A germline-limited *piggyBac* transposase gene is required for precise excision in *Tetrahymena* genome rearrangement

Lifang Feng[1,2,3], Guangying Wang[2], Eileen P. Hamilton[4], Jie Xiong[2], Guanxiong Yan[2], Kai Chen[2], Xiao Chen[1], Wen Dui[1], Amber Plemens[1], Lara Khadr[1], Arjune Dhanekula[1], Mina Juma[1], Hung Quang Dang[5], Geoffrey M. Kapler[5], Eduardo Orias[4], Wei Miao[2,*] and Yifan Liu[1,*]

[1]Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA, [2]Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China, [3]School of Food Science and Biotechnology, Zhejiang Gongshang University, Hangzhou 310018, China, [4]Department of Molecular, Cellular, and Developmental Biology, University of California, Santa Barbara, CA 93106, USA and [5]Department of Molecular and Cellular Medicine, Texas A&M University Health Science Center, College Station, TX 77843, USA

ABSTRACT

Developmentally programmed genome rearrangement accompanies differentiation of the silent germline micronucleus into the transcriptionally active somatic macronucleus in the ciliated protozoan *Tetrahymena thermophila*. Internal eliminated sequences (IES) are excised, followed by rejoining of MAC-destined sequences, while fragmentation occurs at conserved chromosome breakage sequences, generating macronuclear chromosomes. Some macronuclear chromosomes, referred to as non-maintained chromosomes (NMC), are lost soon after differentiation. Large NMC contain genes implicated in development-specific roles. One such gene encodes the domesticated *piggyBac* transposase TPB6, required for heterochromatin-dependent precise excision of IES residing within exons of functionally important genes. These conserved exonic IES determine alternative transcription products in the developing macronucleus; some even contain free-standing genes. Examples of precise loss of some exonic IES in the micronucleus and retention of others in the macronucleus of related species suggest an evolutionary analogy to introns. Our results reveal that germline-limited sequences can encode genes with specific expression patterns and development-related functions, which may be a recurring theme in eukaryotic organisms experiencing programmed genome rearrangement during germline to soma differentiation.

INTRODUCTION

As a manifestation of the dynamic genome, programmed genome rearrangement can occur in development, especially during germline to soma differentiation. This phenomenon, first observed in nematodes over a hundred years ago, is wide-spread in eukaryotes (1). In particular, ciliates feature two types of nuclei with distinct genomic structures and contents—the germline micronucleus (MIC) and somatic macronucleus (MAC), providing unique opportunities for mechanistic studies (2). Programmed genome rearrangement can be used to generate functional genes, as demonstrated by V(D)J recombination for the assembly of immunoglobins and T cell receptors in mammals (3), as well as mating type determination (4) and massive gene unscrambling (5) in the ciliated protozoa *Tetrahymena thermophila* and *Oxytricha trifallax*, respectively. Programmed genome rearrangement may also eliminate large amounts of DNA, sometimes the bulk of the germline genome (1,6,7), and likely evolved as a genome defense mechanism, as many eliminated sequences are transposon-related or simple repeats (5,8,9). Intriguingly, some host genes, expressed and functionally implicated in germline or early development, are also eliminated by this process (5,10). Thus, growing evidence suggests that programmed genome rearrangement may be exapted for gene regulation, as extreme measures for the permanent silencing of genes in soma.

During conjugation, the sexual phase of the life cycle of the ciliated protozoan *T. thermophila*, two major forms of

programmed genome rearrangement accompany the differentiation of the germline MIC to somatic MAC (11,12). The first is chromosome fragmentation at chromosome breakage sequences (CBS), followed by *de novo* telomere addition (Figure 1A). CBS feature a 15 bp consensus sequence (13), which is both sufficient and necessary for fragmenting the five MIC chromosomes into 181 MAC chromosomes (8,14–16). The second major form of programmed genome rearrangement involves excision of internal eliminated sequences (IES), followed by rejoining of flanking sequences (Figure 1A) (17,18). IES, many containing various transposon-related sequences (8), are heterochromatinized by an RNA interference and *Polycomb* repression pathway (19–25). The vast majority of IES are subsequently excised by TPB2, a domesticated *piggyBac* transposase (26,27). Chromosome fragmentation and IES excision occur in the developing MAC, which subsequently undergoes endo-replication to attain the high ploidy of the mature MAC.

It has long been known that some sequences bounded by CBS in the MIC are not maintained as chromosomes in the MAC (28); these non-maintained chromosomes (NMC) are present in the developing MAC, but are absent from the mature MAC (8,16,28). Genes located on several of the larger NMC are highly and specifically expressed from the developing MAC (8,16). Here we focus on NMC-3, which contains *TPB6*, a gene encoding a domesticated *piggyBac* transposase. We establish that this *piggyBac* transposase is required for precise excision of a special class of IES located within gene coding regions. Our work provides a clear demonstration of the exaptation of programmed genome rearrangement to achieve a developmental stage-specific gene expression pattern, which may be a recurring theme in eukaryotic organisms experiencing programmed genome rearrangement during germline to soma differentiation.

MATERIALS AND METHODS

Tetrahymena strains and culture conditions

Wild-type (WT) strains for *T. thermophila* (CU427 and CU428), *T. malaccensis* (436), *T elliotti* (4EA) and *T. borealis* (X4H2) were obtained from the *Tetrahymena* Stock Center (https://tetrahymena.vet.cornell.edu/). $\Delta DCL1$ (22) and $\Delta PDD1$ (29) were homozygous heterokaryon strains kindly provided by Douglas L. Chalker. $\Delta EZL1$ were somatic knockout strains generated by YL (20). Other *T. thermophila* strains were generated in this study (see below). All cells were cultured in SPP medium at 30°C till mid-log phase (2×10^5 cells/ml) before harvesting (30). Conjugation was induced by mixing cells of two different mating types starved in the Dryl's phosphate buffer (1.5 mM CaCl$_2$, 1 mM NaH$_2$PO$_4$, 1 mM Na$_2$HPO$_4$, 2 mM sodium citrate) for at least 16 h.

Preparation of the developing MAC and mature MAC samples

The mature MAC were isolated from vegetatively growing WT cells (CU427 and CU428, respectively) by differential centrifugation following established protocols (31–33),

which contained <1% contaminating MIC DNA as confirmed by sequencing. The developing MAC were isolated from conjugation WT (CU427 and CU428) and ΔNMC-3 parent cells at 24 h after the initiation of conjugation (postmixing). Conjugation is finished by this time and progeny cells contain two developing MAC and one new MIC. The developing MAC were isolated from $\Delta DCL1$, $\Delta EZL1$, and $\Delta PDD1$ cells at 36 h after the initiation of conjugation, to offset potential delay in conjugation progress. In these mutants, conjugation progress is arrested at the development stage with two developing MAC and two new MIC. The developing MAC were separated from the parental MAC (mostly from <5% of cells that failed to undergo conjugation) and new MIC first by differential centrifugation (31). Nuclear pellets collected at 2500 g, containing a large proportion of the developing MAC, were used as input for fluorescence-activated cell sorting (FACS), as previously reported (34). Briefly, nuclear pellets were resuspended in PBS, stained with propidium iodide, sorted with gates optimized for collecting the developing MAC.

Illumina sequencing and data processing

Illumina sequencing libraries for genomic DNA were prepared by NEBNext® kit (New England Biolabs) to provide even coverage across GC%. Genomic DNA sequencing results were mapped back to the *Tetrahymena* reference genome assemblies by Bowtie2 (35) and visualized on GBrowse 2.0 (36). First strand specific libraries for mRNA were constructed using the Illumina TruSeq Stranded mRNA Sample Prep Kit (RS-122-2101). RNA-Seq results were mapped to the *Tetrahymena* reference genome assemblies by TopHat2 (37).

The MAC and MIC reference sequences were obtained from the *Tetrahymena* genome database (TGD: http://ciliate.org) (8,38,39). All sequencing gaps associated with NMC-3 and TPB6-dependent IES were manually closed.

Mapping of telomere-contained reads

The developing and mature MAC genomic DNA sequencing data were mapped back to the *T. thermophila* micronuclear reference genome using BWA version 0.7.10 (40) with default parameters. Telomere-contained reads are defined as: (i) minimal 30 bp mapped to the reference; (ii) containing minimal one untemplated telomeric repeat (C4A2 or G4T2 at the 5′ or 3′ end of reads, respectively); (iii) <2 mismatch/indel after clipping the telomeric repeats; (iv) flagged properly aligned, as 'read mapped in proper pair' in Picard (http://broadinstitute.github.io/picard/).

Coverage of NMC, MDS, and IES during the *Tetrahymena* life cycle

Two *T. thermophila* strains with identical homozygous MIC, derived from genomic exclusion crosses between SB210 and B* BII, were starved overnight and then mixed at 2×10^5 cells/ml to induce conjugation. At 24 h post-mixing, the conjugation culture was refed with SPP medium and propagated asexually for 90 generations. Genomic DNA samples were harvested at 0, 4, 8, 12, 16 and 24 h post-mixing
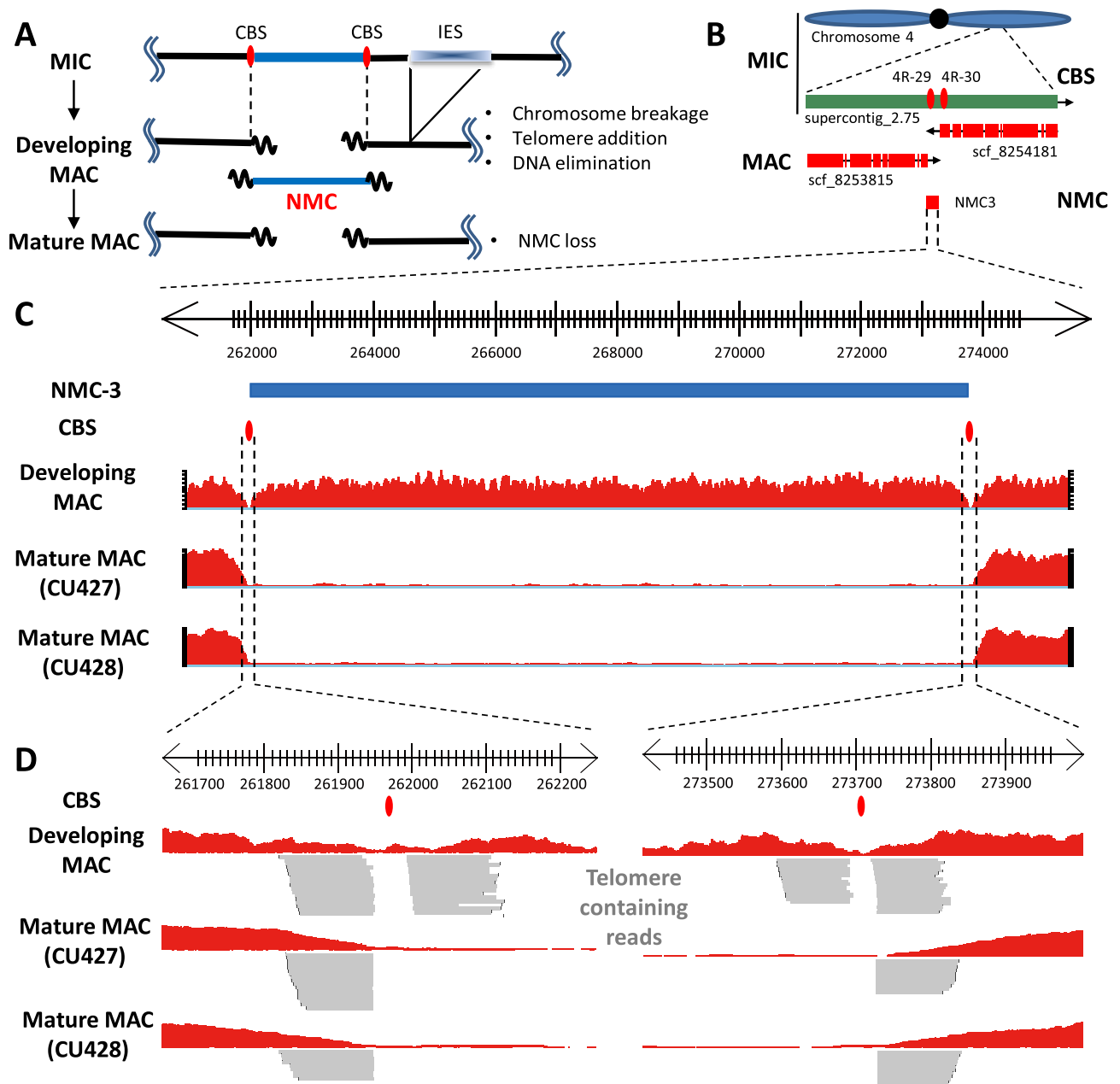
Figure 1. NMC are present in the developing MAC, but absent in the mature MAC. (**A**) Programmed genome rearrangement, including chromosome fragmentation and excision of internally eliminated sequences (IES), occurs as the somatic MAC is differentiated from the germline MIC-derived zygotic nucleus during sexual reproduction. Addition of telomeres (wavy lines) occurs after chromosome fragmentation at chromosome breakage sequences (CBS). Non-maintained chromosomes (NMC) are present in the developing MAC, but lost in the mature MAC during asexual propagation. (**B**) Genomic localization of NMC-3: right arm of MIC chromosome 4→MIC genome sequence supercontig_2.75→ CBS: 4R-29 and 4R-30, giving rise to NMC-3 and two adjacent MAC genome sequence scaffolds (scf_8253815 and scf_8254181). (**C**) DNA sequence coverage of NMC-3 and adjacent regions in the developing MAC and the mature MAC. The developing MAC were isolated from conjugating WT cells at 24 h post-mixing. The mature MAC were purified from the indicated strains of vegetatively growing WT cells. Illumina sequencing results were mapped back to the *T. thermophila* MIC reference genome (8). (**D**) DNA sequence coverage of the regions around CBS 4R-29 and 4R-30. Telomere-containing reads (gray) are aligned to and stacked on the reference genome.

and 10, 30, 50, 70 and 90 asexual generations. The genomic DNA sequencing data were mapped to the *T. thermophila* micronuclear reference genome using BWA, and the relative coverage of NMC, MDS, and IES across different samples were calculated as 'reads per kilobase of the target sequence per million mapped reads' (RPKM). Only the uniquely mapped reads were used for calculating RPKM.

PCR analysis of telomere-capped termini of NMC-3 and adjacent MAC chromosomes

Genomic DNA samples were harvested from WT (CU427 and CU428) and Δ*DCL1* cells, at different time points during conjugation and asexual propagation. Telomere-capped termini of NMC-3 and adjacent MAC chromosomes were amplified with a telomere-specific primer and a locus-specific primer (PCR primers used are listed in Supplementary File S5). Conjugation progress was monitored by a PCR assay (PCR primers used are listed in Supplementary File S5) following the alternative processing of the M element, a well-studied IES. The M element generates short PCR products in conjugating CU427 and CU428 cells, indicative of DNA elimination in the developing MAC (19,41).

Generation of *Tetrahymena* strains

Tetrahymena thermophila were transformed by particle bombardment with the Biolistic® PDS-1000/He Particle Delivery System (Bio-Rad), following the standard procedure (42). For generating the NMC-3 CBS deletion strains, conjugating WT cells were transformed at 8 h post-mixing, to target the newly formed developing MAC. For ΔCBS-L/R strains, the regions surrounding the CBS were replaced by the neo4 cassette, conferring paromomycin resistance. Transformants were selected by paromomycin, and then propagated for at least 50 generations, allowing random assortment of amitotic MAC chromosomes (43). After single-cell isolation, ΔCBS-L/R strains were further confirmed by PCR detecting altered telomere-capped termini associated with NMC-3 and adjacent MAC chromosomes. For generating ΔNMC-3 strains, conjugating WT cells were transformed at 3 h post-mixing, to target the meiotic MIC. For ΔNMC-3 strains, the entire region corresponding to NMC-3, including both of its flanking CBS, was replaced by the neo4 cassette. We first generated strains that completely knocked out the NMC-3 region in the germline MIC, while retaining NMC-3 in the WT MAC (homozygous heterokaryon; ΔNMC-3 parent), by germline transformation and standard genetic manipulations (42,44). Homozygous homokaryon strains (ΔNMC-3 progeny) were generated by crossing two homozygous heterokaryon strains (ΔNMC-3 parent); they are homozygous for the ΔNMC-3 allele in both the MIC and MAC. A *DOP1* mutant in which all WT copies in MAC were replaced with a version retaining IES-10 (*DOP1*::IES-10) was generated by standard somatic transformation and phenotypic assortment, and confirmed by qPCR.

Identification of TPB6-dependent IES

We compared the genomic DNA sequence coverage of the mature MAC from WT (CU427 and CU428, respectively), ΔNMC-3 parent (two independent strains), and ΔNMC-3 progeny cells (three independent strains), as well as the developing MAC from conjugating WT (CU427 and CU428) and ΔNMC-3 parent cells. Illumina sequencing reads of these samples were mapped to the *T. thermophila* MIC genome (TGD: http://ciliate.org) (8). The MIC genome was divided into 150 bp bins. For each bin, sequence coverage was calculated using the Subread package (45), and then normalized (RPKM). We searched for bins (5-fold difference) with consistently high coverage in ΔNMC-3 samples (mature MAC samples from ΔNMC-3 progeny (three strains)), and the developing MAC sample from conjugating ΔNMC-3 parents, but consistently low coverage in WT and ΔNMC-3 parent samples (mature MAC samples from CU427, CU428, and ΔNMC-3 parent cells (two strains) and the developing MAC sample from conjugating WT cells). Those bins with differential coverage were then compared with the known IES to generate a list of candidates, which were individually inspected on GBrowse. All 11 TPB6-dependent IES identified in this way were confirmed by PCR using primers flanking the insertion sites.

Analysis of IES boundaries

Preliminary IES sets were generated by comparing the assembled MAC genome to the MIC genome (8). Here, we focus on 7551 high-confidence IES with no sequencing gaps at their boundaries. To address the heterogeneity in IES boundaries, we sequenced developing MAC samples purified from WT and Δ*DCL1* cells at the end of conjugation. Specifically, Illumina sequencing reads of the WT and Δ*DCL1* samples were first mapped to the MIC genome using Tophat2 (37). Using the split mapped reads, breakpoints potentially corresponding to IES boundaries were extracted according to the CIGAR string (N), requiring an anchor region ≥9 bp. An IES boundary is defined by a breakpoint overlapping with the IES (no 'N' in the 4 bp defining the left boundary and the 4 bp defining the right boundary). The number of reads supporting a breakpoint reflects the frequency for a particular IES boundary in the population. The breakpoint with the highest number of supporting reads is defined as the main boundary for an IES. 117 587 breakpoints (associated with 6755 IES) from the WT sample and 4774 breakpoints (associated with 2273 IES) from the Δ*DCL1* sample (most of which are probably from contaminating MAC) are found within 20 bp of the main boundaries. Distribution of other IES boundaries relative to the main boundary is plotted in a composite analysis, combining all IES, as well as both left and right boundaries.

Transformation assay to test the IES excision efficiency

IES-4 was inserted into a TTAA sequence in the coding region of the *neo4* cassette (46) by fusion PCR and In-Fusion cloning (Clontech) (47) (PCR primers used are listed in Supplementary File S5). The new drug-resistant cassette (*neo4*::IES-4) generated paromomycin-resistant transformants, when delivered into WT conjugating cells during

formation of the developing MAC (8 h post-mixing). Precise IES-4 excision from the *neo4* coding region in the transformants was verified by PCR and sequencing. No transformants were obtained with vegetative cells, in the absence of TPB6 expression.

PCR was used in to introduce mutations into the 5′ half of the TIR in *neo4*::IES-4 (substitution with G in all nine positions; primers used are listed in Supplementary File S5). Similarly, the sequence in between the IES-4 TIR was replaced with heterologous sequences of the same length. The mutated constructs were delivered into WT conjugating cells at 8 h post-mixing. After plating and drug selection, paromomycin-resistant wells were counted and compared with the positive control transformed with the original *neo4*::IES-4 construct. Duplicate experiments were performed for all transformations.

Phenotypic analysis of ΔNMC-3 progeny cells

ΔNMC-3 progeny strains (homozygous homokaryon knockout: knockout in both MIC and MAC) were generated by crossing two ΔNMC-3 parent strains (homozygous heterokaryon: MIC knockout, MAC WT). Osmotic pressure sensitivity of the large contractile vacuole phenotype was tested by growing ΔNMC-3 progeny cells in SPP media with different concentrations of glucose; growth was overnight without shaking. Cells with or without abnormal contractile vacuoles were then counted. WT and ΔNMC-3 progeny cells were grown in a SPP medium with no glucose, without shaking. Their growth curves showed that WT growth was minimally affected by the lack of glucose, while the growth of ΔNMC-3 progeny cells was severely slowed down.

Analysis of differentially expressed genes in WT, ΔNMC-3 parent, ΔNMC-3 progeny, and *TPB6* rescued cells

RNA-Seq results were mapped back to the *Tetrahymena* MAC reference genome and gene expression values (RPKM) were calculated by TopHat2 and Cuffdiff2 ([37,48]). Differentially expressed genes in ΔNMC-3 progeny, compared with WT and ΔNMC-3 parent cells, were identified by DESeq2 ($P^{adj} < 0.01$, RPKM ratio >2 or <0.5) ([49]). Gene expression levels in WT growing and starved cells (Ll, Lm, Lh: logarithmically growing cells at low, medium, and high densities; S3, S6, S9, S12: cells starved in the Tris buffer for 3, 6, 9 and 12 h) are based on previous microarray analysis ([50]). Starvation induced genes are defined as Min(S3, S6, S9, S12) >500, and Max(S3, S6, S9, S12)/Max(Ll, Lm, Lh) ≥2; starvation repressed genes are defined as Max(Ll, Lm, Lh) >500, Min(Ll, Lm, Lh)/Min(S3, S6, S9, S12) ≥2 and Min(Ll, Lm, Lh)>Max(S3, S6, S9, S12). Correlations between expression profiles were evaluated by Pearson's correlation coefficients, and the results were clustered by MeV ([51]).

TPB6 rescue of ΔNMC-3 cells

A DNA fragment containing the *TPB6* coding region with ∼1 kb each of the 5′ and 3′ untranslated regions was amplified by PCR (primers used are listed in Supplementary File S5), using genomic DNA from purified MIC as template. This fragment was delivered into conjugating ΔNMC-3 parent cells at 10 h post-mixing. Transformants were selected by paromomycin in a hypotonic medium (0.5 × SPP, without glucose), after serial cell dilutions. *TPB6* rescue strains emerged as fast growing cells without abnormal contractile vacuoles, after an additional round of replica-plating. After single-cell isolation, *TPB6* rescue strains were validated by PCR showing successful excision of all 11 TPB6-dependent IES (Supplementary Figure S18) and RNA-Seq showing restoration of the transcription profile to that of WT and ΔNMC-3 parent cells (Figure 7D).

Identification of TPB6-dependent IES from four *Tetrahymena* species

Ten genes containing 11 TPB6-dependent IES in four *Tetrahymena* species were retrieved from the *Tetrahymena* genome database (TGD: http://ciliate.org). They were aligned at both the amino acid and nucleotide levels (Supplementary Figure S13), to identify the conserved insertion sites (TTAA) in the MAC genome sequences. Two sets of PCR primers were designed to amplify IES inserts (Supplementary File S5): (i) flanking primers ∼100 bp upstream and downstream of the TTAA sites and (ii) primers annealing specifically to the IES/MDS junctions, with the TIR consensus sequence (TTAACHCTW (H = A/C/T, W = A/T)) at the 3′ termini. Genomic DNA samples from purified MIC of each species were used as template for PCR amplification. PCR products were subsequently cloned and sequenced.

Data deposition

All Illumina DNA and RNA sequencing data were submitted to the NCBI Short Read Archive (SRA); their accession numbers are: SRR5155086–SRR5155096 for genomic DNA of conjugation and asexual propagation cells; SRR5184447 - SRR5184454 for genomic DNA of ΔNMC-3 parent, ΔNMC-3 progeny, Δ*DCL1* and WT cells; and SRR5271088–SRR5271099 for poly(A) transcripts of ΔNMC-3 parent, ΔNMC-3 progeny, *TPB6* rescue and WT cells.

RESULTS

NMC are present in the developing MAC, but missing in the mature MAC

Recent sequencing of the MIC genome of *T. thermophila* revealed 225 instances of CBS with the conserved 15 bp consensus ([8]). They generate 181 MAC chromosomes and 33 potential NMC ([8]); slightly different numbers are reported by Lin *et al.* ([16]). While most NMC are short, generated by local duplication of CBS, several are of substantial size, hereafter referred to as NMC-1, the NMC-2 cluster, and NMC-3 (Supplementary Table S1). Additional observations on these NMC are described in Appendix Section 1 (Supplementary Figure S1). Among the large NMC with gene-coding potential, NMC-3 has the simplest structure, and therefore became the focus of our functional analyses (Figure 1B–D).

The transient nature of these large NMC was confirmed by comparing genomic DNA sequencing results from purified developing MAC (from WT cells late in conjugation) and purified mature MAC (from WT asexually propagating cells). Illumina sequencing results were mapped back to the *T. thermophila* MIC reference genome (8) and visualized on GBrowse (Figure 1C). Direct comparison of sequencing coverage showed that NMC-1, the NMC-2 cluster and NMC-3 were abundantly represented in the developing MAC, often at levels similar to adjacent MAC chromosome regions, but were poorly represented in the mature MAC (Figure 1C and Supplementary Figure S2). In sequences from the developing MAC, a short coverage gap encompasses each CBS flanking an NMC, while in the mature MAC sequences this region of minimal coverage expands to include the entire NMC (Figure 1C and Supplementary Figure S2). Telomere-containing reads from the developing MAC were mapped to the termini of NMC-3, as well as its flanking MAC chromosomes (Figure 1D). Similar results were obtained for NMC-1 and the NMC-2 cluster (Supplementary Figure S3). A PCR assay using a telomere-specific primer and a locus-specific primer showed that at least some, and possibly all, copies of NMC-3 were telomere-capped by late conjugation (Supplementary Figure S4). In the mature MAC samples, no telomere-containing reads were mapped to the termini of NMC-3, while the flanking MAC chromosomes had telomere-containing reads at levels comparable to those seen in the developing MAC sample (Figure 1D and Supplementary Figure S3). Our PCR assay also showed that only the flanking MAC chromosomes were telomere-capped in asexually propagating cells (Supplementary Figure S4). These results indicate that the low level coverage of the NMC sequences in the mature MAC samples comes from unrearranged precursors present in contaminating MIC DNA. We conclude that NMC are completely lost in the mature MAC, despite gaining telomeres in the developing MAC.

NMC are generated and replicated in the developing MAC, but are quickly lost in asexually propagating cells

We next systematically examined changes in NMC levels throughout the *Tetrahymena* life cycle. In this set of experiments, genomic DNA samples, isolated at different time points during conjugation and subsequent asexual propagation of the conjugation progeny, underwent Illumina sequencing. Plots of the normalized coverage of NMC-1, the NMC-2 cluster and NMC-3 revealed very similar profiles (Figure 2A). Focusing on their coverage during conjugation, we found only low levels during early conjugation (Figure 2A). NMC coverage gradually increased during late conjugation (Figure 2A), at time points corresponding to 2C to 4C replication and 4C to 8C endo-replication in the developing MAC. It is important to note that chromosome fragmentation as well as IES excision occurs after 2C to 4C replication, but before, and independent of, 4C to 8C endo-replication (52). For a control, we analyzed the coverage of a MAC chromosome region adjacent to NMC-3. Unlike NMC-3, its coverage stayed at similar levels throughout conjugation (Figure 2B), probably due to continuing presence of the mature MAC in early conjugating cells and

non-maters in the samples. We also analyzed the coverage of two IES near NMC-3. IES increase copy number during 2C to 4C replication, but are later removed from the developing MAC (11). As expected, IES coverage only increased slightly and transiently during the early stages of MAC development (Figure 2C).

We also examined the dynamics of telomere-capped NMC-3 termini during conjugation, using PCR to sample the entire conjugation time course at 2 h intervals (Figure 2D). Each NMC-3 terminus was only detected late in conjugation, corresponding to the time of 4C to 8C endo-replication after chromosome breakage (Figure 2D, green brackets). NMC-3, present as a telomere-capped mini-chromosome (Appendix section 2; Supplementary Figure S4), was still found at high levels at 36 h of conjugation (Figure 2D). In contrast, IES were actively degraded by this time (Figure 4A). As controls, telomere-capped termini of the MAC chromosomes adjacent to NMC-3 were readily detected throughout the conjugation time course (Figure 2D). Conjugation progress was monitored by following the alternative processing of the M element (Figure 2D), a well-studied IES that generates a short PCR product from the developing MAC of conjugating CU427 and CU428 cells, distinct from the longer PCR product amplified from the mature MAC of parental cells (19,41). Excision of the M element had similar temporal profile to generation of NMC-3 (Figure 2D, red brackets). Taken together, our results strongly support that NMC are generated in the developing MAC after chromosome fragmentation, and increase in copy number during the 4C to 8C endo-replication.

Once refed, *Tetrahymena* conjugation progeny complete their endo-replication program and enter asexual propagation by binary fission (2). Illumina sequencing analysis of the subsequent asexual generations showed that NMC coverage dropped precipitously (Figure 2A). As expected, coverage of MAC chromosome regions stayed high, while coverage of IES regions was low (Figure 2B and C). The dynamics of telomere-capped NMC-3 termini after refeeding was also examined by PCR: their levels gradually diminished during asexual propagation, becoming undetectable after about 10 generations (Figure 2E) (16). In contrast, the termini of the MAC chromosomes immediately adjacent to NMC-3 stayed at similar levels throughout the passaging (Figure 2E). Thus, NMC are quickly lost during asexual propagation, which is consistent with passive dilution and/or active degradation. Further experimentation lends strong support to the former scenario, but not the latter.

Retention of NMC-3 sequence in the mature MAC by fusion to a MAC chromosome

To determine whether the NMC-3 sequence can be maintained in the mature MAC, we fused it to each adjacent maintained MAC chromosome, by deleting the intervening CBS (Figure 3A). The CBS deletion constructs were delivered into conjugating *Tetrahymena* cells as the developing MAC began to form, before chromosome fragmentation (Figure 3B). Transformed progeny were selected for paromomycin resistance, conferred by the *neo4* cassette in the CBS deletion constructs, and then propagated asexually. Many cells retained the CBS deletion, as inferred
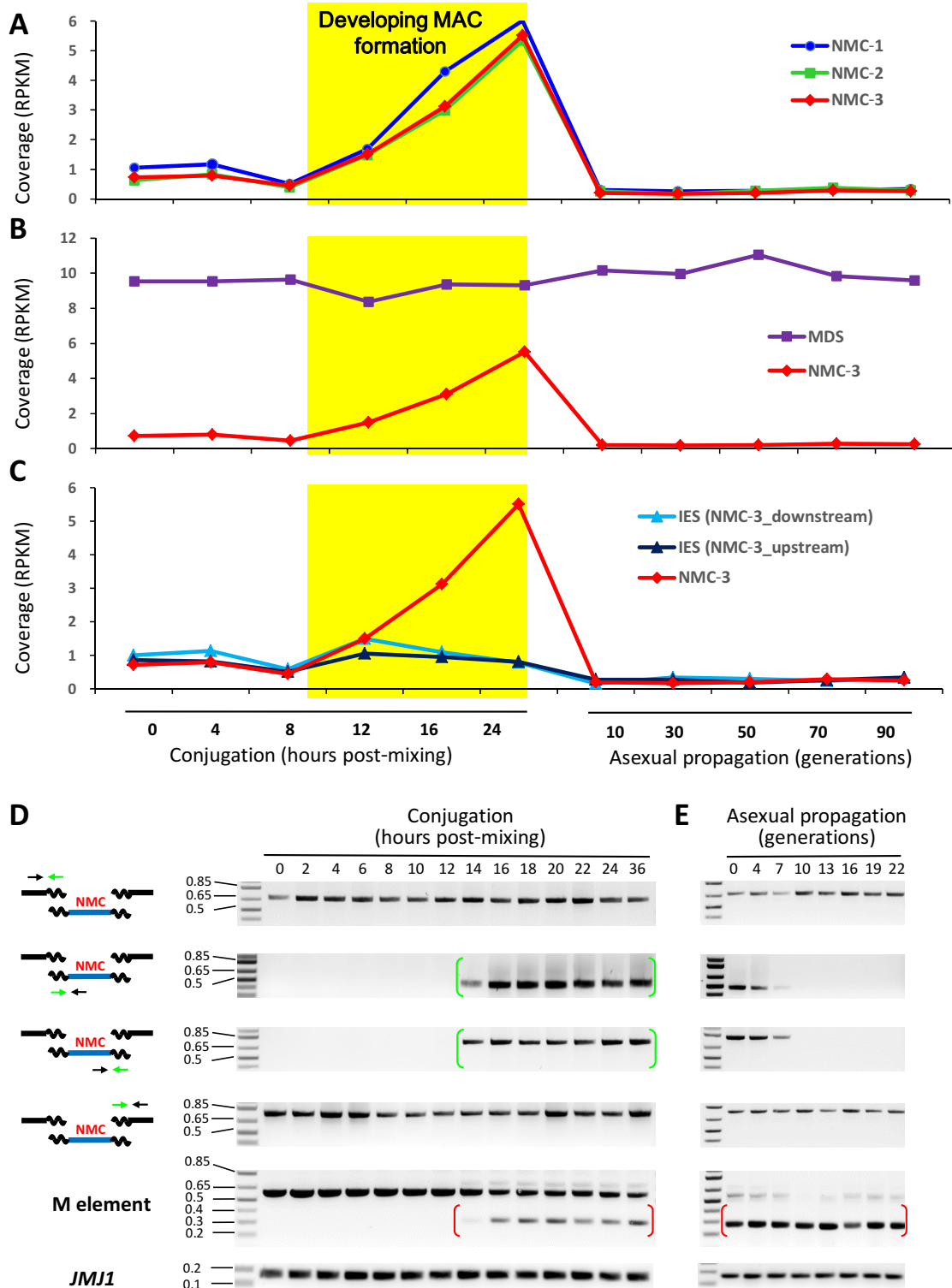
Figure 2. NMC are generated and amplified in the developing MAC during sexual reproduction, but quickly lost in the mature MAC during asexual propagation. (**A**) Normalized DNA sequence coverage (RPKM) of NMC-1, 2 and 3 during sexual reproduction and asexual propagation. (**B**) Comparing coverage of NMC-3 and adjacent MAC-destined sequences (MDS). (**C**) Comparing coverage of NMC-3 and adjacent internal eliminated sequences (IES). (**D**) PCR analysis of NMC-3 during conjugation, employing a telomere-specific primer (green arrows) in combination with locus-specific primers (black arrows). Conjugation progress was monitored by a PCR assay for a well-studied IES, the M element. Its processing gives rise to the short PCR product in conjugating CU427 and CU428 cells (red brackets), distinct from the long PCR product from the mature MAC in parental cells (19,41). A MAC gene, *JMJ1*, was monitored as a loading control. (**E**) PCR analysis of NMC-3 during asexual propagation. The short PCR product corresponding to the processed M element was monitored as a control for the presence of conjugation progeny (red brackets), which remained at constant levels. *JMJ1* was monitored as a loading control.
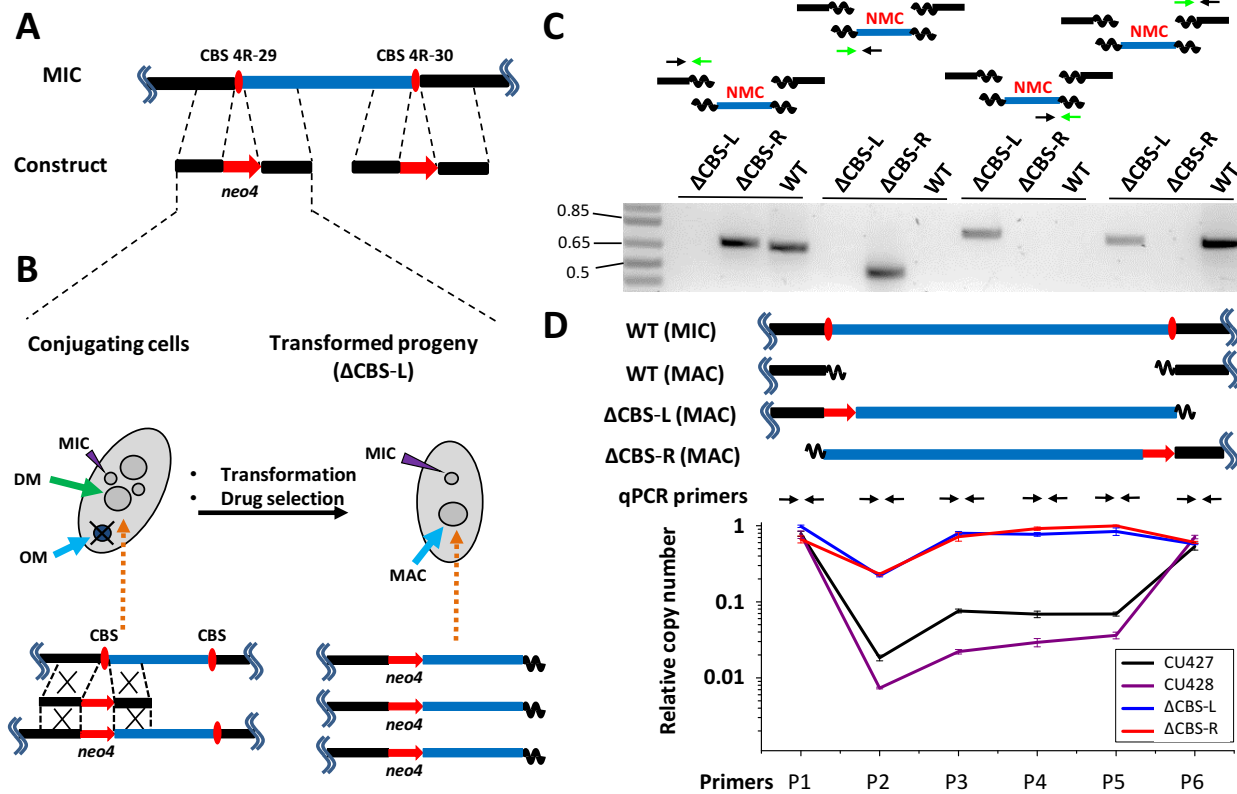
Figure 3. The NMC sequence is retained in the mature MAC when fused to a MAC chromosome. (**A**) A schematic for deletion of either CBS 4R-29 (CBS-L) or 4R-30 (CBS-R). In the constructs, the corresponding CBS was replaced by the neo4 cassette, conferring paromomycin resistance. (**B**) Generation of ΔCBS-L cells. The construct was transformed into WT conjugating cells during formation of the developing MAC, but before chromosome fragmentation. Transformed conjugation progeny were selected with paromomycin, passaged asexually for at least 50 generations, and assayed for the presence of NMC-3. DM: the developing MAC; OM: the old MAC. (**C**) PCR detection of telomere-capped termini of NMC-3 and its adjacent MAC chromosomes in WT, ΔCBS-L, and ΔCBS-R cells. (**D**) qPCR analysis of the copy number of the NMC-3 sequence in WT (CU427 and CU428), ΔCBS-L and ΔCBS-R cells.

from their paromomycin resistance and confirmed by a PCR assay. Unlike WT progeny, the appropriate telomere-capped NMC-3 terminus was present in each of these CBS deletion strains (Figure 3C). In CBS 4R-29 deletion cells (ΔCBS-L), we detected the telomere-containing right terminus of NMC-3 in the mature MAC (Figure 3C). In CBS 4R-30 deletion cells (ΔCBS-R), we detected the telomere-containing left terminus of NMC-3 (Figure 3C). Termini generated by the deleted CBS, for both NMC-3 and the adjacent MAC chromosome, were not detected (Figure 3C), supporting the complete replacement of the WT MAC chromosome flanking NMC-3. In cells with either flanking CBS deleted, quantitative PCR showed a dramatic increase along the entire length of the NMC-3 sequence; levels were comparable to those of the flanking MAC chromosomes (Figure 3D). Thus the NMC-3 sequence can be fully retained in the polyploid MAC. In WT cells, low NMC-3 signals were consistent with amplification from the NMC-3 sequence in the diploid MIC (Figure 3D). We therefore conclude that unlike IES, the NMC sequence can be retained if fused to a MAC chromosome.

TPB6, contained in NMC-3, is required for excision of a special class of IES

Unlike IES, NMC sequences are not heterochromatinized (Appendix section 3; Supplementary Figure S5). RNA-Seq revealed that they produced abundant mRNA (Appendix section 4; Figure 4B, D and Supplementary Figure S5), and many of their gene products may have development-specific roles (Appendix section 5; Supplementary Table S1). *TPB6*, a highly expressed gene located on NMC-3, was chosen for further study (Figure 4B-D and Supplementary Figure S6). *TPB6* encodes a *piggyBac* homologue implicated in IES excision (Appendix section 6; Supplementary Figure S7).

To delete *TPB6*, we generated strains in which NMC-3 and both flanking CBS were deleted in the germline MIC (while keeping the WT MAC), hereafter referred to as ΔNMC-3 parents (Figure 4C). During conjugation between two ΔNMC-3 parent strains, mRNA expression of *TPB6* was completely abolished; nonetheless, viable conjugation progeny were produced, hereafter referred to as ΔNMC-3 progeny. In control cross between WT cells, *TPB6* was specifically expressed at late conjugation (Figure 4D). Prompted by the established role of *TPB2*, another *Tetrahymena piggyBac* homologue, in IES excision (26,27), we sequenced genomic DNA from the develop-
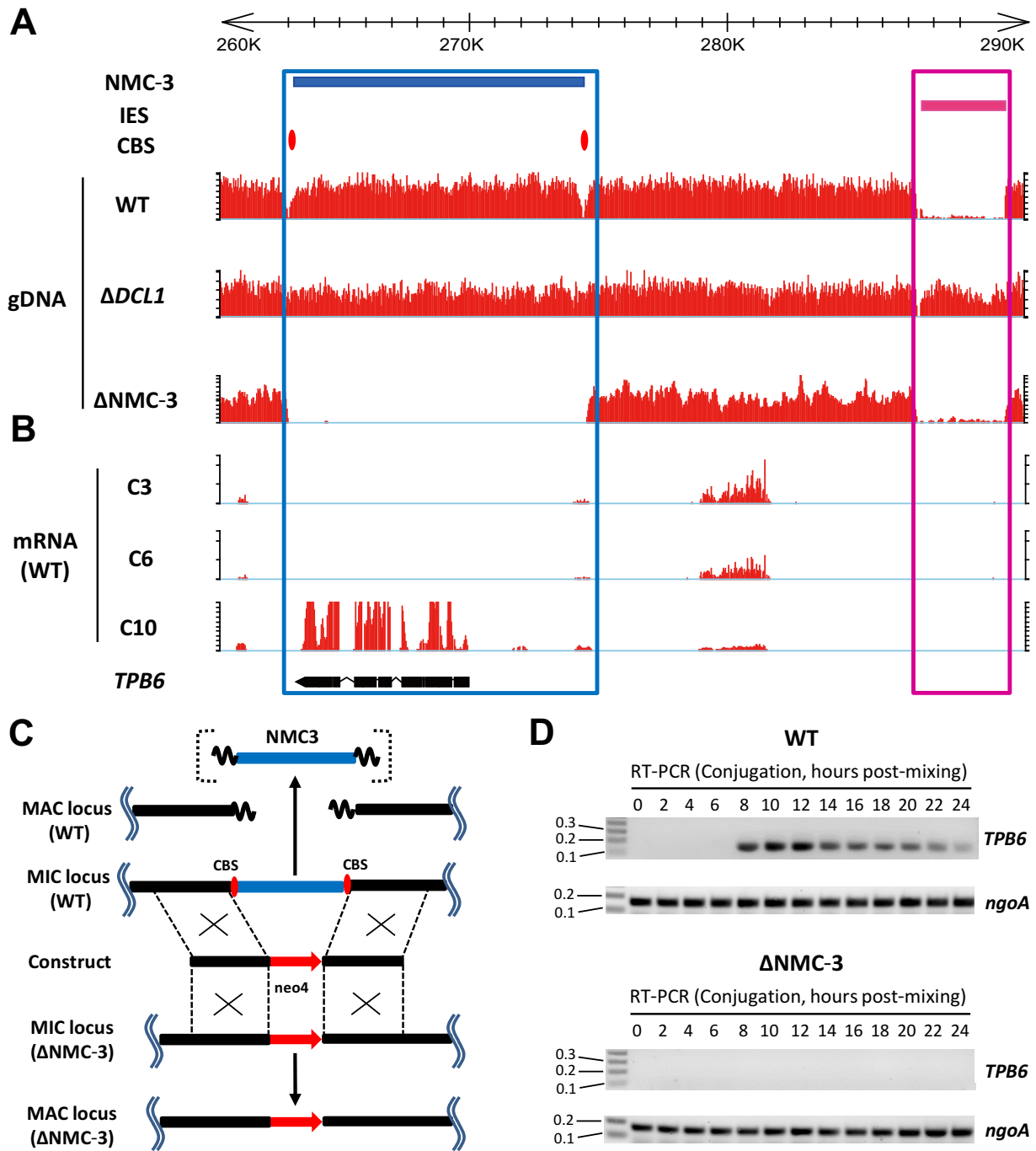
Figure 4. *TPB6* is expressed from the NMC-3 sequence during formation of the developing MAC. (**A**) DNA sequence coverage of NMC-3 and adjacent regions in the developing MAC of WT, *ΔDCL1*, and ΔNMC-3 cells, purified at late conjugation. Positions of NMC-3 (blue box), its flanking CBS (red ovals), and a downstream IES (magenta box) are indicated in the corresponding tracks. (**B**) mRNA sequence coverage of the same region during formation of the developing MAC. RNA-Seq of polyadenylated (poly(A)) transcripts, isolated from conjugating WT cells (CU427 and CU428) at 3, 6 and 10 h post-mixing (C3, C6 and C10, respectively). A gene model for NMC3-contained *TPB6* is shown in a separate track. (**C**) Generation of ΔNMC-3 strains. The entire region corresponding to NMC-3, including both of its flanking CBS, was replaced by the neo4 cassette. Homozygous heterokaryon strains (ΔNMC-3 parent) were generated by germline transformation and standard genetic manipulations (42,44); they are homozygous for the ΔNMC-3 allele in the MIC, but have a WT MAC. Homozygous homokaryon strains (ΔNMC-3 progeny) were generated by crossing two homozygous heterokaryon strains (ΔNMC-3 parent); they are homozygous for the ΔNMC-3 allele in both the MIC and MAC. (**D**) Reverse-transcription PCR (RT-PCR) analysis of *TPB6* expression during conjugation. RNA was isolated at the indicated hours post-mixing. WT: cross between CU427 and CU428; ΔNMC-3: cross between two ΔNMC-3 homozygous heterokaryon strains (ΔNMC-3 parent). *ngoA*, which is highly and uniformly expressed during conjugation (Supplementary Figure S17C), was used as a positive control.

ing MAC of conjugating ΔNMC-3 parent strains, and the mature MAC from three independent ΔNMC-3 progeny strains. As controls, we sequenced genomic DNA from the developing MAC of conjugating WT cells as well as from the mature MAC of both WT parent strains (Figure 5A, B and Supplementary Figure S8). As illustrated in the GBrowse view, a subset of IES, efficiently removed from all WT samples, was consistently retained in ΔNMC-3 samples (Figure 5A, B and Supplementary Figure S8). Systematic search of the over 10 000 IES in the *Tetrahymena* MIC genome revealed only 11 IES affected in this way, hereafter referred to as IES-1 to 11 (Figure 5A and Supplementary Table S2). They include six short IES (IES-1 to IES-6), which range from 140 bp (IES-4) to 457 bp (IES-3), and are among the shortest IES in *T. thermophila* (Figure 5A, C (top row)). The remaining five IES are longer (IES-7 to IES-11), ranging from ∼1 to 4 kb (Figure 5A, C (bottom row)). We independently confirmed each IES by PCR with flanking primers (Figure 5C, D). Short PCR products—corresponding to MAC sequences with the IES excised—were present in WT and ΔNMC-3 parent samples, while long PCR products—corresponding to IES retention—were observed in ΔNMC-3 progeny samples (Figure 5C, D). Our results indicate that removal of all 11 IES is dependent on NMC-3, likely because of *TPB6*. Intriguingly, IES-11 contains a nested regular IES, whose processing is independent of NMC-3 (Figure 5A, C, D), providing built-in confirmation for two distinct classes of IES. Detailed time course analysis of several TPB6-dependent IES revealed that they were all processed late in conjugation, at around the same time for regular IES processing (Appendix Section 7 and Supplementary Figure S9).

### TPB6-dependent IES are precisely excised and flanked by terminal inverted repeats (TIR)

In general, *Tetrahymena* IES are imprecisely excised, with frequent micro-heterogeneity and sometimes even widely separated alternative boundaries (8,16,53,54). We systematically analyzed the boundaries of thousands of well-defined IES (8); the developing MAC were purified from WT cells late in conjugation and their DNA was sequenced (Figure 5E). IES boundaries were identified as deletion breakpoints when the sequencing reads were mapped back to the MIC reference genome (Supplementary File S2). These breakpoints represented numerous independent DNA elimination events. As expected, the vast majority of these breakpoints were absent from ΔDCL1 cells (Supplementary File S3 and Supplementary Figure S10), which are defective in DNA elimination (22,23). Most IES have highly variable boundaries, as illustrated by the dispersion of breakpoints in our composite analysis (Figure 5E) and previous studies (8). Not surprisingly, no strong sequence motif was found at these IES boundaries. In contrast, the 11 TPB6-dependent IES were all precisely removed (Figure 5E). The IES boundaries in the MIC genome were defined by a conserved 9 bp terminal inverted repeat (TIR), with the consensus sequence TTAACHCTW (H = A/C/T, W = A/T) (Figure 5F), as previously reported (8,55). 19 out of 22 conformed to the sequence TTAACACTT, while the three others differed by a single substitution. Indeed, the seven completely conserved

nucleotides in the TIR of TPB6-dependent IES are identical to the corresponding nucleotides in the TIR of the original *piggyBac* transposon from cabbage looper moth *Trichoplusia ni* (Figure 5F) (56). The TTAA sequence is the only part of the TIR retained in the MAC genome (Supplementary File S2), as previously reported (8,54,55). This strongly suggests that these IES are processed like canonical *piggyBac* transposon, resulting in complementary TTAA overhang restored by simple ligation without the need for repair DNA synthesis (57). The physical coupling of the transposase domain with the Ku70/Ku80 β-barrel domain in TPB6 as well as TPB1 may further promote efficient re-ligation of donor sites after IES excision (Supplementary Figure S7) (55,58); a *Tetrahymena* Ku80 homologue has been shown to be required for processing regular IES (59).
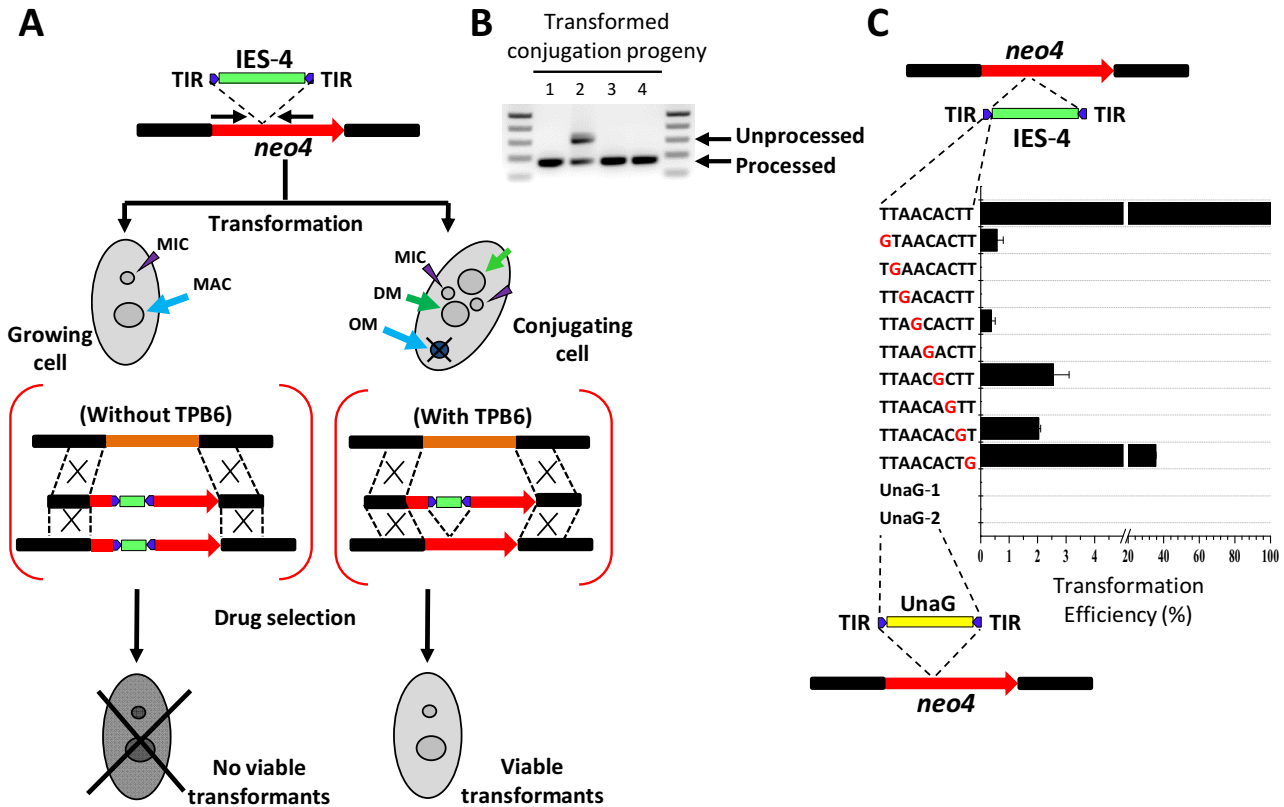
TIR are necessary but not sufficient for excision of TPB6-dependent IES

To test the role of the TIR, we generated a new drug-resistant cassette (*neo4*::IES-4; Figure 6A) by inserting IES-4 into a TTAA sequence in the coding region of the *neo4* cassette (46). This construct was delivered into WT conjugating cells during formation of the developing MAC; drug-resistant transformants were isolated and analyzed (Figure 6A). In the presence of TPB6, the inserted IES, with its intact TIR, was seamlessly and effectively removed from the *neo4* cassette, restoring drug resistance (Figure 6A, B). Thus, it is unlikely that any external flanking sequences are required for the precise excision of TPB6-dependent IES. As expected, no drug-resistant transformants were obtained when the construct was delivered into vegetative cells (Figure 6A). This transformation assay allowed us to rigorously test the degree to which the TIR motif is needed for precise IES excision by systematic mutation of the 5′ half of the TIR in *neo4*::IES-4 (Figure 6C). We scanned all nine positions of the consensus sequence by substitution with G, which is not present in any WT TIR. Mutations at the first eight positions generated no or very few transformants (<3% WT levels), while a mutation at the last position had moderate effect (<40% WT levels; Figure 6C). We conclude that the TIR flanking TPB6-dependent IES is the conserved cis-element required for their precise processing.

The original *piggyBac* transposon from *T. ni* contains a 17 bp terminal inverted repeat and a 19 bp internal inverted repeat, which are separated by short spacers (56) and are necessary and sufficient for its excision (60). The 9 bp TIR flanking TPB6-dependent IES is much shorter. Indeed, the TTAACACTT motif is widespread in the *Tetrahymena* MIC genome, as expected by probability for a sequence of its length (see Supplemental Materials and Methods for details). However, most sequences flanked by a TTAACACTT inverted repeat, even those within the size range of TPB6-dependent IES (<5 kb), are not eliminated as IES, indicating insufficiency for the TIR alone. Furthermore, when the sequence in between the IES-4 TIR was replaced with heterologous sequences of the same length, no transformants were generated (Figure 6C). Consistent with our observations, Cheng *et al.* showed that significant length of internal sequences, in addition to the TIR, is needed for efficient IES processing (55). We conclude that the TIR are necessary, but
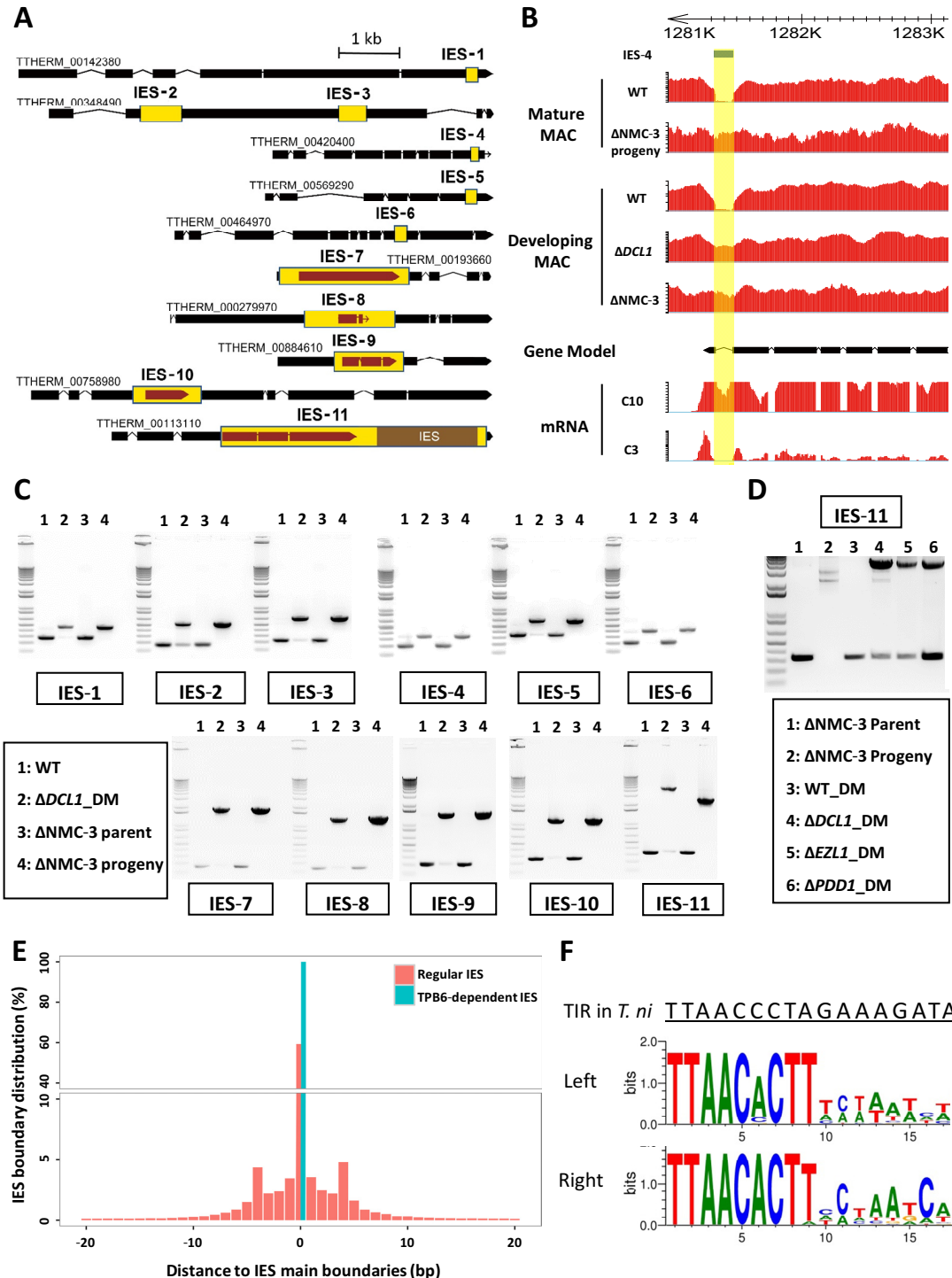
Figure 5. *TPB6* is required for precise excision of a special class of IES. (**A**) A summary of all 11 TPB6-dependent IES (yellow boxes), together with models of associated MAC genes (exons: black boxes; introns: thin lines) and genes within IES (same as for MAC genes, but in brown). Note that IES-2 and IES-3 are inserted into the same gene. A regular IES (beige box) is nested within IES-11. (**B**) Excision of IES-4, associated with the coding region of TTHERM_00420400, is abolished in ΔNMC-3 and ΔDCL1 cells. From top to bottom, the tracks represent DNA sequence coverage of the mature MAC from WT and ΔNMC-3 progeny cells, then the developing MAC from WT, ΔNMC-3 and ΔDCL1 cells; directly below is RNA-Seq coverage of poly(A) RNA isolated from conjugating WT cells, at 3 and 10 h post-mixing (C3 and C10, respectively). (**C**) PCR analysis of excision of all 11 TPB6-dependent IES. Template DNA was prepared from the mature MAC of WT (1), ΔNMC-3 parent (3) and ΔNMC-3 progeny cells (4). It was also prepared from the developing MAC of *ΔDCL1* cells (2). Note that the regular IES nested within IES-11 is excised in ΔNMC-3 progeny but not *ΔDCL1* cells. (**D**) PCR analysis of excision of IES-11. Template DNA was prepared from the mature MAC of ΔNMC-3 parent (1) and ΔNMC-3 progeny cells (2). It was also prepared from the developing MAC of WT (3), *ΔDCL1* (4), *ΔEZL1* (5), and *ΔPDD1* cells (6). (**E**) Highly precise excision of TPB6-dependent IES, compared with imprecise excision of regular IES. (**F**) Sequence logos for the terminal inverted repeats (TIR) of TPB6-dependent IES. The classical TIR in *T. ni piggyBac* transposon is above the logos. The left and right correspond to the upstream and downstream TIR (relative to the gene hosting the IES), respectively.

Figure 6. TIR flanking TPB6-dependent IES are necessary but not sufficient for excision. (**A**) Transformation assay for excision of TPB6-dependent IES. IES-4, with its TIR, is inserted into a TTAA sequence within the coding region of the *neo4* cassette (*neo4*::IES-4). Viable transformants were only obtained when the construct was introduced into WT conjugating cells, allowing TPB6 to seamlessly remove IES-4 and restore the *neo4* cassette. No viable transformants were obtained with growing cells, in which TPB6 is not expressed. (**B**) PCR analysis of the *neo4* cassette in transformed conjugation progeny (four independent strains). The upper band represents the unprocessed *neo4* cassette with IES-4 insertion, while the lower band represents the processed form with IES-4 excised. (**C**) Factors affecting IES-4 excision. Transformation efficiencies of mutated *neo4*::IES-4 cassettes were compared to the original cassette. The left TIR consensus sequence was systematically substituted by G, which is not found in any TIR of TPB6-dependent IES. The IES-4 sequence internal to the original TIR was also replaced by heterologous sequences of the same length (UnaG-1 and UnaG-2, two segments of the coding sequence for the UnaG fluorescent protein (111)).

not sufficient, for precise excision of TPB6-dependent IES; at least some internal sequence information is required. We failed to find any significant internal DNA sequence motif shared by the 11 TPB6-dependent IES, though we could not completely rule out the presence of highly degenerate motifs like those in introns.

RNA interference and *Polycomb* repression are necessary but not sufficient for excision of TPB6-dependent IES

Like regular IES, TPB6-dependent IES were not processed in conjugation progeny of mutants deficient in RNA interference and *Polycomb* repression, including Δ*DCL1*, Δ*EZL1* and Δ*PDD1* (Figure 5C, D and Supplementary Figure S8, S10, and S18). In *Tetrahymena*, *DCL1* encodes a Dicer homologue required for processing double-stranded RNA into small RNA (22,23); *EZL1* encodes an *E(z)* homologue required for histone methylation at H3K27 and H3K9 (20); *PDD1* encodes a chromodomain protein that recognizes both H3K27 and H3K9 methylation (20,61,62). They are critical components in an RNAi-dependent heterochromatin formation pathway required for processing

regular IES. Our result indicates that this pathway is also required for excision of TPB6-dependent IES. We propose that this dependency is direct, even though we cannot rule out the possibility that removal of certain regular IES may be necessary for expressing a (yet unknown) factor important for excision of TPB6-dependent IES. It is worth noting that our result confirms the result of Fass *et al.* with a smaller subset of exonic IES (54), but is in clear disagreement with a recent report concluding that excision of these same IES are independent of this pathway (55). Likely basis for this discrepancy is discussed in Appendix Section 8.

We reexamined published RNA-Seq data for conjugation-specific small RNA, scnRNA, at late conjugation in *Tetrahymena* (63). Regular IES are generally associated with high levels of scnRNA with little strand-bias (Supplementary Figure S5), derived from bi-directional noncoding RNA transcripts (64,65). RNA-Seq also reveals degradation products of mRNA, evident in their strong bias towards the sense strand (Supplementary Figure S11 and S12). Shortly before their elimination, most TPB6-dependent IES were associated with low to moderate levels

of scnRNA, while others accumulated mRNA degradation products (Supplementary Figure S12). Intriguingly, high levels of bi-directional scnRNA were found in the regular IES nested in IES-11, but not in the remainder of IES-11 (Supplementary Figure S12). This regular IES was eliminated, with variable boundaries, in ΔNMC-3 cells, but not ΔDCL1 cells (Figure 5C, D and Supplementary Figure S8). Therefore, high levels of scnRNA, while sufficient for excision of regular IES, are not sufficient for excision of TPB6-dependent IES.

We also reexamined published ChIP-Seq data for PDD1 distribution in the developing MAC (66). Regular IES are generally associated with high levels of PDD1—demonstrated by high ChIP to input ratios (Supplementary Figure S5) (66). Around the time of their elimination, all 11 TPB6-dependent IES were associated with low levels of PDD1 (Supplementary Figure S12). Combined with the scnRNA data, these results suggest that TPB6-dependent IES are not packaged in *bona fide* heterochromatin, although they depend on the heterochromatin formation pathway for their excision. This is consistent with active transcription of genes associated with or contained within TPB6-dependent IES (see below: TPB6-dependent IES determine transcription products limited to the developing MAC formation stage). We conclude that unlike regular IES, the RNAi-dependent heterochromatin formation pathway is necessary but not sufficient for TPB6-dependent IES processing.

Conspicuous phenotypes in ΔNMC-3 progeny are entirely attributable to genes interrupted by retained TPB6-dependent IES

The vast majority of IES in *T. thermophila* reside within either intergenic or intronic regions (8,54). This is attributed to, and has likely co-evolved with, their imprecise excision (Figure 5E), which can disrupt a gene coding region by generating deletions, insertions, and reading frame shifts. In strong contrast, the 11 TPB6-dependent IES are located within the exons of 10 genes (Figure 5A), and all are excised with absolute precision (Figure 5E). Formation of the developing MAC in the absence of TPB6 (ΔNMC-3 progeny) leads to interruption of exons in all 10 genes by the retained IES (Figures 5A, 7A and Supplementary Figure S15). ΔNMC-3 progeny exhibited a severe growth phenotype (Supplementary Figure S17) and RNA-Seq revealed an altered global transcription profile (Figure 7B, D). Genes normally controlled by starvation had aberrant expression in ΔNMC-3 progeny grown in rich medium: many starvation-induced genes were significantly up-regulated, while many starvation-repressed genes were significantly down-regulated (Figure 7B). These included *ngoA* (Supplementary Figure S17C), a well-studied starvation-induced gene in *Tetrahymena* (67). Many of these differentially expressed genes have been implicated in highly conserved metabolic pathways (Supplementary Figure S17D, E). This is consistent with disrupted membrane traffic (see below) affecting intracellular signaling, especially activation of the mTOR pathway to regulate energy and metabolic homeostasis (68).

Conspicuously, many ΔNMC-3 progeny cells were bloated with an extraordinarily large contractile vacuole (Figure 7C), which failed to undergo the normal diastolic-systolic cycle for periodic expulsion of fluid (69). The percentage of abnormal cells increased in hypotonic media and decreased in hypertonic media (Figure 7C). This phenotype is likely due to membrane trafficking defects, as membrane fusion and kiss-and-run exocytosis are required for contractile vacuole function (69,70). Indeed, IES-10 retention results in a truncated product of the *DOP1* gene (Figure 7A), which normally encodes a leucine zipper protein potentially localized in the Golgi and involved in membrane trafficking (71,72). A *DOP1* mutant in which all WT copies in MAC were replaced with a version retaining IES-10 (*DOP1*::IES-10) also featured large contractile vacuoles like ΔNMC-3 progeny (Figure 7C), confirming its direct connection with the contractile vacuole phenotype.

The severe phenotype of ΔNMC-3 progeny was rescued when a DNA fragment containing just WT *TPB6* was transformed into ΔNMC-3 cells during late conjugation (Figure 7D and Supplementary Figure S18). Rescued cells were readily identified and isolated by their growth advantage over ΔNMC-3 progeny cells and their lack of abnormally large contractile vacuoles in hypotonic media (Figure 7D and Supplementary Figure S17). PCR analysis showed that all 11 TPB6-dependent IES were processed in the rescued cells (Supplementary Figure S18). RNA-Seq analysis confirmed that the transcription profile in the rescued cells had been largely restored to that of WT cells (Figure 7D). These results demonstrate that *TPB6* is required for the excision of this special class of IES and functional transcripts, for genes important for normal vegetative growth, require the removal of at least some of these IES. They also demonstrate that *TPB6* is solely responsible for the phenotype of ΔNMC-3 progeny.

TPB6-dependent IES determine transcription products limited to the developing MAC formation stage

All the genes hosting TPB6-dependent IES produce alternative transcription products in WT cells during late conjugation, after initial formation of the developing MAC but before excision of these IES (∼8–12 h post-mixing; Figure 5B, Supplementary Figures S9 and S11, and Supplementary File S1). These same alternative transcription products are also observed in ΔNMC-3 progeny (Figure 5B and Supplementary Figure S11). Additional observations on these transcripts are described in Appendix Section 9 (Figure 7, Supplementary Figure S11 and S15). Furthermore, all of the longer TPB6-dependent IES (IES-7 to IES-11) encode mRNA transcripts with sizable opening reading frames (Figure 5A and Supplementary Figure S11). These transcripts are only expressed during late conjugation (Supplementary Figure S11), and are co-directional with the gene hosting each IES (Figure 5A and Supplementary Figure S11). While the transcript encoded by IES-11 provides an alternative 3′ end for the gene hosting it, the other four IES-encoded transcripts represent entirely independent genes (Figure 5A and Supplementary Figure S11).

Tetrahymena IES often contain transposons and their relics (8), which are effectively silenced by the RNAi-
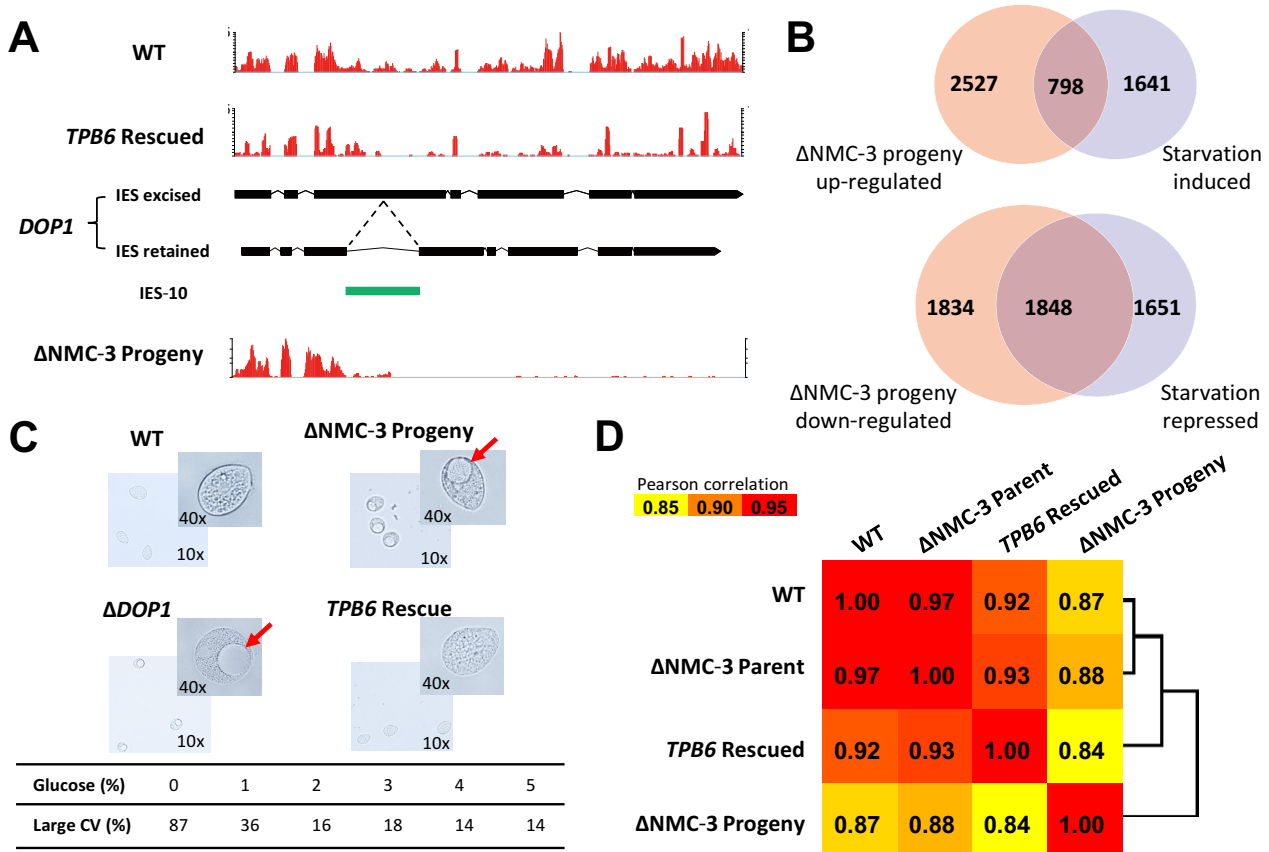
Figure 7. Conspicuous phenotypes in ΔNMC-3 progeny are attributable to genes interrupted by retained TPB6-dependent IES. (**A**) *DOP1*, encoding a leucine zipper protein potentially localized in the Golgi, is truncated in ΔNMC-3 progeny due to failure to remove IES-10. Shown here are tracks representing RNA-Seq coverage of poly(A) RNA isolated from growing cells of WT, *TPB6* rescued, and ΔNMC3 progeny strains. Gene models for both IES-excised and IES-retained forms of *DOP1* are also shown, with the insertion of IES-10 (green box and dashed lines) indicated. (**B**) Overlap between differentially expressed genes from ΔNMC-3 growing cells and starved WT cells. The left circles represent genes up-regulated (top) or down-regulated (bottom) in ΔNMC3 progeny growing cells, compared to WT growing cells. The right circles represent genes up-regulated (top) or down-regulated (bottom) in WT starved cells, compared to WT growing cells. Additional analyses of genes in the overlap are provided in Supplementary Figure S17C-E. (**C**) Extraordinarily large contractile vacuole in ΔNMC3 progeny and *DOP1*::IES-10 cells. This phenotype was aggravated in hypotonic media, alleviated in hypertonic media, and rescued by a DNA fragment containing *TPB6* delivered into conjugating ΔNMC3 parental cells. (**D**) Correlations in gene expression profiles. Expression levels (RPKM) are calculated based on RNA-Seq coverage of poly(A) RNA isolated from growing cells of WT, *TPB6* rescued, ΔNMC3 parent, and ΔNMC3 progeny strains. Pearson's correlation coefficients (PCC) are calculated for pair-wise comparison and illustrated by the color scale. Cluster analysis, based on PCC, is shown on the right. Note the similarities between the gene expression profiles of WT, *TPB6* rescued, and ΔNMC3 parent strains, and their distinction from that of ΔNMC3 progeny.

dependent *Polycomb* repression pathway in WT conjugating cells (6). However, these five IES-encoded open reading frames share no apparent homology with transposon genes (Supplementary Table S2 and Supplementary File S1), and their transcripts are highly abundant in WT conjugating cells (Supplementary Figure S11). Importantly, as shown below, the open reading frames within these five IES, while distinct from one another, have conserved homologues in related *Tetrahymena* species, suggesting that they play functionally significant roles. Indeed, these IES, each containing distinct genes or gene segments, is a natural analogue of engineered *piggyBac* vectors for integrating any genes of interest into target genomes (73).

TPB6-dependent IES are conserved in *Tetrahymena* species

All 11 TPB6-dependent IES are located within conserved genes, found in three other sequenced *Tetrahymena*

species—*T. malaccensis*, *T. elliotti* and *T. borealis* (Figure 8A and Supplementary Figure S13). Intriguingly, the IES insertion sites (donor sites)—TTAA, occur at equivalent positions in each homologue, as revealed by the alignment of amino acid and nucleotide sequences (Figure 8A and Supplementary Figure S13). This is true not only for highly conserved genes hosting TPB6-dependent IES, but also for those genes in which the flanking nucleotide sequences has substantially diverged (Figure 8A and Supplementary Figure S13), strongly suggesting that they also act as functional IES insertion sites (donor sites) in other *Tetrahymena* species. With a few exceptions (see below), we confirmed the presence of IES at all other loci by PCR amplification of MIC-enriched template DNA (Figure 8B-D). Sequencing revealed that these IES had TIR conforming to the TTAACHCTW motif (Figure 8C), but no internal sequence motifs of any significance (Supplementary File S1). Interestingly, the *T. borealis* sequence homologous to
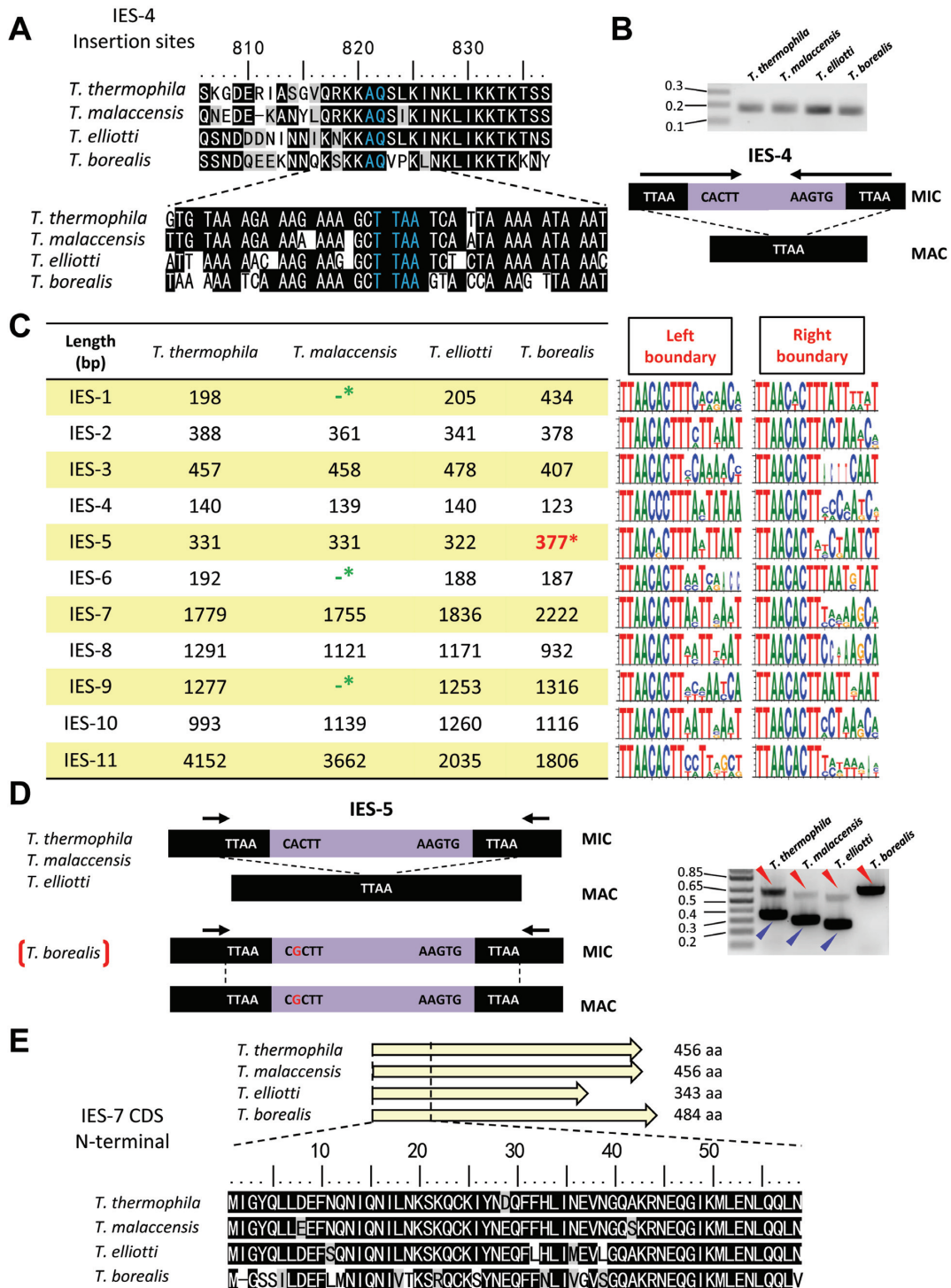
Figure 8. Conservation of TPB6-dependent IES in *Tetrahymena* species. (**A**) Conservation of the IES-4 insertion site in four tetrahymenine species. Amino acid (top panel; black shading: identical; gray shading: similar) and nucleotide sequence alignment (bottom panel; black shading: identical) of the corresponding MAC genes containing the IES-4 insertion site (aqua letters). (**B**) IES-4 equivalent insertions amplified from four *Tetrahymena* species, using specific primers ending at the TIR. The PCR products were subsequently cloned and sequenced. (**C**) Summary of TPB6-dependent IES identified in four *Tetrahymena* species. Left panel: the sizes for all identified IES (their sequences are provided in Supplementary File S4). Note the failure to remove the IES-5 equivalent in the MAC of *T. borealis* (red asterisk), and the missing of the IES-1, IES-6 and IES-9 equivalents in the MIC of *T. malaccensis* (green asterisks). Right panel: sequence logos for the left and right boundaries of TPB6-dependent IES in four *Tetrahymena* species. The left and right boundaries correspond to the upstream and downstream TIR (with adjacent IES sequences), respectively. Note the consensus of the left boundary of IES-5, which deviates from the highly conserved TIR motif, is affected by the mutation in *T. borealis* that prevents excision of that IES. (**D**) A single nucleotide mutation (G substitution) in TIR of the IES-5 equivalent is associated with its retention. Using primers flanking the TIR, both processed (blue arrowhead) and unprocessed forms (red arrowhead) of IES-5 were amplified from the *T. thermophila*, *T. malaccensis* and *T. elliotti* samples, while only the unprocessed form was amplified from the *T. borealis* sample. (**E**) Conservation of the ORF contained in IES-7.

IES-5, unlike its equivalent in the other three *Tetrahymena* species, has been retained in MAC, as shown by both genomic DNA sequencing and PCR (Figure 8D). Retention coincides with a single point mutation in the 5′ half of its TIR (TTAAC**G**CTT). This mutated sequence does not conform to the TIR conserved motif (Figure 5F), and led to dramatically reduced transformation efficiency in our TIR mutation scanning in *T. thermophila* (Figure 6C). This exceptional loss of IES excision provides further support for the rule of TIR conservation, now extended in its phylogenetic range.

Using template DNA enriched for *T. malaccensis* MIC and primers flanking insertion sites of IES-1, IES-6 and IES-9, we were able to PCR amplify short products corresponding to the IES-excised form, but not long products corresponding to the IES-retained form (Appendix Section 10 and Supplementary Figure S14). As *T. malaccensis* is more closely related to *T. thermophila* than *T. elliotti* and *T. borealis* (74), the absence of these IES in its MIC is most likely because they have been excised, instead of never having been present. We speculate that this could result from a rare event leading to mis-localization of the excision machinery in the new MIC instead of in the developing MAC; this event would have occurred in *T. malaccensis* after its divergence from *T. thermophila*.

By comparing TPB6-dependent IES in four *Tetrahymena* species, we found that homologous IES, especially the short ones, were often similar in size (Figure 8C). IES-11 was a notable exception, with a nested regular IES (∼2 kb) present in *T. thermophila* and *T. malaccensis*, but not in *T. elliotti* and *T. borealis* (Figures 5A, C, D and 8C). As mentioned above, we found conserved open reading frames in mRNA transcripts encoded by the five longer TPB6-dependent IES (Figure 8E and Supplementary Figure S16). However, beyond these potential coding regions, TPB6-dependent IES sequences showed considerable variation among the four *Tetrahymena* species (Figure 8C and Supplementary File S1). Apparently, mutations inside these IES, including large insertions, are well tolerated as long as they do not interfere with the sequences required for excision, which are either sparsely or redundantly distributed.

DISCUSSION

Chromosome fragmentation and IES excision, the two major types of programmed genome rearrangement during *Tetrahymena* conjugation, occur as the new somatic MAC develops and differentiates from the germline MIC. NMC, generated by chromosome fragmentation in the developing MAC, but quickly lost in the mature MAC, contain highly expressed genes. One of these genes encodes a domesticated *piggyBac* transposase, *TPB6*, which we show is required for precise excision of a special class of conserved IES residing within gene coding regions. TPB6-dependent IES are flanked by conserved TIR, which are essential for their removal from the MAC genome. Some sequences within the TPB6-dependent IES are also required for their processing; they are most likely targeted by the RNAi-dependent *Polycomb* repression pathway, as mutants in this pathway fail to excise these IES. Retention of TPB6-dependent IES disrupts genes important for normal cell growth (see Table 1

for a comparison between TPB6-dependent IES and regular IES). Our work showcases exaptation of programmed genome rearrangement to address development-related issues. It is in line with the widespread presence of genes with putative development-related functions in germline-limited sequences of various unicellular and multicellular eukaryotes, and may reflect a recurring theme in organisms experiencing programmed genome rearrangement during germline to soma differentiation.

This study was inspired and carried out as an extension of our recently published work (8), describing the sequencing and assembly of the *T. thermophila* MIC genome and its alignment with its cognate somatic genome. As we were writing up this work, two related articles (16,55) came out. While told from different perspectives, the three stories mostly validate one another in areas where they overlap. We agree with Lin *et al.* (16) that NMC generated by chromosome fragmentation in the developing MAC are only transiently present in the mature MAC, and some NMC contain expressed genes. Cheng *et al.* (55) focuses on TPB1, another domesticated *piggyBac* transposase that is shown to be required for the precise excision of exonic IES (this work confirms 11 out of the 12 high-confidence IES identified in their study). Cheng *et al.* (55) briefly investigates TPB6, finds a similar knockout phenotype, and suggests that TPB1/TPB6 may function coordinately, perhaps as a heterodimer. A major discrepancy between their findings and ours concerns the role of the RNAi machinery and *Polycomb* group proteins in TPB1/TPB6-dependent IES excision. Our evidence shows the precise excision of these IES depends on this pathway, in opposition to the conclusion in Cheng *et al.* (55). The likely basis for this discrepancy is discussed in Appendix Section 8. Our results lead us to formulate a different hypothesis concerning the evolution of IES excision: TPB2 and TPB6 represent two separate stages in the recurrent domestication of *piggyBac* transposases in ciliates, involving gradual degeneration and re-tooling of the original *piggyBac* transposition mechanism and concomitant integration with the host RNAi-dependent *Polycomb* repression pathway.

Development-specific genes encoded in germline-limited sequences

There is growing evidence for active transcription of germline-limited sequences during eukaryotic development. In *T. thermophila*, mRNA is transcribed from germline-limited sequences only in the developing MAC, as demonstrated for genes found within some NMC and longer exonic IES ((8,16) and this study). In particular, *TPB6* in NMC-3 is required for precise excision of exonic IES in the developing MAC ((55) and this study). This provides another regulatory mechanism, in addition to, or even in place of the promoter-based mechanism, for development-specific transcription. In *Oxytricha trifallax*, a stichotrichous ciliate distantly related to *Tetrahymena*, germline-limited sequences are found to encode hundreds of proteins expressed at late conjugation (5). In multicellular eukaryotes with programmed DNA elimination, gene expression from germline-limited sequences has been extensively characterized in the nematode *Ascaris suum*

Table 1. Comparison between regular IES and TPB1/TPB6-dependent IES in *Tetrahymena*

	Regular IES	TPB1/TPB6-dependent IES
Location	Intergenic or intronic regions	Gene coding regions/exons
Excision	Imprecise	Precise
Number	>10 000	<20*
piggyBac homologs	TPB2	TPB1 and TPB6**
TIR	No	Yes
mRNA production	Containing minimal amounts of transcripts expressed in WT cells	Located within or containing transcripts highly expressed in developing MAC
Small RNA production	Generally high levels	Generally low levels
Heterochromatin marks	Generally high levels	Generally low levels
Relation to the heterochromatin formation pathway	Necessary and sufficient	Necessary*** but not sufficient
Conservation	Probably not conserved	Highly conserved
Sequence characteristics	Contain transposon-derived and/or repetitive sequences	Contain potentially functional genes with development-specific expression
Loss-of-function phenotype	Conjugation progress arrest; no viable progeny	Severe growth phenotype attributable to a few genes hosting these IES

* Twelve high-confidence IES identified in Cheng *et al.* (55), 11 of which confirmed in this study.
** There is uncertainty as to whether TPB2 is required for TPB6-dependent IES (55).
*** Discrepancy between this study and Cheng *et al.* (55).

(1,75–77), and more recently in the sea lamprey *Petromyzon marinus* (78–80). Although functions for germline-limited genes have long been inferred from their sequence homology, our study of *TPB6* provides direct experimental demonstration. Our result strongly supports a model in which germline-limited sequences can encode genes with development-specific roles, providing a novel alternative mechanism for achieving transient gene expression. Intriguingly, many of these genes have paralogues that are transcribed in the mature MAC, thus providing opportunities for sub-functionalization or neo-functionalization after their compartmentalization in germline-limited sequences.

Molecular mechanisms for generating germline-limited sequences

Chromosome fragmentation can lead to programmed DNA elimination from somatic genomes. In metazoa, this is often attributed to failure to segregate chromosome fragments during mitosis. For monocentric chromosomes, chromosome fragments without centromeres are often lost. For holocentric chromosomes, alteration in localization of centromeric histone H3 variant CENP-A may contribute to selective loss of chromosome fragments (81). However, these mechanisms probably do not apply to NMC in *Tetrahymena*, as the MAC chromosomes segregate by amitosis without spindle formation, and neither have nor need centromeres (8,82,83). Just like maintained MAC chromosomes, telomeres are added to NMC during formation of the developing MAC, precluding telomere deficiency as an underlying cause for their rapid loss during asexual propagation. Our work also shows that NMC are not actively targeted for degradation, as they are maintained when fused to adjacent MAC chromosomes. We strongly suspect that NMC loss is due to replication defects. In general, NMC are short while normal MAC chromosomes are long (8,16). Combined with the sparse distribution of *Tetrahymena* replication origins in MAC chromosomes, estimated at 1 for every 10 kb (47), it is likely that NMC lack efficient replication origins, and this leads to their loss through dilution during asexual propagation. Future studies are warranted to explore this novel mechanism for generating germline-limited sequences.

Programmed DNA excision represents another mechanism for generating germline-limited sequences. In *T. thermophila*, TPB2 is required for excision of the vast majority of >10 000 IES (26,27), while TPB1 and TPB6 are responsible for a special class of IES ((55) and this study). In another oligohymenophorean ciliate, *Paramecium tetraurelia*, excision of ∼45 000 short IES is dependent on a related *piggyBac* DNA transposase (9,84). In the more distantly related stichotrichous ciliate *O. trifallax*, Tc1/*mariner* family DNA transposases are required for IES excision (85). In the sea lamprey *P. marinus*, the presence of short palindromic sequences near break sites of germline-limited sequences suggests site-specific recombination, possibly also catalyzed by a domesticated transposase (80). These programmed genome rearrangement events are reminiscent of V(D)J recombination, which is most likely derived from an ancient *Transib* DNA transposon (86).

Germline-limited sequences are often marked epigenetically before their removal. In particular, heterochromatin formation pathways have been recurrently connected to DNA elimination. In *T. thermophila*, IES are targeted by conjugation-specific small RNA, marked by heterochromatin-specific histone modifications like H3K27 and H3K9 methylation, decorated by chromodomain effector proteins, then condensed into cytologically distinct DNA elimination bodies (19–21,61,62,65). Similar mechanisms are employed in *P. tetraurelia* (87,88). Intriguingly, H3K9 methylation is also enriched in germline-limited sequences in metazoa, including finches (89,90) and sea lamprey (91).

***piggyBac* transposase domestication in ciliates**

In the oligohymenophorean ciliates, *Tetrahymena* and *Paramecium*, recurrent domestication of *piggyBac* transposases reflects that they are readily adapted to perform developmentally programmed genome rearrangement. Self-

removal is probably the *raison d'être* that initiated domestication of various transposons in ciliates ((9,26,27,55,84,85) and this study). In particular, detrimental effects of *piggyBac* transposons are minimized by their seamless removal from the transcriptionally active MAC. This is underpinned by the selective localization of their cognate transposases in the developing MAC but not the germline MIC. Differential nuclear localization may have arisen naturally, as the somatic MAC is the default destination for most nuclear proteins. MIC-specific nuclear pore proteins and importins are highly divergent compared with their MAC orthologues (92–94), and they most likely recognize specialized nuclear localization signals and exclude all but a few nuclear proteins from the transcriptionally silent MIC, including all of the transcriptional machinery and associated chromatin regulatory components (95–98). This localization pattern, confirmed for *Tetrahymena* TPB2 and *Paramecium* Pgm (27,84) but also inferred for *Tetrahymena* TPB1/TPB6, effectively renders IES germline-limited and stably inherited. Since their initial domestication, these *piggyBac* transposases have been highly optimized for IES excision. Remarkably, they all contain the hyperactive D450N mutation, independently revealed by a mutagenesis screen of the original *Trichoplusia ni piggyBac* transposase for optimized excision activity (99). Presumably, they have also lost their transposition activity, evident in the relative stability in genomic positions of IES.

Transposase domestication generally entails degeneration of transposase features and functions, as well as acquisition of host features and functions (100). In particular, there is decreasing TIR targeting, accompanied by increasing dependence on the heterochromatin formation pathway for these domesticated *piggyBac* transposases. Degeneration of TIR over time compromises the transposase's intrinsic capability to target transposon-derived DNA sequences. This has to be compensated for by additional sequence-specific information. Indeed, all these domesticated *piggyBac* transposases require small RNA, the RNAi machinery, and *Polycomb* group proteins for IES processing. From this perspective, as IES excision in *Tetrahymena* and *Paramecium* evolved, the *piggyBac* transposition mechanism was gradually grafted onto the heterochromatin formation pathway.

However, the *piggyBac* transposases in *Tetrahymena* and *Paramecium* have also diversified significantly, revealing the innate plasticity of the domestication process (Table 1). Specifically: (i) *Tetrahymena* TPB1 and TPB6, in coordination, precisely excise IES restricted to exonic regions ((55) and this study); (ii) *Paramecium* PiggyMac (Pgm) can precisely excise IES anywhere in the genome (9,84); (iii) *Tetrahymena* TPB2 imprecisely excises IES restricted to intergenic and intronic regions (26,27).

Like the original *piggyBac* transposase, *Tetrahymena* TPB1 and TPB6 have retained precise excision, as well as the requirement for a TIR of substantial length (9-bp) in their cognate IES. Precise excision is especially important for restoring functions of host genes with IES inserted into coding regions, while degeneration and retention of IES are more easily tolerated in intergenic or intronic regions. Over evolutionary time, this seems to have driven the exclusive association of TPB1/TPB6-dependent IES with gene coding regions, and the conservation of their TIR. Interestingly, the requirement for only moderate levels of small RNA and lack of *bona fide* heterochromatin likely prevent silencing of genes hosting these IES and imprecise excision by the TPB2 transposase, both of which would have disastrous consequences.

A domesticated *piggyBac* transposase in *Paramecium*, PiggyMac (Pgm), is required for excision of ∼45 000 short IES distributed genome-wide (9,84). These IES have retained a minimalist TIR, reduced to an invariable TA dinucleotide (corresponding to the center of the original *piggyBac* TTAA overhang) (101). Although their processing is precise and efficient overall, excision errors are readily detected as IES excision polymorphism (102). Reduction of the TIR length may have allowed *Paramecium* IES to expand their numbers and genomic distribution, as well as made them dependent on the heterochromatin formation pathway.

In contrast, *Tetrahymena* TPB2 has lost excision precision ((8) and this study). Their cognate IESs have no TIR ((8) and this study). Excision boundaries, in the few cases studied, are instead defined by host factors (103). Indeed, small RNA, derived from double-stranded RNA introduced into conjugating *Tetrahymena* cells, are sufficient for eliminating any targeted homologous sequences in an imprecise manner (104), probably through excision by TPB2 (26,27). This allows dramatic and continuous expansion of their cognate IES to include essentially all sequences forming *bona fide* heterochromatin in the developing MAC, which functions as an effective genome defense system. Indeed, many IES are derived from various transposons, some of which are obviously recent invasions (8). Thus *Tetrahymena* TPB2 seemingly represents an extreme case of *piggyBac* transposase domestication.

In conclusion, *piggyBac* transposase domestication in ciliates illustrates their remarkable evolutionary versatility in achieving programmed genome rearrangement. Indeed, the system is so flexible that it allows efficient genome editing programmable by double-stranded RNA (converted into small RNA *in vivo*) (104,105). It will be profitable to further dissect the molecular mechanism for small RNA-targeted excision by *Tetrahymena* and *Paramecium* domesticated *piggyBac* transposases, and explore the feasibility of adapting them to introduce specific genetic changes in other eukaryotes.

Analogy between exonic IES and introns

A striking analogy can be drawn between exonic IES and introns. Exonic IES interrupt coding regions, and need to be excised precisely, albeit at the DNA level during the germline to soma differentiation. *Tetrahymena* exonic IES, among all IES, are the most similar to the original *piggyBac* transposons. Indeed, *Tetrahymena* exonic IES may have descended directly from non-autonomous *piggyBac* transposons. It is unlikely that they arose spontaneously, due to the stringency imposed by the dual requirement for the long TIR and additional internal sequences targeted by the heterochromatin formation pathway. It is even more unlikely that these exonic IES can be removed spontaneously and precisely without TPB1 and TPB6, even by the related

TPB2. Conservation of *Tetrahymena* exonic IES, especially their insertion positions, is therefore analogous to conservation of introns (106). Alternative transcription products are determined by *Tetrahymena* exonic IES, analogous to alternative splicing products. We also note that exonic IES can differentially populate related *Tetrahymena* species. Analogous to intron exonization (107), the IES-5 equivalent is retained in the MAC of *T. borealis*. Analogous to intron loss (108), a few exonic IES have been precisely lost in the MIC of *T. malaccensis*, most likely attributable to mislocalization of TPB1 and TPB6 homologues to the MIC in a rare event. There are many more exonic IES in *Paramecium*. Some may have arisen spontaneously and even recently, due to their minimalist TIR and short length overall, analogous to intronization (109). *Paramecium* exonic IES are under selection for recognition by nonsense-mediated decay, as IES 3n in length that do not contain an in-frame stop codon, like their intron counterparts, are significantly under-represented (9,102).

There is an emerging consensus that the spliceosome and spliceosomal introns evolved from self-splicing group II introns that invaded and populated the genome of the last eukaryotic common ancestor (110). Thus, like domesticated *piggyBac* transposases and IES, they most likely also originated from mobile genetic elements. Comparative studies of these two distinct but analogous systems may provide a new perspective and deep insight for understanding their evolutionary history.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Wang,J. and Davis,R.E. (2014) Programmed DNA elimination in multicellular organisms. *Curr. Opin. Genet. Dev.*, **27**, 26–34.
2. Karrer,K.M. (2012) Nuclear dualism. *Methods Cell Biol.*, **109**, 29–52.
3. Bassing,C.H., Swat,W. and Alt,F.W. (2002) The mechanism and regulation of chromosomal V(D)J recombination. *Cell*, **109**(Suppl.), S45–S55.
4. Cervantes,M.D., Hamilton,E.P., Xiong,J., Lawson,M.J., Yuan,D., Hadjithomas,M., Miao,W. and Orias,E. (2013) Selecting one of several mating types through gene segment joining and deletion in Tetrahymena thermophila. *PLoS Biol.*, **11**, e1001518.
5. Chen,X., Bracht,J.R., Goldman,A.D., Dolzhenko,E., Clay,D.M., Swart,E.C., Perlman,D.H., Doak,T.G., Stuart,A., Amemiya,C.T. *et al.* (2014) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell*, **158**, 1187–1198.
6. Chalker,D.L., Meyer,E. and Mochizuki,K. (2013) Epigenetics of ciliates. *Cold Spring Harb. Perspect. Biol.*, **5**, a017764.
7. Bracht,J.R., Fang,W., Goldman,A.D., Dolzhenko,E., Stein,E.M. and Landweber,L.F. (2013) Genomes on the edge: programmed genome instability in ciliates. *Cell*, **152**, 406–416.
8. Hamilton,E.P., Kapusta,A., Huvos,P.E., Bidwell,S.L., Zafar,N., Tang,H., Hadjithomas,M., Krishnakumar,V., Badger,J.H., Caler,E.V. *et al.* (2016) Structure of the germline genome of Tetrahymenaathermophila and relationship to the massively rearranged somatic genome. *eLife*, **5**, e19090.
9. Arnaiz,O., Mathy,N., Baudry,C., Malinsky,S., Aury,J.M., Denby Wilkes,C., Garnier,O., Labadie,K., Lauderdale,B.E., Le Mouel,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.
10. Streit,A., Wang,J., Kang,Y. and Davis,R.E. (2016) Gene silencing and sex determination by programmed DNA elimination in parasitic nematodes. *Curr. Opin. Microbiol.*, **32**, 120–127.
11. Yao,M.C., Chao,J.L. and Cheng,C.Y. (2014) Programmed genome rearrangements in Tetrahymena. *Microbiol. Spectrum*, **2**, doi:10.1128/microbiolspec.MDNA3-0012-2014.
12. Chalker,D.L. and Yao,M.C. (2011) DNA elimination in ciliates: transposon domestication and genome surveillance. *Annu. Rev. Genet.*, **45**, 227–246.
13. Yao,M.C., Yao,C.H. and Monks,B. (1990) The controlling sequence for site-specific chromosome breakage in Tetrahymena. *Cell*, **63**, 763–772.
14. Fan,Q. and Yao,M.C. (2000) A long stringent sequence signal for programmed chromosome breakage in Tetrahymena thermophila. *Nucleic Acids Res.*, **28**, 895–900.
15. Hamilton,E.P., Williamson,S., Dunn,S., Merriam,V., Lin,C., Vong,L., Russell-Colantonio,J. and Orias,E. (2006) The highly conserved family of Tetrahymena thermophila chromosome breakage elements contains an invariant 10-base-pair core. *Eukaryotic Cell*, **5**, 771–780.
16. Lin,C.G., Lin,I.T. and Yao,M.C. (2016) Programmed minichromosome elimination as a mechanism for somatic genome reduction in Tetrahymena thermophila. *PLoS Genet.*, **12**, e1006403.
17. Yao,M.C. (1982) Elimination of specific DNA sequences from the somatic nucleus of the ciliate Tetrahymena. *J. Cell Biol.*, **92**, 783–789.
18. Yokoyama,R.W. and Yao,M.C. (1982) Elimination of DNA sequences during macronuclear differentiation in Tetrahymena thermophila, as detected by in situ hybridization. *Chromosoma*, **85**, 11–22.
19. Liu,Y., Mochizuki,K. and Gorovsky,M.A. (2004) Histone H3 lysine 9 methylation is required for DNA elimination in developing macronuclei in Tetrahymena. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 1679–1684.
20. Liu,Y., Taverna,S.D., Muratore,T.L., Shabanowitz,J., Hunt,D.F. and Allis,C.D. (2007) RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena. *Genes Dev.*, **21**, 1530–1545.

21. Mochizuki,K., Fine,N.A., Fujisawa,T. and Gorovsky,M.A. (2002) Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. *Cell*, **110**, 689–699.

22. Malone,C.D., Anderson,A.M., Motl,J.A., Rexer,C.H. and Chalker,D.L. (2005) Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of Tetrahymena thermophila. *Mol. Cell Biol.*, **25**, 9151–9164.

23. Mochizuki,K. and Gorovsky,M.A. (2005) A Dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev.*, **19**, 77–89.

24. Aronica,L., Bednenko,J., Noto,T., DeSouza,L.V., Siu,K.W., Loidl,J., Pearlman,R.E., Gorovsky,M.A. and Mochizuki,K. (2008) Study of an RNA helicase implicates small RNA-noncoding RNA interactions in programmed DNA elimination in Tetrahymena. *Genes Dev.*, **22**, 2228–2241.

25. Bednenko,J., Noto,T., DeSouza,L.V., Siu,K.W., Pearlman,R.E., Mochizuki,K. and Gorovsky,M.A. (2009) Two GW repeat proteins interact with Tetrahymena thermophila argonaute and promote genome rearrangement. *Mol. Cell Biol.*, **29**, 5020–5030.

26. Vogt,A. and Mochizuki,K. (2013) A domesticated PiggyBac transposase interacts with heterochromatin and catalyzes reproducible DNA elimination in Tetrahymena. *PLoS Genet.*, **9**, e1004032.

27. Cheng,C.Y., Vogt,A., Mochizuki,K. and Yao,M.C. (2010) A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. *Mol. Biol. Cell*, **21**, 1753–1762.

28. Cassidy-Hanley,D., Bisharyan,Y., Fridman,V., Gerber,J., Lin,C., Orias,E., Orias,J.D., Ryder,H., Vong,L. and Hamilton,E.P. (2005) Genome-wide characterization of Tetrahymena thermophila chromosome breakage sites. II. Physical and genetic mapping. *Genetics*, **170**, 1623–1631.

29. Schwope,R.M. and Chalker,D.L. (2014) Mutations in Pdd1 reveal distinct requirements for its chromodomain and chromoshadow domain in directing histone methylation and heterochromatin elimination. *Eukaryot. Cell*, **13**, 190–201.

30. Gorovsky,M.A., Yao,M.C., Keevert,J.B. and Pleger,G.L. (1975) Isolation of micro- and macronuclei of Tetrahymena pyriformis. *Methods Cell Biol.*, **9**, 311–327.

31. Sweet,M.T. and Allis,C.D. (2006) Isolation and purification of tetrahymena nuclei. *CSH Protoc.*, **2006**, doi:10.1101/pdb.prot4500.

32. Papazyan,R., Voronina,E., Chapman,J.R., Luperchio,T.R., Gilbert,T.M., Meier,E., Mackintosh,S.G., Shabanowitz,J., Tackett,A.J., Reddy,K.L. *et al.* (2014) Methylation of histone H3K23 blocks DNA damage in pericentric heterochromatin during meiosis. *eLife*, **3**, e02996.

33. Xiong,J., Gao,S., Dui,W., Yang,W., Chen,X., Taverna,S.D., Pearlman,R.E., Ashlock,W., Miao,W. and 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.

34. Allis,C.D. and Dennison,D.K. (1982) Identification and purification of young macronuclear anlagen from conjugating cells of Tetrahymena thermophila. *Dev. Biol.*, **93**, 519–533.

35. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

36. Stein,L.D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinform.*, **14**, 162–171.

37. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.

38. Stover,N.A., Punia,R.S., Bowen,M.S., Dolins,S.B. and Clark,T.G. (2012) Tetrahymena Genome Database Wiki: a community-maintained model organism database. *Database*, **2012**, bas007.

39. Eisen,J.A., Coyne,R.S., Wu,M., Wu,D., Thiagarajan,M., Wortman,J.R., Badger,J.H., Ren,Q., Amedeo,P., Jones,K.M. *et al.* (2006) Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. *PLoS Biol.*, **4**, e286.

40. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

41. Duharcourt,S. and Yao,M.C. (2002) Role of histone deacetylation in developmentally programmed DNA rearrangements in Tetrahymena thermophila. *Eukaryotic Cell*, **1**, 293–303.

42. Cassidy-Hanley,D., Bowen,J., Lee,J.H., Cole,E., VerPlank,L.A., Gaertig,J., Gorovsky,M.A. and Bruns,P.J. (1997) Germline and somatic transformation of mating Tetrahymena thermophila by particle bombardment. *Genetics*, **146**, 135–147.

43. Orias,E. and Flacks,M. (1975) Macronuclear genetics of Tetrahymena. I. Random distribution of macronuclear genecopies in T. pyriformis, syngen 1. *Genetics*, **79**, 187–206.

44. Hai,B. and Gorovsky,M.A. (1997) Germ-line knockout heterokaryons of an essential alpha-tubulin gene enable high-frequency gene replacement and a test of gene transfer from somatic to germ-line nuclei in Tetrahymena thermophila. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 1310–1315.

45. Liao,Y., Smyth,G.K. and Shi,W. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, **41**, e108.

46. Mochizuki,K. (2008) High efficiency transformation of Tetrahymena using a codon-optimized neomycin resistance gene. *Gene*, **425**, 79–83.

47. Gao,S., Xiong,J., Zhang,C., Berquist,B.R., Yang,R., Zhao,M., Molascon,A.J., Kwiatkowski,S.Y., Yuan,D., Qin,Z. *et al.* (2013) Impaired replication elongation in Tetrahymena mutants deficient in histone H3 Lys 27 monomethylation. *Genes Dev.*, **27**, 1662–1679.

48. Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.

49. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

50. Miao,W., Xiong,J., Bowen,J., Wang,W., Liu,Y., Braguinets,O., Grigull,J., Pearlman,R.E., Orias,E. and Gorovsky,M.A. (2009) Microarray analyses of gene expression during the Tetrahymena thermophila life cycle. *PLoS One*, **4**, e4429.

51. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

52. Nikiforov,M.A., Smothers,J.F., Gorovsky,M.A. and Allis,C.D. (1999) Excision of micronuclear-specific DNA requires parental expression of pdd2p and occurs independently from DNA replication in Tetrahymena thermophila. *Genes Dev.*, **13**, 2852–2862.

53. Yao,M.C. and Yao,C.H. (1994) Detection of circular excised DNA deletion elements in Tetrahymena thermophila during development. *Nucleic Acids Res.*, **22**, 5702–5708.

54. Fass,J.N., Joshi,N.A., Couvillion,M.T., Bowen,J., Gorovsky,M.A., Hamilton,E.P., Orias,E., Hong,K., Coyne,R.S., Eisen,J.A. *et al.* (2011) Genome-scale analysis of programmed DNA elimination sites in Tetrahymena thermophila. *G3 (Bethesda)*, **1**, 515–522.

55. Cheng,C.Y., Young,J.M., Lin,C.G., Chao,J.L., Malik,H.S. and Yao,M.C. (2016) The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in Tetrahymena thermophila. *Genes Dev.*, **30**, 2724–2736.

56. Cary,L.C., Goebel,M., Corsaro,B.G., Wang,H.G., Rosen,E. and Fraser,M.J. (1989) Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology*, **172**, 156–169.

57. Mitra,R., Fain-Thornton,J. and Craig,N.L. (2008) piggyBac can bypass DNA synthesis during cut and paste transposition. *EMBO J.*, **27**, 1097–1109.

58. Betermier,M., Bertrand,P. and Lopez,B.S. (2014) Is non-homologous end-joining really an inherently error-prone process? *PLoS Genet.*, **10**, e1004086.

59. Lin,I.T., Chao,J.L. and Yao,M.C. (2012) An essential role for the DNA breakage-repair protein Ku80 in programmed DNA rearrangements in Tetrahymena thermophila. *Mol. Biol. Cell*, **23**, 2213–2225.

60. Elick,T.A., Lobo,N. and Fraser,M.J. Jr (1997) Analysis of the cis-acting DNA elements required for piggyBac transposable element excision. *Mol. Gen. Genet.: MGG*, **255**, 605–610.

61. Taverna,S.D., Coyne,R.S. and Allis,C.D. (2002) Methylation of histone h3 at lysine 9 targets programmed DNA elimination in tetrahymena. *Cell*, **110**, 701–711.
62. Madireddi,M.T., Coyne,R.S., Smothers,J.F., Mickey,K.M., Yao,M.C. and Allis,C.D. (1996) Pdd1p, a novel chromodomain-containing protein, links heterochromatin assembly and DNA elimination in Tetrahymena. *Cell*, **87**, 75–84.
63. Noto,T., Kataoka,K., Suhren,J.H., Hayashi,A., Woolcock,K.J., Gorovsky,M.A. and Mochizuki,K. (2015) Small-RNA-mediated genome-wide trans-recognition network in Tetrahymena DNA elimination. *Mol. Cell*, **59**, 229–242.
64. Chalker,D.L. and Yao,M.C. (2001) Nongenic, bidirectional transcription precedes and may promote developmental DNA deletion in Tetrahymena thermophila. *Genes Dev.*, **15**, 1287–1298.
65. Schoeberl,U.E., Kurth,H.M., Noto,T. and Mochizuki,K. (2012) Biased transcription and selective degradation of small RNAs shape the pattern of DNA elimination in Tetrahymena. *Genes Dev.*, **26**, 1729–1742.
66. Kataoka,K. and Mochizuki,K. (2015) Phosphorylation of an HP1-like protein regulates heterochromatin body assembly for DNA elimination. *Dev. Cell*, **35**, 775–788.
67. Shen,X. and Gorovsky,M.A. (1996) Linker histone H1 regulates specific gene expression but not global transcription in vivo. *Cell*, **86**, 475–483.
68. Zhang,C.S., Jiang,B., Li,M., Zhu,M., Peng,Y., Zhang,Y.L., Wu,Y.Q., Li,T.Y., Liang,Y., Lu,Z. *et al.* (2014) The lysosomal v-ATPase-Ragulator complex is a common activator for AMPK and mTORC1, acting as a switch between catabolism and anabolism. *Cell Metab.*, **20**, 526–540.
69. Plattner,H. (2015) The contractile vacuole complex of protists—new cues to function and biogenesis. *Crit. Rev. Microbiol.*, **41**, 218–227.
70. Cotter,K., Stransky,L., McGuire,C. and Forgac,M. (2015) Recent insights into the structure, regulation, and function of the V-ATPases. *Trends Biochem. Sci.*, **40**, 611–622.
71. Gillingham,A.K., Whyte,J.R., Panic,B. and Munro,S. (2006) Mon2, a relative of large Arf exchange factors, recruits Dop1 to the Golgi apparatus. *J. Biol. Chem.*, **281**, 2273–2280.
72. Barbosa,S., Pratte,D., Schwarz,H., Pipkorn,R. and Singer-Kruger,B. (2010) Oligomeric Dop1p is part of the endosomal Neo1p-Ysl2p-Arl1p membrane remodeling complex. *Traffic*, **11**, 1092–1106.
73. Yusa,K. (2015) piggyBac transposon. *Microbiol. Spectrum*, **3**, MDNA3–0028–2014.
74. Coyne,R.S. and Yao,M.C. (1996) Evolutionary conservation of sequences directing chromosome breakage and rDNA palindrome formation in tetrahymenine ciliates. *Genetics*, **144**, 1479–1487.
75. Etter,A., Aboutanos,M., Tobler,H. and Muller,F. (1991) Eliminated chromatin of Ascaris contains a gene that encodes a putative ribosomal protein. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 1593–1596.
76. Etter,A., Bernard,V., Kenzelmann,M., Tobler,H. and Muller,F. (1994) Ribosomal heterogeneity from chromatin diminution in Ascaris lumbricoides. *Science*, **265**, 954–956.
77. Wang,J., Mitreva,M., Berriman,M., Thorne,A., Magrini,V., Koutsovoulos,G., Kumar,S., Blaxter,M.L. and Davis,R.E. (2012) Silencing of germline-expressed genes by DNA elimination in somatic cells. *Dev. Cell*, **23**, 1072–1080.
78. Smith,J.J., Antonacci,F., Eichler,E.E. and Amemiya,C.T. (2009) Programmed loss of millions of base pairs from a vertebrate genome. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 11212–11217.
79. Bryant,S.A., Herdy,J.R., Amemiya,C.T. and Smith,J.J. (2016) Characterization of somatically-eliminated genes during development of the Sea Lamprey (Petromyzon marinus). *Mol. Biol. Evol.*, **33**, 2337–2344.
80. Smith,J.J., Baker,C., Eichler,E.E. and Amemiya,C.T. (2012) Genetic consequences of programmed genome rearrangement. *Curr. Biol.*, **22**, 1524–1529.
81. Kang,Y., Wang,J., Neff,A., Kratzer,S., Kimura,H. and Davis,R.E. (2016) Differential chromosomal localization of centromeric histone CENP-A contributes to nematode programmed DNA elimination. *Cell Rep.*, **16**, 2308–2316.
82. Cui,B. and Gorovsky,M.A. (2006) Centromeric histone H3 is essential for vegetative cell division and for DNA elimination during conjugation in Tetrahymena thermophila. *Mol. Cell. Biol.*, **26**, 4499–4510.
83. Cervantes,M.D., Xi,X., Vermaak,D., Yao,M.C. and Malik,H.S. (2006) The CNA1 histone of the ciliate Tetrahymena thermophila is essential for chromosome segregation in the germline micronucleus. *Mol. Biol. Cell*, **17**, 485–497.
84. Baudry,C., Malinsky,S., Restituito,M., Kapusta,A., Rosa,S., Meyer,E. and Betermier,M. (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. *Genes Dev.*, **23**, 2478–2483.
85. Nowacki,M., Higgins,B.P., Maquilan,G.M., Swart,E.C., Doak,T.G. and Landweber,L.F. (2009) A functional role for transposases in a large eukaryotic genome. *Science*, **324**, 935–938.
86. Kapitonov,V.V. and Jurka,J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.*, **3**, e181.
87. Lhuillier-Akakpo,M., Frapporti,A., Denby Wilkes,C., Matelot,M., Vervoort,M., Sperling,L. and Duharcourt,S. (2014) Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements. *PLoS Genet.*, **10**, e1004665.
88. Lepere,G., Nowacki,M., Serrano,V., Gout,J.F., Guglielmi,G., Duharcourt,S. and Meyer,E. (2009) Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. *Nucleic Acids Res.*, **37**, 903–915.
89. del Priore,L. and Pigozzi,M.I. (2014) Histone modifications related to chromosome silencing and elimination during male meiosis in Bengalese finch. *Chromosoma*, **123**, 293–302.
90. Goday,C. and Pigozzi,M.I. (2010) Heterochromatin and histone modifications in the germline-restricted chromosome of the zebra finch undergoing elimination during spermatogenesis. *Chromosoma*, **119**, 325–336.
91. Timoshevskiy,V.A., Herdy,J.R., Keinath,M.C. and Smith,J.J. (2016) Cellular and molecular features of developmentally programmed genome rearrangement in a vertebrate (Sea Lamprey: Petromyzon marinus). *PLoS Genet.* **12**, e1006103.
92. Malone,C.D., Falkowska,K.A., Li,A.Y., Galanti,S.E., Kanuru,R.C., LaMont,E.G., Mazzarella,K.C., Micev,A.J., Osman,M.M., Piotrowski,N.K. *et al.* (2008) Nucleus-specific importin alpha proteins and nucleoporins regulate protein import and nuclear division in the binucleate Tetrahymena thermophila. *Eukaryot. Cell*, **7**, 1487–1499.
93. Iwamoto,M., Koujin,T., Osakada,H., Mori,C., Kojidani,T., Matsuda,A., Asakawa,H., Hiraoka,Y. and Haraguchi,T. (2015) Biased assembly of the nuclear pore complex is required for somatic and germline nuclear differentiation in Tetrahymena. *J. Cell Sci.*, **128**, 1812–1823.
94. Iwamoto,M., Mori,C., Kojidani,T., Bunai,F., Hori,T., Fukagawa,T., Hiraoka,Y. and Haraguchi,T. (2009) Two distinct repeat sequences of Nup98 nucleoporins characterize dual nuclei in the binucleated ciliate tetrahymena. *Curr. Biol.*, **19**, 843–847.
95. Mochizuki,K. and Gorovsky,M.A. (2004) RNA polymerase II localizes in Tetrahymena thermophila meiotic micronuclei when micronuclear transcription associated with genome rearrangement occurs. *Eukaryot. cell*, **3**, 1233–1240.
96. Stargell,L.A. and Gorovsky,M.A. (1994) TATA-binding protein and nuclear differentiation in Tetrahymena thermophila. *Mol. Cell Biol.*, **14**, 723–734.
97. Stargell,L.A., Bowen,J., Dadd,C.A., Dedon,P.C., Davis,M., Cook,R.G., Allis,C.D. and Gorovsky,M.A. (1993) Temporal and spatial association of histone H2A variant hv1 with transcriptionally competent chromatin during nuclear development in Tetrahymena thermophila. *Genes Dev.*, **7**, 2641–2651.
98. Strahl,B.D., Ohba,R., Cook,R.G. and Allis,C.D. (1999) Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 14967–14972.
99. Li,X., Burnight,E.R., Cooney,A.L., Malani,N., Brady,T., Sander,J.D., Staber,J., Wheelan,S.J., Joung,J.K., McCray,P.B. Jr *et al.* (2013) piggyBac transposase tools for genome engineering. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E2279–E2287.
100. Sinzelle,L., Izsvak,Z. and Ivics,Z. (2009) Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell. Mol. Life Sci.: CMLS*, **66**, 1073–1093.

101. Betermier,M. and Duharcourt,S. (2014) Programmed rearrangement in ciliates: Paramecium. *Microbiol. Spectrum*, **2**, doi:10.1128/microbiolspec.MDNA3-0035-2014.

102. Duret,L., Cohen,J., Jubin,C., Dessen,P., Gout,J.F., Mousset,S., Aury,J.M., Jaillon,O., Noel,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.

103. Carle,C.M., Zaher,H.S. and Chalker,D.L. (2016) A parallel G quadruplex-binding protein regulates the boundaries of DNA elimination events of Tetrahymena thermophila. *PLoS Genet.*, **12**, e1005842.

104. Yao,M.C., Fuller,P. and Xi,X. (2003) Programmed DNA deletion as an RNA-guided system of genome defense. *Science*, **300**, 1581–1584.

105. Garnier,O., Serrano,V., Duharcourt,S. and Meyer,E. (2004) RNA-mediated programming of developmental genome rearrangements in Paramecium tetraurelia. *Mol. Cell. Biol.*, **24**, 7370–7379.

106. Rogozin,I.B., Wolf,Y.I., Sorokin,A.V., Mirkin,B.G. and Koonin,E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.: CB*, **13**, 1512–1517.

107. Schmitz,J. and Brosius,J. (2011) Exonization of transposed elements: a challenge and opportunity for evolution. *Biochimie*, **93**, 1928–1934.

108. Mourier,T. and Jeffares,D.C. (2003) Eukaryotic intron loss. *Science*, **300**, 1393.

109. Irimia,M., Rukov,J.L., Penny,D., Vinther,J., Garcia-Fernandez,J. and Roy,S.W. (2008) Origin of introns by 'intronization' of exonic sequences. *Trends Genet.*, **24**, 378–381.

110. Irimia,M. and Roy,S.W. (2014) Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.*, **6**, a016071.

111. Kumagai,A., Ando,R., Miyatake,H., Greimel,P., Kobayashi,T., Hirabayashi,Y., Shimogori,T. and Miyawaki,A. (2013) A bilirubin-inducible fluorescent protein from eel muscle. *Cell*, **153**, 1602–1611.