistability means the frequent switching between monostable active and silent states (Sneppen and Ringrose 2019). The polyploidy of the *Paramecium* MAC introduces here an additional layer of complexity. Similar to ChIPs of different cell states from a culture of metazoan cell cultures, which cannot dissect different cell states of a mixture from a real bivalent domain, we cannot be sure that the 800 copies of a gene in the MAC are coregulated.

If *Paramecium*, for instance, would use gene dosage to regulate gene expression level, one would expect different ratios of marks: some copies silent, some copies active. This is what we can observe to some extent, because the random forest analysis suggests that the K4/K27me3 ratio explains gene expression levels better than the H3K27me3 alone. In a previous study, increased H3K27me3 levels in association with decreased levels of H3K4me3 at an endogenous reporter gene have been shown to go along with siRNA mediated silencing (Götz et al. 2016), which supports the K4/K27me3 ratio hypothesis for controlling gene expression levels. In addition, the finding that we see increasing partial correlation values of K4/K27me3 in genes that show high regulation dynamics could be called poised as such. This suggests that the bivalency of K4/K27me3 in chromatin poising could be an ancient and general mechanism rather than an invention of metazoans.

In *Paramecium*, the Polycomb group methyltransferase Ez11 was shown to mediate both H3K9me3 and H3K27me3 during development: Loss of these marks is accompanied by loss of transposon repression and elimination and, in addition, a transcriptional up-regulation of early developmental genes (Frapporti et al. 2019). As Ez11 shows also low expression during vegetative growth, it remains to be elaborated whether Ez11 or another SET-domain-containing enzyme catalyzes the replicative maintenance of H3K27me3 during vegetative cell divisions. In addition, it remains to note that a putative repressive function of H3K27me3 could, in principle, be blocked by a phospho-switch by a neighboring serine-residue as this was initially shown for loss of binding of HP1 to H3K9me9 in context with H3K10 serine phosphorylation (Fischle et al. 2005). However, this is unlikely for *Paramecium* H3K27me3 as all H3 variants miss the conserved serine 28 in *Paramecium* (Supplemental Fig. S1; Lhuillier-Akakpo et al. 2016).

From an evolutionary point of view, this could imply that although *Paramecium* is unicellular, the epigenomic repertoire already has the capacity to manifest vegetative gene expression regulation during development, meaning to place histone marks for poising genes. The inheritance of gene expression pattern was previously shown also for the multigene family of surface antigen genes as transcription of a single gene follows the expression pattern of its cytoplasmic parent (Baranasic et al. 2014; Simon and Plattner 2014), but we would need to analyze the genome-wide extent of such an inheritance and/or whether such a mechanism is coupled with other genomic parameters like, for instance, subtelomeric localization of the respective genes.

ChIP-seq reveals broad domains instead of narrow peaks

When looking for the distribution of marks along genes, the absence of narrow peaks becomes apparent as all histone mark distributions are more comparable to broad domains instead of local

and narrow peaks, which explains the failure of peak calling. Broad domains were also found in mammals. For instance, H3K27me3 was shown in mammalian chromatin to be distributed along ORFs (Zhou et al. 2011). Also in mammals, broad H3K4me3 was shown for tumor-suppressor genes with exceptionally high expression, where this mark has also been attributed to transcriptional elongation (Chen et al. 2015). In addition to tumor-suppressors, broad H3K4me3 domains have been implicated with genes for cellular identity and transcriptional consistency; as the broadest domains show increased Pol II pausing, the investigators suggest the broad mark as a buffer domain to ensure the robustness of the transcriptional output (Benayoun et al. 2014). This model could also fit to our observations, which suggest not only that H3K4me3 is the key regulator of transcription but that H3K4me3 appears in broad domains along ORFs highly covered with Pol II. Concerning the different patterns of Pol II along ORFs compared with other species, either for poised or nonpoised genes, the buffer domain model could hold true for the majority of *Paramecium* genes.

Nucleosome positioning and GC content

Paramecium has an exceptional genome composition with an average GC content of 28%, including the even more AT-rich intergenic regions. It is known that GC content favors nucleosome positioning (Tillo and Hughes 2009). Our data show that nucleosome occupancy is mostly restricted to ORFs, which would correlate to increased GC levels but also correlated to gene expression levels as higher expressed genes show higher occupancy of promoter proximal- and intron-associated nucleosomes. It is difficult to reason how much the sequence content of the *Paramecium* genome itself encodes the deposition of nucleosomes from our data. There is ample discussion about the DNA sequence preferences of nucleosomes (Meyer and Liu 2014), and also MNase-seq can generate a signature of higher occupancy at GC-rich regions on naked as well as occupied DNA (Chung et al. 2011). One may conclude that this bias explains the large drop of MNase-seq read occupancy at intergenic regions. However, analysis of ChIP-seq data shows a similar drop at intergenic regions and similar phasing patterns in our data, and Supplemental Figure S12 suggests that our procedure and the applied PCR amplification have minimized GC biases. We argue that it is unlikely to observe these trends exclusively owing to methodological biases in AT content.

Our results of nucleosome positioning fit to observations in *Tetrahymena*, where well-positioned nucleosomes in the MAC match GC oscillations but are also affected by *trans*-factors, for example, the transcriptional landscape (Xiong et al. 2016). In addition, studies in *Tetrahymena* revealed that N_6 -methyladenine (6mA) is preferentially found at the AT-rich linker DNA of well-positioned nucleosomes of Pol II transcribed genes (Wang et al. 2017; Luo et al. 2018). Also, in *Paramecium*, 6mA sites enriched between well-positioned nucleosomes are positively correlated with gene expression (Hardy et al. 2021). The latter finding would fit our observations: The more nucleosomes, the more 6mA, the more transcription.

Qualitative aspects of gene expression

To understand the relation between epigenomic data and gene expression, throughout this study we categorized genes based on their expression levels (high, low, silent). Although this categorization helps, it should be treated with a grain of salt as the cut-offs are rather arbitrary. Another aspect that requires cautious interpreta-

tion is the analyses presented in Figure 7. Specifically, Figure 7A shows the linear relation between epigenetic signals and mRNA expression in a qualitative manner. The random forests analysis, presented in Figure 7, B and C, reveals both the linear and nonlinear relationships inherent in the epigenetic data while calculating the probabilities to predict/classify a gene as highly or lowly expressed. For example, we can observe that H3K9ac is directly proportional to the different expression groups in Figure 7A. However, Figure 7C suggests genes with low H3K9ac are associated with high expression. Although this may seem counter-intuitive, both results are correct owing to the high colinearity of epigenetic marks (Supplemental Fig. S10). Hence, the random forests model relies on the H3K9ac signal only when the H3K4me3 signal is not sufficient to increase the probability of predicting a gene as highly expressed.

A divergent mechanism of transcriptional elongation

How can the highly regulated CTD phosphorylation and interaction with the different RNA modification and elongation complexes of metazoans be compared to our data? *Paramecium* Pol II does not show the serine-rich heptamer repeats. Thus, it would be surprising if a regulated and patterned phosphorylation of individual serines would be possible. As the *Paramecium* CTD is still rich in serines, although not organized in a repeat structure, it still seems likely that phosphorylation could be an activating mark. This needs to be discovered, and we need to note here that our polyclonal serum against one peptide, including unphosphorylated serines, could miss CTD variants being phosphorylated. An argument against this would be that we can detect Pol II, for example, in the center and 3'-regions of genes, where most serines are phosphorylated in mammalian CTDs. It seems quite tempting to speculate that Pol II of *Paramecium* does not need to be that highly regulated compared with mammals. First of all, alternative splicing is extremely limited; no single example of exon skipping has been reported (Jaillon et al. 2008); and, therefore, the well-positioned nucleosomes do not need to control this. In addition, the data of Gnan et al. (2022) support the idea that the GC content, not nucleosome positioning, contributes to splice efficiency. Introns are recognized by intron definition, and even artificially introduced introns in GFP are efficiently spliced (Jaillon et al. 2008). Our data suggest that introns serve in nucleosome positioning that may permit more intron accumulation in genes, increasing transcription. This would be supported by our data showing that genes with higher intron frequency show higher transcript levels.

Concerning the issues of pausing and elongation, our data suggest pausing to occur, but the pattern is different to other species because we find high levels of Pol II associated with nucleosomes along the entire ORF not only restricted to +1 nucleosomes. Given the fact that +1 nucleosomes are quite prominent, the question raises whether the stops of Pol II at +1 nucleosomes are mechanistically different from stops at all nucleosomes inside the ORF or whether this is a general phenomenon of *Paramecium* Pol II to stop at nucleosomes, maybe by less efficient elongation. For instance, the tiny introns of *Paramecium* do not contribute to a significant enlargement of transcriptional units compared with other species with introns, which are often much larger than the exons. It is therefore the question whether Pol II elongation has the need to be highly supported. *Paramecium* and *Tetrahymena* miss homologs of NELF, and two recent studies showed the mediator complex, a key regulator of Pol II interaction

with transcription and elongation factors, to be highly divergent in *Tetrahymena* (Garg et al. 2019; Tian et al. 2019). Additionally, in *Paramecium* we cannot identify all components of the Paf complex regulating elongation, 3'-end processing, and histone modification (Jaehning 2010). Especially, the subunit Paf1, involved in serine phosphorylation of the CTD of Pol II, is missing, which fits to the missing serine repeats of the CTD. Because of the lack of canonical elongation systems going along with a lack of conserved serine residues, we conclude that transcriptional elongation in *Paramecium* is regulated differently. As discussed above, broad H3K4me3 going along with increased occupancy of Pol II in ORFs might be an alternative control of transcription by buffer domains. It seems tempting to speculate this strange form of Pol II buffering represents an alternative or maybe an ancient form of elongation control.

This is the first description of the *Paramecium* vegetative chromatin landscape, which appears to be quite different to that of other unicellular eukaryotes and multicellular species. Broad domains along the gene bodies regulate transcription, whereas the noncoding and nonexpressed regions are devoid of epigenetic information. Paradoxically, our data also indicate silent genes to be devoid of epigenetic information, and it has to be clarified if and how the cell prevents spurious Pol II activity at these unoccupied regions. The Pol II distribution we observe is also quite different to other species; the process of transcriptional initiation and elongation appears to be controlled without sophisticated control of CTD phosphorylation and canonical complexes, like NELF, Paf, and Mediator, that assist Pol II in generating mature mRNA. However, this work here attributes to the vegetative nucleus only. We have to keep in mind that the transcriptional machinery needs to switch its mode of action to lncRNA transcription from the meiotic micronuclei during development. As such, functional and temporal dynamics require more alterations of the polymerase complex than in other species. There are plenty of challenges left, especially about the control of Pol II without or with limited CTD phosphorylation. Our study shows the unusual pattern of Pol II in expressed genes and in the light of so many missing interaction partners of Pol II; it is not a surprise that the epigenome looks different from other species in addition to the fact that no mitotic condensation is necessary in the MAC. Concerning Pol II interaction complexes, future studies will need to show whether some components are absent or whether they are too divergent such that reverse genetics cannot identify them. Their identification and contribution to Pol II activity and modulation will shed light on the mechanisms controlling mRNA and lncRNA transcription and the epigenetic marks in support of them. The comparison of the divergent mRNA transcription in *Paramecium* might unravel new basic principles of how, for example, a gene can be silenced in absence of repressive marks, and these principles might be applicable to understand the regulation of individual genes in other species.

Methods

Cell culture and RNA isolation

P. tetraurelia cells (strain 51) of serotype A were cultured as described before using *Klebsiella planticola* for regular food in wheat grass powder (WGP) (Simon et al. 2006). All cultures for this study were grown at 31°C. To ensure the vegetative state of the MAC, cells were stained with DAPI.

Genomic annotations

The genomic features shown in Figure 2B are captured from the annotations of the respective organisms, namely, from *Paramecium*DB (strain 51, version 2), *Tetrahymena* Genome Database (version 2014) (Stover et al. 2006), PomBase (version 2020) (The Gene Ontology Consortium 2019), and the ensemble database for *D. melanogaster* (release 98), and *Homo sapiens* (release 100) (Yates et al. 2020).

Antibodies

ChIP-seq-grade antibodies directed against histone modifications were purchased from Diagenode: H3K9ac C15410004, H3K27me3 C15410195, and H3K4me3 C15410003. For the antibody against *P. tetraurelia* RPB1, the peptide SPHYTSHNTN SPSPSYRSS-C was used for immunization. Purification and testing of specificity by western blots and immunostaining were performed as described recently (Drews et al. 2021). Because there are some amino acid differences in the N-terminal tail of the *Paramecium* H3P1 to Human H3 (Supplemental Fig. S1A), the peptide PtH3K27me3 TKAARK(me3)TAPAVG was synthesized, and binding affinity of the purchased H3K27me3 antibody to the PtH3K27me3 peptide was verified by dot-blots and competition assays. For details, see Supplemental Methods.

Fixation of cells

Isolation of intact MACs from fixed cells was performed using an adapted NEXSON protocol (Arrigoni et al. 2016). Two to 3 million cells were washed twice in Volvic and starved for 20 min at 31°C. After harvesting (2500 rpm, 2 min), the cell pellet without remaining media was resuspended in 2 mL fixative solution (20 mM Tris-HCL at pH 8, 0.5 mM EGTA, 1 mM EDTA, 10 mM NaCl, 1% methanol-free formaldehyde). After incubation (15 min, room temperature), the reaction was quenched by adding glycine to a final concentration of 125 mM. Cells were centrifuged (3300g, 3 min, 4°C), and the supernatant was discarded. The pellet was washed once in ice-cold PBS buffer and once in PBS buffer supplemented with cOmplete protease inhibitor cocktail, EDTA-free (PIC; Roche 11873580001). Cell suspension was split in half and centrifuged (3300g, 5 min, 4°C), and cell pellets were flash-frozen in liquid nitrogen.

MNase-seq

One aliquot was thawed on ice, resuspended in 2 mL Farnham laboratory buffer (5 mM PIPES at pH 8, 85 mM KCl, 0.5% NP-40), and evenly split into precooled 1.5-mL Bioruptor tubes (Diagenode). After sonication (15 sec on/30 sec off, five cycles, 4°C) using Bioruptor 300 (Diagenode) 5 µL was stained with DAPI to verify isolation of intact MACs. Cell suspension was centrifuged twice (3000g, 5 min, 4°C) with washing of the pellet in Farnham laboratory buffer in between. The following isolation of DNA covered by mononucleosomes was isolated as described previously (Xiong et al. 2016). One aliquot of isolated nuclei was resuspended in 1× MNase buffer (50 mM Tris-HCL at pH 8.0, 5 mM CaCl₂) and split into portions of 20,000 nuclei per reaction. After centrifugation (3000g, 5 min, 4°C) nuclei pellets were resuspended in 500 µL MNase reaction buffer (50 mM Tris-HCL at pH 8.0, 5 mM CaCl₂, 10 mM β-mercaptoethanol, 1% NP-40, 500 ng BSA). To each reaction, 10 or 128 U of MNase (NEB M0247S) was added, and after incubation (10 min, 37°C, 450 rpm), the reaction was stopped (10 mM EGTA, 1 mM EDTA, 5 min, 450 rpm). DNA corresponding to the size of mononucleosomes (100–200 bp) was isolated from a 3% agarose gel using a MinElute gel extraction kit (Qiagen 28604).

As input, nuclei were treated with Proteinase K, extracted as described, and treated with 0.1 U or 1.5 U MNase (5 min, 28°C) and extracted again. DNA library preparation was performed using NEBNext Ultra DNA library prep kit for Illumina (NEB E7370) with 10 ng input, 11 PCR cycles, and KAPA Taq HotStart DNA polymerase (Kapa Biosystems KK1512). The MNase-seq read count correlation of four independent replicates, each, used for subsequent analyses can be found in Supplemental Figure S13.

ChIP-seq

Nuclei pellets were resuspended in shearing buffer (10 mM Tris-HCl at pH 8, 0.1% SDS, 1 mM EDTA) and transferred in fresh, pre-cooled Bioruptor tubes. The suspension was sonicated (30 sec on/30 sec off, five cycles, 4°C). After centrifugation (16,000g, 10 min, 4°C), the supernatant was aliquoted in 100- μ L portions and stored at -80°C . To control shearing efficiency, 50 μ L was decrosslinked using Proteinase K (20 mg/mL), followed by phenol/chloroform/isoamylalcohol extraction, which was repeated after RNase A (10 mg/mL) digestion. Aliquots of 2 μ g were run on a 1.5% agarose gel. Eight micrograms of adequately sheared chromatin was subjected to immunoprecipitation using an iDeal ChIP-seq kit for histones (Diagenode C01010050) with 2 μ g of antibodies against histone modifications or 10 μ g of custom RPB1 antibody. Input was generated by putting 1 μ L of chromatin aside without mixing to antibodies. After overnight IP and elution from the magnetic beads, precipitated chromatin and the input kept aside were decrosslinked, RNase A-treated, and extracted as described above. DNA library preparation was performed using a NEBNext Ultra DNA Library Prep Kit for Illumina for serine-rich heptad repeats (NEB E7370) with 10 ng input, 11 PCR cycles, and KAPA Taq HotStart DNA polymerase (Kapa Biosystems KK1512). Precipitated DNA and input DNA were equally handled. ChIP-seq read count correlation of four independent replicates of H3K4me3, H3K27me3, and H3K9ac IP each, used for subsequent analyses, can be found in Supplemental Figure S14.

Sequencing and preprocessing

DNA libraries resulting from MNase digestion and ChIP were sequenced on an Illumina HiSeq 2500 in high-output run mode, and reads were adapter and quality trimmed. For details, see Supplemental Methods. All MNase, Pol II, and histone ChIP-seq reads were aligned to the MAC genome *P. tetraurelia* (strain 51, version 2) (Arnaiz et al. 2012) after quality control. For details, see Supplemental Methods. We used deepTools2 (Ramírez et al. 2016) to investigate the quality of replicates (*multiBamSummary*, *plotFingerprint*, and *plotCorrelation* tools) with subsequent down-sampling of some histone ChIP replicates, which had rather high coverage (see Supplemental Table S1). We used the DANPOS2 (Chen et al. 2013) software for position or peak calling with default parameters. We used the *dpos* functionality to call the positions of MNase and Pol II peaks and the *dpeak* functionality for histone ChIP peak calling. MNase-seq data were normalized to naked DNA inputs, whereas ChIP-seq data were normalized to the respective input files listed in Supplemental Table S1. Further, we made use of the *profile* functionality of DANPOS2 to visualize how a chromatin feature is distributed in a genomic annotation of interest (see Figs. 3, 4).

Segmentation analysis of chromatin marks

We used ChromHMM (Ernst and Kellis 2012) to perform genome-wide segmentation using the histone marks (H3K27me3, H3K4me3, H3K9ac) and MNase data. The genome was binarized into 200-bp bins based on a Poisson background model using

the *BinarizeBam* function. This was used to learn a CS model with five states using the *LearnModel* function. We used the *plotProfile* and *plotHeatmap* functionality of deepTools2 to create scaled enrichment plots of different chromatin features.

Gene expression and intron data

We used the mRNA expression data of strain 51 wild-type serotype A from our previous work (European Nucleotide Archive [ENA, <http://www.ebi.ac.uk/ena>] accession number PRJEB9464) (Cheaib et al. 2015). We quantified the expression using Salmon (v0.8.2) (Patro et al. 2017) default parameters for all replicates and used the mean of replicates in all downstream analyses. We used the transcript annotation from the MAC genome of *P. tetraurelia* (version 2; strain 51) (Arnaiz et al. 2017). For intron profiles, we created a 20-bp window centered on the first and last intron base of the 5'-exon-intron junction and the 3'-intron-exon junction. We plotted the nucleosome profile for 1500 bp around this window with the center of *x*-axis representing the junctions (see Fig. 4F).

Comparative Pol II analysis and PI

We used the data sets mentioned in Supplemental Table S1 for the comparative Pol II analysis of different organisms shown in Figure 6. We calculated the PI, after applying a threshold on the number of reads in the TSS region of genes (see Supplemental Fig. S8), depending on the distribution of read counts of individual data sets. The thresholds are mentioned in Figure 6C. mRNA quantification was performed using the default parameters of Salmon with transcripts obtained from the respective genomic annotations mentioned above (mean of replicates). We defined a region starting at 30 bp upstream of the TSS until 300 bp downstream from the TSS as the *TSS region*, and a region starting at 300 bp downstream from the TSS until the TTS as the *gene body*. The PI is calculated as a ratio of reads (in TPM) in the TSS region compared with reads in the gene body. Genes with a PI greater than 1.5 were considered as paused.

Classification of gene expression using random forests

After removing 1369 silent genes ($TPM=0$), we split the remaining genes into 19,090 high ($TPM>2$) and 20,001 low expressed genes ($TPM>2$). Cut-offs were determined using the first quartile of the distribution of wild-type 51A serotype mRNA expression. For these gene sets, gene body normalized read counts were calculated for H3K27me3, H3K4me3, H3K9ac, Pol II, and MNase, as well as the ratio of H3K4me3 and H3K27me3. We also obtained three genetic features: gene length, intron frequency, and intergenic length. We built a random forests classifier in Python (version 3) using the default parameters available with the scikit-learn package (Pedregosa et al. 2011). We used all available data to train the model using a 40-fold cross-validation (CV) method, and the CV-based PR-AUC was used to evaluate the performance of different models. A PR-AUC of one would represent a perfect model, which 100% of the time would correctly predict whether a gene is highly or lowly expressed. Further, we used the shap package (Lundberg et al. 2020) to calculate the global and local feature importance.

Partial correlation networks

We investigated the partial correlation of any two epigenetic marks of interest after removing the effects of other measured epigenetic marks by using the sparse partial correlation networks method (Lasserre et al. 2013). We used the gene body normalized signals of all the epigenetic marks in this study and the mRNA expression for this analysis.

Analyses of gene expression plasticity

The mean TPM for each gene over different conditions (expression data from serotype A, B, D, and H as well as heat shock conditions) (Cheaib et al. 2015) was calculated. The absolute deviation from the mean for each gene was calculated. We refer to genes with a large fluctuation as plastic genes. For the random forests analysis of plastic genes, we grouped all genes in four groups of roughly similar gene numbers. We performed random down-sampling (five times) of highly or lowly expressed genes such that there is an equal number of genes in both groups for classification.

Data access

All raw read data generated in this study have been submitted to the European Nucleotide Archive (ENA; <https://www.ebi.ac.uk/ena/browser/home>) under accession number PRJEB46233.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Karl Nordstöm for the help with the initial analyses and Salmon DNA sequencing and Laura Arrigoni and Ulrike Boenisch for support with NEXSON and ChIP-seq. We thank Sandra Duhaucourt and Melody Matelot for sharing unpublished MNase data sets. This work was supported by grants from the German Research Council (DFG) to M.S. (SI1379/3-1), M.H.S. (3140/1-1), and M.J. (CRC894). A.S. was supported by the German Federal Ministry of Research and Education grant for de.NBI (031L01 01D). We acknowledge the support of the Freiburg Galaxy Team, University of Freiburg (Germany) funded by the Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC.

References

Allen SE, Nowacki M. 2020. Roles of noncoding RNAs in ciliate genome architecture. *J Mol Biol* **432**: 4186–4198. doi:10.1016/j.jmb.2019.12.042

Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, Wilkes CD, Garnier O, Labadie K, Lauderdale BE, Le Mouél A, et al. 2012. The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet* **8**: e1002984. doi:10.1371/journal.pgen.1002984

Arnaiz O, Van Dijk E, Bétermier M, Lhuillier-Akakpo M, de Vanssay A, Duhaucourt S, Sallet E, Gouzy J, Sperling L. 2017. Improved methods and resources for paramecium genomics: transcription units, gene annotation and gene expression. *BMC Genomics* **18**: 483. doi:10.1186/s12864-017-3887-z

Arrigoni L, Richter AS, Betancourt E, Bruder K, Diehl S, Manke T, Bönisch U. 2016. Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Res* **44**: e67. doi:10.1093/nar/gkv1495

Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**: 171–178. doi:10.1038/nature05230

Baranasik D, Oppermann T, Cheaib M, Cullum J, Schmidt H, Simon M. 2014. Genomic characterization of variable surface antigens reveals a telomere position effect as a prerequisite for RNA interference-mediated silencing in *Paramecium tetraurelia*. *mBio* **5**: e01328. doi:10.1128/mBio.01328-14

Batsché E, Yaniv M, Muchardt C. 2006. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat Struct Mol Biol* **13**: 22–29. doi:10.1038/nsmb1030

Benayoun BA, Pollina EA, Ucar D, Mahmoudi S, Karra K, Wong ED, Devarajan K, Daugherty AC, Kundaje AB, Mancini E, et al. 2014. H3k4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**: 673–688. doi:10.1016/j.cell.2014.06.027

Bétermier M, Duhaucourt S. 2014. Programmed rearrangement in ciliates: *Paramecium*. *Microbiol Spectr* **2**: MDNA3-0035-2014. doi:10.1128/microbiolspec.MDNA3-0035-2014

Böhm S, Östlund Farrants A. 2011. Chromatin remodelling and RNA processing. In *RNA processing* (ed. Grabowski P), p. 1. IntechOpen, London. doi:10.5772/20998

Buratowski S. 2009. Progression through the RNA polymerase II CTD cycle. *Mol Cell* **36**: 541–546. doi:10.1016/j.molcel.2009.10.019

Cheaib M, Dehghani Amirabad A, Nordström KJ, Schulz MH, Simon M. 2015. Epigenetic regulation of serotype expression antagonizes transcriptome dynamics in *Paramecium tetraurelia*. *DNA Res* **22**: 293–305. doi:10.1093/dnares/dsv014

Chen K, Xi Y, Pan X, Li Z, Kaestner K, Tyler J, Dent S, He X, Li W. 2013. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res* **23**: 341–351. doi:10.1101/gr.142067.112

Chen K, Chen Z, Wu D, Zhang L, Lin X, Su J, Rodriguez B, Xi Y, Xia Z, Chen X, et al. 2015. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* **47**: 1149–1157. doi:10.1038/ng.3385

Chung H-R, Dunkel I, Heise F, Linke C, Krobitch S, Ehrenhofer-Murray AE, Sperling SR, Vingron M. 2011. The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One* **5**: e15754. doi:10.1371/journal.pone.0015754

Drews F, Karunanithi S, Götz U, Marker S, deWijn R, Pirritano M, Rodrigues-Viana AM, Jung M, Gasparoni G, Schulz MH, et al. 2021. Two Piwis with Ago-like functions silence somatic genes at the chromatin level. *RNA Biol* **18**: 757–769. doi:10.1080/15476286.2021.1991114

Duret L, Cohen J, Jubin C, Dessen P, Gouët J-F, Mousset S, Aury J-M, Jaillon O, Noël B, Arnaiz O, et al. 2008. Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: a somatic view of the germline. *Genome Res* **18**: 585–596. doi:10.1101/gr.074534.107

Egloff S, Murphy S. 2008. Cracking the RNA polymerase II CTD code. *Trends Genet* **24**: 280–288. doi:10.1016/j.tig.2008.03.008

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216. doi:10.1038/nmeth.1906

Fischle W, Tseng BS, Dormann HL, Ueberheide BM, Garcia BA, Shabanowitz J, Hunt DF, Funabiki H, Allis CD. 2005. Regulation of hp1–chromatin binding by histone h3 methylation and phosphorylation. *Nature* **438**: 1116–1122. doi:10.1038/nature04219

Frapporti A, Pina CM, Arnaiz O, Holoch D, Kawaguchi T, Humbert A, Eleftheriou E, Lombard B, Loew D, Sperling L, et al. 2019. The Polycomb protein Ezh1 mediates H3K9 and H3K27 methylation to repress transposable elements in *Paramecium*. *Nat Commun* **10**: 2710. doi:10.1038/s41467-019-10648-5

Furrer DI, Swart EC, Kraft MF, Sandoval PY, Nowacki M. 2017. Two sets of Piwi proteins are involved in distinct sRNA pathways leading to elimination of germline-specific DNA. *Cell Rep* **20**: 505–520. doi:10.1016/j.celrep.2017.06.050

Garg J, Saettone A, Nabeel-Shah S, Cadornin M, Ponce M, Marquez S, Pu S, Greenblatt J, Lambert J-P, Pearlman RE, et al. 2019. The Med31 conserved component of the divergent mediator complex in *Tetrahymena thermophila* participates in developmental regulation. *Curr Biol* **29**: 2371–2379.e6. doi:10.1016/j.cub.2019.06.052

The Gene Ontology Consortium. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**: D330–D338. doi:10.1093/nar/gky1055

Gnan S, Matelot M, Weiman M, Arnaiz O, Guérin F, Sperling L, Bétermier M, Thermes C, Chen C-L, Duhaucourt S. 2022. GC content but not nucleosome positioning directly contributes to intron-splicing efficiency in *Paramecium*. *Genome Res* (this issue) **32**: 699–709. doi:10.1101/gr.276125.121

Götz U, Marker S, Cheaib M, Andresen K, Shrestha S, Durai DA, Nordström KJ, Schulz MH, Simon M. 2016. Two sets of RNAi components are required for heterochromatin formation *in trans* triggered by truncated transgenes. *Nucleic Acids Res* **44**: 5908–5923. doi:10.1093/nar/gkw267

Guérin F, Arnaiz O, Boggetto N, Wilkes CD, Meyer E, Sperling L, Duhaucourt S. 2017. Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements. *BMC Genomics* **18**: 327. doi:10.1186/s12864-017-3713-7

Hardy A, Matelot M, Touzeau A, Klopp C, Lopez-Roques C, Duhaucourt S, DeFrance M. 2021. DNAModAnnot: a R toolbox for DNA modification filtering and annotation. *Bioinformatics* **37**: 2738–2740. doi:10.1093/bioinformatics/btab032

Huang H, Smothers JF, Wiley EA, Allis CD. 1999. A nonessential HP1-like protein affects starvation-induced assembly of condensed chromatin and gene expression in macronuclei of *Tetrahymena thermophila*. *Mol Cell Biol* **19**: 3624–3634. doi:10.1128/MCB.19.5.3624

Ignarski M, Singh A, Swart EC, Arambasic M, Sandoval PY, Nowacki M. 2014. *Paramecium tetraurelia* chromatin assembly factor-1-like protein

- PtCAF-1 is involved in RNA-mediated control of DNA elimination. *Nucleic Acids Res* **42**: 11952–11964. doi:10.1093/nar/gku874
- Ishihara S, Sasagawa Y, Kameda T, Yamashita H, Umeda M, Kotomura N, Abe M, Shimono Y, Nikaido I. 2021. Local states of chromatin compaction at transcription start sites control transcription levels. *Nucleic Acids Res* **49**: 8007–8023. doi:10.1093/nar/gkab587
- Jaehning JA. 2010. The Paf1 complex: platform or player in RNA polymerase II transcription? *Biochim Biophys Acta* **1799**: 379–388. doi:10.1016/j.bbagr.2010.01.001
- Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, Soudemont B, Nowacki M, Serrano V, Porcel BM, Ségurens B, et al. 2008. Translational control of intron splicing in eukaryotes. *Nature* **451**: 359–362. doi:10.1038/nature06495
- Jimeno-González S, Reyes JC. 2016. Chromatin structure and pre-mRNA processing work together. *Transcription* **7**: 63–68. doi:10.1080/21541264.2016.1168507
- Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**: 207–220. doi:10.1038/s41576-018-0089-8
- Lassere J, Chung H-R, Vingron M. 2013. Finding associations among histone modifications using sparse partial correlation networks. *PLoS Comput Biol* **9**: e1003168. doi:10.1371/journal.pcbi.1003168
- Lhuillier-Akakpo M, Guérin E, Frapporti A, Duharcourt S. 2016. DNA deletion as a mechanism for developmentally programmed centromere loss. *Nucleic Acids Res* **44**: 1553–1565. doi:10.1093/nar/gkv1110
- Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**: 56–67. doi:10.1038/s42256-019-0138-9
- Luo G-Z, Hao Z, Luo L, Shen M, Sparvoli D, Zheng Y, Zhang Z, Weng X, Chen K, Cui Q, et al. 2018. N⁶-Methyldeoxyadenosine directs nucleosome positioning in *Tetrahymena* DNA. *Genome Biol* **19**: 200. doi:10.1186/s13059-017-1381-1
- Meyer CA, Liu XS. 2014. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* **15**: 709–721. doi:10.1038/nrg3788
- Nalabothula N, McVicker G, Maiorano J, Martin R, Pritchard JK, Fondufe-Mittendorf YN. 2014. The chromatin architectural proteins HMGD1 and H1 bind reciprocally and have opposite effects on chromatin structure and gene regulation. *BMC Genomics* **15**: 92. doi:10.1186/1471-2164-15-92
- Nordström KJ, Schmidt F, Gasparoni N, Salhab A, Gasparoni G, Kattler K, Müller F, Ebert P, Costa IG, DEEP consortium, et al. 2019. Unique and assay specific features of NOME-, ATAC- and DNase I-seq data. *Nucleic Acids Res* **47**: 10580–10596. doi:10.1093/nar/gkz799
- Pan G, Tian S, Nie J, Yang C, Ruotti V, Wei H, Jonsdottir GA, Stewart R, Thomson JA. 2007. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1**: 299–312. doi:10.1016/j.stem.2007.08.003
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417–419. doi:10.1038/nmeth.4197
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res* **12**: 2825–2830.
- Ramírez F, Ryan DP, Grünig B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–W165. doi:10.1093/nar/gkw257
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26. doi:10.1038/nbt.1754
- Russell CB, Fraga D, Hinrichsen RD. 1994. Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia*. *Nucleic Acids Res* **22**: 1221–1225. doi:10.1093/nar/22.7.1221
- Rzeszutek I, Maurer-Alcala X, Nowacki M. 2020. Programmed genome rearrangements in ciliates. *Cell Mol Life Sci* **77**: 4615–4629. doi:10.1007/s00018-020-03555-2
- Samuel C, Mackie J, Sommerville J. 1981. Macronuclear chromatin organization in *Paramecium primaurelia*. *Chromosoma* **83**: 481–492. doi:10.1007/BF00328274
- Schulman IG, Cook RG, Richman R, Allis CD. 1987. *Tetrahymena* contain two distinct and unusual high mobility group (HMG)-like proteins. *J Cell Biol* **104**: 1485–1494. doi:10.1083/jcb.104.6.1485
- Simon M, Plattner H. 2014. Unicellular eukaryotes as models in cell and molecular biology: critical appraisal of their past and future value. *Int Rev Cell Mol Biol* **309**: 141–198. doi:10.1016/B978-0-12-800255-1.00003-X
- Simon MC, Marker S, Schmidt HJ. 2006. Posttranscriptional control is a strong factor enabling exclusive expression of surface antigens in *Paramecium tetraurelia*. *Gene Expr* **13**: 167–178. doi:10.3727/000000006783991809
- Singh A, Vancura A, Woycicki RK, Hogan DJ, Hendrick AG, Nowacki M. 2018. Determination of the presence of 5-methylcytosine in *Paramecium tetraurelia*. *PLoS One* **13**: e0206667. doi:10.1371/journal.pone.0206667
- Sneppen K, Ringrose L. 2019. Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent. *Nat Commun* **10**: 2133. doi:10.1038/s41467-019-10130-2
- Stover NA, Krieger CJ, Binkley G, Dong Q, Fisk DG, Nash R, Sethuraman A, Weng S, Cherry JM. 2006. *Tetrahymena* genome database (TGD): a new genomic resource for *Tetrahymena thermophila* research. *Nucleic Acids Res* **34**: D500–D503. doi:10.1093/nar/gkj054
- Tian M, Mochizuki K, Loidl J. 2019. Non-coding RNA transcription in *Tetrahymena* meiotic nuclei requires dedicated mediator complex-associated proteins. *Curr Biol* **29**: 2359–2370.e5. doi:10.1016/j.cub.2019.05.038
- Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **10**: 442. doi:10.1186/1471-2105-10-442
- Wang Y, Chen X, Sheng Y, Liu Y, Gao S. 2017. N⁶-Adenine DNA methylation is associated with the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in *Tetrahymena*. *Nucleic Acids Res* **45**: 11594–11606. doi:10.1093/nar/gkx883
- Xiong J, Gao S, Dui W, Yang W, Chen X, Taverna SD, Pearlman RE, Ashlock W, Miao W, Liu Y. 2016. Dissecting relative contributions of cis- and trans-determinants to nucleosome distribution by comparing *Tetrahymena* macronuclear and micronuclear chromatin. *Nucleic Acids Res* **44**: 10091–10105. doi:10.1093/nar/gkw684
- Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2020. Ensembl 2020. *Nucleic Acids Res* **48**: D682–D688. doi:10.1093/nar/gkz966
- Zhao X, Liu Y. 2019. Transcription regulation: tales of a divergent mediator. *Curr Biol* **29**: R685–R688. doi:10.1016/j.cub.2019.06.033
- Zhao XD, Han X, Chew JL, Liu J, Chiu KP, Choo A, Orlov YL, Sung W-K, Shahab A, Kuznetsov VA, et al. 2007. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* **1**: 286–298. doi:10.1016/j.stem.2007.08.004
- Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Gene* **12**: 7–18. doi:10.1038/nrg2905

Received August 20, 2021; accepted in revised form March 4, 2022.



Broad domains of histone marks in the highly compact *Paramecium* macronuclear genome

Franziska Drews, Abdulrahman Salhab, Sivarajan Karunanithi, et al.

Genome Res. 2022 32: 710-725 originally published online March 9, 2022
Access the most recent version at doi:[10.1101/gr.276126.121](https://doi.org/10.1101/gr.276126.121)

Supplemental Material <http://genome.cshlp.org/content/suppl/2022/03/30/gr.276126.121.DC1>

Related Content **GC content, but not nucleosome positioning, directly contributes to intron splicing efficiency in *Paramecium***
Stefano Gnan, Mélody Matelot, Marion Weiman, et al.
[Genome Res. April , 2022 32: 699-709](https://doi.org/10.1101/gr.276126.121)

References This article cites 64 articles, 5 of which can be accessed free at:
<http://genome.cshlp.org/content/32/4/710.full.html#ref-list-1>

Articles cited in:
<http://genome.cshlp.org/content/32/4/710.full.html#related-urls>

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

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

Affordable, Accurate
Sequencing.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>



Two Piwis with Ago-like functions silence somatic genes at the chromatin level

Franziska Drews, Sivarajan Karunanithi, Ulrike Götz, Simone Marker, Raphael deWijn, Marcello Pirritano, Angela M. Rodrigues-Viana, Martin Jung, Gilles Gasparoni, Marcel H. Schulz & Martin Simon

To cite this article: Franziska Drews, Sivarajan Karunanithi, Ulrike Götz, Simone Marker, Raphael deWijn, Marcello Pirritano, Angela M. Rodrigues-Viana, Martin Jung, Gilles Gasparoni, Marcel H. Schulz & Martin Simon (2021): Two Piwis with Ago-like functions silence somatic genes at the chromatin level, RNA Biology, DOI: [10.1080/15476286.2021.1991114](https://doi.org/10.1080/15476286.2021.1991114)

To link to this article: <https://doi.org/10.1080/15476286.2021.1991114>



View supplementary material [↗](#)



Published online: 18 Oct 2021.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

Two Piwis with Ago-like functions silence somatic genes at the chromatin level

Franziska Drews^{a,b}, Sivarajan Karunanithi^{c,d}, Ulrike Götz^b, Simone Marker^b, Raphael deWijn^b, Marcello Pirritano^{a,b}, Angela M. Rodrigues-Viana^b, Martin Jung^e, Gilles Gasparoni^f, Marcel H. Schulz^{c,d}, and Martin Simon^{a,b*}

^aMolecular Cell Biology and Microbiology, Wuppertal University, Wuppertal, Germany; ^bMolecular Cell Dynamics, Centre for Human and Molecular Biology, Saarland University, Saarbrücken, Germany; ^cCluster of Excellence, Multimodal Computing and Interaction, Saarland University and Department for Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany; ^dInstitute for Cardiovascular Regeneration, Goethe-University Hospital, Frankfurt, Germany; ^eSchool of Medicine, Medical Biochemistry and Molecular Biology, Saarland University, Homburg, Germany; ^fGenetics/Epigenetics, Centre for Human and Molecular Biology, Saarland University, Saarbrücken, Germany

ABSTRACT

Most sRNA biogenesis mechanisms involve either RNase III cleavage or ping-pong amplification by different Piwi proteins harbouring slicer activity. Here, we follow the question why the mechanism of transgene-induced silencing in the ciliate *Paramecium* needs both Dicer activity and two Ptiwi proteins. This pathway involves primary siRNAs produced from non-translatable transgenes and secondary siRNAs from targeted endogenous loci. Our data does not indicate any signatures from ping-pong amplification but Dicer cleavage of long dsRNA. Ptiwi13 and 14 prefer different sub-cellular localizations and different preferences for primary and secondary siRNAs but do not load them mutually exclusive. Both Piwis enrich for antisense RNAs and show a general preference for uridine-rich sRNAs along the entire sRNA length. In addition, Ptiwi14-loaded siRNAs show a 5'-U signature. Our data indicates both Ptiwis and 2'-O-methylation contributing to strand selection of Dicer cleaved siRNAs. This unexpected function of the two distinct vegetative Piwis extends the increasing knowledge of the diversity of Piwi functions in diverse silencing pathways. We describe an unusual mode of action of Piwi proteins extending not only the great variety of Piwi-associated RNAi pathways but moreover raising the question whether this could have been the primordial one.

ARTICLE HISTORY

Received 28 May 2021
Revised 2 September 2021
Accepted
27 September 2021

KEYWORDS

Argonaute; piwi; sRNA loading; RNA interference; siRNA; dicer; transgene-induced silencing; secondary siRNAs

Introduction



RNA silencing is a term describing a broad variety of mechanisms that use short RNA molecules to regulate gene expression. These can either target already transcribed mRNAs post-transcriptionally (PTGS) or they can interfere in transcription via co-transcriptional targeting of nascent transcripts (CTGS), thus recruiting chromatin modifying complexes [1,2]. An important component of RNAi (RNA interference) mechanisms is Argonaute proteins (Ago), which load small RNAs (sRNAs), thus creating functional complexes. Agos themselves can be phylogenetically dissected into two clades: Agos and Piwis (P-element-induced wimpy testes), the latter being discovered in *Drosophila* germline stem cells. Agos form the RISC (RNA induced silencing complex), with miRNAs and siRNAs both being ubiquitously expressed, whereas Piwis and piRNAs were believed to be expressed in germline cells only [3].


siRNAs are a distinct class of regulatory RNAs produced by the dsRNA (double stranded RNA)-specific ribonuclease Dicer: these have been shown across kingdoms to act either in PTGS and CTGS of protein-coding genes and structural elements such as centromeres through the life cycle [1]. In many systems, secondary siRNAs have been shown to be

produced involving activity of RNA-dependent RNA polymerases (RDR).

In contrast to Dicer-cleaved siRNAs, piRNAs were mainly described to silence transposable elements during gametogenesis. However, increasing data on different piRNA mechanisms reveal an unexpected diversity of those. This diversity does not only concern piRNA targets but also their temporal/spatial occurrence and, most importantly, the piRNA biogenesis mechanisms. In mouse and *Drosophila*, similar mechanisms were described: single-stranded precursor 5'-U RNAs are loaded into Piwi proteins, and this is followed by subsequent 3'-trimming of the piRNA end. This initiation is then followed by Dicer-independent amplification of piRNAs by the ping-pong mechanism involving the reciprocal cleavage of complementary ssRNA, thus generating an internal single nucleotide A preference (reviewed in [4]).

3'-Nucleotides of mature piRNAs usually carry a 2'-O-methylation, which is added after 3'-processing. Some mechanistic diversity becomes apparent as *Drosophila* and mouse piRNAs become 3'-trimmed after Piwi loading, but this was not reported in *Caenorhabditis elegans* where mature piRNAs are generated from precursors by 5'- and 3'-processing, and then the mature piRNA is loaded into Piwi

*CONTACT Martin Simon  masimon@uni-wuppertal.de  Molecular Cell Biology and Microbiology, Wuppertal University, Wuppertal, Germany; Molecular Cell Dynamics, Centre for Human and Molecular Biology, Saarland University

 Supplemental data for this article can be accessed [here](#).

© 2021 Informa UK Limited, trading as Taylor & Francis Group

[5]. In all systems, 3'-methylation of piRNAs occurs after Piwi loading.

Piwi-mediated piRNA biogenesis differs to siRNAs and miRNAs maturation by the action of Agos and Piwis [6]. Agos load duplexes of Dicer cuts and select for guide and passenger strand before screening for targets. In contrast, Piwis do not show strand selection from duplexes and show a tendency for 5'-U containing ssRNAs, the latter has been shown to be due to both biased piRNA biogenesis and selectivity of Piwis [7].

As mentioned, piRNA pathways differ extremely between species and show a lack of conservation of involved genes [8]. Studies in *Drosophila* demonstrated that piRNA pathway genes evolve rapidly, indicating an arms race between transposons and their cellular defence [9,10]. Also downstream mechanisms differ between species as *C. elegans* shows an absence of ping-pong amplification using RDR-dependent siRNAs for amplification of the initial piRNAs [11]. Moreover, a screen of non-model species revealed the absence of the piRNA system in nematode lineages other than the most prominent *C. elegans*: these organisms apparently use RDR-dependent siRNAs to account for transposon control [12].

Consequently, piRNA biogenesis pathways are highly diverse, and increasing evidence indicates that piRNAs are not restricted to the germline but are present in low abundance also in somatic tissues, e.g. piRNA-like sRNAs have been identified in various somatic tissues by their ping-pong signature, and increasing reports also show piRNAs regulating expression of endogenous genes in somatic cells too [13,14]. However, purely descriptive studies reporting somatic piRNA expression need to be handled with care because of miss-annotations in piRNA databases containing piRNA-sized fragments of longer RNAs showing high RNA levels in somatic tissues [15].

As a result, an ongoing discussion asks for the evolution of these multiple functions of piRNAs, and one possibility is the co-option of transposon-derived piRNAs to regulate genomic functions [16]. A piRNA analysis of several arthropod species revealed somatic piRNAs targeting transposons and mRNAs across all species, and the authors consequently scrutinize that the ancestral role of piRNAs was to protect the germline from transposons [17].

In the context of changing dogmas about Piwis, ciliates provide an excellent model. They use several vegetative and developmental sRNA pathways [18], and they do not harbour any Agos. *Paramecium tetraurelia*, for instance, contains 15 distinct Piwi proteins called Ptiwis [19]. These unicellular eukaryotes undergo sexual recombination of meiotic nuclei in order to develop somatic macronuclei (MAC) in the same cell with germline micronuclei (MIC). For further evaluation of *Paramecium* Piwi functions in the evolutionary context, one should be aware that the ciliate nuclear dimorphism cannot be seen as an ancestral state of multicellular species as germ-soma nuclear differentiation evolved at least twice in unicellular species [20]. As such, vegetative cells comprise functions of germline and somatic cells, rather than being exclusively somatic or germline.

Here, we characterize two Ptiwi proteins (Ptiwi13 and Ptiwi14) expressed during vegetative growth of *Paramecium*

tetraurelia. Both are involved in transgene-induced silencing. In this mechanism, transformation of non-expressible transgenes silence endogenous gene loci and this has been shown to involve dynamic chromatin remodelling [19,21,22]. This is at first glance similar to other mechanisms in which small RNA-mediated interaction of two different genetic loci have been reported. Next to co-suppression in plants where endogenous and exogenous homologous genes are post-transcriptionally silenced by siRNAs [23], also the paramutation represents an epigenetic interaction between two different genetic loci [24]. Interestingly, paramutations in *Drosophila* and *C. elegans* can involve both piRNA and siRNA elements, respectively [25,26].

The mechanism of transgene-induced silencing in *Paramecium* differs from those phenomena as only truncated, non-expressible transgenes can trigger silencing [27]. Transcription of translatable and intact mRNA from transgenes appears to repress silencing and, in addition, also the deletion of these genes in F1 progeny [28]. Such transgenerational manifestations can also be observed in *C. elegans* and *Drosophila*; however, inheritance there concerns only gene silencing, not gene deletion [26,29].

In *Paramecium*, the precise characterization of sRNAs of transgene-induced silencing and especially of their bio-accumulation is missing. Two distinct Ptiwi proteins are involved, which is in conflict with the fact that Dicer is involved in the mechanism [22,30]. The sRNA specificity of the individual Ptiwi proteins remains unknown, and as such, their role and the origin and function of their associated sRNAs also remains elusive. The aim of this study is to dissect transgene-induced sRNAs by their loading into Ptiwi proteins to clarify about the the role and the mechanism of the two distinct Ptiwis in the mechanism in *Paramecium*.

Materials and methods

Cell culture, RNAi, microinjection

Paramecium tetraurelia cells (stock 51 and d4-2) were cultured as described before using *Klebsiella planticola* for regular food in wheat grass powder (WGP) [31]. All cultures for this study were grown at 31°C. RNAi by feeding of dsRNA-producing bacteria was carried out as described before [32,33] using the double T7 vector L4440 in the RNase III-deficient *E. coli* HT115DE3. Microinjection of the pTI-/- and FLAG fusion transgenes was carried out as described before [34].

Phylogenetic analysis

The evolutionary history was inferred using the Neighbour-Joining method with 1000 bootstrap replicates [35,36]. The optimal tree is shown in Fig. 1A. Evolutionary distances were computed using the Poisson correction method [37] and are in the units of the number of amino acid substitutions per site. Ambiguous positions were removed by the pairwise deletion option. There were a total of 1.703 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [38]. Proteins were aligned with Muscle using default parameters. Ptiwi1-15 sequences were described in [19], and we

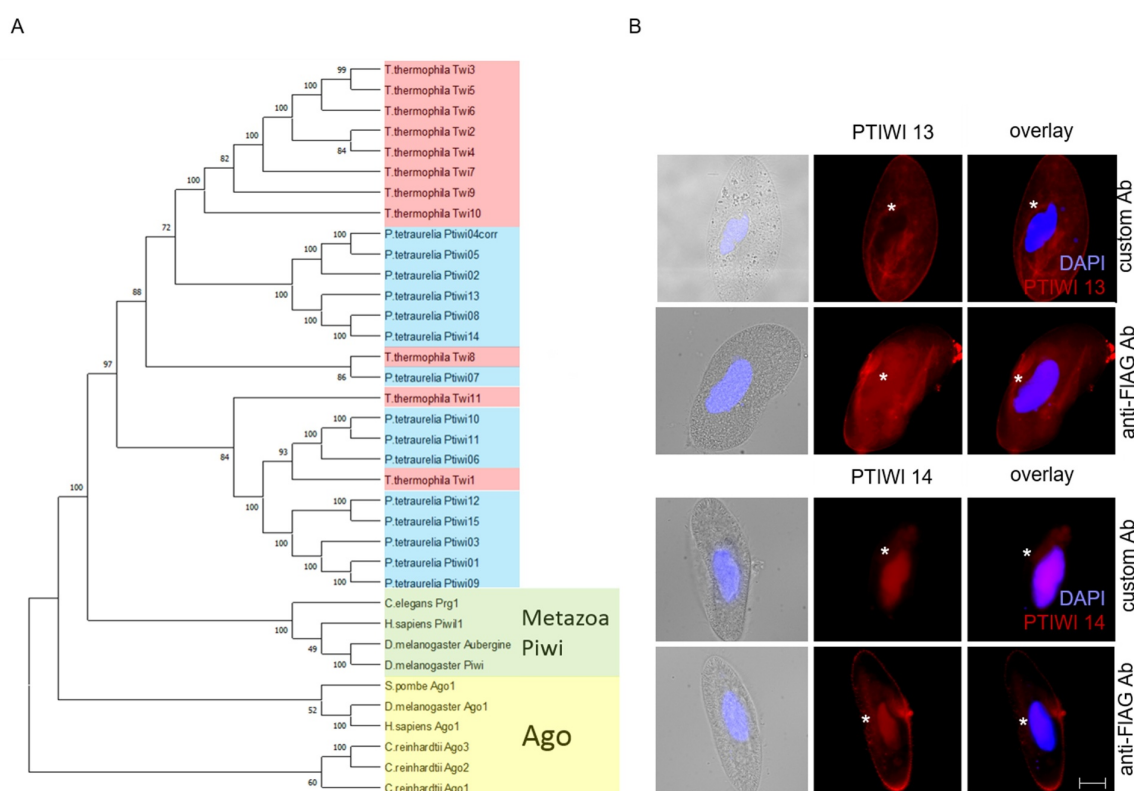


Figure 1. Relationship of Ptiwi proteins and their localization in vegetative cells. A: Phylogenetic tree of *Paramaecium tetraurelia* Ptiwi proteins (blue) in relationship to *Tetrahymena* Piwis (Twi, red), metazoan Piwi proteins (green) and Agos (yellow). Support values are given at nodes, see Methods for details. The amino acid sequence of the putative pseudogene Ptiwi04 was corrected using its paralog Ptiwi05. B: Localization of Ptiwi proteins in vegetative *Paramaecium* cells injected with Ptiwi13-FLAG (top) or Ptiwi14-FLAG (bottom). Cells were analysed by indirect immunofluorescent staining using custom antibodies directed against Ptiwi13 or Ptiwi14 labelled with secondary Alexa594-conjugated antibody (red). Additionally, the cells were stained using anti-FLAG antibody. Representative overlays of Z-stacks of magnified views are presented. Other panels show DAPI (in blue), brightfield and overlay of DAPI and Alexa594 signal. White asterisk indicates the position of the macronucleus. Scale bar is 10 μ m and exposure is 2 s.

also used a curated amino acid sequence for the putative pseudogene Ptiwi04 using its paralog Ptiwi05 as a template.

RNA isolation and treatment

Total RNA was isolated with TriReagent (Sigma). Integrity was checked by denaturing gel electrophoresis after DNase I (Invitrogen) digestion and subsequent purification with acid phenol. For dissection of 3'-modifications by periodate oxidation, 20 μ g RNA were dissolved in 17.5 μ l 4.375 mM borax, 50 mM boric acid, pH 8.6, and 2.5 μ l 200 mM sodium periodate were added. After 10 min incubation in the dark, 2 μ l glycerol were added with another 10 min incubation. After drying in the speed-vac, the pellet was dissolved in 50 μ l 33.75 mM Borax; 50 mM boric acid; pH 9.5 and incubated for 90 min at 45°C. The RNA was subsequently purified with Sephadex G-25 columns (GE).

sRNA sequencing and analyses

For siRNA sequencing, 17–25 nt small RNA fractions were isolated by denaturing PAGE and subjected to standard small RNA library preparation using the NEB Next small

RNA sequencing Kit (NEB, Frankfurt a.M., Germany). The procedure includes 3'-OH and 5'-monophosphate-specific ligation steps, and we tried to lower 3'-2'-O-me biases by 18 hours 3'-ligation at 16°C. After 10 PCR cycles, the libraries were gel-purified and sequenced on the HiSeq 2500 using the Rapid Mode with 28 cycles. Reads were de-multiplexed, and adapter sequences were trimmed using Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) that uses Cutadapt [39] with a stringency cut-off of 10. For analysis of reads, we used normalized counts and converted these values to transcripts per million (TPM), which we also refer to as sRNA accumulation. For the analyses specific to endogenous clusters, a TPM value greater than one was termed to be present in Ptiwi IPs. We used the RAPID pipeline to obtain the normalized counts, implementing the KnockDown Corrected Scaling (KDSCS) method [40]. Hierarchical clustering of data sets was performed with complete linkage using an Euclidean distance measure, and heatmaps were created using R/Bioconductor package gplots (v3.0.1.1). Shown statistical analyses were performed on average of reads of dupli- or triplicates of experiments, including error bars based on calculated variance. Data are deposited at

the European Nucleotide archive, ENA, Acc. Nr. PRJEB38766, and information of sequencing depth and mapping statistics are provided in Table S1.

sRNA signatures

Sequence logos of 23nt sRNAs were generated using WebLogo3 [41] with error bars twice the height of the correction for small sample size. Probabilities for overlapping reads from aligned sRNA reads were calculated using the small RNA signature analysis tool in Galaxy [42]. sRNAs from 17 to 25 nt were mapped to each region of interest, allowing no mismatches or multimapper in bowtie [43], and overlaps from 1 to 25 nucleotides were calculated. Plots for read length distribution and coverage were created using Geneious Prime 2020.1.2.

Antibodies, western blots, immunostaining

Peptides corresponding to the amino acids 684–698 and 449–463 of *Paramecium* Ptiwi13 (C-DDAPPQARKNNKSPY) and Ptiwi14 (C-QNWMQRLTAEIGDK), respectively, were used for immunization of rabbits. Purification of antibodies from serum was performed by coupling the respective peptides to SulfoLink coupling resin (Thermo Scientific) and following the manual instructions. Purified antibodies were tested by dotblot assays (Fig S2). Western blots were carried out as described previously [44] using indicated antibodies diluted 1:250 in 5% milk/TBST. Indirect immunofluorescence staining was carried out as previously described [45]. Cells were permeabilized in 2.5% Triton X-100 and 1% formaldehyde for 30 min followed by fixation in 4% formaldehyde and 1.2% Triton X-100 for 10 min. After blocking in 3%BSA/TBST, the cells were incubated in primary antibody diluted 1:200 in 3% BSA/TBST under mild agitation overnight at 4°C. After washing and incubation with 1:2500 Alexa Fluor 568 F(ab')₂ fragment of goat anti-Rabbit IgG (H + L) (Thermo Scientific # A-21069), the cells were stained with DAPI and mounted in VECTASHIELD (VectorLaboratories). Images were acquired using Zeiss Axio Observer with ApoTome. For expression of tagged Ptiwis, the respective orf was cloned into *Paramecium* FLAG-Vectors pPXV containing three FLAG sequences either at the N or C terminus (kind gift of M. Valentine and J. Van Houten, Vermont, USA) as described in [46]. Injected clones were screened by single-cell PCR, and positives were grown for cell fixation and protein isolation. Macronuclei were isolated as described [47], and protein was isolated by adding preheated Laemmli sample buffer with subsequent boiling for 5 min.

Ptiwi-immunoprecipitation

Transgenic Ptiwi lines harbouring the pTI^{-/-} transgene and a single Ptiwi-FLAG fusion construct (described above) were used for Ptiwi IPs using monoclonal anti-FLAG M2 (Sigma). Our procedure follows the protocol by [48] for developmental Ptiwis with the following modifications. 500 k cells of a single transgenic line were grown and harvested by snap freezing in

2 ml lysis buffer. 1 ml of the lysate was broken in a Dounce homogenizer and 1 ml was sonified until also MACs were destroyed. After addition of 50 µl Anti-FLAG M2 Magnetic Beads (Sigma) and incubated over night by gentle agitation. After washing beads with wash buffer and re-suspended in 100µl. 10µl were used for western controls by addition of 2.5µl Laemmli sample buffer and subsequent boiling for 2 min. 90µl were extracted with TriReagent LS (Sigma) according to the manufacturer's recommendation.

Results

Ptiwi13 and 14 prefer different subcellular localizations

Fig. 1A shows the evolutionary relationship between the 15 *Paramecium tetraurelia* Ptiwi proteins. The phylogenetic tree reveals ciliate Piwi proteins clustering with metazoan Piwis, which are clearly separated from Agos. The two Piwis involved in transgene-induced silencing, Ptiwis 13 and 14, show some degree of similarity but no close relationship. They do not appear to be the result of one most recent genome duplication of which the *Paramecium tetraurelia* genome has undergone at least three. Ptiwi 14 has an ohnolog of the most recent WGD, Ptiwi08, which shows developmental expression [49,50]. In addition, a recent study identified orthologs of Ptiwi 13 and 14 in most of the species of the *Paramecium aurelia* complex and in addition in *Paramecium caudatum* and *Paramecium bursaria* [51]. An analysis of the catalytic domain (Fig S1) shows that the catalytic DEDH tetrad [52] is present in Ptiwi13 and 14, suggesting that both are capable of slicer activity.

To clarify the subcellular localization, we raised antibodies against specific peptides corresponding to both Ptiwis for immunolocalization (Methods and Fig S2) and additionally used FLAG-tagged Ptiwi transgenes, which we used for later Ptiwi IPs as well. Figure 1B indicates clear cytosolic Ptiwi13 signals in stainings with specific and FLAG antibodies. This cytosolic signal appears a bit structured, likely due to fixation-induced binding of soluble proteins to ER membranes. Ptiwi13 custom antibodies reveal additional MAC signals in ca. 20% of cells as shown in Fig S2, and also the FLAG antibodies do not show an absence of Ptiwi13 signal in the MAC. We conclude that Ptiwi13 has a predominant localization in the cytosol but can also appear in the MAC. Ptiwi14 staining with custom- and FLAG-Abs shows mainly MAC signals and only faint signals in the cytoplasm. We conclude that both Ptiwis have different sub-cellular localization preferences in MAC and cytosol, but for both Ptiwis, also less intense signals in the respective other compartment are apparent. This is supported by the comparison of total and MAC protein Western blots shown in Fig S2 and by an *in silico* analysis of the amino acid sequences by ngLoc method, which is a Bayesian classification method to predict localization of proteins [53]. According to the multi-localization confidence score (MLCS above 20), the algorithm predicts that both Ptiwis shuttle between nucleus and cytosol, where the evidence for Ptiwi13 is higher (Fig S3). Slight differences in the ratio of Mac and cytosolic signals

between FLAG and custom Abs may be due to increased levels of antigenic sites, resulting from the over-expression of FLAG-Ptiwi constructs. As these data exclude that the over-expression or the FLAG tag causes false positive localization, we proceeded with Ptiwi-IPs of these fusion proteins.

Ptiwi13 and 14 have different loading preferences for endogenous and exogenous sRNAs

For Ptiwi IPs, we injected FLAG-tagged Ptiwi13 and 14 transgenes, respectively, into a transgenic RNAi-strain harbouring the pTI^{-/-} transgene (Fig. 2A). The latter contains a GFP marker and, additionally, a truncated version of the endogenous ND169 gene causing silencing of the endogenous locus. Western blots of aliquots of the lysates/pull-downs verified the successful immunoprecipitation and the absence of soluble proteins present in the supernatant (Fig. 2B). As a first insight, the trimmed read length distribution of Ptiwi IP reads shown in Fig. 2C reveal a clear 23nt peak, which is the predominant siRNA read length in *Paramecium*. We then mapped reads to different classes of RNA templates and quantified them relative to the respective abundance in Ptiwi overexpression lines to limit the effects of an individual Ptiwi overexpression to stabilization of individual RNA species. Please note here that Ptiwi overexpression may cause unspecific binding of abundant RNA species.

Fig. 2D indicates that Ptiwi13 enriches for sRNAs of exogenous precursors such as food bacteria and mitochondria. This is in agreement with a previous report that Ptiwi13 is also involved in exogenously triggered RNAi when paramecia are fed with dsRNA producing bacteria [19]: it has later been

shown that *Paramecium* also converts exogenous ssRNA of the food bacteria such as rRNA and mRNA into siRNAs [19,54]. In addition, Ptiwi13 also enriches for fragments of rRNA and snRNA.

In contrast, Ptiwi14 IPs show accumulation of small RNAs produced from all protein coding genes (all MAC genes) and a subset of previously characterized siRNA producing genes (SRCs, small RNA clusters) of the *Paramecium* genome [55]. Small RNAs from ncRNAs are clearly underrepresented in Ptiwi14 IPs as well as fragments of snoRNA and tRNAs. The latter are also not found in Ptiwi13 IPs.

In summary, our data indicates both Ptiwis not to be redundant but with different localization and loading preferences.

2° siRNAs are enriched but not exclusively found in Ptiwi14

Both Ptiwis have been earlier shown to be necessary for efficient transgene-induced silencing [19]. Figure 3A shows the genomic structure of the endogenous ND169 gene involved in trichocyst discharge. This gene becomes silenced on the chromatin level when cells are injected with a truncated form of this gene shown below: the pTI^{-/-} transgene shows two deletions: one on the 5'-coding region (ND-1) and the 3'-coding region including the 3'-UTR (ND-2) [22].

Mapping sRNA reads to the endogenous ND169 is shown in (Fig. 3B): siRNAs mapping to the regions called ND-1 and ND-2 result from the endogenous ND169 gene only, as these regions are not present in the transgene. siRNAs mapping to ND-1 and ND-2 therefore represent 2° siRNAs.

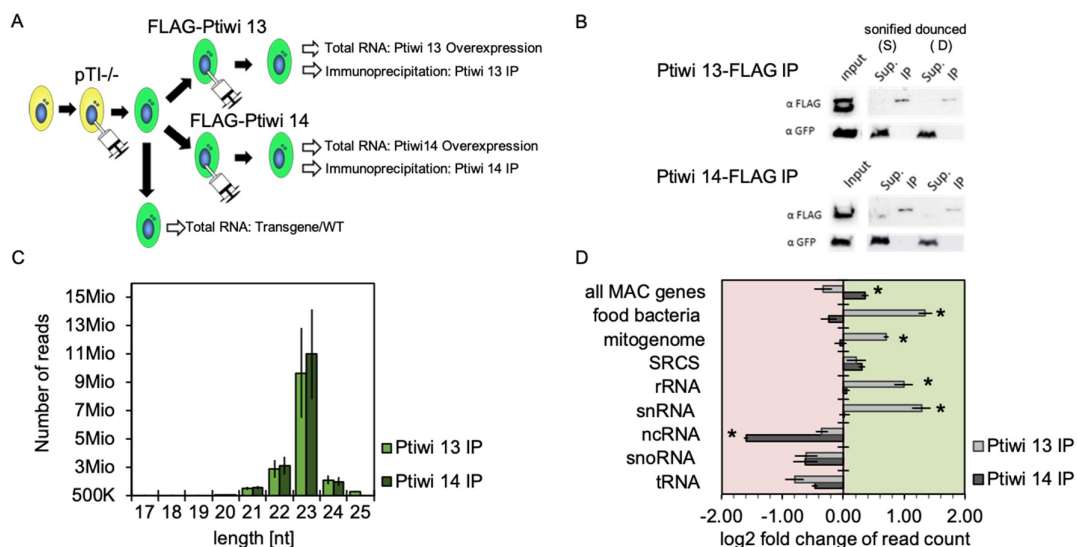


Figure 2. Analysis of sRNAs in Ptiwi immunoprecipitations. A: Experiment overview. A single cell was injected with the pTI^{-/-} transgene. After establishment of a stable line, the cells were injected with FLAG-Ptiwi13/14 constructs, respectively (green). B: Control Western blots for the IPs using anti-FLAG Abs for Ptiwi detection and anti-GFP (Sup.-Supernatant, IP-Immunoprecipitation). Two different setups of the IPs used sonication (S) and douncing (D) for cell lysis, the latter remains MAC structure but permeabilized. C: Total read length distribution of Ptiwi IPed reads after adapter trimming. Average of reads from three IP replicates is shown. D: Relative enrichment of RNA reads in Ptiwi IPs mapping to different categories of genomic templates. Average of reads from three IP replicates was calculated, and the enrichment in reference to individual Ptiwi overexpressing lines is shown. * p-value < 0.005.

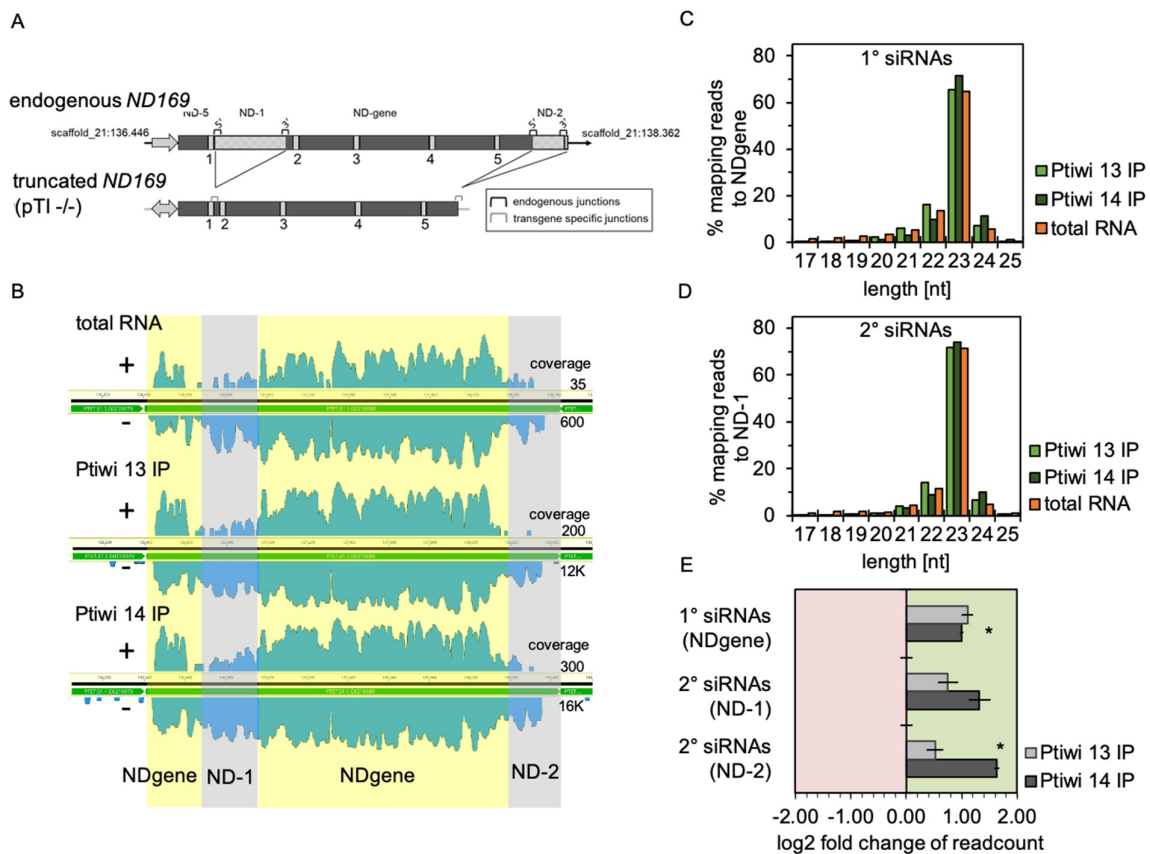


Figure 3. Ptiwi13 and 14 load transgene associated sRNAs. **A:** Detailed scheme of the endogenous *ND169* locus (top) and the truncated transgene (bottom). Introns are numbered and brackets symbolize specific junctions. Shaded regions are not part of the transgene (ND-1 and ND-2). **B:** Coverage tracks of siRNAs mapped to the endogenous *ND169* locus. siRNAs were separated by their direction (sense/antisense). Regions accounted for 1° siRNAs (NDgene, yellow) and 2° siRNAs (ND-1 and ND-2, grey) are highlighted. Coverage track in log scale is shown for one replicate each, while numbers on the right indicate untransformed sense and antisense coverage. **C:** Read length distribution of 1° siRNAs and 2° siRNAs from Ptiwi IPs and total RNA from pTI^{-/-} injected cells. Data are shown as proportion of reads mapping to the NDgene locus and **D:** the ND-1 locus. **E:** Relative enrichment of RNA reads in Ptiwi IPs mapping to different regions of the transgene. Enrichment was calculated for the average of reads of three replicates in reference to individual Ptiwi overexpressing lines. * p-value < 0.005.

They are therefore 2° siRNAs produced from the targeted gene after triggering by 1° siRNAs produced from the transgene [22]. Those regions existing in the transgene and the endogenous gene (called NDgene in the following) consist of both 1° and 2° siRNA; however, as the abundance of 2° siRNAs is more than 10-fold lower compared to 1°, the NDgene regions show predominantly 1° RNAs [22]. Figure 3B also shows the reads from Ptiwi IPs. These maps already rebut one of our first hypothesis on the question why two different Ptiwis may be involved in this mechanism: both Ptiwi IPs show reads mapping to the NDgene and to the ND-1/2 region. As such, they do not mutually exclusively load 1° siRNAs and 2° siRNAs. Analysing quality and quantity of these sRNAs, Fig. 3C and Fig. 3D show that both Ptiwi bound sRNAs are of predominant 23nt length. Ptiwi14 significantly enriches more 2° siRNAs (Fig. 3E). The finding that silencing of the *ND169* gene by the transgene was shown to occur on the chromatin level [22] may make sense in this context as 2° siRNAs may then be produced from nascent transcripts in the nucleus and loaded by nuclear Ptiwi14.

Ptiwi13 and 14 specifically load 23nt antisense siRNAs

To characterize the nature of Ptiwi loaded sRNAs, we had first a look at the ratio of 23nt reads to other read lengths. Comparing total RNA from pTI^{-/-} transgenic cells to IPs, Fig. 4A shows that both Ptiwis specifically load 23nt sRNAs; however, the ratio of 23nt to other lengths varies between different RNA species. For several RNAs, e.g. food bacteria, rRNA etc., one can see that Ptiwis specifically select 23nt sRNAs among many other RNAs. It seems likely that fragments of these RNAs are produced by different mechanisms creating several lengths of sRNA of which Ptiwis enrich for 23nt sRNAs.

This appears different for transgene-associated 1° and 2° siRNAs, which show almost identical ratio of 23nt siRNAs in total RNA and IPs, suggesting that distinct biogenesis mechanisms contributes to more precise sRNA cleavage. Going more into detail with these transgene associated siRNAs, both Ptiwis load predominantly antisense RNAs as shown in Fig. 4B. We compared this to Ptiwi knockdowns, in which the individual Ptiwis are silenced by introduction of dsRNA by feeding bacteria. In the transgene (pTI^{-/-})

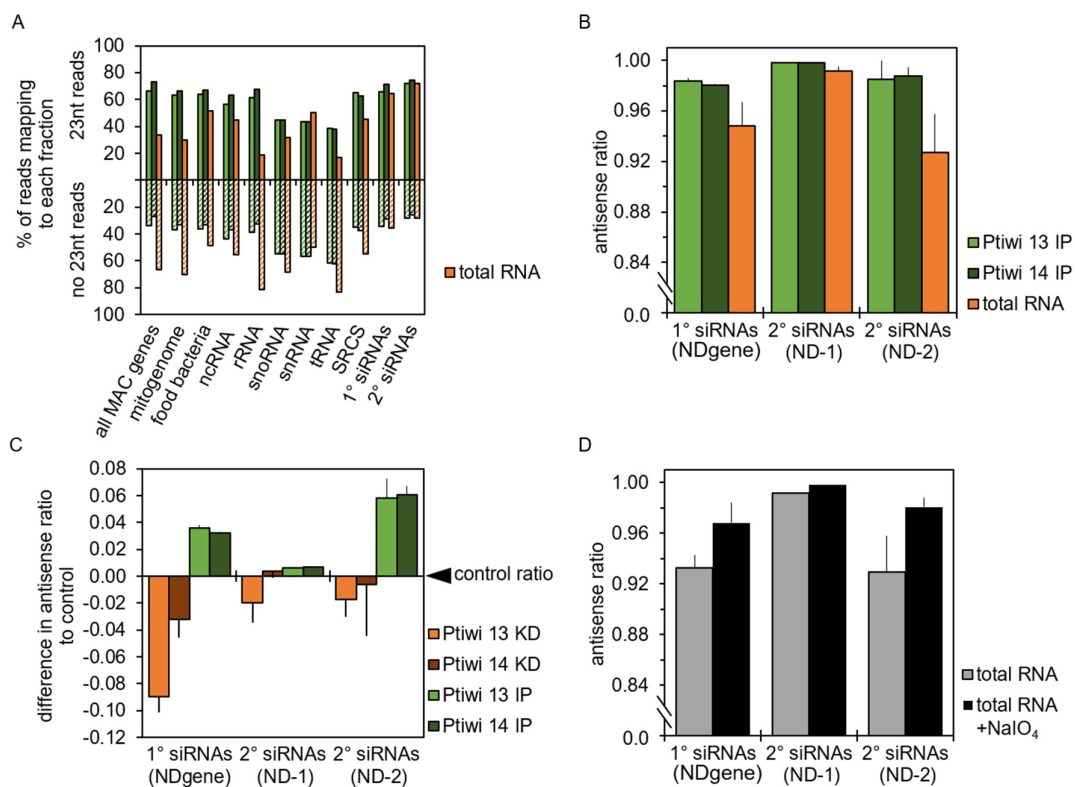


Figure 4. Asymmetric modification and Ptiwi selectivity contributes to accumulation of antisense siRNAs. A: Number of 23nt reads mapping to each indicated genomic feature and transgene regions accounting for 1° and 2° siRNAs were related to the total number of reads of other sizes. Calculation is shown for the mean of Ptiwi IPs and RNA from pTI^{-/-} injected cells as control (total RNA). B: Antisense ratio of reads from Ptiwi IPs calculated by merging three replicates each. C: Difference in the antisense ratio of reads to the antisense ratio of respective control was calculated. Data is shown for reads from knockdown of Ptiwis in duplicates and IPs (triplicates) mapping to the indicated transgene regions. D: Antisense ratio of small RNAs from pTI^{-/-} transgene samples (total RNA, untreated) and small RNAs treated with sodium periodate (+NaIO₄). Average of reads from two replicates is shown.

background, silencing of Ptiwi13 or 14 causes rescue of the *ND169* silencing phenotype, meaning that cells can eject trichocysts again [22]. Comparing the antisense ratio of Ptiwi knockdown and IPs to each other, our data indicate a certain decrease of the antisense ratio in knockdowns and an increase in IPs (Fig. 4C). It's worth to note here that Ptiwi13 is recursive: being involved in the dsRNA feeding pathway may be less efficient compared to Ptiwi14 silencing but has been shown to be efficient for reporter RNAi rescue in several instances [19,22]. These changes in the antisense ratio are only moderate, and it is either possible that both Ptiwis are redundant or that also other factors contribute to strand selection. In many systems, 2'-O-methylation was shown to occur in context of Piwi-associated sRNAs [56]. Using periodate oxidation of RNA and subsequent library preparation, we can show that both 1° and 2° are resistant to periodate thus likely to be methylated at the 3'-end (Fig S4). Moreover, our data indicate that predominantly antisense sRNAs are modified in this manner (Fig. 4D), suggesting that this modification contributes to strand selection and stabilization.

Ptiwi14 loaded siRNAs have a 5'-uridine preference

As the current data implicates that both Ptiwis select strands from Dicer cuts rather than amplify sRNAs in a ping-pong

manner, we followed this idea by analysing the sequence logos of transgene-associated sRNAs. Figure 5A shows logos of 1° and 2° sRNAs obtaining 5'-uridine preference. To decide whether these RNAs should result from a Dicer cut, one should see an A preference at position 21 for the non 5'-U reads. These are shown in Fig. 5A, but one cannot identify such a Dicer signature nor a ping-pong signature (an A at position 10 of the non 5'-U reads). Ptiwi IPs (Fig. 5B) reveal that the 5'-U preference of total RNA is mainly due to Ptiwi14, whose RNAs show a much stronger 5'-U preference compared to Ptiwi13. Unfortunately, lack of Dicer or ping-pong logos do not allow for further conclusions about the biogenesis mechanisms.

We therefore additionally analysed reads for their overlapping signature: transgene 1° siRNAs show a peak at 21nt overlaps, which fit to 23nt Dicer cuts (Fig. 5C). The 21nt overlaps are prominent but not dominant in Fig. 5C, which is likely due to the strand selection by Ptiwis, which causes degradation of the passenger strand. Fig. 5C in addition shows that we cannot identify any 21nt read overlaps in periodate-treated samples in agreement with the hypothesis of strand-specific methylation.

Interpreting this as an argument for Dicer cleavage, this is contrary to the missing Dicer signature in sequence logos, which would have been an A-preference at position 21 in non-5'-U reads. We have to consider that the observed 5'-U preference is not that strong compared to 5'-Us in *Paramecium*

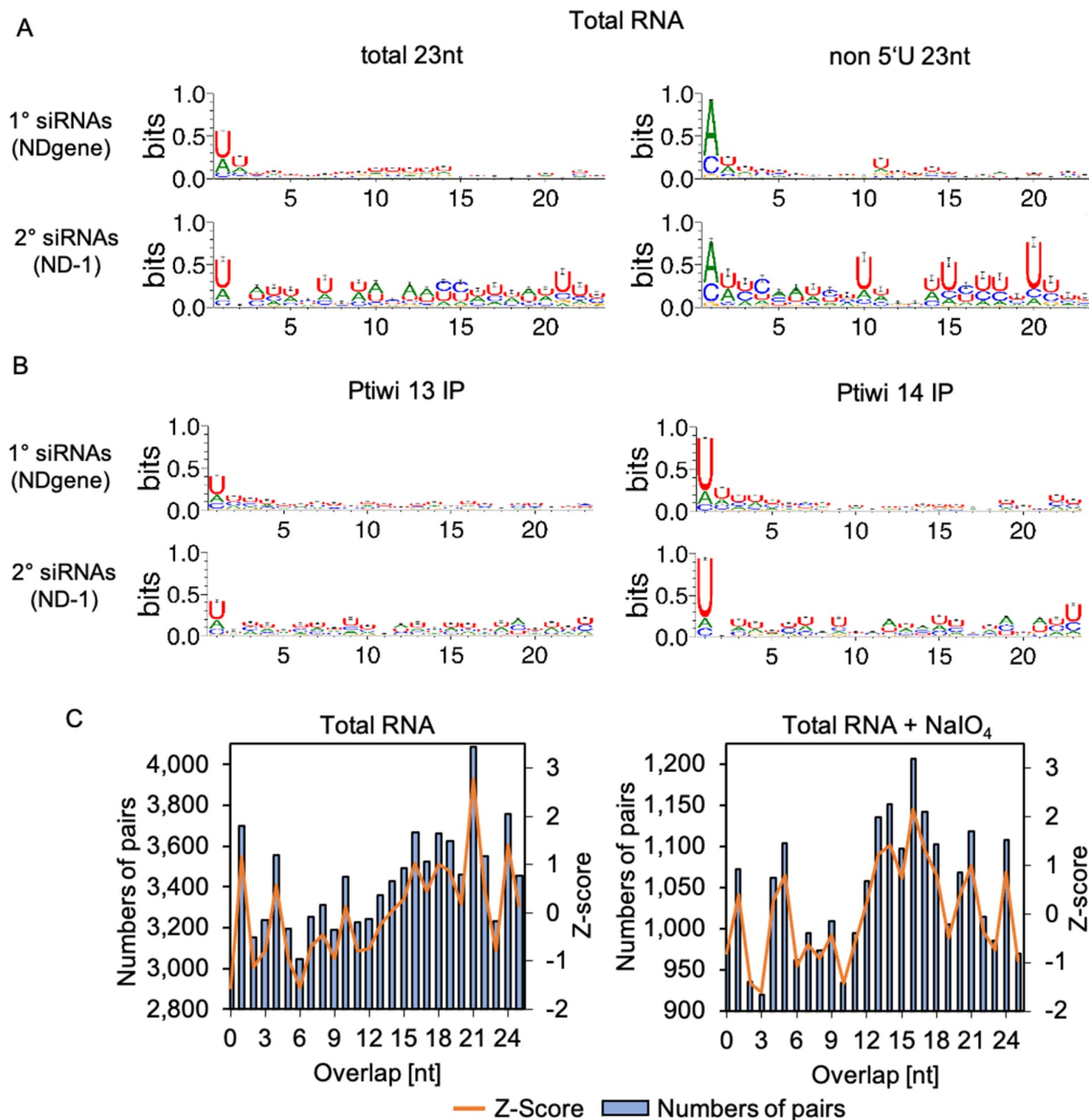


Figure 5. Sequence logos of 1° and 2° siRNAs. A: Sequence logos of 23nt antisense reads from pTI^{-/-} injected cell lines mapping to transgene regions. Logos for either all sequences or the ones without 5'-U are shown. B: Sequence logos of 23nt antisense reads from Ptiwi13 IP (left) and Ptiwi14 IP (right). C: Overlap predictions of small RNAs from 17 to 25nt of untreated, total RNA and the same small RNAs treated with sodium periodate (+NaIO₄). Z scores for overlapping pairs are included.

scnRNAs for instance [30,57], and thus the complementary 21-As on the passenger strand might not be detectable. In addition, Fig S5 shows that the 5'-U preference is still pronounced in periodate-treated RNAs, thus indicating that nucleotide preference and methylation co-occur on the same molecules.

siRNA uridine content contributes to strand selection

When analysing the sequence logos, not only the 5'-U preference was observed but some logos suggested that stabilized strands are rich in uridines (Fig S5, Fig S6). We followed this by analysis of the antisense ratio along the transgene and endogenous ND169. Figure 6A shows again that most areas show dominant antisense preference for 1° and 2° siRNAs: an

exception is the promoter proximal region (called ND-5), which shows almost 50/50 strand distribution. We consequently calculated uridine and adenosine content of these regions (Fig. 6B), revealing that the ND-5 region is different from the other regions as it shows a much higher uridine content on the sense strand. We therefore asked whether this could be seen in sRNAs, too. Also in sRNAs, the ND-5 region shows a different behaviour compared to other regions (Fig. 6C), and we consequently calculated the U-content of 23nt sRNAs of (i) *in silico* diced RNA, (ii) total transgene siRNAs and (iii) siRNAs of Ptiwi IPs. As demonstrated in Fig. 6D, the exceptional sense bias of the promoter proximal ND-5 region correlates to the enrichment of U-rich sRNAs, mainly by Ptiwi14. The analysis further reveals first that for all regions, the U-content of the more abundant antisense RNAs is higher than the sense siRNAs,

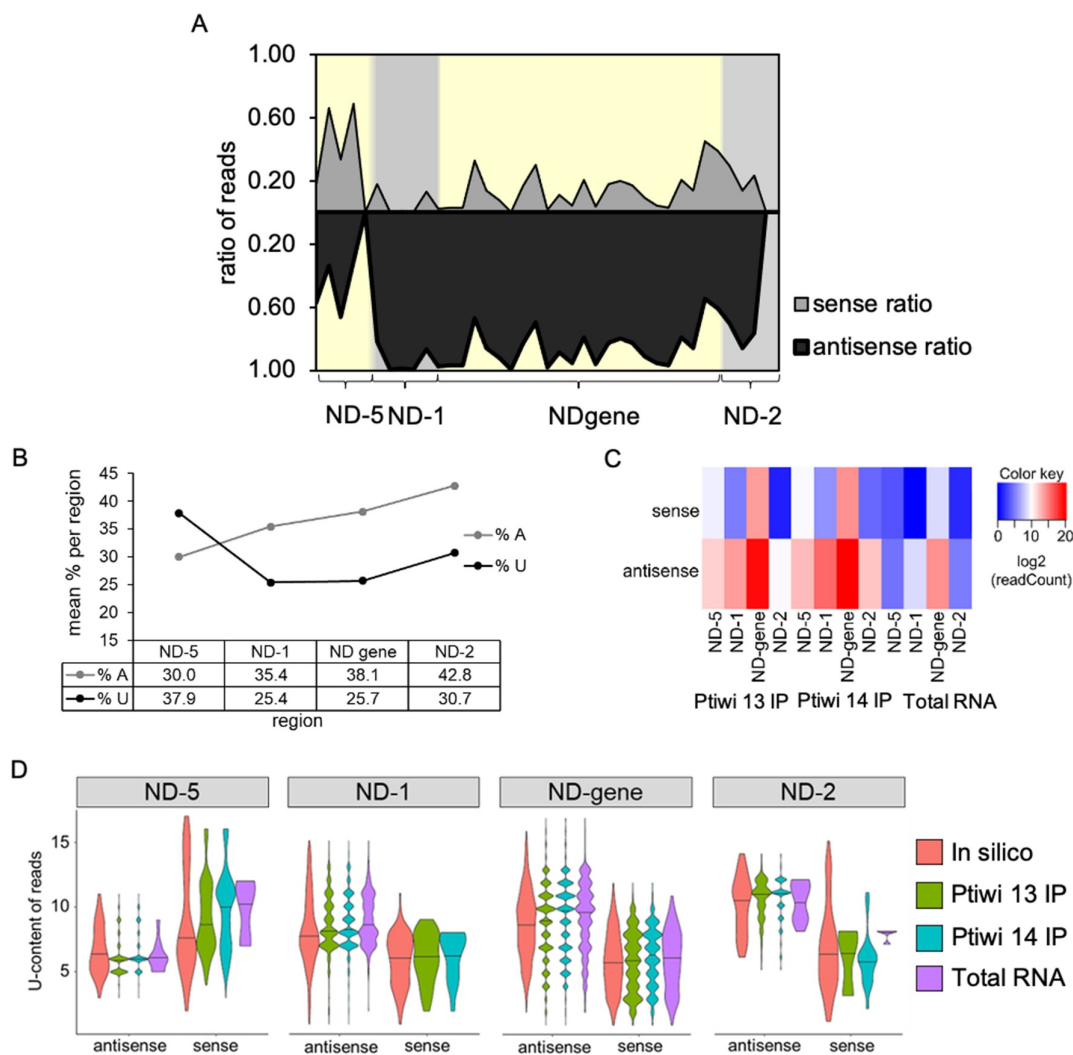


Figure 6. U-content analysis of 23nt sRNAs in Ptiwi IPs. **A:** Sense/antisense reads from pTI^{-/-} injected cells mapping to indicated regions in 50nt windows. **B:** Percentage of adenine and uridine of the sense RNA transcript of each region. **C:** Heatmap of reads mapping to the indicated regions separated by their direction. **D:** X axis shows U-content of reads found in Ptiwi IPs, while density represents number of reads. In silico data is generated by counting U-content of 23mers of the DNA sequence for each region.

and second for almost all regions for 1° and 2° , the U-content of Ptiwi IPed sRNAs is above *in silico* diced sRNAs. Thus, strand selection by Ptiwi13 and Ptiwi14 seems to include a general selection for uridine-rich sRNAs.

Transgene-induced silencing mimics endogenous siRNA accumulation

We finally had a closer look at endogenous siRNAs. We have recently described siRNA-producing loci in the *Paramecium* genome, showing read length preference of 23nt [55]. **Figure 7A** shows also for these endogenous clusters a predominant overlap of 21nt, indicating Dicer to be involved at least in the majority of them. Surprisingly, most of the endogenous clusters can be identified also in IPs of Ptiwi13 and 14 (**Fig. 7B, C** and **Fig S7**). As shown in **Fig S8**, those SRC-derived, Ptiwi-loaded siRNAs with high antisense ratios also show a higher

U-content in antisense reads. This again leads to the hypothesis of loading preferences of Ptiwis for uridine-rich siRNAs. **Supplement Fig. S9** shows that both Ptiwis load SRC small RNAs independent of the expression level of genes: as *Paramecium* SRCs correlate with both, silent and high expressed genes, the function of these sRNAs is hardly understood, but the Ptiwi-IP data here suggest that Ptiwis do not dissect between sRNAs from silent or high expressed genes. In these analyses, we will miss any trans-acting mechanisms as mapping with only one mismatch allowed will likely result in *cis* acting correlations only. It is likely that also *trans* actions of SRC-produced siRNAs could occur, which is difficult to analyse as we do not know about the target recognition of Ptiwi-bound sRNAs. Correlating gene expression level with sRNA antisense ratio of SRCs located in protein coding genes, we also cannot identify a correlation between mRNA and antisense ratio of small RNAs in SRCs (**Suppl. Fig S10**). Few silent genes, however, show a strict bias of siRNAs, which

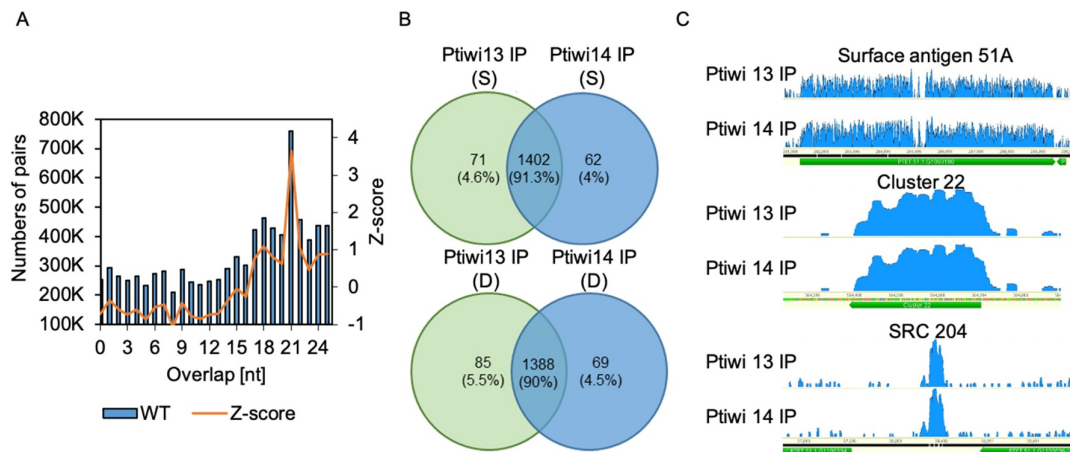


Figure 7. Endogenous sRNAs in Ptiwi IPs. A: Calculation of overlaps of endogenous 17–25nt sRNAs isolated from wildtype RNA including Z-score for overlapping pairs. B: sRNAs mapping to SRCs (small RNA clusters) in the *Paramecium* genome are analysed by their presence in Ptiwi IPs. Venn diagrams of the two IPs of sonified (S) and dounced (D) lysates with numbers of SRCs detected and the proportion of each fraction of the total found SRCs. C: Three examples of endogenous loci shown by coverage tracks of unnormalized data (surface antigen 51A, cluster22 and SRC 204).

could be examples of classical sRNA silenced genes but as this analysis was only possible for few genes (Suppl. Fig. S10). As a result of the comparison of transgene-associated silencing and endogenous small RNA producing loci, both appear to have common genetic requirements with endogenous siRNA accumulation pathways and appears therefore as a suitable model to study endogenous siRNA accumulation. Transgene-induced silencing appears therefore as a suitable model to study endogenous siRNA accumulation.

Discussion

The dissection between Agos and Ptiwis was originally not only based on sequence similarity but on their spatial and temporal activity in germline and somatic cells. The most important distinction between both was due to their action in strand selection: Agos load RNase III generated duplexes, whereas Ptiwis load longer ssRNA and generate their own sRNAs. Although many recent reports were published describing non-canonical functions of Piwis in somatic cells targeting non-transposable elements, the latter aspect of strand selection activity remains an important difference between the two groups.

Ciliates belong to the SAR-supergroup, which are equidistant to animals, plants and fungi [58] and especially *Paramecium* is different to many other species, not only metazoans, as the genome does not contain Agos but 15 Ptiwis [19]. Thus, *Paramecium* offers exciting possibilities for evolutionary comparison of RNAi mechanisms and the mode of action of individual components.

Transgene-associated siRNAs are Dicer products

We started this work here based on the surprising finding that two different Ptiwis are involved in a process where a non-expressible transgene silences a genetic locus at the chromatin level [19,22]. There are a couple of possible hypotheses why two distinct Ptiwis are necessary for this mechanism. First,

a logical idea would be the action of ping-pong amplification. Although this would have been supported because Ptiwi13 and Ptiwi14 likely have slicer activity, it is in conflict with the fact that Dicer1 is necessary to produce at least the 1° siRNAs (Fig S11). A second hypothesis would have been that both Ptiwis distinguish between 1° and 2° siRNAs. Our data clearly shows that both hypotheses are not true because we do not see any ping-pong signature but a 21nt overlap of reads, thus strongly suggesting that Dicer cuts at least the majority of siRNAs. Ptiwi IPs disprove also the second hypothesis because both Ptiwis load both 1° and 2° siRNAs, however, in different quantities. But what is the function of both Ptiwis then? Could they have the very same function being redundant? This seems not very likely because our data shows some discrete differences between Ptiwi13 and Ptiwi14: (i) Ptiwi13 loads more sRNAs from exogenous templates, (ii) both have different sub-cellular localization preferences, (iii) Ptiwi14 loads more 2° siRNAs and finally (iv) Ptiwi14 shows a much stronger preference for 5'-U RNAs. It seems therefore more likely that both Ptiwis have indeed distinct and specialized functions in this mechanism. Concerning the subcellular localization, our data resulting from IF, Western and *in silico* analysis support the idea that both Ptiwis can be found in both cytosol and MAC but with different preferences. Our data indicate to be more Ptiwi13 in the cytosol and more Ptiwi14 in the MAC. It seems likely that both are involved in a shuttling process between cytosol and nucleus, reminiscent of the nematode Ago NRDE-3, which localizes in the cytosol and redistributes to the nucleus when bound to 2° siRNAs from the feeding pathway [59]. Also *Arabidopsis* Ago4 assembles with siRNAs in the cytosol and is then transported in the nucleus [60]. Among these examples for nuclear import of sRNA loaded Ago/Piwis, this would make sense if we think about the trigger for transgene-induced silencing. It has been shown that explicitly non-expressible transgenes induce RNAi [28,33], tempting that a quality control mechanism is involved in this process to dissect which transgene can produce

translatable mRNA. Such processes, e.g. nonsense-mediated RNA decay, works in the cytosol as translation is an efficient way to dissect between defect and translatable mRNAs. As we have previously shown that transgene-induced silencing works on the chromatin level, this cytosolic signal needs to be transported into the nucleus.

Ptiwis select strands from dicer products

Our data indicate that both Ptiwis select for strand-specific sRNAs from Dicer cut duplexes, which represents a non-canonical function of Piwi proteins. This finding is fostered by several aspects. Dicer knockdown reduces all sRNAs [22,30], and in addition we have shown here that bulk transgene siRNAs show 21nt overlaps of 23nt RNAs. Our study also brings strand-asymmetry in association with individual properties of sRNAs, apparently contributing to strand selection and stabilization by Ptiwis: 5'-U preferences, U-content and 3'-methylation. 5'-U preferences have been frequently described, e.g. for the Piwi lacking *Arabidopsis* Ago1 [61]. It is also quite reminiscent to the strong 5'-U preference of *Drosophila* Piwi (and weaker in Aubergine), which act together with Ago3 in the ping-pong amplification of piRNAs [62]. However, we cannot identify any sRNAs with an A-preference at position 10, which would result from such a mechanism. 5'-Nucleotide preferences were also reported for the developmental Ptiwis in *Paramecium* showing a strong 5'-UNG signature [30,48,57]. Recent evidence from *in vitro* dicing experiments show that this signature is due to cleavage preference of the involved Dicer-like enzymes rather than due to preferential Ptiwi loading [63]. The authors speculate about a co-evolution of Dicer-like enzymes to produce sRNAs targeting germline-specific DNA with strong sequence bias at its ends. This seems likely for the particular need to target conserved sequence-ends, but for the control of endogenous gene expression, accumulation of such conserved sequence features in siRNAs would not make sense. Interestingly, the scnRNA mechanism holds another interesting aspect to discuss as it involves also a special kind of 2° sRNAs (iesRNAs) transcribed from already excised and circularized IESs. However, 1° and 2° developmental sRNAs have been shown to be loaded in distinct Ptiwis. This is contrary to our vegetative mechanism; this makes sense because scnRNAs and iesRNAs have different biogenesis mechanisms and different properties [57,63]. Vice versa, we may conclude from this that 1° and 2° siRNAs in transgene-induced silencing have identical biogenesis mechanisms, meaning that the same mechanism act on the transgene and the endogenous gene. Similar to transgene-associated Ptiwis, the knockdown of Ptiwi10 and Ptiwi09 during meiosis resulted in accumulation of duplexes and the authors concluded that these Ptiwis could be responsible for strand selection in reminiscence of Agos [48]. This is further supported by the involvement of Dicer/Dicerl-like proteins in the biogenesis of these two sRNA classes [57], thus indicating that the two developmental Piwi proteins act also more like Agos, which is similar to our finding here. Ciliates may in general use Piwis in an Ago manner, which is surprising not only for the vegetative Ptiwis described here but even more for the massive

elimination of transposons and transposon-derived sequences during development, which is the hall mark of Dicer-independent piRNA in other organisms.

Nucleotides as a biochemical reason and the resulting thermodynamic behaviour of a sRNA duplex are only individual aspects for strand selection. It has been demonstrated that many protein factors in addition to Agos contribute to strand selection, allowing for dynamic adaptation of the miRNA system in response to challenges to adapt gene expression [64]. Two aspects need to be taken into account: phosphorylation by *de novo* RDR initiation and availability of RISC targets. Concerning the latter, it has first been shown in plants that the alteration of the target RNA binding quality alters miRNA abundance [65]. It seems likely that such a parameter also contributes to the strand selection in our example, as for instance the ND-5 region still does not show a clear sense bias, which could be explained if indeed antisense strands are preferred due to available targets by a sense (m)RNA. As mentioned above, also phosphorylation of RDR transcripts may play a role in strand selection: *Tetrahymena* Dicer2 was shown to be physically coupled with the RDRC: *in vitro*, Dicer2 cleaves discrete siRNAs from the 5'-triphosphorylated ends of dsRNA only [66]. As also the transgene-induced silencing here employs RDR activity, cyclic and phased *de novo* RDR activity on the sense transcript could produce duplexes with triphosphorylated antisense ends. If Ptiwis would select for those, the RDR products would be preferentially loaded.

Conclusion

Our study is another evidence for the extreme diversification of small RNA amplification not only in ciliates and the data raises the question whether ciliates use their Piwis totally different to other species or whether this function of Piwis could be the primordial one and later split into Agos and Piwi mode of actions. Vice versa to the ciliate Piwis in strand selection, yeast Ago has been demonstrated to load single-stranded RNAs, which become trimmed into pri-RNAs which are Dicer independent [67]. This means that *S. pombe* Ago can also process longer ssRNA into functional small RNAs. Such a Piwi-like function may fit to the position of *S. pombe* Ago between mammalian Agos and Piwis in our phylogenetic reconstruction in Fig. 1A. This non-canonical Ago function in yeast together with our data in ciliate Piwi function, let us assume that Piwis and Agos in unicellular eukaryotes are more diverse in their activity than expected. One may hypothesize that yeast and ciliates owning only Agos or Piwis, respectively, use those in a more flexible way compared to species harbouring both Agos and Piwis. This flexibility along with the absence of miRNAs in ciliates [55] may also be compared with the highly conserved miRNA loading Agos in vertebrates, which show a much higher degree of conservation compared to siRNA loading Agos of e.g. nematodes and insects, which are still in an arms race with viral adaption [68]. Although this comparison of conserved miRNA loading Agos in mammals and evolutionary flexible antitransposon and antiviral Agos/Piwis in plants, nematodes and single-celled organisms make sense at first glance, recent studies also demonstrate antiviral RNAi in interferon deficient mammalian cells [69], so depending on the extent of this, also the

mammalian RNAi mechanisms still needs to adopt to new pathogens.

As until now, no ping-pong amplification has been demonstrated in any ciliate, further research has to clarify whether this is absent in ciliates at all. From the evolutionary point of view, Ago-like usage of Piwis seems surprising in unicellular eukaryotes as Agos have been demonstrated on bacteria already [70], although their function is less understood. Further studies need to clarify whether Agos may have been depleted in ciliates or, on the other hand, if strand selection of dsRNA by Piwis may have been the primordial function.

Acknowledgments

This work was supported by grants from the German research Council (DFG) to MS (SI1379/3-1), MHS SCHU (3140/1-1) and MJu MJ (CRC894). We are grateful to Dominique Furrer and Mariusz Nowacki for sharing RIP experience, Kaz Mochizuki for advice in Ptiwi IPs, and Megan Valentine and Judy VanHouten for sharing the FLAG fusion vector. Finally thanks to Christoph Kellner for scripts on U-counts and Anke Behnke for help in phylogenetic analyses.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by grants from the German research Council (DFG) to MS (SI1379/3-1), MHS SCHU (3140/1-1) and MJu MJ (CRC894).

References

- [1] Svoboda P. Introduction to RNAi and miRNA pathways. *Nakladatelstv Karolinum*. 2020
- [2] Bhattacharjee S, Roche B, Martienssen RA. RNA-induced initiation of transcriptional silencing (RITS) complex structure and function. *RNA Biol*. 2019;16(9):1133–1146.
- [3] Yamashiro H, Sioni MC. PIWI-interacting RNA in *Drosophila*: biogenesis, transposon regulation, and beyond. *Chem Rev*. 2017;118(8):4404–4421.
- [4] Czech B, Munafo M, Ciabrelli F, et al. piRNA-guided genome defense: from biogenesis to silencing. *Annu Rev Genet*. 2018;52:131–157.
- [5] Weick EM, Miska EA. piRNAs: from biogenesis to function. *Development*. 2014;141(18):3458–3471.
- [6] Czenik ES, Zamore PD. Argonaute proteins. *Curr Biol*. 2011;21(12):R446–R449.
- [7] Stein CB, Genzor P, Mitra S, et al. Decoding the 5' nucleotide bias of PIWI-interacting RNAs. *Nat Commun*. 2019;10(1):828.
- [8] Parhad SS, Theurkauf WE. Rapid evolution and conserved function of the piRNA pathway. *Royal Soc Open Biol*. 2019;9(1):180181
- [9] Obbard DJ, Gordon KH, Buck AH, et al. The evolution of RNAi as a defence against viruses and transposable elements. *Philos Trans R Soc B*. 2009;364(1513):99–115.
- [10] Blumenstiel JP, Erwin AA, Hemmer LW. Focus: epigenetics: what Drives Positive Selection in the *Drosophila* piRNA Machinery? The Genomic Autoimmunity Hypothesis. *Yale J Biol Med*. 2016;89(4):499.
- [11] Gu W, Shirayama M, Conte JD, et al. Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. *Mol cell*. 2009;36(2):231–244.
- [12] Sarkies P, Selkirk ME, Jones JT, et al. Ancient and novel small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages. *PLoS Biol*. 2015;13(2):e1002061.
- [13] Ross RJ, Weiner MM, Lin H. PIWI proteins and PIWI-interacting RNAs in the soma. *Nature*. 2014;505(7483):353.
- [14] Rojas-Ros P, Simonelig M. piRNAs and PIWI proteins: regulators of gene expression in development and stem cells. *Development*. 2018;145(17):dev161786.
- [15] Tosar JP, Rovira C, Cayota A. Non-coding RNA fragments account for the majority of annotated piRNAs expressed in somatic non-gonadal tissues. *Commun Biol*. 2018;1(1):1–8.
- [16] Sarkar A, Volff JN, Vaury C. piRNAs and their diverse roles: a transposable element-driven tactic for gene regulation? *FASEB J*. 2017;31(2):436–446.
- [17] Lewis SH, Quarles KA, Yang Y, et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. *Nat Ecol Evol*. 2018;2(1):174–181.
- [18] Nekrasova IV, Potekhin AA. Diversity of RNA interference pathways in regulation of endogenous and exogenous sequences expression in ciliates *Tetrahymena* and *Paramecium*. *Ecol Gene*. 2019;17(2):113–125
- [19] 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):4249–4264.
- [20] Cheng CY, Orias E, Leu JY, et al. The evolution of germ–soma nuclear differentiation in eukaryotic unicells. *Curr Biol*. 2020;30(10):R502–R510.
- [21] Ruiz F, Vayssié L, Klotz C, et al. Homology-dependent gene silencing in *Paramecium*. *Mol Biol Cell*. 1998;9(4):931–943.
- [22] Götz U, Marker S, Cheaib M, et al. Two sets of RNAi components are required for heterochromatin formation in trans triggered by truncated transgenes. *Nucleic Acids Res*. 2016;44(12):5908–5923.
- [23] Rajeev Kumar S, Anunanthini P, Ramalingam S. Epigenetic silencing in transgenic plants. *Front Plant Sci*. 2015;6:693.
- [24] Hollick JB. Paramutation and related phenomena in diverse species. *Nat Rev Genet*. 2017;18(1):5.
- [25] Hermant C, Boivin A, Teyssset L, et al. Paramutation in *Drosophila* requires both nuclear and cytoplasmic actors of the piRNA pathway and induces cis-spreading of piRNA production. *Genetics*. 2015;201(4):1381–1396.
- [26] Sapetschnig A, Sarkies P, Lehrbach NJ, et al. Tertiary siRNAs mediate paramutation in *C. elegans*. *PLoS Genet*. 2015;11(3):e1005078.
- [27] Galvani A, Sperling L. Transgene-mediated post-transcriptional gene silencing is inhibited by 3' non-coding sequences in *Paramecium*. *Nucleic Acids Res*. 2001;29(21):4387–4394.
- [28] Garnier O, Serrano V, Duharcourt S, et al. RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. *Mol Cell Biol*. 2004;24(17):7370–7379.
- [29] De Vanssay A, Bougé AL, Boivin A, et al. Paramutation in *Drosophila* linked to emergence of a piRNA-producing locus. *Nature*. 2012;490(7418):112–115.
- [30] Lepere G, Nowacki M, Serrano V, et al. Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. *Nucleic Acids Res*. 2008;37(3):903–915.
- [31] Cheaib M, Dehghani Amirabad A, Nordström KJ, et al. Epigenetic regulation of serotype expression antagonizes transcriptome dynamics in *Paramecium tetraurelia*. *DNA Res*. 2015;22(4):293–305.
- [32] Simon MC, Marker S, Schmidt HJ. Inefficient serotype knock down leads to stable coexistence of different surface antigens on the outer membrane in *Paramecium tetraurelia*. *Eur J Protistol*. 2006;42(1):49–53.
- [33] Galvani A, Sperling L. RNA interference by feeding in *Paramecium*. *Trends Genet*. 2002;18(1):11–12.
- [34] Pirritano M, Götz U, Karunanithi S, et al. Environmental temperature controls accumulation of transacting siRNAs involved in heterochromatin formation. *Genes (Basel)*. 2018;9(2):117.
- [35] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4(4):406–425.

- [36] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *evolution*. 1985;39(4):783–791.
- [37] Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: *Evolving genes and proteins*. New York: Elsevier; 1965. p. 97–166.
- [38] Kumar S, Stecher G, Li M, et al. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35(6):1547.
- [39] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*. 2011;17(1):10–12.
- [40] Karunanithi S, Simon M, Schulz MH. Automated analysis of small RNA datasets with RAPID. *PeerJ*. 2019;7:e6710.
- [41] Crooks GE, Hon G, Chandonia JM, et al. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188–1190.
- [42] Antoniewski C. Computing siRNA and piRNA overlap signatures. *Methods Mol Biol*. 2014;1173:135–146.
- [43] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
- [44] Klöppel C, Müller A, Marker S, et al. Two isoforms of eukaryotic phospholipase C in *Paramecium* affecting transport and release of GPI-anchored proteins in vivo. *Eur J Cell Biol*. 2009;88(10):577–592.
- [45] Frapporti A, Pina CM, Arnaiz O, et al. The Polycomb protein Ezh1 mediates H3K9 and H3K27 methylation to repress transposable elements in *Paramecium*. *Nat Commun*. 2019;10(1):1–15.
- [46] Valentine MS, Rajendran A, Yano J, et al. *Paramecium* BBS genes are key to presence of channels in cilia. *Cilia*. 2012;1(1):16.
- [47] Preer LB, Hamilton G, Preer JRJR. Micronuclear DNA from *Paramecium tetraurelia*: serotype 51 A gene has internally eliminated sequences. *J Protozool*. 1992;39(6):678–682.
- [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):505–520.
- [49] Aury JM, Jaillon O, Duret L, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*. 2006;444(7116):171–178.
- [50] 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(suppl_1):D632–D636.
- [51] Jenkins BH, Maguire F, Leonard G, et al. Characterization of the RNA-interference pathway as a tool for reverse genetic analysis in the nascent phototrophic endosymbiosis, *Paramecium bursaria*. *R Soc Open Sci*. 2021;8(4):210140.
- [52] Nakanishi K, Ascano M, Gogakos T, et al. Eukaryote-specific insertion elements control human ARGONAUTE slicer activity. *Cell Rep*. 2013;3(6):1893–1900.
- [53] King BR, Vural S, Pandey S, et al. ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes. *BMC Res Notes*. 2012;5(1):351.
- [54] Carradec Q, Götz U, Arnaiz O, et al. Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate *Paramecium tetraurelia*. *Nucleic Acids Res*. 2015;43(3):1818–1833.
- [55] Karunanithi S, Oruganti V, Marker S, et al. Exogenous RNAi mechanisms contribute to transcriptome adaptation by phased siRNA clusters in *Paramecium*. *Nucleic Acids Res*. 2019;47(15):8036–8049.
- [56] Ji L, Chen X. Regulation of small RNA stability: methylation and beyond. *Cell Res*. 2012;22(4):624–636.
- [57] 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):174–188.
- [58] Keeling PJ, Burger G, Durnford DG, et al. The tree of eukaryotes. *Trends Ecol Evol*. 2005;20(12):670–676.
- [59] Guang S, Bochner AF, Pavelec DM, et al. An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science*. 2008;321(5888):537–541.
- [60] Ye R, Wang W, Iki T, et al. Cytoplasmic assembly and selective nuclear import of *Arabidopsis* Argonaute4/siRNA complexes. *Mol Cell*. 2012;46(6):859–870.
- [61] Mi S, Cai T, Hu Y, et al. Sorting of small RNAs into *Arabidopsis* argonaute complexes is directed by the 5' terminal nucleotide. *Cell*. 2008;133(1):116–127.
- [62] Brennecke J, Aravin AA, Stark A, et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*. 2007;128(6):1089–1103.
- [63] Hoehener C, Hug I, Nowacki M. Dicer-like enzymes with sequence cleavage preferences. *Cell*. 2018;173(1):234–247.
- [64] Meijer HA, Smith EM, Bushell M. Regulation of miRNA strand selection: follow the leader? *Biochem Soc Trans*. 2014;42(4):1135–1140.
- [65] Todesco M, Rubio-Somoza I, Paz-Ares J, et al. A collection of target mimics for comprehensive analysis of microRNA function in *Arabidopsis thaliana*. *PLoS Genet*. 2010;6(7). doi:10.1371/journal.pgen.1001031
- [66] Lee SR, Collins K. Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs. *Nat Struct Mol Biol*. 2007;14(7):604–610.
- [67] Marasovico. Argonaute and Triman generate dicer-independent priRNAs and mature siRNAs to initiate heterochromatin formation. *Mol Cell*. 2013;52(2):173–183.
- [68] Wynnant N, Santos D, Broeck JV. The evolution of animal Argonautes: evidence for the absence of antiviral AGO Argonautes in vertebrates. *Sci Rep*. 2017;7(1):1–13.
- [69] Maillard PV, Van der Veen AG, Deddouche-Grass S, et al. Inactivation of the type I interferon pathway reveals long double-stranded RNA-mediated RNA interference in mammalian cells. *EMBO J*. 2016;35(23):2505–2518.
- [70] Jinek M, Doudna JA. A three-dimensional view of the molecular machinery of RNA interference. *Nature*. 2009;457(7228):405–412.

Paramecium epigenetics in development and proliferation

Franziska Drews¹ | Jens Boenigk^{2,3} | Martin Simon¹ 

¹Molecular Cell Biology and Microbiology, School of Mathematics and Natural Sciences, University of Wuppertal, Wuppertal, Germany

²Centre for Water and Environmental Research (ZWU), University of Duisburg-Essen, Essen, Germany

³Biodiversity, University of Duisburg-Essen, Duisburg, Germany

Correspondence

Martin Simon, Molecular Cell Biology and Microbiology, School of Mathematics and Natural Sciences, University of Wuppertal, Wuppertal, Germany.
Email: masimon@uni-wuppertal.de

Funding information

DFG, Grant/Award Number: SII397/3-1

Abstract

The term epigenetics is used for any layer of genetic information aside from the DNA base-sequence information. Mammalian epigenetic research increased our understanding of chromatin dynamics in terms of cytosine methylation and histone modification during differentiation, aging, and disease. Instead, ciliate epigenetics focused more on small RNA-mediated effects. On the one hand, these do concern the transport of RNA from parental to daughter nuclei, representing a regulated transfer of epigenetic information across generations. On the other hand, studies of *Paramecium*, *Tetrahymena*, *Oxytricha*, and *Stylonychia* revealed an almost unique function of transgenerational RNA. Rather than solely controlling chromatin dynamics, they control sexual progeny's DNA content quantitatively and qualitatively. Thus epigenetics seems to control genetics, at least genetics of the vegetative macronucleus. This combination offers ciliates, in particular, an epigenetically controlled genetic variability. This review summarizes the epigenetic mechanisms that contribute to macronuclear heterogeneity and relates these to nuclear dimorphism. This system's adaptive and evolutionary possibilities raise the critical question of whether such a system is limited to unicellular organisms or binuclear cells. We discuss here the relevance of ciliate genetics and epigenetics to multicellular organisms.

KEYWORDS

ciliate, genome rearrangement, RNA, RNA interference, transgenerational inheritance

CILIATES, particularly *Paramecium*, served as model organisms in genetics and epigenetics long before the latter term was even used for the first time. From the historical point of view, ciliate genetics had its first's heydays from 1940 to 1960, when many important discoveries were made, resulting in a detailed description of epigenetic phenomena. As a result, most textbooks for undergraduates dedicated individual chapters to ciliate cell biology and genetics in the 1970s, but these chapters disappeared from textbooks in modern times (Preer, 1997).

This situation has now changed (Boenigk, 2021). Ciliate epigenetic research experiences a renaissance, although this wording might not be entirely precise as research is not simply making a replica of the former work. We are now able to describe epigenetic phenomena discovered phenotypically in the early 1940s on the molecular level and identify small RNAs, histone modifications, and DNA modifications that are responsible for these phenomena. Still, genetics research is primarily based on yeast, *C. elegans*, *Drosophila*, zebrafish, and

[Correction added on 30 May 2022, after first online publication: Projekt DEAL funding statement has been added.]

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Journal of Eukaryotic Microbiology* published by Wiley Periodicals LLC on behalf of International Society of Protistologists.

mouse models. Ciliate genetics is quite different at first glance. The occurrence of two different kinds of nuclei, that is, somatic and germline genome, makes ciliates behave like a germ cell and like a somatic cell simultaneously. This aspect complicates many research fields; for instance, it's still complicated to isolate pure micronucleus DNA from *Paramecium* (Guérin et al., 2017), which is why genomics analysis of the germline genome started relatively late. In this review, we want to introduce the genetics and epigenetics of ciliates and, in particular of *Paramecium*. Genetics and epigenetics are indeed closely related in this species as, for instance, the DNA content of the somatic macronucleus is regulated by different epigenetic mechanisms. We aim to describe nuclear dimorphism from an evolutionary point of view, which makes this unusual feature seem even more interesting.

NUCLEAR DIMORPHISM: TWO DISTINCT GENOMES IN A SINGLE CELL

Multicellular organisms can be dissected into somatic tissues and germ cells, reflecting different requirements. Germ cells store genetic information without dynamic gene expression and low epigenetic predetermination while DNA is protected from damage. Somatic cells, in contrast, arise from epigenetically variable cells to differentiate into different tissues forming different characteristic transcriptomic states while still harboring a certain degree of transcriptional dynamics to react to changing environments. Transcriptional activity in somatic cells makes DNA susceptible to damage. Thus, the germline represents a protected backup of an individual's genetic information.

Ciliates belong to the SAR clade, a supergroup consisting of Rhizaria, Alveolata, and Stramenopiles (Burki et al., 2020). Especially the Alveolata, including the ciliate phylum, have been described as morphological and ecologically diverse; currently, ~8000 ciliate species have been described among 11 classes (Adl et al., 2012; Grattepanche et al., 2018). Ciliates show a separation of germline and somatic genomes, but within a single cell. They contain one or more germline micronuclei and one or more somatic macronuclei. Both types of nuclei are drastically different in terms of their genomic and epigenetic content and their mode of division.

Micronuclei (MICs) are diploid and contain chromosomes that show similarities to those in metazoans as they are large and have centromeres, telomeres, and transposable elements (Figure 1A). Micronuclei are silent and consist of condensed chromatin. During asexual vegetative cell divisions, they divide by classical mitosis, in which spindle formation by microtubules guarantees for controlled segregation of chromosomes.

Sexual recombination in ciliates occurs either by a self-fertilization process called autogamy or by mating

two cells of compatible mating types with reciprocal exchange of meiotic nuclei deriving from the micronuclei (Sonneborn, 1937) (Figure 1B). Ciliate meiosis seems to be a bit less complex than in other organisms. The lack of synaptonemal complexes indicates a reduced capacity for recombination regulation, which may be partly counterbalanced by a substantial elongation of meiotic nuclei, possibly substituting for a physical linkage of homologs (Loidl, 2021). Figure 1B illustrates that the fusion of two identical gametes during autogamy leads to isozygotic progeny, being homozygous for all genetic loci, whereas conjugation leads to heterozygous individuals. The parental macronucleus fragments into the so-called "skein" stage in both processes. The fragments of the old macronucleus remain in the cells, still being actively transcribed for the following few cells divisions before they are eventually diluted out.

Macronuclei (MACs) differ strongly from micronuclei. MACs are actively transcribed, thereby regulating the vegetative cell metabolism. They are devoid of classical heterochromatin, and recent chromatin analyses of nucleosome occupancy suggest that DNA is highly accessible (Drews et al., 2022). MAC chromosomes show different polyploidy levels, ranging from classical diploid macronuclei in the Karyorelictea up to ~15,000n in Spirotrichea. MAC chromosomes are much shorter than micronucleus chromosomes as they are fragmented versions of those. This high degree of polyploidy and the short length of the MAC chromosomes means that the ciliate MAC contains hundreds of times more chromosome ends, and therefore telomeres, compared to mammalian cells. Not surprisingly, some basic elucidations of chromosome structure are based on ciliate model organisms. This comprises the first description and identification of telomere structure and telomerase activity in *Tetrahymena* (Blackburn & Gall, 1978; Greider & Blackburn, 1985) and the exciting finding that DNA replication and cell division in *Paramecium* does not include any telomere shortening (Gilley & Blackburn, 1994).

The division of the MACs deviates from that of the MICs: Most ciliates undergo amitotic MAC divisions, which means that the MAC elongates and chromosomes are distributed by chance. The MAC S-Phase itself is relatively long compared to the MIC, and DNA is amplified during more than 50% of the cell cycle (Berger, 1988). As heterochromatin is unknown in MACs, chromosomes are not condensed during amitosis, and transcription continues during the division. This may accelerate the cell division process and enhance the maximal division rate.

The segregation of MAC chromosomes is random, conflicting with the need for precise gene dosages required for regulated cell metabolism. Random segregation may be compensated by the high degree of polyploidy as strong unbalances of MAC chromosomes may only occur with exceptionally high numbers of cell divisions. Further, analyses of the DNA content in *Tetrahymena*

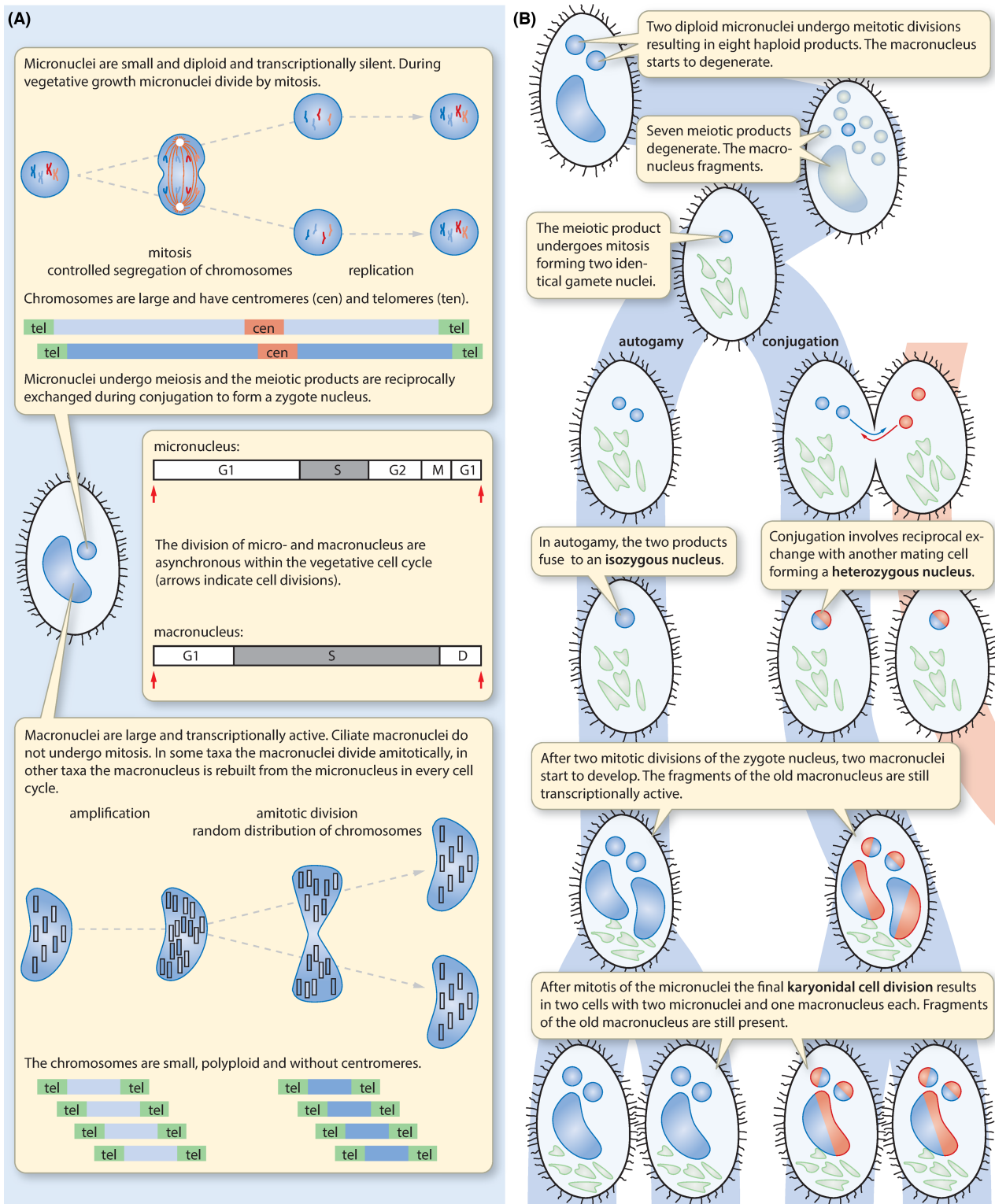


FIGURE 1 (A) Difference between *Paramecium* MIC and MAC division and chromosomal structure; (B) Autogamy and Conjugation in *Paramecium aurelia*

suggest mechanisms accounting for the maintenance of copy numbers (Doerder & DeBault, 1978). For instance, transgenes injected into the *Paramecium* macronucleus are autonomously replicated (Gilley et al., 1988; Godiska et al., 1987). Further studies demonstrated that

transgenes, either at low or high copy numbers, maintain their ploidy during amitotic divisions (Garnier et al., 2004; Götz et al., 2016). The molecular mechanism behind this copy number maintenance remains to be analyzed.

Apart from the total copy number of an individual MAC chromosome, allele frequency in heterozygous cells also develops strikingly differently. Imbalances of homologous chromosomes can occur by unequal distribution of chromosomes during amitosis (Figure 2). With increasing numbers of mitotic cell divisions, the macronuclei of daughter cells can become increasingly homozygous while the micronuclei remain heterozygous. This phenotypic assortment is best studied in *Tetrahymena* (Orias & Flacks, 1975). Exclusively stochastic distribution of alleles would rarely lead to a homozygous MAC and require many cell divisions. This is different when assuming selection pressure to act on the emerging new individuals: for instance, clones carrying high numbers of alleles with a germline-encoded mutation, or a MAC nuclear genome variant as described below, will be negatively selected while clones having low numbers of such variations will be positively selected (Maurer-Alcalá & Nowacki, 2019). This can be interpreted as a natural somatic selection and should be lost after the subsequent sexual recombination. However, the genome architecture of ciliates may limit this effect: on the one hand, species with high levels of polyploidy distribute many more gene copies. Thus, a complete phenotypic assortment becomes unlikely with increasing levels of ploidy. While ciliates with gene-sized nanochromosomes (*Stylonychia*, *Oxytricha*) could indeed counter-select individual alleles, this may be of minor importance in species with several thousand genes per MAC chromosome. Selection will affect a potentially deleterious allele and all other alleles on the same chromosome. Such MAC chromosomes carry thousands of genes, but only a few adverse mutations may be tolerated at low copy numbers

to account for the heterozygous genotype of other alleles on this chromosome (Maurer-Alcalá & Nowacki, 2019). *Paramecium*, however, shows a phenomenon of heterogenous MAC chromosomes. Recent data suggest that artificially induced deletions in some of the MAC chromosomes can segregate into phenotypically wild-type and mutant lines with increasing amitotic divisions: thus, phenotypic assortments may also occur on variants of large chromosomes in *Paramecium* (Nekrasova et al., 2019).

DIVIDING AND NONDIVIDING MACs

The current phylogeny affiliates all ciliates with one of three clades: Intramacronucleata, Heterotrichea, and Karyorelictea. These three clades also differ systematically by the MAC cell division processes. There is no single report of MAC mitosis in any ciliate species nor any report of chromatin condensation during amitotic divisions.

The Intramacronucleata show amitotic MAC divisions with the help of intranuclear microtubules, which assist the MAC to elongate and divide but do not form any classical spindle for chromosome segregation (Tucker et al., 1980). Most species studied on the molecular level to date belong to two classes of the Intramacronucleata, the Oligohymenophorea (*Paramecium*, *Tetrahymena*) and the Spirotrichea (*Oxytricha*, *Stylonychia*, and *Euplotes*) (Katz, 2001).

Although members of the Heterotrichea also show amitotic MAC divisions, these differ mechanistically

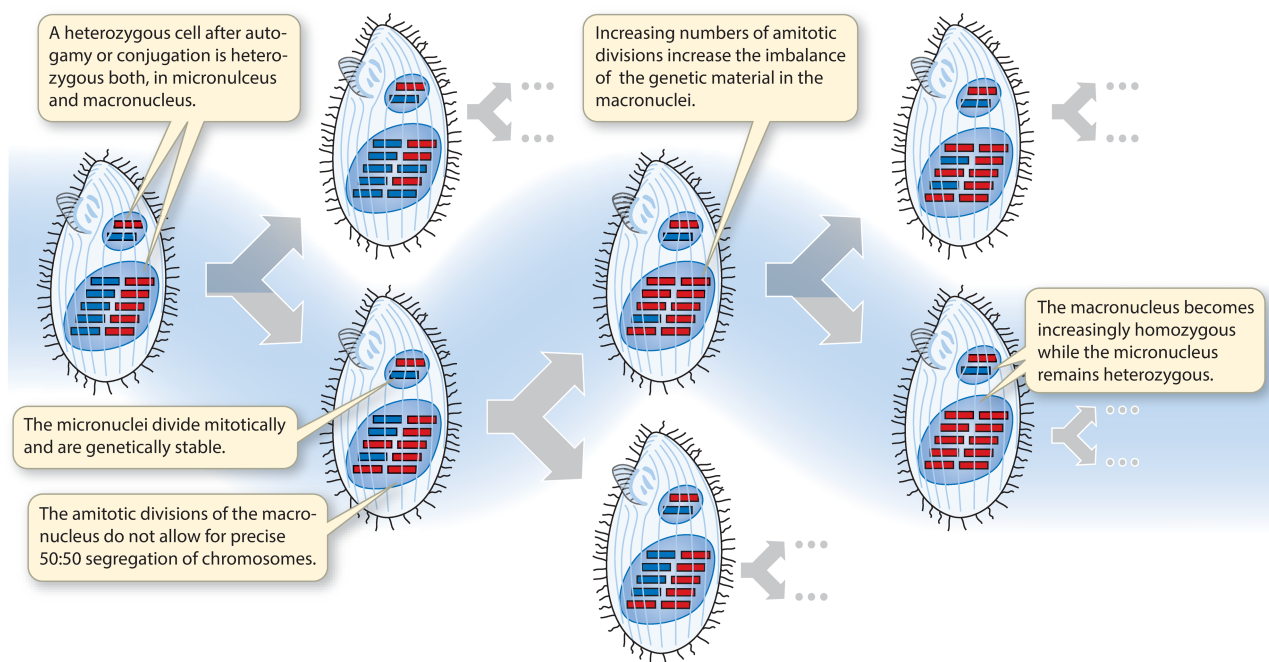


FIGURE 2 Phenotypic assortment during vegetative fissions illustrated in *Tetrahymena*

from those of the Intramacronucleata. Microtubules here control amitosis outside the MAC in parallel orientation to elongation (Jenkins, 1977). As the localization of microtubules is different in both clades, it has been speculated that the ability for amitosis evolved independently at other times (Herrick, 1994; Orias, 1991a, 1991b).

In this context, the third ciliate clade, the Karyorelictea, also needs consideration. Igor B. Raikov extensively studied Karyorelictea (Raikov, 1985). Members of this clade differ from Intramacronucleata and Heterotrichea. They do not show any MAC divisions and have a relatively low DNA content, implying a ploidy similar to the diploid MICs (Kovaleva & Raikov, 1978; Ovchinnikova & Selivanova, 1965). Their inability to divide requires the generation of new MACs from MICs with every cell division, not only linked to sexual events like conjugation or autogamy as in other ciliates (Figure 3).

The current phylogenetic data suggest that Karyorelictea, which have nondividing MACs, and Heterotrichea, exhibiting amitotic MACs, together form a sister group to all other ciliates (Baroin-Tourancheau et al., 1992; Katz, 2001). As such, the Karyorelictea show a closer phylogenetically relationship to Heterotrichea and are separated from the Intramacronucleata, which further supports the hypothesis that amitosis evolved at least twice.

Most Karyorelictea indeed possess numerous MACs and MICs, while an individual MAC's life span of three to seven cell divisions is controlled by so far unknown mechanisms (Yan et al., 2017). Karyorelict nuclei form nuclear groups consisting most frequently of one MIC and two MACs, for example, *Loxodes rostrum* has a

single group of one MIC and two MACs in a single cell. In other species, nuclear groups can occur multiple times so that many MICs and MACs can co-exist in a cell. Among the Karyorelictea, the nuclear grouping (number of MICs and MACs per group) and the total number of nuclei vary considerably between species within the Loxodidae.

We cannot assess the extent of genome rearrangements occurring during MAC differentiation from MICs. Based on the decrease of chromatin granules during MAC development, Raikov suggested that DNA elimination may appear to some extent (Raikov, 1994). He described the MACs as "paradiploid," which implies a minimal degree of replication only. On the other hand, an increase of DNA content was related to aging macronuclei in *Loxodes magnus* (Raikov et al., 1963). To compare the mechanisms of MAC development, genome data are necessary to analyze which DNA sequences undergo amplification and which DNA elements from the MIC may be eliminated. However, Raikov's data clearly show differences in Karyorelict MACs in DNA processing and, most strikingly, in the lack of MAC divisions which let him hypothesize that nondividing MACs might have been the ancestral state of an early nuclear dimorphic ciliate (Raikov, 1976, 1982).

EVOLUTION OF THE NUCLEAR DIMORPHISM AND AMITOSIS

Speculating about the evolutionary scenario that may have evolved the recent ciliate clades, we start with an

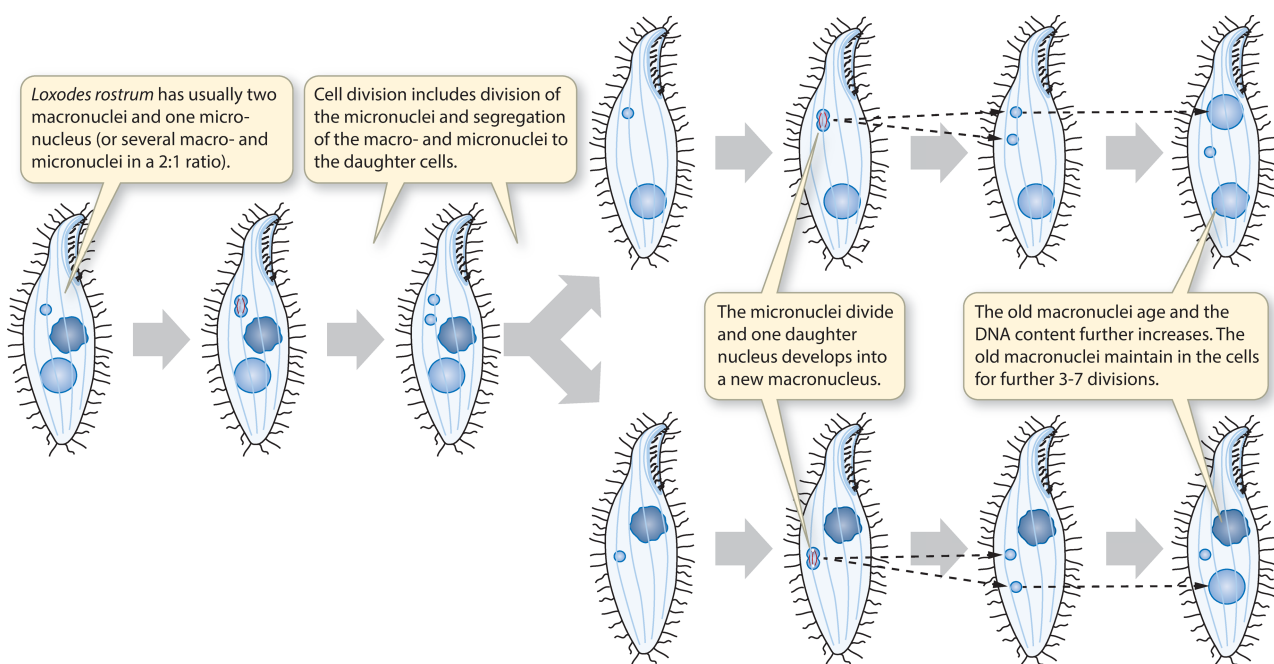


FIGURE 3 Vegetative cell divisions in Karyorelictea with nondividing MACs require generating a new MAC in every cell division. Please note that the figure shows on one nuclear set (1MAC + 2MICs); most Karyorelictea consist of multiple of these sets

ancestral eukaryotic protist, with a single nucleus undergoing mitotic divisions and capable of meiotic reductions (Figure 4).

It seems unlikely that a mononuclear cell evolved genome rearrangements and high polyploidy levels,

as the latter would be problematic in meiotic chromosome pairing and segregation. It seems more likely that an early event was the occurrence of two distinct nuclei, maybe by erroneous mitosis of a zygotic nucleus without cell division (Cheng et al., 2020).

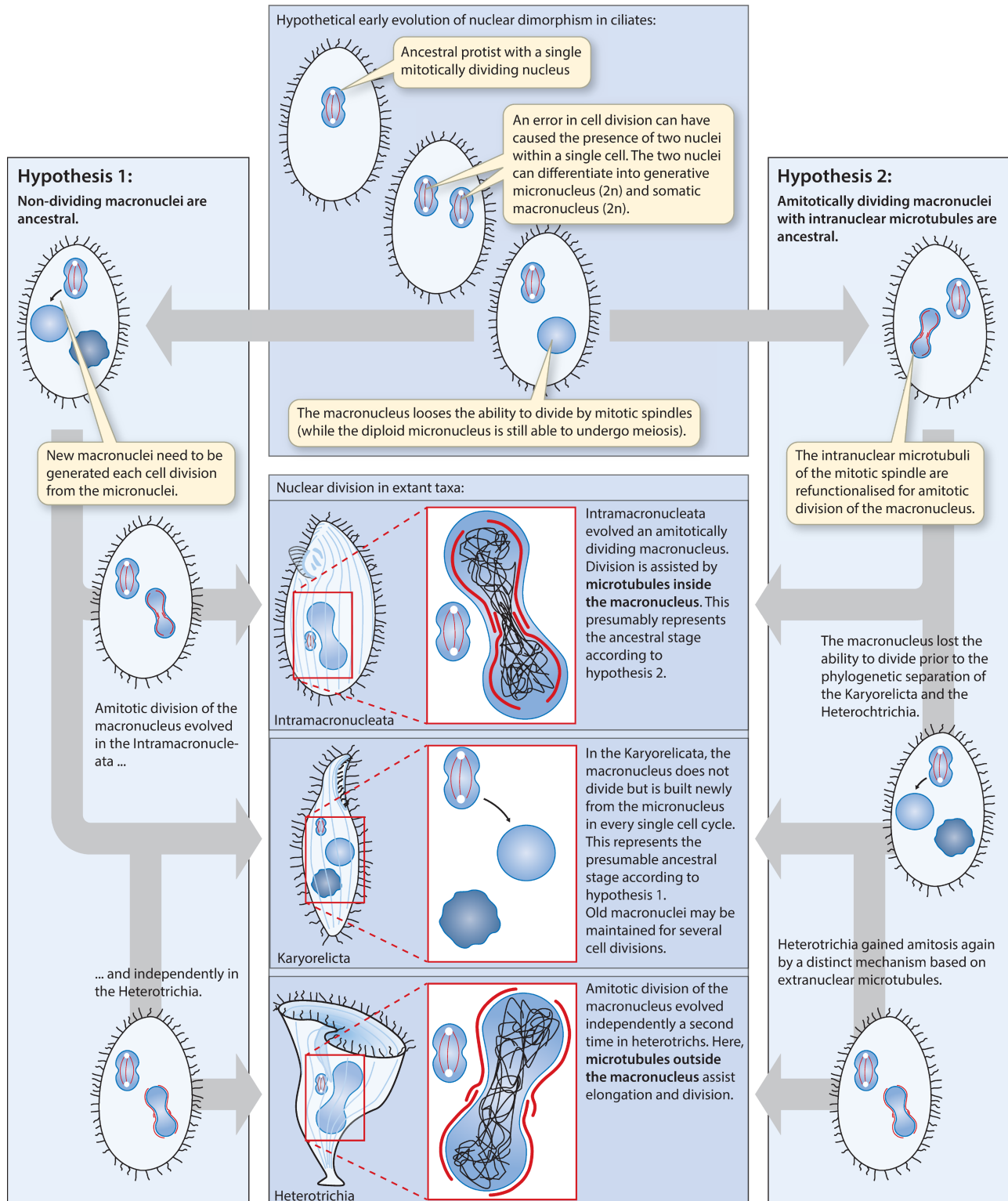


FIGURE 4 Hypothetical models for the evolution of recent ciliates based on the phylogenetic relationship of the three clades Intramacronucleata, Karyorelictea, and Heterotrichia

Soon after the occurrence of the second nucleus, differentiation of both germline and somatic nucleus may have occurred: the advantage is likely the same reason which favors germline/soma differentiation in metazoans: a protected silent germline whereas the soma is actively transcribed but sensitive for environmentally induced mutations. This could imply that the heterochromatic nature of the MIC and the euchromatic nature of the MAC were manifested soon after nuclear dualism occurred.

The next step could have been the loss of mitosis of the MAC. There are plenty of ideas for why this may have occurred. Amitotic dividing MACs show two characteristics: polyploidy and the failure to condensate chromosomes by heterochromatin. The latter would be a prerequisite for mitosis, whereas a certain polyploidy level might still be feasible for mitotic divisions. Polyploidy, in general, would have been of great advantage as gene dosage is the easiest way to increase protein levels. Thus, it would also be a conceivable scenario to increase cell size. As higher polyploidy levels are problematic for the spindle apparatus, the gain of polyploidy may have caused the loss of mitotic divisions.

There are two competing hypotheses for the evolution of ciliate MACs, which are mainly based on the ideas of Hammerschmidt et al. (1996), later outlined by Katz (2001) and Cheng et al. (2020). The first hypothesis outlines that the loss of mitotic division led to a Karyorelict-like ancestral ciliate (Hypothesis 1, Figure 4 left). As a consequence of this scenario, amitotic divisions would have been invented two times independently of each other, first by intranuclear microtubules in the Intramacronucleata and later by extranuclear microtubules in Heterotrichea. The second hypothesis suggests that the ancestral ciliate is more similar to recent Intramacronucleata: intranuclear spindle-microtubules could have been re-functionalized directly to assist the elongation of the MAC by intranuclear fibers. In this scenario, the situation in Karyorelicta is interpreted as derived, that is, amitotic dividing MACs would have lost their ability to divide to a Karyorelict-similar ancestor. Later, the Heterotrichea would have evolved from this Karyorelict-similar ancestor by the neofunctionalization of extranuclear microtubules.

A driving force for the evolution of polyploidy may be its ability to compensate for the increased probability of losing genetic material in the uncontrolled segregation of the amitotic MAC. At first glance, polyploidy seems to be linked to amitotic division. This would be right if Raikov's description of paradiplod MACs in the Karyorelictea would be accurate (Kovaleva & Raikov, 1978; Raikov & Karadzhan, 1985). Some recent single-cell genome data from *Loxodes* shows that different MAC chromosomes indeed have different amplification levels (Maurer-Alcalá et al., 2018). This could indicate an individual and controlled chromosome amplification or the effect of aging MACs with biased or uncontrolled

amplification. It seems clear that we need more molecular data in Karyorelictea MACs, also individual MACs to analyze, on the one hand, the degree of genome rearrangements and amplification.

One major conclusion of these evolutionary scenarios is that the amitotic division is not a primitive version of mitosis: the fact that this evolved independently twice clarifies that amitosis is crucial for ciliate genetics. Independent of whether polyploidy was the reason to evolve amitosis or vice versa, it seems clear that phenotypic assortments depend on amitotic divisions and polyploidy. Both together could provide a powerful mechanism for adaptation.

To understand the evolution of the recent ciliate clades and weigh the above hypotheses, we need to consider current genome rearrangements. Several hypotheses proposed that the evolution of the nuclear dimorphism occurred with the mechanisms in which the parental MAC affects the development of the new MAC (Bracht et al., 2013; Katz, 2001; Klobutcher & Herrick, 1997). These epigenetic mechanisms occurring during MAC development allow the transfer of information from the parental MAC to the new one. This is mediated by non-coding RNAs, resulting in the MAC genome sequence variability. The following chapters will introduce these rearrangements and their epigenetic control.

GENOME REARRANGEMENTS DURING MAC DEVELOPMENT

The genome structures between MIC and MAC chromosomes differ considerably. One can summarize the events contributing to the development of MAC chromosomes from MIC chromosomes by (i) DNA elimination, (ii) chromosome fragmentation, (iii) unscrambling, and (iv) amplification (Figure 5). The DNA elimination process involves the removal of a large fraction of the MIC genome in the form of transposable elements (TEs), repetitive DNA such as microsatellites, and internal eliminated sequences (IESs) (Figure 5A). The extent of elimination depends on the species, for example, 25% of the MIC genome is eliminated in *Paramecium*, 34% in *Tetrahymena*, and a massive wave of DNA elimination eliminates 95% of the *Oxytricha* MIC genome (Allen & Nowacki, 2020).

DNA elimination of IESs

Figure 5B compares the elimination and fragmentation events between *Paramecium* and *Tetrahymena*. IESs are short elements removed from the MIC DNA to create functional MAC chromosomes. In *Paramecium*, IESs are precisely eliminated. This is indeed necessary due to their localization inside and outside coding genes (Arnaiz et al., 2012). IESs of *Tetrahymena* instead are

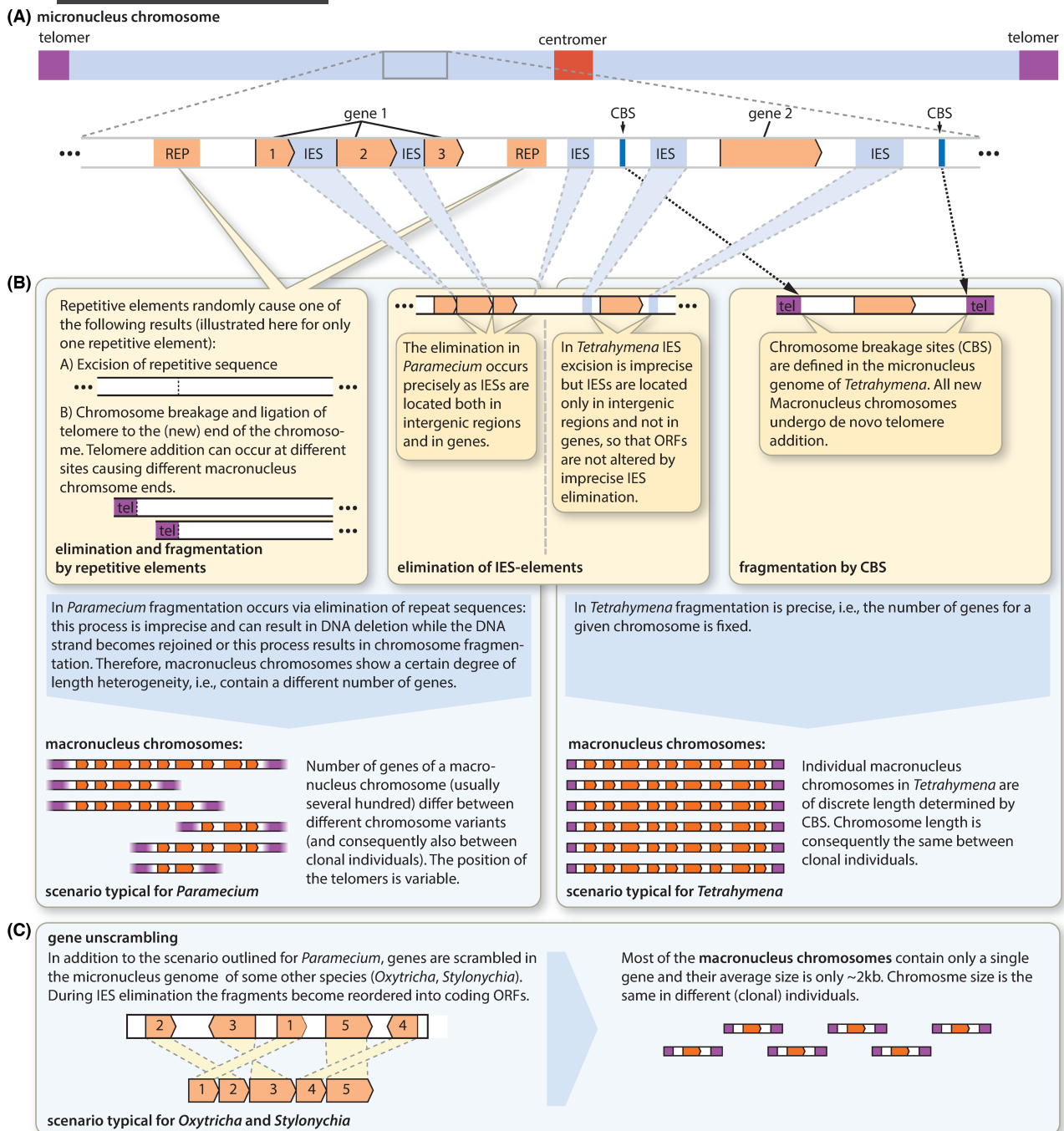


FIGURE 5 Summary of the mechanism of genome rearrangements occurring during MAC development in different ciliates

located in intergenic regions only. Thus, their elimination allows for an imprecise elimination (Chalker & Yao, 2011). All ~45,000 IESs in *Paramecium* are flanked by two 5'-3' TA dinucleotides. After the cleavage, one TA dinucleotide remains in the MAC DNA after rejoining neighbor DNA elements (Gratias & Bétermier, 2003).

One of the greatest mysteries concerns the evolutionary significance of IESs: ten thousands of elimination events need to occur precisely and efficiently in every single conjugation or autogamy. Concerning the origin

of IESs, Larry Klobutcher and Glenn Herrick proposed that IESs are remnants of transposons (Klobutcher & Herrick, 1997). Their model involves an initial MIC invasion of an autonomous transposon, providing a niche for the blooming of these transposons without harming their host: precise removal from MAC DNA would allow colonization in the silent MIC (Schoeberl & Mochizuki, 2011). In the next step of the model, the host would domesticate excision activity, for example, by bringing the transposase under control of a host promoter to control transposon excision during MAC

development. This would then dramatically decrease the selection pressure on the transposons themselves. They would lose their protein-coding ability and autonomy. As a result, they would shorten to the smaller size of recent IESs.

Since MIC DNA in *Paramecium* is difficult to purify (due to the high copy number of chromosomes in the MAC), the MIC genome was sequenced just recently. By comparing the MIC genomes of nine different *Paramecium* species, Sellis et al. could identify families of mobile IESs resulting from recent insertions, thus supporting the above-described model by Klobutcher and Herrick (Sellis et al., 2021). The authors identified several waves of IES insertions: older IES are much shorter than younger ones, agreeing with the model.

What could be the evolutionary advantage of such a transposon invasion for the host? The comparison of MAC DNA to MIC DNA revealed MAC heterogeneity for so-called "TA-indels." These are TA-dinucleotide-flanked remnants of MIC DNA that were found in only a few MAC chromosomes and could represent: (i) inefficiently eliminated IESs, (ii) IESs with alternative TA boundaries, (iii) or so-called cryptic IES, which are MAC destined sequences mimicking IESs. The analysis of the genomic localization of these TA-indels revealed that they are underrepresented in coding sequences and TA-indels that are multiple of 3nt occur at a significantly lower frequency as expected (Duret et al., 2008). This implies that IES excision is not 100% efficient and might represent an evolutionary driver to vary the MAC genome composition while the two MIC alleles remain unaffected.

Their excision requirements can further classify IESs. Epigenetic control was described for a subset of IESs: the presence of an IES in the parental MAC controls whether the IES is excised in the new Mac: IES elimination occurs only if an IES is absent in the parental MAC, otherwise not (Duharcourt et al., 1995, 1998; Meyer & Keller, 1996). The parental MAC should not influence F1 progeny's genotype according to Mendelian rules. IES excision, therefore, represents an exciting example of epigenetic control by the parental somatic nucleus. The underlying mechanisms involve a process that could be considered as "genome scanning" between the parental MAC genome content and the MIC: only the genetic material of the "successful" parent should be able to enter the F1 somatic nucleus (see below).

However, this epigenetic control is not the case for all IESs. Old IESs are less often under parental control than new ones. We will later discuss that the epigenetic excision of (young) IESs involves small RNA, inducing heterochromatic marks at IESs for excision by the PiggyMac transposase in *Paramecium*. In contrast, old IESs shorten but acquire sequence characteristics that make small RNAs and histone modifications dispensable for efficient excision (Sellis et al., 2021). This means

that *Paramecium* requires fresh insertions of transposable elements in the MIC to allow for new epigenetically controlled MAC variability.

Imprecise DNA elimination of repetitive elements in *Paramecium* and chromosome fragmentation

In *Tetrahymena*, the fragmentation of MIC chromosomes during MAC formation is determined by ~280 chromosome breakage sites (CBS) in the MIC, representing a conserved 15-bp signal (Cassidy-Hanley et al., 2005; Hamilton et al., 2005). Telomere addition occurs in a small range of 14–34 bp in the margin of the CBS, thus causing a small degree of heterogeneity in MAC chromosomes (Hamilton et al., 2005).

Chromosome fragmentation by imprecise elimination of repetitive elements is different in *Paramecium*. After eliminating repetitive elements, such as transposons or minisatellites, ends can re-join, or the elimination event can be accomplished by de novo telomere addition, which causes the fragmentation of chromosomes. Both, re-joining or fragmentation, can occur at an individual chromosome because these elimination processes occur simultaneously to DNA amplification: as a result, the ~800 copies of the individual MAC chromosomes exist in different length variants (Figure 5B). A further level of complexity is introduced because telomere addition occurs at various sites: there are heterogeneous chromosome ends, where TARs (telomere addition sites) are spread over several kb. One of the best-studied examples is the MAC chromosome containing the 51A surface antigen gene: this chromosome has three different TARs. The resulting three alternative macronuclear ends have a distance between the telomere and the 51A gene of 8, 13, and 26 kb (Forney & Blackburn, 1988).

As a result of this heterogeneity, the MAC genome assembly only represents a consensus sequence chromosomes: several shorter versions exist, as shown in Figure 5B. In addition to the two telomere-containing scaffolds, the MAC genome assembly also comprises many shorter scaffolds with a lower copy number: these could represent gaps or alternative ends of MAC contigs (Aury et al., 2006; Duret et al., 2008). It is not easy to estimate the contribution of this heterogeneity to the regulation of gene expression. An exception is the understanding of the regulation of surface antigens, where a preferential subtelomeric localization has been implicated for *Paramecium* (Baranasic et al., 2014).

In addition to the scanning of IESs in the parental MAC, alternative MAC chromosome ends are also under epigenetic control: the presence of parental DNA sequences determines their fate in the developing MAC, stabilizing these alternative chromosomal structures

across sexual generations while the MIC remains wild-type (Epstein & Forney, 1984; Forney & Blackburn, 1988; Meyer, 1992).

Gene unscrambling

In Spirotrichea, an additional mechanism occurs. Individual fragments of a gene are not linearly arranged in the MIC but scrambled across the chromosomes and interrupted by IESs. Thus, a precise unscrambling of MIC sequences into functional ORFs is required to create functional genes (Figure 5C). In *Oxytricha* and *Stylonychia*, macronuclear destined sequences, MDS are interrupted by IESs. IES elimination and unscrambling coincide, and MDS originating from different loci are fused to functional genes (Prescott, 1999).

Two additional characteristics distinguish the MAC chromosome development in Spirotrichea from that in Oligohymenophorea: Spirotrichea produces gene-sized nanochromosomes, that is, most genes are located on individual MAC chromosomes are only 1–5 kb in size. Second, these gene-sized nanochromosomes can have enormously different copy numbers (Swart et al., 2013; Xu et al., 2012) in contrast to the Oligohymenophorea, which show nearly identical MAC chromosome copy numbers.

Unscrambling is also epigenetically controlled by the parental MAC, providing a lncRNA (long non-coding RNA) serving as a template to sort the individual MDS of a scrambled gene (Nowacki et al., 2008).

IES excision, as outlined above, shows some similarities to Intron/Exon splicing. This comparison may be misleading due to mechanistic differences, but the general comparison of IESs to introns seems reasonable. Following this line, unscrambling could be compared to an event of trans-splicing in which individual MDS of a gene can be located from far and unlinked loci (Swart et al., 2013; Xu et al., 2012). A further similarity relates to the evolutionary significance of splicing, that is, the occurrence of alternative scrambling creates new combinations of functional domains (Gao et al., 2015).

Amplification

As mentioned, most ciliates show a certain degree of polyploidy in the MAC. DNA elimination and amplification are not separated in time as several rounds of replication already occur in the zygotic genome (Betermier et al. 2000). Consequently, the whole chromosomes already exist in several copies before DNA elimination starts. Amplification can produce alternative MAC variants contributing to the heterogeneity mentioned above in chromosome fragmentation and IES excision in individual cells. Until now, almost no

single-cell genomic data are available for ciliates; thus, the individual extent of MAC heterogeneity still needs to be analyzed. This may become even more pressing, taking the evolutionary consequences of MAC heterogeneity into account. As will be discussed in detail later, phenotypic assortment and epigenetic inheritance might contribute to a dynamic and rapid adaptation of the MAC genotype. Single-cell data would undoubtedly help untangle the relative importance of the described mechanisms to contribute to MAC heterogeneity. However, they should not be limited to stable lab cultures but include cultures under selection pressure to different environmental stressors.

SELECTIVE HYBRIDIZATION IS USED DIFFERENTLY FOR GENOME SCANNING IN OLIGOHYMENOPHOREA AND SPIROTRICHEA

The creation of a functional genome during sexual development involves different classes of RNAs, although the underlying mechanisms differ substantially between *Paramecium* and *Oxytricha* (Figure 6). This chapter focuses on the IES excision of *Paramecium*, and molecular aspects of unscrambling in *Oxytricha* since other mechanisms leading to a functional MAC genome were already described above (see Figure 5).

Starting with *Paramecium* meiosis, long double-stranded RNA is transcribed by RNA-polymerase in association with the MIC specific transcription factor complex SPT4-SPT5 from MIC chromosomes (Gruchota et al., 2017; Owsian et al., 2022). Small dsRNAs are produced by two Dicer-like proteins, Dcl2/3, with characteristics in a 5' UNG signature, 3' 2nt overhangs, and the precise length of 25nt (Lepère et al., 2009; Sandoval et al., 2014). Subsequently, single-stranded RNAs are selected by Ptiwi01 and Ptiwi09, two ohnologs belonging to the PIWI subclade of Piwi/Ago proteins (Bouhouche et al., 2011; Furrer et al., 2017). Ptiwi-sRNA complexes shuttle into the old, parental MAC. The scanning process occurs: long transcripts from MAC chromosomes (Lepère et al., 2008) are scanned by 25nt Piwi bound sRNAs (further called *scnRNAs*) in a homology-dependent manner. ScnRNAs that match their target are degraded. Since chromosomes of the parental MAC do not contain IES, scnRNAs complementary to those sequences, do not find a matching target and remain intact while unmatched scnRNAs are degraded. Subsequently, surviving scnRNAs are transported into the new developing MAC. Here, scnRNAs mark IES for their excision. However, only a minor fraction of IESs depend on the scnRNA pathway (Furrer et al., 2017; Lhuillier-Akakpo et al., 2014). The elimination process in the new developing MAC is initiated either by the interaction of the scnRNA with a nascent transcript or the DNA itself (Pina

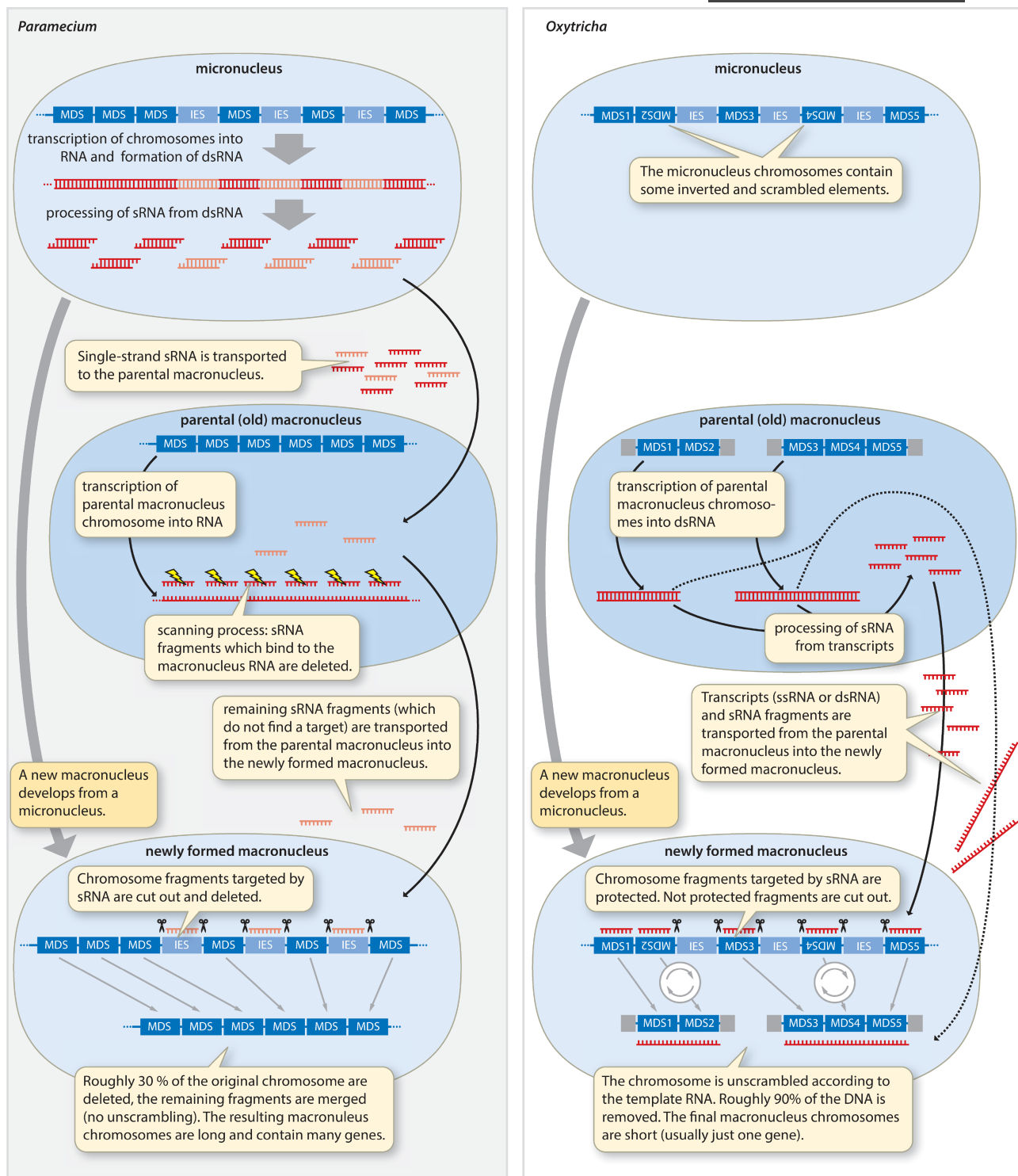


FIGURE 6 Comparison of the scnRNA models between *Paramecium* and *Oxytricha*

et al., 2021). In brief, IES targeted by scnRNAs become excised by the domesticated piggyBac transposase called piggyMac. The remaining MDS segments are ligated to generate intact chromosomes, followed by final rounds of endoreplication. On the other hand, excised IESs are circularized and enter an sRNA amplification pathway shown in Figure 8.

Unscrambling in *Oxytricha*

In contrast to Oligohymenophorea, genome rearrangements in *Oxytricha* do not involve transcription of the MIC. Bidirectional transcription of all chromosomes occurs in the parental MAC by RNA Polymerase II (Khurana et al., 2014; Lindblad et al., 2017; Nowacki

et al., 2008) which subsequently followed by accumulation of 27nt sRNA (Fang et al., 2012). In addition, long RNAs are shuttled in the new MAC, serving as a template for subsequent genome rearrangements. The accumulation of 27nt sRNAs with a 5' U preference peaks exclusively early in development, probably when long RNAs are synthesized. The relationship between sRNAs and template RNAs remains obscure so far. Long RNAs could be precursors of 27nt small RNAs, but since sRNAs do not cover telomere sequences as long RNAs from chromosome-wide transcription do, it is tempting to speculate that another RNA class serves as a precursor for sRNAs (Lindblad et al., 2017). Additionally, it is unclear if 27nt sRNAs are derived from dsRNA or ssRNA precursors as described for Ptiwi bound scnRNAs (Fang et al., 2012). Similar to the *Paramecium* mechanism, *Oxytricha* sRNAs are shuttling from the parental MAC to the zygotic MAC, bound to Otiwil, one of 13 *Oxytricha* PIWI-proteins.

In the zygotic nucleus, chromosomes first undergo some pre-amplification rounds that lead to giant polytene chromosomes followed by the excision of IESs (Spear & Lauth, 1976; Yerlici & Landweber, 2014). In contrast to the mechanism described for Oligohymenophorea, Otiwil bound sRNAs protect MDSs. IESs remain unprotected, which consequently leads to their excision. *Oxytricha* manages genome rearrangements differently from *Paramecium*: Instead of using one domesticated piggyBac transposase, *Oxytricha* uses thousands of transposable elements exclusively expressed from the MIC. Those *telomeric repeats bearing transposable elements* (TBEs) encode for three different transposases, which are expressed solely during development (Nowacki et al., 2009). The need for thousands of TBEs is likely linked to the massive scrambling events occurring in Spirotrich MAC, resulting in the mutualistic toleration of transposon accumulation while guaranteeing sufficient transposase activity (Vogt et al., 2013). More recently, Swart et al., 2013 identified two types of domesticated transposases encoded on MAC nanochromosomes likely to be involved in genome rearrangements, as well.

Little is known about the involvement of additional factors in IES excisions, such as guiding the sRNA-Otiwil complexes and the formation of heterochromatin. In *Stylonychia*, another Spirotrich ciliate, excision involves the deposition of histone H3 modifications: MDS become associated with marks for open chromatin while sequences to be eliminated are likely found in regions that become marked by de-novo lysine methylation (H3K27me3/H3K9me3) (Postberg et al., 2008).

Once the IES is excised, the remaining MDS must be sorted to generate a functional MAC genome. MDS from far distant locations in MIC chromosomes are found to be in the direct vicinity in MAC chromosomes which is achieved by *unscrambling*. The order of and

orientation of MDSs is usually the same as encoded in the MIC. Still, at least 12% of all MAC contigs in *Oxytricha* show scrambling (Burns et al., 2016), and 8% of all MAC contigs show a nested structure, where an MDS for one chromosome is positioned inside the MDS coding for another chromosome (Braun et al., 2018). MDSs from the same MIC loci can be *shuffled* to generate divergent nanochromosomes, even with the potential to build new genes with functional diversification (Chen et al., 2015). This was recently shown for *Euplotes* pheromone gene, whose 5'-region derives from an MDS of an unrelated gene (Ricci et al., 2021). This process can be compared to alternative splicing on the DNA level and leads to the chance of higher gene diversity that can even be transmitted to the next generation (Chen et al., 2014; Maurer-Alcalá & Nowacki, 2019).

The exact order of MDS is controlled by two different genetic tools: (i) pointer sequences that are 2- to 20-bp long repeats at each MDS-IES junction and (ii) long template RNAs synthesized in the parental MAC. Homologous recombination between identical repeats at pointer sequences leads to the formation of a streamlined MDS-MDS order by leaving one pointer copy in the final MAC chromosome sequence (Nowacki et al., 2008; Yerlici & Landweber, 2014). Unscrambling is probably also facilitated by the length of pointer sequences since longer pointers are found at MDS to be scrambled (Chen et al., 2014). The recombination process can probably happen as a *cis*—and/or *trans* process since polytenization before IES excision likely would allow for *trans* recombination of MDS from zygotic chromosomes (Chen et al., 2014).

In addition to pointers, long RNA templates generated from each parental MAC chromosome, even from those not showing any scrambling (Lindblad et al., 2017; Nowacki et al., 2008), serve as a template for the order of MDSs. Long RNAs are probably stabilized as dsRNA in the zygotic MAC by Rpb2-a, a paralog of the second-largest polymerase II subunit, although sequence-based evidence is missing (Khurana et al., 2014). Based on the long RNA encoded blueprint, MDS segments are joined by homologous recombination while excised IESs have different fates: at least IES segments from unscrambled loci become circularized, and since transcripts from those circles could be detected, similar to *Paramecium*, likely, circularized IES segments are not exclusively prone to be degradation products (Yerlici et al., 2019).

The mechanism of unscrambling was thought to be unique to Spirotrichea. Still, at least in *Tetrahymena*, seven complex MAC chromosomes are generated by the site-specific joining of non-contiguous segments of germline DNA (Hamilton et al., 2016), and a handful of scrambled loci have been detected recently (Sheng et al., 2020).

IES excision upon heterochromatin formation in *Paramecium*

In multicellular species, transposon silencing is realized by Piwi-proteins which act in a complex with piRNAs in two ways: targeting RNA of transposons post-transcriptionally in the cytoplasm or via co-transcriptional targeting of nascent transcripts in the nucleus (Czech et al., 2018). This seems similar in ciliates, as DNA elimination after placement of repressive histone marks can be seen as an extreme form of transcriptional silencing, namely DNA elimination. Figure 7 partially summarizes the current idea of IES excision upon RNA-mediated heterochromatin formation in the new developing MAC.

In *Paramecium*, recruitment of scnRNAs bound to Ptiwi01/09 to their IES targets in the zygotic nucleus is thought to be achieved by a long nascent transcript, likely to be synthesized by RNA Polymerase II depending and elongation factor TFIIS4 (not shown, Maliszewska-Olejniczak et al., 2015). Upon targeting the IESs by Ptiwi bound scnRNAs, tri-methylation at the histone H3 at lysine K9 and K27 co-occurs. Both H3K9me3 and H3K27me3 are introduced by the Enhancer-of-zeste-like

protein Ezh1, a histone methyltransferase of the polycomb repressive complex 2 (PRC2): knock-down of Ezh1 causes the retention of 70% of all IES (Frapporti et al., 2019; Lhuillier-Akakpo et al., 2014). Ezh1 is probably associated with the chromatin assembly factor 1 (PtCaf1), which guides Ezh1 for methylation by its histone-binding domain. Since PtCaf1 is also involved in the upstream scanning process, there is accumulating evidence for the RNA-guided DNA elimination linked to changes in chromatin conformation (Ignarski et al., 2014).

This chromatin modification needs to occur in a strict local manner since the majority of IES is shorter than the DNA wrapped around a nucleosome (<150 bp), and only nucleosomes that cover IES should be targeted (Lhuillier-Akakpo et al., 2014).

Interestingly both, PtCaf1 and Ezh1 are not strictly limited to the developing MAC: both GFP-fusions also show a signal in the parental MAC in the early stages of autogamy before its fragmentation (Ignarski et al., 2014; Lhuillier-Akakpo et al., 2014). Their function in the old MAC remains unclear, maybe the regulation of developmentally regulated genes or lncRNAs transcription.

Ezh1 is likely to be found in a PRC2 multiprotein complex with PtCaf1 and a potpourri of other proteins associated, while the composition of the *Paramecium* PRC2 complex is still obscure. H3K9 and H3K27 trimethylation further recruit or activate an excisase complex comprised of the piggyBac transposase called piggyMac (Pgm) and five associated transposases, termed piggyMac likes (PgmLs) (Bischerour et al., 2018). Pgm is required to excise all IESs and transposable elements (Arnaiz et al., 2012) and probably acts as a homooligomer (Dubois et al., 2017). The Ku70/80 heterodimer anchors the Pgm-PgmL complex (Marmignon et al., 2014) and a histone chaperone of the FACT complex, Spt16-1, that probably mediates chromatin rearrangements, allowing the Pgm-PgmL complex to access the DNA for excision (de Vanssay et al., 2020).

Pgm induces DSB at the conserved TA dinucleotide at IES boundaries, resulting in overhangs centered around the TA dinucleotide sequence and now associated PgmLs from a large bridging complex to fine-tune the positioning of the Pgm transposase (Bischerour et al., 2018). Anchoring of Pgm-PgmLs complexes on Ku70/80 heterodimers probably ensures the efficient coupling of DNA excision and DSB repair since Ku proteins are DSB repair factors recruiting proteins of the non-homologous end joining (NHEJ) pathway (Abello et al., 2020). Upon IES excision, remaining MDSs are ligated in a streamlined fashion by Ligase IV and Xrcc4 while one of the two TA dinucleotides remains in the final MAC genome sequence (Kapusta et al., 2011).

As already mentioned, there are still open questions concerning IES recognition and excision. Only a minor fraction of IES (~ 5%) depend on the scnRNA pathway, and at least 30% are independent of any chromatin conformation changes. Further, although the IES excision

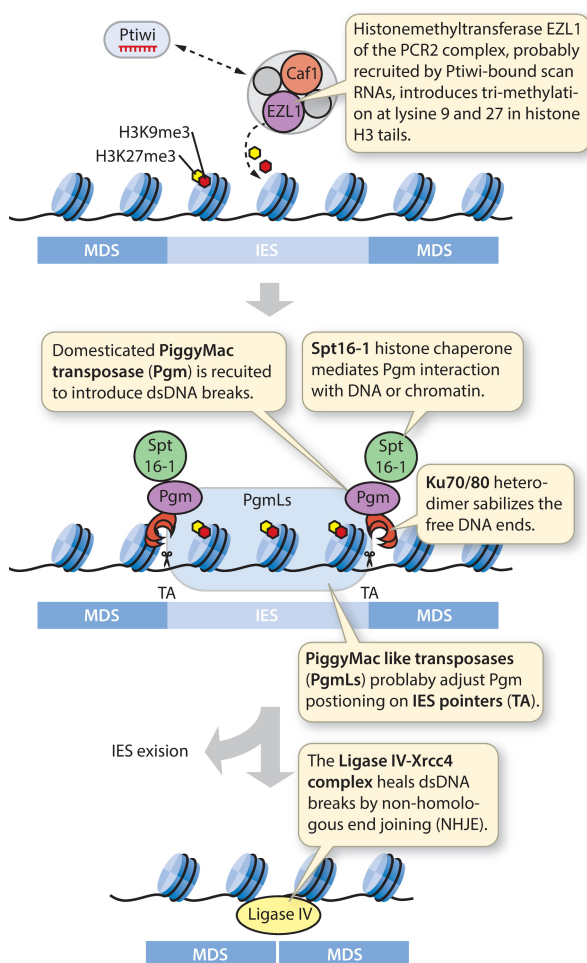


FIGURE 7 A model for heterochromatin related excision of IESs in *Paramecium*

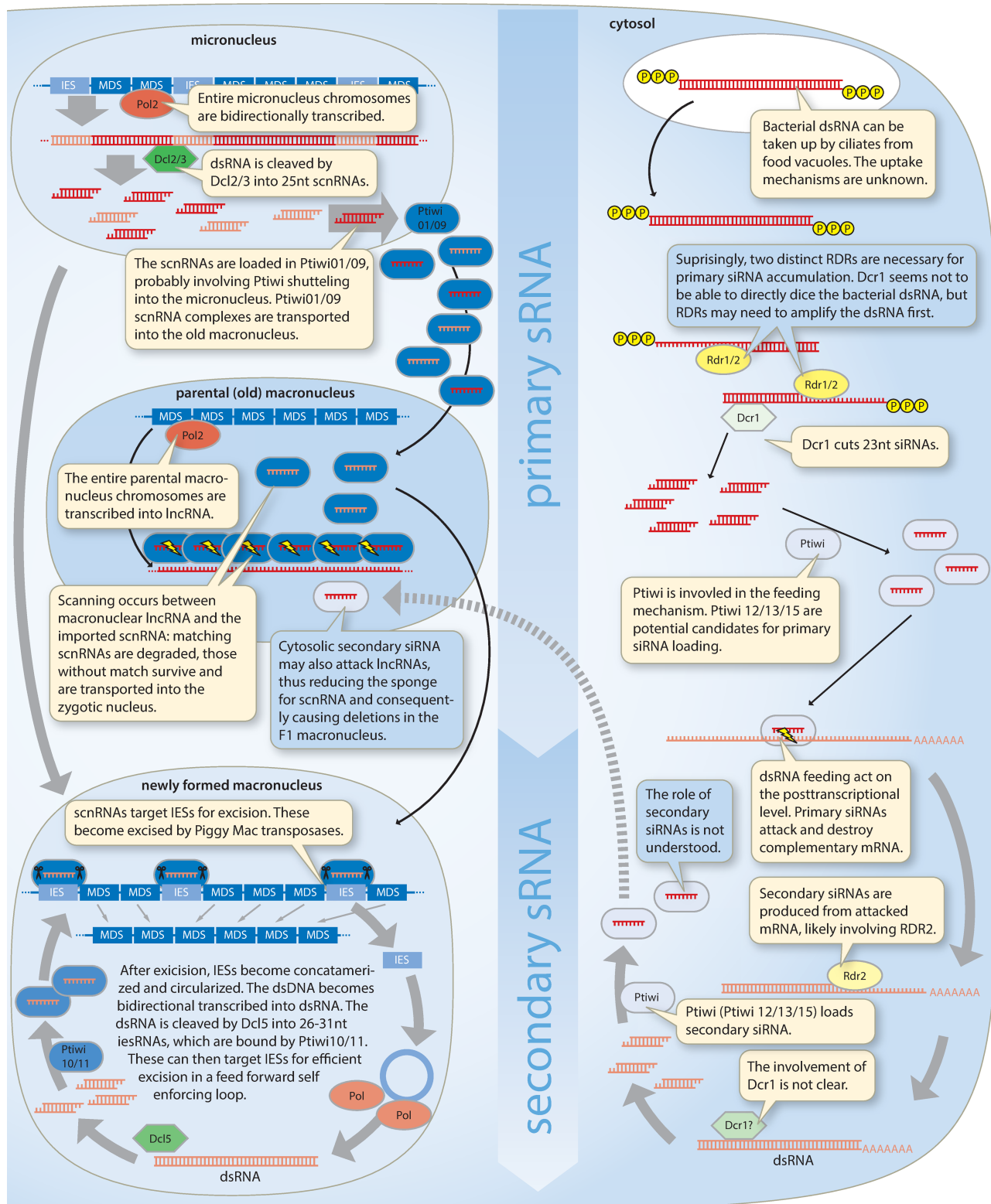


FIGURE 8 Comparison of developmental and vegetative RNAi mechanisms involving 1° and 2° sRNAs in *Paramecium*

machinery is exact, ~7000 sites with excision variability have been identified: events like occasional IES retention, excision of IES with alternative boundaries, and cryptic IES, that is, the excision of MAC destined sequences at TA dinucleotides, contribute to the variability of the

MAC genome (Duret et al., 2008; Swart et al., 2014). An interesting example has been described for the epigenetic control of mating type inheritance in *P. tetraurelia*. The promoter of the mtA gene becomes excised, similar to the excision of a cryptic IES, during MAC development

by the scnRNA and PiggyMac machinery, thus preventing the mtA expression resulting in the production of mating-type O cells. By this mechanism, the mating-type can be inherited by the co-option of the genome rearrangement machinery to regulate gene expression (Sawka-Gądek et al., 2021; Singh et al., 2014).

Heterozygosity also needs to be taken into account into the evaluation of the genome scanning mechanism. In contrast to *Tetrahymena*, only a limited amount of cytoplasm is exchanged between the mates in *Paramecium*. Accidentally increased cytoplasmic exchange between mating pairs altered gene expression in exconjugants (Sonneborn & Lesuer, 1948). Considering the need to have scnRNAs for both maternal and paternal IESs, a cross of different strains with IES insertion polymorphisms (one strain contains IESs that are not present in the mating partner) would cause a problem for scnRNA dependent IESs. If a cell during conjugation receives a (paternal) gamete nucleus with an IES not present in the maternal MIC genome, this IES should not be excised because of the lack of homologous scnRNAs. However, paternal scnRNAs were shown to efficiently program the excision of such IESs on the maternal side. Although the effect of paternal scnRNAs was documented only for a few IESs (intra-species insertion polymorphisms are rare), the authors see no reason why this should not be the case for all scnRNA-dependent IESs. It remains unclear whether the demonstrated action of paternal scnRNAs depends on their physical transfer between cells in sufficient amounts or whether they may somehow program excision in gametic nuclei locally before the latter are exchanged cross-fertilization (Pellerin G, Nekrasova I, Potekhin A, Meyer E personal communication).

CROSS-OVER BETWEEN DEVELOPMENTAL AND VEGETATIVE RNAi IN *PARAMECIUM*

In *Paramecium*, components of developmental chromosome rearrangements show a transcriptional activation during sexual recombination. Most show low or no RNA transcripts during vegetative growth (Arnaiz et al., 2010). Also, somatic RNAi pathways have been described, induced either by the injection of truncated transgenes or by ingesting bacterially produced dsRNA. Both pathways differ in the active RNAi components (Marker et al., 2010). In Figure 8, we compare these developmental and vegetative RNAi pathways and their possible overlap.

scnRNA and iesRNAs are involved in IES excision

After bidirectional transcription of the meiotic MIC, two Dicer-like enzymes, Dcl2/3, cleave the scnRNA duplexes

(Figure 8 left). These cuts do not occur randomly or phased, but Dcl2 shows a sequence preference to cleave at the conserved ends of IESs, thus enriching the scnRNA pool for IES targeting scnRNAs (Hoehener et al., 2018). The process then follows the mechanism outlined above. ScnRNAs are loaded into Ptiwi01/09, and scanning occurs against a transcribed lncRNA in the parental MAC. After excision in the developing MAC, IESs are not directly degraded but ligated into circles: small IES form concatamers to build larger circles and large IESs are directly ligated into circular DNA (Allen et al., 2017). This is the beginning of the second round of sRNA accumulation: so-called iesRNAs are produced from transcripts of the circular DNA by Dcl5 (Sandoval et al., 2014); these secondary (2°) sRNAs are then loaded into Ptiwi10/11 (Furrer et al., 2017). Likely, 2° sRNAs represent a feed-forward loop to guarantee efficient IES excision: as amplification starts before DNA elimination, many IESs already exist in hundreds of copies. Also, in *Tetrahymena*, 2° scnRNAs have been described in a trans-acting network: early scnRNAs can act to other IESs *in trans* triggering accumulation of 2° sRNAs which can in turn target IESs *in trans* (Noto et al., 2015).

Primary and secondary siRNAs are involved in feeding induced silencing

RNA interference to silence protein-coding genes can be triggered in *Paramecium* by feeding dsRNA-producing bacteria (Figure 8, right). The protocol is similar to nematodes and allows for rapid and easy reverse genetics analyses (Galvani & Sperling, 2002). In contrast to the developmental scnRNA/iesRNA pathway, several papers demonstrated the involvement of RNA-dependent RNA polymerases (RDR) (reviewed in Nekrasova & Potekhin, 2019). One surprising finding was that two distinct RDRs are necessary for 1° siRNA accumulation, implicating that the dsRNA trigger needs to be amplified by RDRs before Dicer cleavage (Marker et al., 2010). The reason for this remains unknown. It may be related to a preference for 5'-tri-phosphorylated siRNAs, which was described for *Tetrahymena* Dcr2 being physically coupled with the RDR (Lee & Collins, 2007). However, this remains to be analyzed for *Paramecium*. In particular, it remains unclear why two distinct RDRs are involved: mutants and knock-downs of both show a loss of 1° siRNAs, so they do not seem to be redundant (Carradec et al., 2015; Marker et al., 2010, 2014). One could imagine that RDRs and CID (cytoplasmic uridylyltransferase) are the initial components of a mechanism dissecting endogenous and exogenous RNAs (self vs. non-self). In nematodes, for instance, viral RNAs are 3'-uridylylated as part of a conserved antiviral mechanism (Le Pen et al., 2018). However, this is unlikely, as both RDRs involved in the feeding mechanisms are also necessary for a certain number of endogenous siRNA clusters, so there

seems to be a mechanistic overlap between exogenous and endogenous RNAi (Karunanithi et al., 2019, 2020).

1° siRNAs trigger both post-transcriptional mRNA degradation and the production of 2° siRNAs produced from the mRNA by RDR2 (Carradec et al., 2015). It is also unclear whether 2° siRNAs are direct RDR products as described in nematodes or require another cut by Dicer. Three different Ptiwis are involved in this mechanism (Ptiwi12,13,15) (Bouhouche et al., 2011). Their precise role remains to be elaborated: they could be mutually exclusive specific for 1° or 2° siRNAs, which would then be similar to the developmental RNAi pathway. Ptiwi13 has been shown to load siRNAs matching to food bacterial genomes (Drews et al., 2021); thus, it may also be responsible for loading 1° siRNAs in this mechanism here. However, transgene-induced RNAi recently reported that the two Ptiwis involved do not show a mutual exclusive loading of 1° and 2° siRNAs: their function may be in RNA shuttling and transport (Drews et al., 2021; Götz et al., 2016).

Comparison and overlap of vegetative and developmental RNAi

Both pathways seem to occur independently, and both pathways involve a mechanism to create 2° sRNAs. A striking mechanistic difference is that no RDR activity was reported until now for the scn/iesRNA pathway. Progeny of RDR mutants is fully viable (Marker et al., 2014). Also, in *Tetrahymena*, the activity of RDR1, the only RDR in *Tetrahymena*, seems dispensable for scnRNA production (Noto et al., 2015). *Paramecium* owns three RDRs, and the scnRNA/iesRNA pathway only operates with bidirectional RNA transcripts rather than dsRNA produced by RDRs, for both 1° and 2° sRNAs.

Most interestingly, overlaps have been described for the developmental and the vegetative RNAi pathways. Injection of non-expressible transgenes and feeding of dsRNA induces macronuclear deletions in F1 progeny. Although the precise mechanism is unclear, 23nt siRNAs from feeding could target the parental macronuclear lncRNA for degradation. As a result, these sequences could not pair with scnRNAs, therefore targeting a gene deletion in the developing MAC (Garnier et al., 2004). This surprising finding may be attributed to 2° siRNAs because the deletion does not only cover the dsRNA region but occurs mainly at the transcribed region.

The same study demonstrated that the injection of transgenes causes MAC deletions only if the transgene is silent: expressed transgenes cannot cause F1 deletions (Garnier et al., 2004). This is quite reminiscent of the classical paramutation, although occurring not on the silencing- but the DNA elimination level. The biological significance of this mechanism is not clear in *Paramecium*; no endogenous siRNA-producing locus has been identified, which could cause a deletion.

Similarly, vegetative silencing decreases F1 copy number in *Oxytricha* and *Stylonychia*, and *vice versa*, overexpression of genes in the parental MAC increases F1 copy number. Maternal RNA was implicated in also transporting quantitative information to the new MAC (Heyse et al., 2010; Nowacki et al., 2010). Recent data indeed suggest that coding mRNA could serve as a template for 27nt scanning RNAs in *Stylonychia*, as sRNA and mRNA abundance correlate to each other (Postberg et al., 2019). It seems likely that Spirotrichea could use this mechanism to adapt the copy number of individual nanochromosomes to the most advantageous gene dosage for an environmental condition: without the need for mutational changes (Yao, 2010).

In contrast, the biological significance of the 2° siRNA shortcut in the *Paramecium* RNAi mechanisms leading to MAC deletions remains to be elucidated. It seems risky at first glance that environmental RNA can interfere with the genome content. But this does not happen directly. It looks tempting to speculate that the detour via the 2° siRNAs could introduce a threshold to avoid harmful deletions triggered by exogenous RNA. Many studies demonstrated that the vegetative silencing phenotype correlates with the abundance of 1° siRNAs (Lepère et al., 2009) and 2° siRNAs are much lower abundant (Carradec et al., 2015; Götz et al., 2016).

SUMMARY OF CILIATE GENETICS AND EPIGENETICS: CONTROLLED GENOME INSTABILITY FOR RAPID MAC ADAPTATION?

We discussed here many different aspects of MAC genome heterogeneity: chromosome fragmentation, alternative chromosome ends, inefficient IES excision, alternative scrambling, induced macronuclear deletions. Most of these aspects are well studied in lab cultures, although some only with individual genetic loci. The extent of these parameters of MAC variability in the wildtype under different evolutionary forces remains largely unknown. However, it becomes clear that the combination of nuclear dimorphism, amitosis, and parental controlled DNA rearrangements allow for a powerful adaptive capacity. This is, in principle, the outline of this review: in Figure 9, we summarize the potential evolutionary consequences of the particular ciliate features we introduced here.

The nuclear dimorphism allows for the differentiation of a somatic MAC and amitosis. Individual alleles can be pre-selected and relatively enriched by phenotypic assortments. Depending on the species and its genome characteristics, the genotype would be capable of environmental adaptation within a few asexual cell divisions. Following Mendelian rules, all of these genetic adaptations in the MAC would be lost after sexual reproduction and the regeneration of a new MAC.

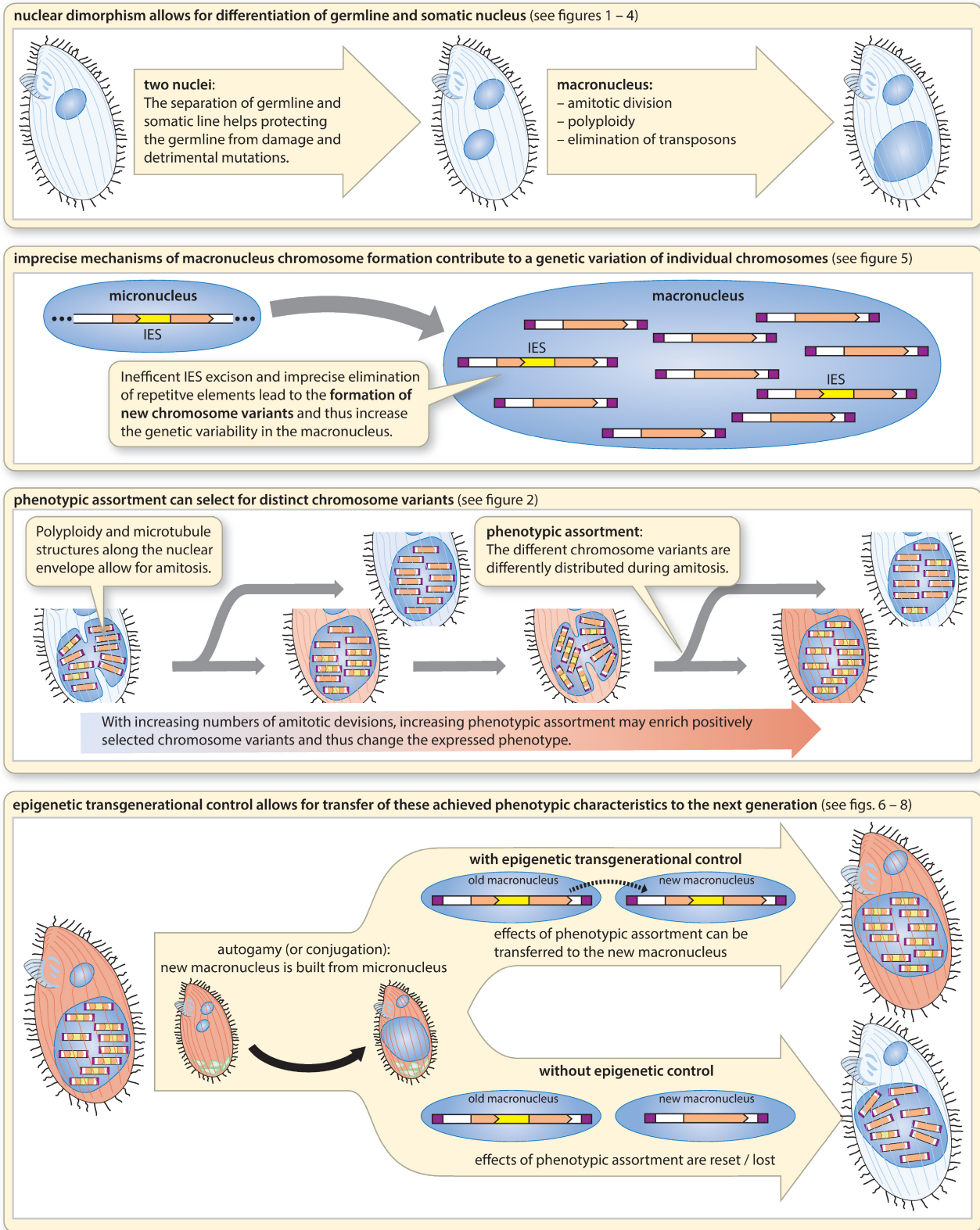


FIGURE 9 A summary of the distinct models of ciliate genetics and epigenetics described in this review in association with their putative evolutionary advantages

However, the transgenerational RNA phenomena discussed above would provide a chance to pass on epigenetic information to the new MAC. This would represent

an exciting example of the inheritance of acquired characters in the Lamarckian manner. As a result, ciliates could epigenetically manifest adapted MAC genotypes.

IESs can then be seen as a toolbox allowing for error-prone elimination enabling a permanent try and fail variability of new MAC genotypes. At the same time, the germline remains protected and safe. Indeed, few recent studies have started to elaborate on the variable IES retention in lines under alternating selection pressure, suggesting that such a mechanism may contribute to rapid phenotypic adaptation (Catania et al., 2021; Vitali et al., 2019). In addition to IES variation, alternative unscrambling of genes contributes to gene family evolution by alternative processing of scrambled micronuclear loci (Gao et al., 2014; Katz & Kovner, 2010).

Several studies indicate gene families in ciliates evolve faster compared to other eukaryotes. On the one hand, this appeared due to an increased ratio of nonsynonymous/synonymous substitution rates of individual loci (Katz et al., 2004; Zufall et al., 2006). Otherwise, the genome architecture, meaning the degree of macronuclear processing, was implicated in gene-fragment rearrangements because the most divergent proteins have been identified in ciliate lineages with highly processed genomes, thus supporting the idea of a programmed MAC variability (Zufall et al., 2006). Still, most of these conclusions result from the interpretation of single protein families. Genome data, especially MAC and MIC, is still missing for many lineages. An exception is a recent analysis of transcript diversity in single-cell transcriptomes from different non-model ciliates. Karyorelictea show the lowest paralog diversity, which is higher in ciliates with extensively re-arranged genomes; in addition, the study supports the idea that the loss of gene linkage by gene-sized chromosomes enhances gene evolution (Yan et al., 2019).

Coming back to one of the initial questions: what might have been the driving force to evolve the nuclear dimorphism in ciliates? We summarized here the genetic and epigenetic features that need to be considered. The main features of ciliates (except the Karyorelictea) are the amitotic division of the MAC and the paternally controlled DNA rearrangements during MAC development. Unfortunately, data about these mechanisms are available only from the Intramacronucleata. We need comparable data of the Karyorelictea and Heterotrichea to evaluate transgenerational epigenetics as a driving force in the ciliated evolution of the nuclear dimorphism as discussed above.

Considering the evolution of this system, we need to extent the evolutionary scenario of chapter 3 by the concept of transgenerational epigenetics. Phylogenetic analysis of RNAi key components (RDR, Dicer, and Piwi) suggests that the last common eukaryotic ancestor was already capable of RNAi by post-transcriptional silencing and transcriptional regulation by histone modification (Cerutti & Casas-Mollano, 2006). The authors suggest that RNAi's ancestral role was the silencing of transposable elements. Presumably, the very early ciliates already silenced transposable elements by small RNAs. Transposon silencing in other species is usually

realized by heterochromatinization, and recent literature suggests that, in particular, H3K9 methylation represents an ancestral mark for repeats and transposable elements (Kabi & Filion, 2021). Taken together, the literature so far indicates that the last common ciliate ancestor already used transcriptional silencing by sRNAs for transposon inactivation. As the general hardware was already present, two events may have been necessary to build the modern ciliates. First, the exchange of RNA between paternal and zygotic nucleus: for these mechanisms, it is clearly of great advantage that both nuclei are present in the same cytoplasm. The dimorphism may be seen as a prerequisite for this massive transgenerational transfer of non-coding RNAs. Second, the coupling of RNAi with the domesticated transposase allows for DNA excision instead of silencing.

In general, transposon integration represents a mechanism to generate genetic diversity. Transposon integration is mostly disadvantageous or neutral. Co-opting of advantageous transposon integrations was also described in vertebrates (Bourque, 2009) and believed to contribute to lineage-specific characters (Warren et al., 2015). For instance, the Gibbon lineage appears to be diverged from the Hominidae by massive genome rearrangements to a gibbon-specific retrotransposon (LAVA) (Carbone et al., 2006, 2014). However, this example of transposon-mediated genome variability and speciation appears to follow classical Darwinian evolution, meaning an uncontrolled, stochastic genotype alteration with subsequent natural selection. The described mechanisms here suggest that ciliates have evolved a system of transposon domestication by parental control, using this for short time adaptation by inheritance of acquired characters in a Lamarckian manner. Both micronuclear genome stability and MAC polyploidy provide an evolutionary buffer for either a genetic reset or the progressive adaptation of MAC genotypes.

Also, in mammals, V(D)J recombination represents a form of genomic instability/variability by somatic recombination. This instability is essential for the maturation of B- and T-cells and involves the creation of new immunoglobulin genes: the RAG1 gene, which initiates the recombination evolved by transposase domestication (Huang et al., 2016). Comparing both systems, V(D)J recombination of IES excision, mammals need to limit this to specific gene families, whereas ciliates could use the entire genome as a playground.

THE EXCEPTIONAL QUALIFICATION OF CILIATES FOR EPIGENETIC RESEARCH AND ITS INTERPRETATION IN AN ECO-EVOLUTIONARY CONTEXT

Current estimates calculate with up to 40,000 ciliate species, of which ~4500 morphospecies have been described

(Foissner et al., 2009). Individual ciliate species and strains were able to adapt to almost all ecosystems, ranging from freshwater and marine to terrestrial habitats (Finlay et al. 1998) and parasitic life strategies (Clark & Forney, 2003).

An apparent feature of ciliates is their relatively large cell size compared to flagellates in the same environments. In particular, a selective advantage may have been provided by an increase in the quantity of gene products, which is linked to an increase in ploidy. The indirect consequence, that is, increasing cell size, may have allowed for widening the prey spectrum to a greater variance of prey organisms (Cheng et al., 2020) and at the same time may have altered (reduced) grazing pressure on ciliates by unicellular predators.

A second characteristic feature to be discussed here is the combination of nuclear dimorphism with transgenerational inheritance. This massive investment in genome rearrangements and, in particular, the transgenerational mechanisms described in this review require some (energetic) resources. The increasing direction of resources towards optimizing the F_1 genomes (instead of growth rate) can be interpreted as a shift towards a K-strategy: ciliates invest more energy in their sexual progeny. Their "parental care" achieves offspring which is highly adapted to the current environment. Thus, a considerable fraction of energy is channeled to the survival of individual cells rather than too high growth rates as characteristic for K-strategists (Figure 10).

Many of the epigenetic mechanisms that allow for this "parental care" have been resolved in the past decades while some are still uncertain or unknown. Ciliate epigenetics mainly refers to DNA rearrangements and eliminations that occur to an extent unknown in other organisms. It seems clear that the development of the new MAC inside the cell, which still contains the (old) parental macronucleus, logistically facilitates RNA transfer between generations.

This would support the model of Katz, 2001 which hypothesizes that the nuclear dimorphism evolved basically as a mechanism allowing the parental MAC to influence the F_1 MAC and thus the F_1 phenotype.

TRANSGENERATIONAL RNA IN MULTICELLULAR ORGANISMS

Intergenerational epigenetic regulation appears more difficult in multicellular organisms but not impossible. In *C. elegans*, for instance, feeding of dsRNA by mechanisms similar to those described in Figure 8 has been described to induce secondary siRNAs likely involved in transcriptional silencing, which are stably maintained in subsequent generations (Luteijn et al., 2012). The striking difference to ciliates is that transgenerational epigenetics of other eukaryotes rarely involves genome rearrangements.

RNA is right now in the focus of transgenerational epigenetics in multicellular organisms too. piRNAs (piwi-interacting RNAs) are active components of a conserved pathway in gonadal cells to inactivate transposons (Czech et al., 2018). In *C. elegans*, piRNAs and nuclear RNAi mechanisms cause an epigenetic memory that lasts more than 20 generations (Ashe et al., 2012). Mammalian oocytes exhibit active piRNA pathways; thus, a transgenerational impact of maternal control seems not unrealistic, but we know much less about this than the model systems. Further, a soma-germline transfer of small RNAs (mainly miRNAs (microRNAs) and tRNA fragments) from the epididymis to sperm represents a previously unexpected paternal somatic RNA transfer (Sharma et al., 2018). The extent of these transgenerational effects remains to be studied. Nevertheless, recent reviews summarize the emerging knowledge on transgenerational sRNAs as a redefinition of the inheritance concept in animals (Cecere, 2021), that is, similar to the terminology applied for ciliates.

To date, research on transgenerational inheritance seems to be more advanced in ciliates. But due to their model character, it may be easier to observe epigenetic phenomena phenotypically in ciliates than multicellular organisms. A thrilling question is: could future research uncover a similar extent of epigenetic transfer of parental information to sexual progeny in mammals?

This seems not likely due to the more limited contact between parental and F_1 cell lineages and the extent of epigenetic mechanisms in cell and tissue differentiation. However, as outlined above, transgenerational RNA indeed occurs in germ cells, and future research may uncover exciting functions of transgenerational RNA contributing to Lamarckian inheritance in mammals.

We have summarized in this review that ciliates evolved special genetic hardware for transgenerational RNA transfer, coupled with a genomic system with pre-determined breaking points, providing a modular design allowing for genomic variability. This is in strong contrast to the needs of long-living and differentiated mammalian somatic cells, reliant on genome stability, allowing genetic instability at few sites, only, for example, immunoglobulin genes. It is tempting to speculate that the degree of transgenerational epigenetics could be associated with the complexity of the species in terms of tissue differentiation. Therefore, it may not be accidental that indications for intergenerational epigenetic control in metazoan were mostly reported for animals with a low degree of cell (and tissue) differentiation. Most reports so far concern nematodes, whose adults consist of only a few hundred cells: Here, several heritable epigenetic manifestations mediated by small RNAs have been described (Rechavi et al., 2014; Shirayama et al., 2012), which could lead to individual epigenetic variations (Hourri-Zeevi et al., 2020).

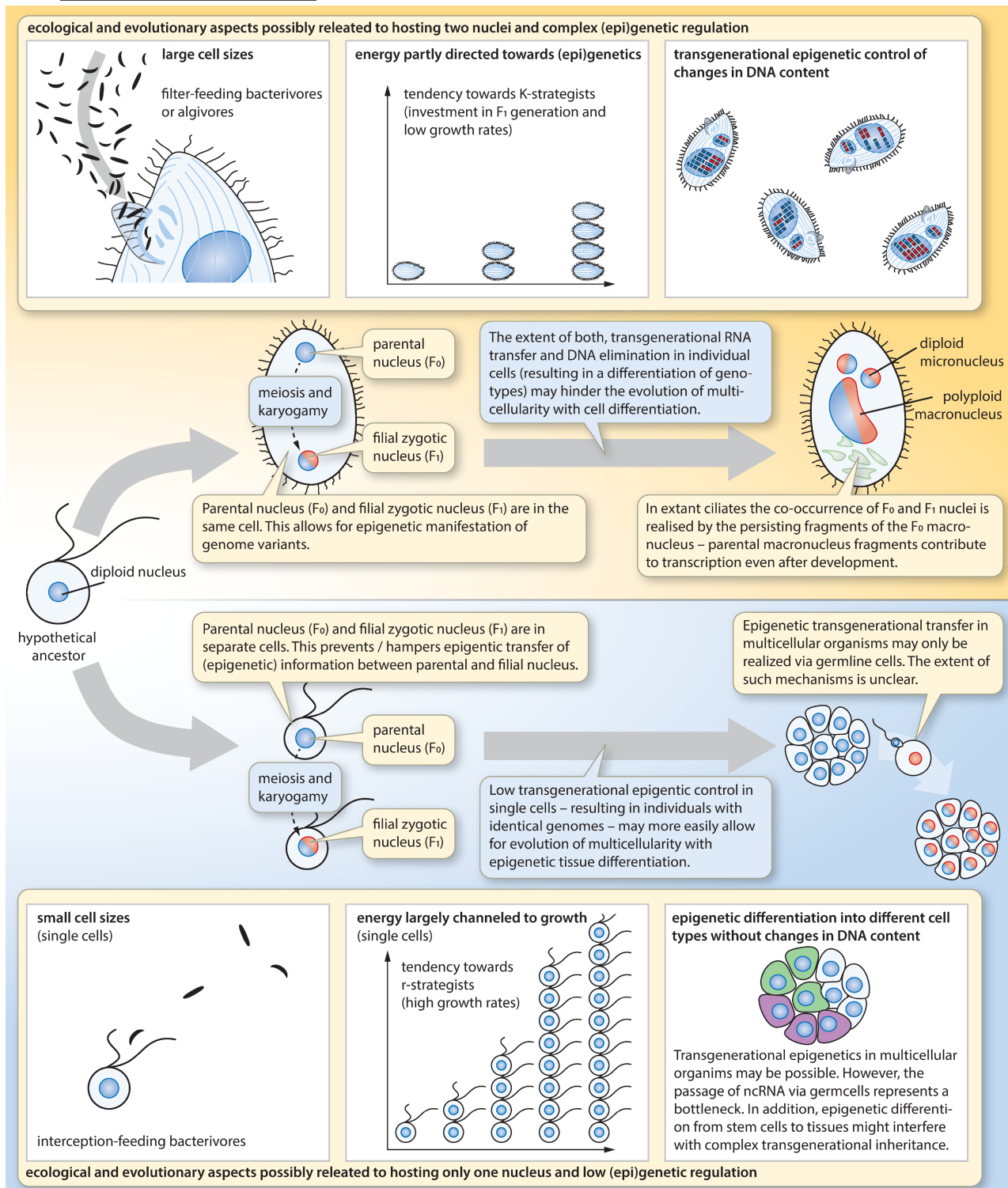


FIGURE 10 Evolutionary dissection of individual ciliate features in relation to mono-nuclear protists and multicellular organisms

WHAT DO WE LEARN FROM CILIATE EPIGENETICS?

We summarized here the current knowledge about ciliate epigenetics mechanisms: shortly outlined, they evolved genetic hardware in the MIC, which allows for a dynamic MAC composition that is to some extent

controlled by regulatory RNA. Interestingly, the latter's composition involves parental information: Acquired characters can be passed through generations, enabling the adaptation of MAC genotype and phenotype in a Lamarckian manner. Still, the real contribution of these mechanisms to adaptation in free-living ciliates remains to be elaborated.

Next to this, we describe that the amitotic MAC division is not a primitive mechanism as one could think at first glance compared to mitotic divisions. Re-functionalizing microtubules to divide the MAC indicates a critical tool that allows for phenotypic assortments and thus genetic variability in a single generation.

In principle, ciliate epigenetics uses common mechanisms found in many kingdoms: regulatory RNA, RNA interference, and RNA-induced heterochromatin formation. Also, the domestication of TE elements was recently described not to be unique to ciliates but to occur unexpectedly often (Jangam et al., 2017). As a result, these individual mechanisms seem to be specialized in ciliates but not unique to them. Acknowledging this, it seems tempting to speculate that ciliate epigenetics uncovers some mechanistic principles relevant to other species. As discussed, ciliates have an advantage for the transgenerational RNA transfer as parental and zygotic nuclei are in the same cytoplasm. Therefore, ciliates remain the preferred organisms to analyze the phenomenon of RNA transfer, which is the general challenge of transgenerational epigenetics, not only in ciliates. Specificity, timing, and transport mechanisms for RNA transfer are, in general, hardly understood, although being a common and essential factor of epigenetics across kingdoms.

ACKNOWLEDGMENTS

FD was supported by grant DFG SI1397/3-1 to MS. We thank Marcello Pirritano and Alexey Potekhin for their helpful discussions on the manuscript and Eric Meyer for discussion on the paternal scnRNAs. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Martin Simon  <https://orcid.org/0000-0002-0962-7788>

REFERENCES

- Abello, A., Régner, V., Arnaiz, O., Le Bars, R., Bétermier, M. & Bischerour, J. (2020) Functional diversification of *Paramecium* Ku80 paralogs safeguards genome integrity during precise programmed DNA elimination. *PLoS Genetics*, 16, e1008723.
- Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S. et al. (2012) The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, 59, 429–493.
- Allen, S.E., Hug, I., Pabian, S., Rzeszutek, I., Hoehener, C. & Nowacki, M. (2017) Circular concatemers of ultra-short DNA segments produce regulatory RNAs. *Cell*, 168, 990–999.e7.
- Allen, S.E. & Nowacki, M. (2020) Roles of noncoding RNAs in ciliate genome architecture. *Journal of Molecular Biology*, 432, 4186–4198.
- Arnaiz, O., Goût, J.-F., Bétermier, M., Bouhouche, K., Cohen, J., Duret, L. et al. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate *Paramecium tetraurelia*. *BMC Genomics*, 11, 547.
- Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.-M., Denby Wilkes, C. et al. (2012) The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genetics*, 8, e1002984.
- Ashe, A., Sapetschnig, A., Weick, E.-M., Mitchell, J., Bagijn, M.P., Cording, A.C. et al. (2012) piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans*. *Cell*, 150, 88–99.
- Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M. et al. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, 444, 171–178.
- Baranasic, D., Oppermann, T., Cheaib, M., Cullum, J., Schmidt, H. & Simon, M. (2014) Genomic characterization of variable surface antigens reveals a telomere position effect as a prerequisite for RNA interference-mediated silencing in *Paramecium tetraurelia*. *MBio*, 5, e01328.
- Baroin-Tourancheau, A., Delgado, P., Perasso, R. & Adoutte, A. (1992) A broad molecular phylogeny of ciliates: identification of major evolutionary trends and radiations within the phylum. *Proceedings of the National Academy of Sciences of the United States of America*, 89, 9764–9768.
- Berger, J.D. (1988) The cell cycle and regulation of cell mass and macronuclear DNA content. In: Görtz, H.-D. (Ed.) *Paramecium*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 97–119.
- Bétermier, M., Duharcourt, S., Seitz, H. & Meyer, E. (2000) Timing of developmentally programmed excision and circularization of *Paramecium* internal eliminated sequences. *Molecular and cellular biology*, 20.5, 1553–1561.
- Bissherour, J., Bhullar, S., Denby, W.C., Régner, V., Mathy, N., Dubois, E. et al. (2018) Six domesticated PiggyBac transposases together carry out programmed DNA elimination in *Paramecium*. *eLife*, 7, e37927.
- Blackburn, E.H. & Gall, J.G. (1978) A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *Journal of Molecular Biology*, 120, 33–53.
- Boenigk, J. (2021) *Boenigk, Biologie – Arbeitsbuch für Studium und Oberstufe*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Bouhouche, K., Gout, J.-F., Kapusta, A., Bétermier, M. & Meyer, E. (2011) Functional specialization of Piwi proteins in *Paramecium tetraurelia* from post-transcriptional gene silencing to genome remodelling. *Nucleic Acids Research*, 39, 4249–4264.
- Bourque, G. (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics & Development*, 19, 607–612.
- Bracht, J.R., Fang, W., Goldman, A.D., Dolzhenko, E., Stein, E.M. & Landweber, L.F. (2013) Genomes on the edge: programmed genome instability in ciliates. *Cell*, 152, 406–416.
- Braun, J., Nabergall, L., Neme, R., Landweber, L.F., Saito, M. & Jonoska, N. (2018) Russian doll genes and complex chromosome rearrangements in *Oxytricha trifallax*. *G3 (Bethesda)*, 8, 1669–1674.
- Burki, F., Roger, A.J., Brown, M.W. & Simpson, A.G.B. (2020) The new tree of eukaryotes. *Trends in Ecology & Evolution*, 35, 43–55.
- Burns, J., Kukushkin, D., Chen, X., Landweber, L.F., Saito, M. & Jonoska, N. (2016) Recurring patterns among scrambled genes in the encrypted genome of the ciliate *Oxytricha trifallax*. *Journal of Theoretical Biology*, 410, 171–180.
- Carbone, L., Alan Harris, R., Gnerre, S., Veeramah, K.R., Lorente-Galdos, B., Huddleston, J. et al. (2014) Gibbon genome and the fast karyotype evolution of small apes. *Nature*, 513, 195–201.
- Carbone, L., Vessere, G.M., ten Hallers, B.F.H., Zhu, B., Osogawa, K., Mootnick, A. et al. (2006) A high-resolution map of synteny disruptions in gibbon and human genomes. *PLoS Genetics*, 2, e223.
- Carradec, Q., Götz, U., Arnaiz, O., Pouch, J., Simon, M., Meyer, E. et al. (2015) Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate *Paramecium tetraurelia*. *Nucleic Acids Research*, 43, 1818–1833.
- Cassidy-Hanley, D., Bisharyan, Y., Fridman, V., Gerber, J., Lin, C., Orias, E. et al. (2005) Genome-wide characterization of *Tetrahymena thermophila* chromosome breakage sites. II. Physical and genetic mapping. *Genetics*, 170, 1623–1631.

- Catania, F., Rothering, R. & Vitali, V. (2021) One cell, two gears: extensive somatic genome plasticity accompanies high germline genome stability in *Paramecium*. *Genome Biology and Evolution*, 13(12), evab263.
- Cecere, G. (2021) Small RNAs in epigenetic inheritance: from mechanisms to trait transmission. *FEBS Letters*, 595, 2953–2977.
- Cerutti, H. & Casas-Mollano, J.A. (2006) On the origin and functions of RNA-mediated silencing: from protists to man. *Current Genetics*, 50, 81–99.
- Chalker, D.L. & Yao, M.-C. (2011) DNA elimination in ciliates: transposon domestication and genome surveillance. *Annual Review of Genetics*, 45, 227–246.
- Chen, X., Bracht, J., Goldman, A., Dolzhenko, E., Clay, D., Swart, E. et al. (2014) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell*, 158, 1187–1198.
- Chen, X., Jung, S., Beh, L.Y., Eddy, S.R. & Landweber, L.F. (2015) Combinatorial DNA rearrangement facilitates the origin of new genes in ciliates. *Genome Biology and Evolution*, 7, 2859–2870.
- Cheng, C.-Y., Orias, E., Leu, J.-Y. & Turkewitz, A.P. (2020) The evolution of germ-soma nuclear differentiation in eukaryotic unicells. *Current Biology*, 30, R502–R510.
- Clark, T.G. & Forney, J.D. (2003) Free-living and parasitic ciliates. In Craig, A., & Scherf, A. eds, *Antigenic variation*. London: Academic Press, pp. 375–402.
- Czech, B., Munafò, M., Ciabrelli, F., Eastwood, E.L., Fabry, M.H., Kneuss, E. et al. (2018) piRNA-guided genome defense: from biogenesis to silencing. *Annual Review of Genetics*, 52, 131–157.
- Doerder, F.P. & DeBault, L.E. (1978) Life cycle variation and regulation of macronuclear DNA content in *Tetrahymena thermophila*. *Chromosoma*, 69, 1–19.
- Drews, F., Karunanithi, S., Götz, U., Marker, S., deWijn, R., Pirritano, M. et al. (2021) Two Piwis with Ago-like functions silence somatic genes at the chromatin level. *RNA Biology*, 18, 757–769.
- Drews, F., Salhab, A., Karunanithi, S., Cheaib, M., Jung, M., Schulz, M. et al. (2022) Broad domains of histone marks in the highly compact *Paramecium* macronuclear genome. *Genome Research*, 32, 710–725.
- Dubois, E., Mathy, N., Régner, V., Bischerour, J., Baudry, C., Trouslard, R. et al. (2017) Multimerization properties of PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements. *Nucleic Acids Research*, 45, 3204–3216.
- Duharcourt, S., Butler, A. & Meyer, E. (1995) Epigenetic self-regulation of developmental excision of an internal eliminated sequence on *Paramecium tetraurelia*. *Genes & Development*, 9, 2065–2077.
- Duharcourt, S., Keller, A.M. & Meyer, E. (1998) Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in *Paramecium tetraurelia*. *Molecular and Cellular Biology*, 18, 7075–7085.
- Duret, L., Cohen, J., Jubin, C., Dessen, P., Goût, J.-F., Mousset, S. et al. (2008) Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: a somatic view of the germline. *Genome Research*, 18, 585–596.
- Epstein, L.M. & Forney, J.D. (1984) Mendelian and non-mendelian mutations affecting surface antigen expression in *Paramecium tetraurelia*. *Molecular and Cellular Biology*, 4, 1583–1590.
- Fang, W., Wang, X., Bracht, J.R., Nowacki, M. & Landweber, L.F. (2012) Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement. *Cell*, 151, 1243–1255.
- Foissner, W., Chao, A. & Katz, L.A. (2009) Diversity and geographic distribution of ciliates (Protista: Ciliophora). In: Foissner, W. & Hawksworth, D.L. (Eds.) *Protist diversity and geographical distribution. Vol. 8. Topics in biodiversity and conservation*. Dordrecht, Netherlands: Springer, pp. 111–129.
- Finlay, B.J., Esteban, G.F. & Fenchel, T. (1998) Protozoan diversity: converging estimates of the global number of free-living ciliate species. *Protist*, 149(1), 29–37.
- Forney, J.D. & Blackburn, E.H. (1988) Developmentally controlled telomere addition in wild-type and mutant paramecia. *Molecular and Cellular Biology*, 8, 251–258.
- Frapporti, A., Miró, P.C., Arnaiz, O., Holoch, D., Kawaguchi, T., Humbert, A. et al. (2019) The Polycomb protein Ez11 mediates H3K9 and H3K27 methylation to repress transposable elements in *Paramecium*. *Nature Communications*, 10, 2710.
- Furrer, D.I., Swart, E.C., Kraft, M.F., Sandoval, P.Y. & Nowacki, M. (2017) Two sets of Piwi proteins are involved in distinct sRNA pathways leading to elimination of germline-specific DNA. *Cell Reports*, 20, 505–520.
- Galvani, A. & Sperling, L. (2002) RNA interference by feeding in *Paramecium*. *Trends in Genetics*, 18, 11–12.
- Gao, F., Roy, S.W. & Katz, L.A. (2015) Analyses of alternatively processed genes in ciliates provide insights into the origins of scrambled genomes and may provide a mechanism for speciation. *MBio*, 6 (1), e01998–14.
- Gao, F., Song, W. & Katz, L.A. (2014) Genome structure drives patterns of gene family evolution in ciliates, a case study using *Chilodonella uncinata* (Protista, Ciliophora, Phyllopharyngea). *Evolution*, 68, 2287–2295.
- Garnier, O., Serrano, V., Duharcourt, S. & Meyer, E. (2004) RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. *Molecular and Cellular Biology*, 24, 7370–7379.
- Gilley, D. & Blackburn, E.H. (1994) Lack of telomere shortening during senescence in *Paramecium*. *Proceedings of the National Academy of Sciences of the United States of America*, 91, 1955–1958.
- Gilley, D., Preer, J.R., Aufderheide, K.J. & Polisky, B. (1988) Autonomous replication and addition of telomere-like sequences to DNA microinjected into *Paramecium tetraurelia* macronuclei. *Molecular and Cellular Biology*, 8, 4765–4772.
- Godiska, R., Aufderheide, K.J., Gilley, D., Hendrie, P., Fitzwater, T., Preer, L.B. et al. (1987) Transformation of *Paramecium* by microinjection of a cloned serotype gene. *Proceedings of the National Academy of Sciences of the United States of America*, 84, 7590–7594.
- Götz, U., Marker, S., Cheaib, M., Andresen, K., Shrestha, S., Durai, D.A. et al. (2016) Two sets of RNAi components are required for heterochromatin formation in trans triggered by truncated transgenes. *Nucleic Acids Research*, 44, 5908–5923.
- Gratias, A. & Bétermier, M. (2003) Processing of double-strand breaks is involved in the precise excision of *paramecium* internal eliminated sequences. *Molecular and Cellular Biology*, 23, 7152–7162.
- Grattepanche, J.-D., Walker, L.M., Ott, B.M., Paim Pinto, D.L., Delwiche, C.F., Lane, C.E. et al. (2018) Microbial diversity in the eukaryotic SAR clade: illuminating the darkness between morphology and molecular data. *BioEssays*, 40, e1700198.
- Greider, C.W. & Blackburn, E.H. (1985) Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. *Cell*, 43, 405–413.
- Gruchota, J., Denby, W.C., Arnaiz, O., Sperling, L. & Nowak, J.K. (2017) A meiosis-specific Spt5 homolog involved in non-coding transcription. *Nucleic Acids Research*, 45, 4722–4732.
- Guérin, F., Arnaiz, O., Boggetto, N., Denby, W.C., Meyer, E., Sperling, L. et al. (2017) Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements. *BMC Genomics*, 18, 327.
- Hamilton, E., Bruns, P., Lin, C., Merriam, V., Orias, E., Vong, L. et al. (2005) Genome-wide characterization of *tetrahymena thermophila* chromosome breakage sites. I. Cloning and identification of functional sites. *Genetics*, 170, 1611–1621.

- Hamilton, E.P., Kapusta, A., Huvos, P.E., Bidwell, S.L., Zafar, N., Tang, H. et al. (2016) Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. *eLife*, 5, e19090.
- Hammerschmidt, B., Schlegel, M., Lynn, D.H., Leipe, D.D., Sogin, M.L. & Raikov, I.B. (1996) Insights into the evolution of nuclear dualism in the ciliates revealed by phylogenetic analysis of rRNA sequences. *Journal of Eukaryotic Microbiology*, 43, 225–230.
- Herrick, G. (1994) Germline-soma relationships in ciliated protozoa: the inception and evolution of nuclear dimorphism in one-celled animals. *Seminars in Developmental Biology*, 5, 3–12.
- Heyse, G., Jönsson, F., Chang, W.-J. & Lipps, H.J. (2010) RNA-dependent control of gene amplification. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 22134–22139.
- Hoehener, C., Hug, I. & Nowacki, M. (2018) Dicer-like enzymes with sequence cleavage preferences. *Cell*, 173, 234–247.e7.
- Houri-Zeevi, L., Korem, K.Y., Antonova, O. & Rechavi, O. (2020) Three rules explain transgenerational small RNA inheritance in *C. elegans*. *Cell*, 182, 1186–1197.e12.
- Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H. et al. (2016) Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell*, 166, 102–114.
- Ignarski, M., Singh, A., Swart, E.C., Arambasic, M., Sandoval, P.Y. & Nowacki, M. (2014) *Paramecium tetraurelia* chromatin assembly factor-1-like protein PtCAF-1 is involved in RNA-mediated control of DNA elimination. *Nucleic Acids Research*, 42, 11952–11964.
- Jangam, D., Feschotte, C. & Betrán, E. (2017) Transposable element domestication as an adaptation to evolutionary conflicts. *Trends in Genetics*, 33, 817–831.
- Jenkins, R.A. (1977) The role of microtubules in macronuclear division of *blepharisma* *. *The Journal of Protozoology*, 24, 264–275.
- Kabi, M. & Filion, G.J. (2021) Heterochromatin: did H3K9 methylation evolve to tame transposons? *Genome Biology*, 22, 325.
- Kapusta, A., Matsuda, A., Marmignon, A., Ku, M., Silve, A., Meyer, E. et al. (2011) Highly precise and developmentally programmed genome assembly in *Paramecium* requires ligase IV-dependent end joining. *PLoS Genetics*, 7, e1002049.
- Karunanithi, S., Oruganti, V., Marker, S., Rodriguez-Viana, A.M., Drews, F., Pirritano, M. et al. (2019) Exogenous RNAi mechanisms contribute to transcriptome adaptation by phased siRNA clusters in *Paramecium*. *Nucleic Acids Research*, 47, 8036–8049.
- Karunanithi, S., Oruganti, V., de Wijn, R., Drews, F., Cheaib, M., Nordström, K. et al. (2020) Feeding exogenous dsRNA interferes with endogenous sRNA accumulation in *Paramecium*. *DNA Research*, 27, 1–10.
- Katz, L.A. (2001) Evolution of nuclear dualism in ciliates: a reanalysis in light of recent molecular data. *International Journal of Systematic and Evolutionary Microbiology*, 51, 1587–1592.
- Katz, L.A., Bornstein, J.G., Lasek-Nesselquist, E. & Muse, S.V. (2004) Dramatic diversity of ciliate histone H4 genes revealed by comparisons of patterns of substitutions and paralog divergences among eukaryotes. *Molecular Biology and Evolution*, 21, 555–562.
- Katz, L.A. & Kovner, A.M. (2010) Alternative processing of scrambled genes generates protein diversity in the ciliate *Chilodonella uncinata*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 314, 480–488.
- Khurana, J.S., Wang, X., Chen, X., Perlman, D.H. & Landweber, L.F. (2014) Transcription-independent functions of an RNA polymerase II subunit, Rpb2, during genome rearrangement in the ciliate, *Oxytricha trifallax*. *Genetics*, 197, 839–849.
- Klobutcher, L.A. & Herrick, G. (1997) Developmental genome reorganization in ciliated protozoa: the transposon link. *Progress in Nucleic Acid Research and Molecular Biology*, 56, 1–62.
- Kovaleva, V.G. & Raikov, I.B. (1978) Diminution and re-synthesis of DNA during development and senescence of the? diploid? Macronuclei of the ciliate *Trachelonema sulcata* (Gymnostomata, Karyorelictida). *Chromosoma*, 67, 177–192.
- Le Pen, J., Jiang, H., Di Domenico, T., Kneuss, E., Kosalka, J., Leung, C. et al. (2018) Terminal uridylyltransferases target RNA viruses as part of the innate immune system. *Nature Structural & Molecular Biology*, 25, 778–786.
- Lee, S.R. & Collins, K. (2007) Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs. *Nature Structural & Molecular Biology*, 14, 604–610.
- Lepère, G., Bétermier, M., Meyer, E. & Duharcourt, S. (2008) Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in *Paramecium tetraurelia*. *Genes & Development*, 22, 1501–1512.
- Lepère, G., Nowacki, M., Serrano, V., Gout, J.-F., Guglielmi, G., Duharcourt, S. et al. (2009) Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. *Nucleic Acids Research*, 37, 903–915.
- Lhuillier-Akakpo, M., Frapporti, A., Denby, W.C., Matelot, M., Vervoort, M., Sperling, L. et al. (2014) Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements. *PLoS Genetics*, 10, e1004665.
- Lindblad, K.A., Bracht, J.R., Williams, A.E. & Landweber, L.F. (2017) Thousands of RNA-cached copies of whole chromosomes are present in the ciliate *Oxytricha* during development. *RNA*, 23, 1200–1208.
- Loidl, J. (2021) *Tetrahymena* meiosis: simple yet ingenious. *PLoS Genetics*, 17, e1009627.
- Luteijn, M.J., van Bergeijk, P., Kaaij, L.J.T., Almeida, M.V., Roovers, E.F., Berezikov, E. et al. (2012) Extremely stable Piwi-induced gene silencing in *Caenorhabditis elegans*. *EMBO Journal*, 31, 3422–3430.
- Maliszewska-Olejniczak, K., Gruchota, J., Gromadka, R., Denby, W.C., Arnaiz, O., Mathy, N. et al. (2015) TFIIIS-dependent non-coding transcription regulates developmental genome rearrangements. *PLoS Genetics*, 11, e1005383.
- Marker, S., Carradec, Q., Tanty, V., Arnaiz, O. & Meyer, E. (2014) A forward genetic screen reveals essential and non-essential RNAi factors in *Paramecium tetraurelia*. *Nucleic Acids Research*, 42, 7268–7280.
- Marker, S., Le Mouël, A., Meyer, E. & Simon, M. (2010) Distinct RNA-dependent RNA polymerases are required for RNAi triggered by double-stranded RNA versus truncated transgenes in *Paramecium tetraurelia*. *Nucleic Acids Research*, 38, 4092–4107.
- Marmignon, A., Bischerour, J., Silve, A., Fojcik, C., Dubois, E., Arnaiz, O. et al. (2014) Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *PLoS Genetics*, 10, e1004552.
- Maurer-Alcalá, X.X. & Nowacki, M. (2019) Evolutionary origins and impacts of genome architecture in ciliates. *Annals of the New York Academy of Sciences*, 1447, 110–118.
- Maurer-Alcalá, X.X., Yan, Y., Pilling, O.A., Knight, R. & Katz, L.A. (2018) Twisted tales: insights into genome diversity of ciliates using single-cell 'omics. *Genome Biology and Evolution*, 10, 1927–1939.
- Meyer, E. (1992) Induction of specific macronuclear developmental mutations by microinjection of a cloned telomeric gene in *Paramecium primaurelia*. *Genes & Development*, 6, 211–222.
- Meyer, E. & Keller, A.-M. (1996) A mendelian mutation affecting mating-type determination also affects developmental genomic rearrangements in *Paramecium tetraurelia*. *Genetics*, 143, 191–202.
- Nekrasova, I., Nikitashina, V., Bhullar, S., Arnaiz, O., Singh, D.P., Meyer, E. et al. (2019) Loss of a fragile chromosome region

- leads to the screwy phenotype in *Paramecium tetraurelia*. *Genes (Basel)*, 10(7), 513.
- Nekrasova, I.V. & Potekhin, A.A. (2019) Diversity of RNA interference pathways in regulation of endogenous and exogenous sequences expression in ciliates Tetrahymena and Paramecium. *Ecological Genetics*, 17, 113–125.
- Noto, T., Kataoka, K., Suhren, J.H., Hayashi, A., Woolcock, K.J., Gorovsky, M.A. et al. (2015) Small-RNA-mediated genome-wide trans-recognition network in tetrahymena DNA elimination. *Molecular Cell*, 59, 229–242.
- Nowacki, M., Haye, J.E., Fang, W., Vijayan, V. & Landweber, L.F. (2010) RNA-mediated epigenetic regulation of DNA copy number. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 22140–22144.
- Nowacki, M., Higgins, B.P., Maquilan, G.M., Swart, E.C., Doak, T.G. & Landweber, L.F. (2009) A functional role for transposases in a large eukaryotic genome. *Science*, 324, 935–938.
- Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T.G. & Landweber, L.F. (2008) RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature*, 451, 153–158.
- Orias, E. (1991a) Evolution of amitosis of the ciliate macronucleus: gain of the capacity to divide. *The Journal of Protozoology*, 38, 217–221.
- Orias, E. (1991b) On the evolution of the karyorelict ciliate life cycle: heterophasic ciliates and the origin of ciliate binary fission. *BioSystems*, 25, 67–73.
- Orias, E. & Flacks, M. (1975) Macronuclear genetics of Tetrahymena I. Random distribution of macronuclear gene copies in *T. pyriformis*, Syngen 1. *Genetics*, 79, 187–206.
- Ovchinnikova, L.P. & Selivanova, G.V. (1965) Photometric study of the DNA content in the nuclei of *Spirostomum ambiguum* (Ciliata, Heterotricha). *Acta Protozoologica*, 3, 1–8.
- Owsian, D., Gruchota, J., Arnaiz, O. & Nowak, J.K. (2022) The transient Spt4-Spt5 complex as an upstream regulator of non-coding RNAs during development. *Nucleic Acids Research*, 50(5), 2603–2620. <https://doi.org/10.1093/nar/gkac106>
- Pina, C.M., Kawaguchi, T., Charmant, O., Michaud, A., Cohen, I., Humbert, A. et al. (2021) *Paramecium* Polycomb Repressive Complex 2 physically interacts with the small RNA binding PIWI protein to repress transposable elements. *BioRxiv*.
- Postberg, J., Heyse, K., Cremer, M., Cremer, T. & Lipps, H.J. (2008) Spatial and temporal plasticity of chromatin during programmed DNA-reorganization in Stylonychia macronuclear development. *Epigenetics Chromatin*, 1, 3.
- Postberg, J., Weil, P.P. & Pembaur, A. (2019) Biogenesis of developmental master regulatory 27nt-RNAs in stylonychia-can coding RNA turn into non-coding? *Genes (Basel)*, 10, 940.
- Preer, J.R. (1997) Whatever happened to paramecium genetics? *Genetics*, 145, 217–225.
- Prescott, D.M. (1999) The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates. *Nucleic Acids Research*, 27, 1243–1250.
- Raikov, I.B. (1976) Evolution of macronuclear organization. *Annual Review of Genetics*, 10, 413–440.
- Raikov, I.B. (1982) The protozoan nucleus–morphology and evolution. *The protozoan nucleus–morphology and evolution*.
- Raikov, I.B. (1985) Primitive never-dividing macronuclei of some lower ciliates. *International Review of Cytology*, 95, 267–325.
- Raikov, I.B. (1994) The nuclear apparatus of some primitive ciliates, the karyorelictids: structure and divisional reorganization. *Bolletino Di Zoologia*, 61, 19–28.
- Raikov, I.B., Cheissin, M. & Buze, E.G. (1963) A photometric study of DNA content of macro- and micronuclei in *Paramecium caudatum*, *Nassula ornata* and *Loxodes magnus*. *Acta Protozoologica*, 1, 25–30.
- Raikov, I.B. & Karadzhan, B.P. (1985) Fine structure and cytochemistry of the nuclei of the primitive ciliate *Tracheloraphis crassus* (Karyorelictida). *Protoplasma*, 126, 114–129.
- Rechavi, O., Houri-Ze'evi, L., Anava, S., Goh, W.S.S., Kerk, S.Y., Hannon, G.J. et al. (2014) Starvation-induced transgenerational inheritance of small RNAs in *C. elegans*. *Cell*, 158, 277–287.
- Ricci, F., Luporini, P., Alimenti, C. & Vallesi, A. (2021) Functional chimeric genes in ciliates: an instructive case from *Euplotes raikovi*. *Gene*, 767, 145186.
- Sandoval, P.Y., Swart, E.C., Arambasic, M. & Nowacki, M. (2014) Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting. *Developmental Cell*, 28, 174–188.
- Sawka-Gądek, N., Potekhin, A., Singh, D.P., Grevtseva, I., Arnaiz, O., Penel, S. et al. (2021) Evolutionary plasticity of mating-type determination mechanisms in *Paramecium aurelia* sibling species. *Genome Biology and Evolution*, 13, evaa258.
- Schoeberl, U.E. & Mochizuki, K. (2011) Keeping the soma free of transposons: programmed DNA elimination in ciliates. *Journal of Biological Chemistry*, 286, 37045–37052.
- Sellis, D., Guérin, F., Arnaiz, O., Pett, W., Lerat, E., Boggetto, N. et al. (2021) Massive colonization of protein-coding exons by selfish genetic elements in *Paramecium* germline genomes. *PLoS Biology*, 19, e3001309.
- Sharma, U., Sun, F., Conine, C.C., Reichholz, B., Kukreja, S., Herzog, V.A. et al. (2018) Small RNAs are trafficked from the epididymis to developing mammalian sperm. *Developmental Cell*, 46, 481–494.e6.
- Sheng, Y., Duan, L., Cheng, T., Qiao, Y., Stover, N.A. & Gao, S. (2020) The completed macronuclear genome of a model ciliate *Tetrahymena thermophila* and its application in genome scrambling and copy number analyses. *Science China Life Sciences*, 63, 1534–1542.
- Shirayama, M., Seth, M., Lee, H.-C., Gu, W., Ishidate, T., Conte, D. et al. (2012) piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. *Cell*, 150, 65–77.
- Singh, D.P., Saudemont, B., Guglielmi, G., Arnaiz, O., Goût, J.-F., Prajer, M. et al. (2014) Genome-defence small RNAs exapted for epigenetic mating-type inheritance. *Nature*, 509, 447–452.
- Sonneborn, T.M. (1937) Sex, sex inheritance and sex determination in *Paramecium aurelia*. *Proceedings of the National Academy of Sciences of the United States of America*, 23, 378–385.
- Sonneborn, T.M. & Lesuer, A. (1948) Antigenic characters in *Paramecium aurelia*, variety 4; determination, inheritance and induced mutations. *American Naturalist*, 82, 69–78.
- Spear, B.B. & Lauth, M.R. (1976) Polytene chromosomes of Oxytricha: biochemical and morphological changes during macronuclear development in a ciliated protozoan. *Chromosoma*, 54, 1–13.
- Swart, E.C., Bracht, J.R., Magrini, V., Minx, P., Chen, X., Zhou, Y.I. et al. (2013) The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biology*, 11, e1001473.
- Swart, E.C., Wilkes, C.D., Sandoval, P.Y., Arambasic, M., Sperling, L. & Nowacki, M. (2014) Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion. *Nucleic Acids Research*, 42, 8970–8983.
- Tucker, J.B., Beisson, J., Roche, D.L. & Cohen, J. (1980) Microtubules and control of macronuclear “amitosis” in *Paramecium*. *Journal of Cell Science*, 44, 135–151.
- de Vanssay, A., Touzeau, A., Arnaiz, O., Frapporti, A., Phipps, J. & Duharcourt, S. (2020) The *Paramecium* histone chaperone Spt16-1 is required for Pgm endonuclease function in programmed genome rearrangements. *PLoS Genetics*, 16, e1008949.
- Vitali, V., Hagen, R. & Catania, F. (2019) Environmentally induced plasticity of programmed DNA elimination boosts somatic variability in *Paramecium tetraurelia*. *Genome Research*, 29, 1693–1704.

- Vogt, A., Goldman, A.D., Mochizuki, K. & Landweber, L.F. (2013) Transposon domestication versus mutualism in ciliate genome rearrangements. *PLoS Genetics*, 9, e1003659.
- Warren, I.A., Naville, M., Chalopin, D., Levin, P., Berger, C.S., Galiana, D. et al. (2015) Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Research*, 23, 505–531.
- Xu, K., Doak, T.G., Lipps, H.J., Wang, J., Swart, E.C. & Chang, W.-J. (2012) Copy number variations of 11 macronuclear chromosomes and their gene expression in *Oxytricha trifallax*. *Gene*, 505, 75–80.
- Yan, Y., Maurer-Alcalá, X.X., Knight, R., Kosakovsky Pond, S.L. & Katz, L.A. (2019) Single-cell transcriptomics reveal a correlation between genome architecture and gene family evolution in ciliates. *MBio*, 10, e02524–19.
- Yan, Y., Rogers, A.J., Gao, F. & Katz, L.A. (2017) Unusual features of non-dividing somatic macronuclei in the ciliate class Karyorelictea. *European Journal of Protistology*, 61, 399–408.
- Yao, M.-C. (2010) Modulating somatic DNA copy number through maternal RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 21951–21952.
- Yerlici, V.T. & Landweber, L.F. (2014) Programmed genome rearrangements in the ciliate oxytricha. *Microbiology Spectrum*, 2, 2–6.
- Yerlici, V.T., Lu, M.W., Hoge, C.R., Miller, R.V., Neme, R., Khurana, J.S. et al. (2019) Programmed genome rearrangements in Oxytricha produce transcriptionally active extrachromosomal circular DNA. *Nucleic Acids Research*, 47, 9741–9760.
- Zufall, R.A., McGrath, C.L., Muse, S.V. & Katz, L.A. (2006) Genome architecture drives protein evolution in ciliates. *Molecular Biology and Evolution*, 23, 1681–1687.

How to cite this article: Drews, F., Boenigk, J. & Simon, M. (2022) *Paramecium* epigenetics in development and proliferation. *Journal of Eukaryotic Microbiology*, 00, e12914. Available from: <https://doi.org/10.1111/jeu.12914>

Acknowledgments

The last years in the lab that led to the finalization of my doctorate felt like a long, exciting, and sometimes challenging journey that would not even have started (and now ends) without the help and support of several amazing people. Therefore, I want to take the opportunity to thank everyone who has supported me in his/her very own way.

The greatest thanks go to my supervisor, Prof. Dr. Martin Simon, who supported me from the very beginning and offered me the opportunity to learn so much about epigenetics and chromatin biology. I thank you for believing in my expertise, although I crawled out of a cheap bus after a 18-hour trip to sign my contract in Saarland. Martin, ich danke dir auch für dein herzliches Willkommen im fernen Saarland und dein offenes Ohr, auch in Fragen abseits des Laborgeschehens. Ich danke dir auch für dein Vertrauen in mich, deinen stetem Anschlag und vor allem für die vielzähligen Möglichkeiten, mich in der wissenschaftlichen Community zu integrieren und auf nationalen wie internationalen Meetings unsere Ergebnisse zu präsentieren. Jedes Event, jedes Seminar ist eine kleine Herausforderung für mich und ich danke dir dafür, dass du mir all das zutraust.

I am also very grateful to Prof. Dr. Jörn Walter for reviewing this thesis, and furthermore, I would like to thank him for continuous input on our *Paramecium* epigenetics model and fruitful discussions in the kitchen after the Thursday seminar. Einen Berliner im Saarland zu treffen und gemeinsam Feuerzangenbowle auf der Weihnachtsfeier zu trinken war mir eine Freude.

I want to add some more words on my short stop in Saarbrücken, where my PhD journey started. I thank all the people from the Genetics Department for their warm welcome, funny lunch breaks, and lab outings. Special thanks go to Dr. Gilles Gasparoni and Dr. Abdulrahman Salhab who took care of my precious samples and especially in my early days explained the bioinformatics pipelines to me. Special thanks go to Dr. Martin Jung, who supported me with his expertise in antibody purification and warmly welcomed me in his lab in Homburg.

Many thanks go to Steffi and Andrea, former members of the Cell Biology Department: I thank you for your support in western blotting and especially Steffi for her help in all the organizational stuff that came along with finalizing this thesis. Ich danke euch beiden für die tolle Zeit in Saarbrücken und die Samstag-Abend Events. I thank Lara Sohn (Katzenstimme Grenzenlos) who takes care of the Uni-Kittens and rescued Fred in Croatia thus giving me the opportunity to become a weird cat lady. Kein Dank geht an Fred und Toni. Nutzloses Schlafen auf der Tasta war mir keine Hilfe.

I also owe a great thanks to all the collaborators and my co-authors who essentially contributed to these studies. I am thankful to Dr. Torsten Möhlmann and Dr. Daniel Hickl for our great collaboration on the *Arabidopsis* vacuolar RNAome. It's nice to keep in touch with botanists, discussing all the stuff I learned (and forgot) during my master's studies. Thanks also go to Prof. Dr. Jens Boenigk for fruitful discussions and beautiful images that contributed to our manuscript. I thank Dr. Marcel Schulz and Dr. Sivarajan Karunanithi for scientific discussions and Siva for his continuous bioinformatic support. I would like to thank all members of the Swart lab for their nice welcome in Tübingen and Dr. Aditi Singh for fruitful discussions and support in *Paramecium* chromatin biology.

My PhD journey included moving from Saarland to Wuppertal and I thank Melanie Möller and Natalie Fabis for their warm welcome. It's fun working with you. Thanks also go to the members of the Bornhorst-Team, who welcomed all of us in Wuppertal.

Several people supported me without exactly knowing what I'm doing in the lab and why finishing this thesis took ages. But without their support, I would have gone running into the woods. Meine Lisa, du zeigst grenzenlose Unterstützung, ungefragt zu jeder Zeit. Ich danke dir von Herzen für deine jahrelange Freundschaft und die tollen Abende in Frankfurt. Sandy und Vici, danke für euer Verständnis, die schönen Mädeltreffen und eure stets offenen Ohren. Ich danke meinen Berliner Freundinnen, die keine Ahnung haben, was ich in Wuppertal tue, aber mich immer wissen lassen, dass Sie an mich denken. Danke auch an die Kendricks, die mir immer wieder die Möglichkeit geben, meine Akkus aufzutanken.

There is one person I could list hundreds of things he did to support me and contributed to my personal and scientific development from the very first day in Saarbrücken. Marcello, I don't even know where to begin :) You are the best colleague one could ask for; you are supportive, selfless, and empathic, and sharing an office with you was probably one of the best parts of my PhD journey. I owe you a huge thanks and many apologies for always talking to myself, forcing you to celebrate my birthday and listening to my cat stories. Ich danke dir für deine Freundschaft und Unterstützung, nicht nur im Labor und am PC sondern auch abseits des Uni-Alltags bei einem deiner grandiosen Drinks.

I'm very grateful to my favorite Kellner for always supporting me with an open bottle of wine when things got serious. Christoph, ich danke dir für deine Informatik-Expertise und deinen Enthusiasmus, wenn es um das "Skripten" geht. Ich lerne gern von und mit dir. Auch wenn wir oft viele Kilometer voneinander getrennt sind, weiß ich, dass du für mich da bist. Du machst mich glücklich. Danke.

My sincere thanks go to my mum and dad, Manuela and Ralf. Ich danke euch für euren Rückhalt und euer stetes Verständnis dafür, dass ich nicht zu jeder Familienfeier anwesend sein kann. Ihr habt mich mit dem Ehrgeiz ausgestattet, der mich dazu veranlasst hat, diese Promotion überhaupt zu beginnen. Mama, ich finde es großartig, wie viel ich von dir lernen kann und unseren Austausch über SchülerInnen und StudentInnen bringt mich immer zum Lachen. In deiner Art zu Lehren und in deiner Gelassenheit bist du mir ein großes Vorbild.

Declaration of Authorship

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Nutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die vorliegende Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Ich erkläre darüber Hinaus mit meiner Unterschrift, dass ich

- keine im Merkblatt "Hinweise zur Vermeidung von Plagiaten" der Naturwissenschaftlich - Technischen Fakultät beschriebenen Form des Plagiats begangen habe,
- alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu dokumentiert habe,
- keine Daten manipuliert habe.

Unterschrift:

Datum:
