Transcribed germline-limited coding sequences in *Oxytricha trifallax*

Richard V. Miller[1,2], Rafik Neme[1,3], Derek M. Clay[1,2], Jananan S. Pathmanathan[1], Michael W. Lu[1,6], V. Talya Yerlici[1,4], Jaspreet S. Khurana[1,5], and Laura F. Landweber[1,6*]

[1] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

[2] Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

[3] Current address: Department of Chemistry and Biology, Universidad del Norte, Barranquilla, Colombia

[4] Current address: Department of Laboratory Medicine and Pathobiology, Faculty of Medicine, University of Toronto, Toronto, ON M5G 1M1, Canada

[5] Current address: Strand Therapeutics, Cambridge, MA 02139, USA

[6] Department of Biological Sciences, Columbia University, New York, NY 10027, USA

Running Head: *Oxytricha* transcribed germline-limited ORFs

Keywords: germline; genome rearrangement; DNA elimination; noncoding RNA; ciliate; micronucleus

*To whom correspondence should be addressed:

Laura F. Landweber

Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

Phone: +1 212 305-3898

Email: Laura.Landweber@columbia.edu

Abstract

The germline-soma divide is a fundamental distinction in developmental biology, and different genes are expressed in germline and somatic cells throughout metazoan life cycles. Ciliates, a group of microbial eukaryotes, exhibit germline-somatic nuclear dimorphism within a single cell with two different genomes. The ciliate *Oxytricha trifallax* undergoes massive RNA-guided DNA elimination and genome rearrangement to produce a new somatic macronucleus (MAC) from a copy of the germline micronucleus (MIC). This process eliminates noncoding DNA sequences that interrupt genes and also deletes hundreds of germline-limited open reading frames (ORFs) that are transcribed during genome rearrangement. Here, we update the set of transcribed germline-limited ORFs (TGLOs) in *O. trifallax*. We show that TGLOs tend to be expressed during nuclear development and then are absent from the somatic MAC. We also demonstrate that exposure to synthetic RNA can reprogram TGLO retention in the somatic MAC and that TGLO retention leads to transcription outside the normal developmental program. These data suggest that TGLOs represent a group of developmentally regulated protein coding sequences whose gene expression is terminated by DNA elimination.

Introduction

Ciliates are a lineage of microbial eukaryotes characterized by functional nuclear differentiation. Each ciliate cell has one or more somatic macronuclei (MAC) and one or more germline micronuclei (MIC). The somatic MAC of the ciliate *Oxytricha trifallax* contains a genome of over 17,000 gene-sized nanochromosomes that are transcribed throughout the organism's life cycles (Swart et al. 2013; Lindblad et al. 2019). The germline genome is a fragmented and scrambled version of the somatic genome that undergoes a complex process of DNA deletion and rearrangement during sexual reproduction (Chen et al. 2014).

Previous studies have shown that *Oxytricha*'s sexual rearrangement cycle is guided by several noncoding RNA pathways. In the early stages of the sexual life cycle, bidirectional transcription across the length of nanochromosomes produce thousands of long template RNAs from the parental MAC (Lindblad et al. 2017). These transcripts then have the ability to program the rearrangement of retained macronuclear destined sequences (MDSs) during development. Each nanochromosome, itself, forms by joining specific MDSs, with telomere addition at the ends. Previous experiments have showed that injection of synthetic template RNAs can also reprogram aberrant MDS rearrangements (Nowacki et al. 2008; Bracht et al. 2017; Nowacki et al. 2011). Millions of 27-nucleotide long PIWI-associated small RNAs (piRNAs) are abundant during early *Oxytricha* rearrangement and interact with the *Oxytricha* PIWI ortholog Otiwi-1. These piRNAs also derive from the parental MAC. Their role is to mark and protect retained MDS sequences against DNA deletion during development of the zygotic MAC. Injection of synthetic piRNA sequences that target internal eliminated sequences (IESs) that interrupt MDSs in the MIC can prevent their deletion during rearrangement and reprogram their retention in the MAC instead (Fang et al. 2012). Programmed IES retention is now used as a genetic tool to create somatic knockout strains in *Oxytricha* (Khurana et al. 2018; Beh et al. 2019).

Besides IESs and transposons that are eliminated during development, *Oxytricha* has other classes of germline-specific MIC DNA sequences (Chen et al. 2016). Analysis of the germline MIC genome together with transcriptome-guided gene prediction previously uncovered 810 germline-limited protein coding genes encoded in the MIC genome (Chen et al. 2014). These germline-limited genes are specifically transcribed during rearrangement, and 26% of them had demonstrated translation of peptides present in a survey of one developmental time point.

Other lineages also have germline-limited protein coding sequences, including the ciliate *Tetrahymena thermophila* (Hamilton et al. 2016; Lin et al. 2016; Feng et al. 2017), the parasitic roundworm *Ascaris suum* (Wang et al. 2012; Wang et al. 2017), and the sea lamprey *Petromyzon marinus* (Bryant et al. 2016; Smith et al. 2009; Smith et al. 2012; Timoshevskiy et al. 2016; Timoshevskiy et al. 2017). Protein coding sequences are discarded in all these cases, and genes eliminated from somatic lineage cells are typically predicted to have functions in the germline and embryogenesis (Smith et al. 2012; Bryant et al. 2016). The songbird *Taeniopygia guttata* has a germline-limited chromosome that is deleted from somatic lineage cells (Pigozzi and Solari 1998; Pigozzi and Solari 2005; Itoh et al. 2009; Biederman et al. 2018; Kinsella et al. 2019; Torgasheva et al. 2019).

Here, we introduce the set of transcribed germline-limited ORFs (TGLOs) in *Oxytricha*, which update and expand the previously annotated germline-limited genes (Chen et al. 2014). We also provide functional experiments that reprogram the somatic retention of a small number of TGLOs to test the hypothesis that developmental deletion is the main mechanism to repress their gene expression during asexual growth. Like the previous set of germline-limited genes, we show that TGLOs contain several predicted functions and conserved domains that could be involved in *Oxytricha* development. This work also identified a locus, g111288, that is retained in the somatic MAC of a subset of progeny cells, revealing an example of a strain-specific macronuclear chromosome.

Materials and methods

Illumina library preparation and sequencing

Genomic DNA was collected from mated *O. trifallax* cells at various developmental time-points using the Nucleospin genomic DNA spin column (Machery-Nagle). Illumina DNA sequencing libraries were prepared using the NEBNext Ultra II library preparation kit (New England Biolabs). 2 x 250 bp paired end sequencing reads were obtained using an Illumina HiSeq 2500, and remaining adapter sequences were trimmed using Trim Galore! software in the Galaxy cloud computing environment.

Total RNA was extracted from mated *O. trifallax* cells at various developmental time-points using Trizol reagent (Thermo Fisher, Waltham, MA, USA). Contaminating DNA was removed using a Turbo DNase kit (Thermo Fisher, Waltham, MA, USA). Poly-adenylated transcripts were enriched using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, Ipswich, MA, USA). RNA sequencing libraries were prepared using the ScriptSeq version 2 kit (Illumina, San Diego, CA, USA). 2 x 75 bp paired end sequencing reads were obtained using an Illumina HiSeq 2500, and remaining adapter sequences were trimmed using Trim Galore! software in the Galaxy cloud computing environment.

TGLO computational prediction

We predicted TGLOs using a previously published pipeline for germline-limited gene prediction with some modifications (Chen et al. 2014). We predicted coding sequences with AUGUSTUS (version 3.3.0) (Stanke et al. 2006) using a gene prediction model trained on *O. trifallax* somatic MAC genes and transcripts as hints. We generated hint files for the gene prediction software by mapping RNA-seq data from cells collected at various time points to the germline MIC genome using HISAT2 (version 2.0.5). We ran AUGUSTUS with the options --

UTR=on and --alternatives-from-evidence=true. We filtered AUGUSTUS gene predictions to keep only models supported by hints including at least four supporting RNA-seq reads and greater than 80% of the coding sequence covered by RNA-seq reads to obtain the high transcription dataset. We kept only models supported by hints including at least two supporting RNA-seq reads and required greater than 20% of the coding sequence be covered by RNA-seq reads to obtain the low transcription dataset. We also removed candidate sequences with more than a minimal number of whole cell genomic DNA reads mapped from asexually growing cultures of either parental genotype or a pool of F1 cells to ensure that MAC encoded candidates were removed while accounting for the fact that some MIC encoded sequences will be present in whole cell sequencing reads.

DNA sequencing analysis

Genomic DNA sequencing reads were aligned to the *O. trifallax* MIC genome assembly using BWA-MEM (version 0.7.17) with the -M option to mark short split alignments as supplementary alignments. Alignment files were processed using the Samtools software package (version 0.1.20) (Li et al. 2009). FeatureCounts software (version 2.0.0) (Liao et al. 2014) was used to assess the raw number of reads mapping to *O. trifallax* genome features (Burns et al. 2016). Relative DNA copy number changes for each genome feature were normalized using the R/Bioconductor package DESeq2 (version 1.26.0) (Love et al. 2014). Heat maps showing normalized DNA copy number during the developmental life cycle were generated using the log2 normalized copy number values and the pheatmap R package (version 1.0.12).

Transcriptome sequencing analysis

Poly(A)-selected RNA sequencing reads were aligned to the *O. trifallax* MAC genome assembly and MIC genome assembly using HISAT2 (version 2.0.4) and Bowtie2 in the local alignment mode, respectively. Relative DNA copy number changes were normalized using the R/Bioconductor package DESeq2. Alignment files were processed using the Samtools software package (version 0.1.20) (Li et al. 2009). FeatureCounts software (version 2.0.0) (Liao et al. 2014) was used to assess the raw number of reads mapping to *O. trifallax* genome features (Burns et al. 2016). Relative RNA expression changes for each genome feature were normalized using the R/Bioconductor package DESeq2 (version 1.26.0) (Love et al. 2014). Heat maps showing normalized RNA expression during the developmental life cycle were generated using the log2 normalized copy number values and the pheatmap R package (version 1.0.12). Two timepoints of triplicate RNA-seq reads (12 hr and 36 hr) from the late time-course were previously uploaded to the European Nucleotide Archive under the project number PRJEB32087.

Small RNA sequencing analysis

Previously sequenced Otiwi-1-dependent piRNAs (Fang et al. 2012) were aligned to the *O. trifallax* MIC genome assembly using Bowtie2 (version 2.3.4.1) in the local alignment mode. Alignment files were processed using the Samtools software package (version 0.1.20) (Li et al. 2009), and alignments were viewed in the context of the MIC genome using the Integrative Genomics Viewer (version 2.7.2) (Robinson et al. 2011).

Mass spectrometry analysis

Raw data were analyzed using MaxQuant (version 1.6.3.4) to search against a combined database containing previously published macronuclear-encoded and MIC-limited genes in addition to either highly-transcribed or lowly-transcribed TGLOs (Chen et al. 2014). Searches were performed using Trypsin/P as the enzyme with a maximum of two missed cleavages, methionine oxidation and protein N-terminal acetylation as variable modifications, cysteine carbamidomethylation as a fixed modification, precursor mass tolerances of 20 ppm for the first search and 4.5 ppm for the main search, and a maximum FDR of 1% for both peptides and proteins.

Cell culture

Oxytricha trifallax cells were cultured in Petri dishes or large Pyrex dishes containing Pringsheim medium (0.11 mM Na_2HPO_4, 0.08mM $MgSO_4$, 0.85 mM $Ca(NO_3)_2$, 0.35 mM KCl, pH 7.0) and fed *Chlamydomonas reinhardtii* and *Klebsiella pneumoniae* according to previously published methods (Khurana et al. 2014). Matings were performed by starving the compatible parental mating types 310 and 510, mixing the mating types, and diluting to a concentration of about 5,000 cells per milliliter in Pringsheim medium and plating the cells in 10 cm plastic Petri dishes. Matings were assessed several hours after mixing mating types by calculating the percentage of paired cells per total cells.

Reverse transcription PCR (RT-PCR)

Cell cultures or mating time-courses were concentrated by centrifugation and total RNA was extracted using Trizol. Turbo DNase (Thermo Fisher, Waltham, MA, USA) was used to digest DNA before extracting RNA again. Eluted DNA-free total RNA was reverse transcribed using oligo (dT) and AMV reverse transcriptase (New England Biolabs, Ipswich, MA, USA). PCR was performed using cDNA template and Phusion High Fidelity DNA polymerase (New England Biolabs, Ipswich, MA, USA).

Nanochromosome assembly

Pooled F1 cells were sequenced using Illumina sequencing. Short reads were mapped to the germline MIC genome. Reads mapping to g111288 were isolated. Next, we searched for the

5’ and 3’ end of an arbitrary read mapping to g111288 in the other reads. We iterated the process of searching for the 5’ or 3’ end of each read in the remaining reads until we found a read terminating with a telomere repeat (C_4A_4). We manually assembled the sequences of the reads into an g111288 nanochromosome.

In vitro transcription

To prepare long single-stranded RNA (ssRNA) transcripts for microinjection, PCR primers were first designed to use Phusion High-Fidelity DNA polymerase (New England Biolabs, Ipswich, MA, USA) to amplify the coding sequence of the desired TGLO and add a T7 promoter to the gene. The T7-flanked product was cloned using the TOPO TA cloning kit (Thermo Fisher, Waltham, MA, USA) and Sanger sequenced (Genewiz, South Plainfield, NJ, USA) to verify the insert. In vitro transcription was done using the HiScribe T7 High Yield RNA Synthesis Kit according to the manufacturer’s instructions (New England Biolabs, Ipswich, MA, USA).

RNA injection

In vitro transcribed RNA was extracted using Trizol and resuspended to a concentration of 3 micrograms per microliter. ssRNA was microinjected into mating cells at 12 hours post-mixing according to previously published protocols (Fang et al. 2012). Post-injected cells were allowed to recover in Volvic water for two days before picking single cells and plating them in Volvic to establish clonal lines.

5’ rapid amplification of cDNA ends (5’ RACE)

We used a published 5’ RACE protocol (Scotto-Lavino et al. 2006) with minor changes. Briefly, total RNA was extracted in Trizol (Thermo Fisher, Waltham, MA, USA) and treated with Turbo DNase (Ambion). One microgram of DNase-treated total RNA was reverse transcribed using AMV reverse transcriptase (New England Biolabs, Ipswich, MA, USA) and a gene-specific primer for either the germline-limited gene or actin II control. cDNA was poly(A) tailed using terminal transferase (New England Biolabs, Ipswich, MA, USA). The A-tailed cDNA was amplified using two rounds of PCR amplification using Phusion High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA, USA). The first round of amplification was done over 15 cycles, the first round product was diluted 1:1000, the diluted first round product was amplified over 35 cycles, and the products from the second round of amplification were resolved on an agarose gel and stained with ethidium bromide (Bio-Rad, Hercules, CA).

RT-qPCR

As we did previously, we reverse transcribed total RNA from two different times during the organism's life cycle using random hexamer primers. This cDNA was used as template in a series of RT-qPCR experiments to detect the expression of either germline-limited ORF candidate or actin. We used Power Sybr Green qPCR Master Mix (Thermo Fisher, Waltham, MA, USA) and custom qPCR primers (Integrated DNA Technologies, Coralville, IA, USA) and performed the reaction using a CFX384 Touch Real-Time PCR Detection System (Bio-Rad, Hercules, CA, USA). We analyzed the Cq values using a standard curve method and compared the number of transcripts in each sample to the number of small subunit mitochondrial rRNA.

Southern hybridization

1 µg of genomic DNA was resolved on a 1% agarose gel, and ethidium bromide was used for visualization. MAC DNA was purified according to previously published methods (Swart et al. 2013). Dilute PCR products were used as a control to approximate the expected copy number in the genomic DNA lanes. The 1 Kb Plus DNA ladder (Thermo Fisher, Waltham, MA, USA) was used as a size standard. After gel electrophoresis, DNA was blotted onto a nylon membrane, detected using a digoxigenin-labeled DNA probe, and detected using chemiluminescence according to a previously published protocol (Yerlici et al. 2019)

Primers

The following primers were synthesized by Integrated DNA Technologies (Coralville, IA, USA) for use in this study.

g104149 retention fwd: 5'-CGATGATGATGCAGAGCAGTGGAGGCTTAG-3'

g104149 retention rev: 5'-CATATCGTGTTCATTCATGTAAGATAACTACTGCTTG-3'

g67186 retention fwd: 5'-CAATTCACATAATCCTCTATTTCTGCAACTTTTTCTAGAC-3'

g67186 retention rev: 5'-GAATTATTTGTAAATACTTGACTGACTCATTGTTGATAAAATGATTTAC-3'

QT RACE: 5'-CCAGTGAGCAGAGTGACGAGGACTCGAGCTCAAGC-3' (Scotto-Lavino 2006)

QO RACE: 5'-CCAGTGAGCAGAGTGACG-3' (Scotto-Lavino 2006)

QI RACE: 5'-GAGGACTCGAGCTCAAGC-3' (Scotto-Lavino 2006)

Actin II RT: 5'-GTGGTGAAGTTATATCCTCTCTTGGCCAATAATG-3'

Actin II GSP 1: 5'-TGGCATGAGGAATTGCGTAACCTTCATAGA-3'

Actin II GSP 2: 5'-TCCATCTCCAGAGTCAAGCACAACACC-3' g104149 RT: 5’-TTGGGTAAATTCTGGCCAACTCCCTTG-3’

g104149 GSP 1: 5’-CCAAGCTTCTCTGCACCTCATCCGTGAACA-3’

g104149 GSP 2: 5’-GTCTGCCCATCCACGATTTCACTGACC-3’

g67186 RT: 5’-AGCCTTGGTCCCTTCTGAGGCAG-3’

g67186 GSP 1: 5’-CCTGGCAAGAGCAACTTGACAGCAC-3’

g67186 GSP 2: 5’-GAGAGGCCAGAGGCTTCATTGCATACC-3’

g104149 gene qPCR fwd: 5’-CCAAGCTTCTCTGCACCTCATCCGTGAACA-3’

g104149 gene qPCR rev: 5’-AAGGTCAGTGAAATCGTGGATGGGCAGACT-3’

g67186 gene qPCR fwd: 5’-TGCAATGAAGCCTCTGGCCTCTCA-3’

g67186 gene qPCR rev: 5’-CCTGGCAAGAGCAACTTGACAGCAC-3’

g67186 upstream qPCR fwd: 5’-CAATTCAATAGCACCGAATAGAAAGCTTATTTTATACAAGGATTAG-3’

g67186 upstream qPCR fwd: 5’-CTAGATTTAATTAAAACTTGAAATGTCTACAGCCCATTAATAATTCG-3’

Actin II qPCR fwd: 5’-GGTGTTGTGCTTGACTCTGGAGATGGA-3’

Actin II qPCR rev: 5’-TGGCATGAGGAATTGCGTAACCTTCATAGA-3’

Mitochondrial 23S rDNA qPCR fwd: 5’-GATAGGGACCGAACTGTCTCACG-3’ (Nowacki et al. 2009)

Mitochondrial 23S rDNA qPCR rev: 5’-CATATCCTGGTTGTGAATAATCTTCCAAGGG-3’ (Nowacki et al. 2009)

Telomere primer 1: 5’-ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTCCCCAAAACCCCAAAACCCCAAAA -3’ (Nowacki et al. 2008)

Telomere primer 2: 5’-ACTATAGGGCACGCGTGGT-3’ (Nowacki et al. 2008) g43073 TSP 1: 5’-GCCAGGTAGTTGCAAGCGCTCTCGAGAG-3’

g43073 TSP 2: 5’-GCTCAAAGTTTTAACTACTTGATTGAAGTGTAGATTTGGCAATC-3’

g104149 TSP 1: 5’-GTAAATTCTGGCCAACTCCCTTGAGTTCCAAGCTTC-3’

g104149 TSP 2: 5’-CAAAGTCTGCCCATCCACGATTTCACTGACCTTTG-3’

g93797 TSP 1: 5’-GCCCAATTCATATGCTGCTTCTTTGAGCCACTTG-3’

g93797 TSP 2: 5’-GATCTGGTTTTCACAGTTGAGGTAGTAGTAGTAG-3’

g111288 fwd PCR: 5’-CTCTACTCTCTTAGGTCTCCCTCTGCCATT-3’

g111288 rev PCR: 5’-AGCGGCCTGAAACTTTGTAAGGAGTAAGAT-3’

Actin II fwd PCR: 5’-GACTCAAATTATGTTTGAAGTCTTCAATGTACCTTGCC-3’

Actin II rev PCR: 5’-GTGGTGAAGTTATATCCTCTCTTGGCCAATAATG-3’

g111288 nanochromosome gene fwd qPCR: 5’-CAGGCCGCTTTAACTGCAACCATAGTTG-3'

g111288 nanochromosome gene rev qPCR: 5’-GGAAATTGAGCCAACTTTACAGTTAGAGCC-3'

g111288 nanochromosome MDS2 fwd qPCR: 5’-CTTTCCTACAAATCCCCTTAAATTTCCAGTCTTGTAC-3'

g111288 nanochromosome MDS2 rev qPCR: 5’-GTACCATGCTAGGATGTTATTGAAATCATAGAAGAC-3'

g111288 nanochromosome MDS4 fwd qPCR: 5’-CGTCAAATTCAGTAACTAGCTCAGGTACGTC-3'

g111288 nanochromosome MDS4 rev qPCR: 5’-CTACCCTCCCGAGGAAAATACCTGG-3'

g111288 nanochromosome MDS7 fwd qPCR: 5’-CTGAAATGGCTGTATCTATGGTTATTATAAAGAATTAGTG-3' g111288 nanochromosome MDS7 rev qPCR: 5’-CAATCATCACTCTCCCTAACCGTACCTC-3'

g111288 nanochromosome IES6 fwd qPCR: 5’-GGGAAGTTATTTTATTATGAGTTTAGGTTGCATTCATTC-3'

g111288 nanochromosome IES6 rev qPCR: 5’-GAATGAAAATGAGTGAATTAAGAATTTTAATGAAGTATGATATAACATTC-3'

Bioinformatic analyses

Short read DNA sequences were locally aligned to reference sequences using Bowtie 2 (Langmead and Salzberg 2012) or BWA-MEM. Short read RNA sequences were aligned to reference sequences using HISAT2 (Kim et al. 2019). Sanger sequencing DNA reads were aligned to reference sequences using the Geneious aligner in the Geneious software package (version 5.9) (Biomatters, Ltd., Auckland, New Zealand) with default parameters.

Results

Thousands of transcribed germline-limited open reading frames (TGLOs) are expressed during development.

We examined potential germline-limited coding sequences in the *Oxytricha trifallax* MIC genome by searching for transcribed germline-limited open reading frames, which we refer to as TGLOs. Here we adapted a computational pipeline originally used to identify 810 germline-limited protein coding genes expressed during *Oxytricha trifallax* development (Figure 1A, left) (Chen et al. 2014) to expand and update this set. First, we used Augustus gene prediction (Stanke et al. 2006) and RNA sequencing hints from throughout the organism's life cycle to predict 217,805 potential coding sequences in the germline genome. To exclude potential coding sequences that are present in the somatic MAC genome or are not transcribed at significant levels, we restrict TGLOs to computationally predicted ORFs with virtually no DNA sequencing coverage in the MAC genome of both parental strains. Another requirement is that they have RNA expression in at least one time-point during the organism's life cycle. To set read mapping thresholds appropriate for the variable sequencing depth of individual RNA and DNA libraries, we used a Monte Carlo approach in which the predicted 217,805 candidate loci were randomly shuffled 100 times throughout the germline-limited portion of the MIC genome, while recording the distribution of the number of DNA and RNA reads mapped to the random loci. The distributions of DNA or RNA reads mapped to randomly shuffled TGLO loci were treated as the background germline-limited coverage. We required that TGLOs have a number of DNA sequencing reads mapping to them from either parent or the F1 progeny that is no greater than the fifth percentile from the background germline-limited coverage simulation (i.e. no reads mapped per TGLO). On the other hand, highly expressed TGLOs should have RNA sequencing coverage equal to at least the 95th percentile from the random distribution (i.e. four reads mapped per TGLO). We also used a lower RNA sequencing threshold (i.e. a minimum of two reads mapped per TGLO) because at least one experimentally confirmed TGLO was not present in the high transcription TGLO dataset. CD-HIT (Fu et al. 2012) and RepeatMasker (Smit et al. 2013) were used to cluster similar sequences and to remove sequences associated with repetitive elements. The final mutually exclusive datasets contained 4342 and 6296 TGLOs with high and low transcription levels, respectively (Figure 1A, center). Like the previously reported germline-limited gene dataset, TGLOs tend to be intron-poor, with 8.8% and 6.4% of high and low transcription TGLOs, respectively, containing introns compared to 64.9% of MAC encoded genes. These datasets update our previous estimates and contain 279 (213 and 66, respectively, in the high and low transcription sets) of the 810 germline-limited genes predicted in Chen et al. (2014) (Figure 1A, right) (Chen et al. 2014), with some of the reduction attributed to strain-specific differences described below.

We investigated the presence of conserved domains and predicted gene functions using the functional annotation tool eggNOG-mapper (version 2) (Huerta-Cepas et al. 2017). 111 high transcription and 245 low transcription TGLOs mapped to conserved eggNOG orthology clusters (version 5.0) (Figure 1B) (Huerta-Cepas et al. 2019). 54 of these 356 TGLOs corresponded to previously-predicted germline-limited genes (Chen et al. 2014) (42 high and 12 low transcription TGLOs, respectively). Predicted functions and conserved domains included several potentially involved in DNA rearrangement and epigenetic regulation, including MT-A70, miRNA methylation, DNA helicase, PHD zinc finger, and high mobility group.

Protein expression of TGLOs could also suggest a function role for a subset of predicted coding sequences. One quarter (26%) of the original 810 germline-limited genes had peptides identified in a nuclear proteome extracted from mid-rearrangement cells at a single timepoint (Chen et al. 2014), and we queried the new TGLO datasets against this previously published peptide dataset. 144 high and 48 low transcription TGLOs (101 and 42 newly discovered, respectively) were present in this limited 40 hour proteomic survey. Several peptides from the developmental survey were also mapped to TGLOs with eggNOG functional predictions (Figure 1B, blue text).

The previously published set of germline-limited genes was limited to developmental gene expression, with most germline-limited genes transcribed beginning 40 hours after mixing of parental cells (Chen et al. 2014). We assessed the transcription profiles of TGLOs throughout the organism's developmental life cycle using a deeply sequenced set of developmental RNA sequencing libraries. Two partially overlapping triplicate RNA sequencing time-courses across post-zygotic development showed that RNA expression from both high (Figure 1C) and low transcription TGLOs also clustered toward the later stages of rearrangement. Conversely, a random sample of one thousand somatic MAC-encoded genes had a diverse set of RNA expression profiles during the same time-course, suggesting that TGLOs are enriched in developmental expression.

TGLO genes are eliminated after gene expression.

By definition, TGLO DNA sequences are restricted to the germline MIC. Since the germline genome is diploid, two copies of each TGLO locus are present in each micronucleus. Since DNA copy number changes significantly throughout MAC development (Spear and Lauth 1976), we studied DNA copy number changes and elimination of TGLOs during development. A preliminary copy-number study indicated that most TGLOs are eliminated by the end of the developmental life cycle, with most of the rows corresponding to TGLO loci trending toward blue in the late time-course DNA heatmap data (Figure 2A). However, there is notable heterogeneity but the DNA copy number profiles of TGLOs, with some showing very little copy number variation throughout development, leaving it unclear whether the loci are eliminated from the developing somatic MAC by the end of the sexual life cycle (Figure 2A). Two clusters of high transcription TGLOs have either relatively high DNA copy numbers from early rearrangement (24 hr) until the end of the time-course (Figure 2A top left), and one cluster of low transcription TGLOs also exhibits a similar effect (Figure 2A bottom left).

Since we previously reported that telomeres are permissive to transcription in *O. trifallax*, unlike in other lineages (Beh et al. 2019), we amplified several TGLO loci via telomere suppression PCR (Chang et al. 2004) to determine whether telomeres are added upstream of these loci before DNA elimination. We found that three out of six sampled TGLOs—representing both high and low transcription TGLOs—had telomeres added near the ORF during mid to late development and before their elimination from the developing somatic MAC (Figure 2B), consistent with their transcriptional pattern.

Strain-specific germline-limited ORFs

Our studies uncovered one case of a germline-encoded ORF that was also present at a low copy level in the somatic MAC of one parent. The protein coding locus, OXYTRIMIC_220 (“g111288”), was included in the previously reported set of 810 MIC-limited genes, but it does not encode any conserved functional domains nor was it detected in a developmental mass spectrometry survey (Chen et al. 2014). The initial Augustus gene prediction identified this ORF. However, it was later excluded from the pipeline after incorporating new DNA sequencing libraries from the parent strains and F1 progeny, which suggested that g111288 is present in the somatic MAC of at least one parental strain.

We used PCR to amplify g111288 from parental genomic DNA to test whether the locus is present in the somatic genome of either parent strain. We found that the coding sequence was abundant in strain JRB510, which was not the reference strain used for genome sequencing (Swart et al. 2013; Chen et al. 2014). In addition, we found that several cell lines derived from either single F1 progeny or genetically manipulated F1 lines also contained g111288 at detectable DNA copy levels (Figure 3A). In addition, the g111288 locus varied in DNA copy level in individual F1 lines derived from different parental crosses.

Since g111288 appeared to be present in the MAC genome of only parental strain, JRB510, and germline limited in the original reference strain JRB310, we investigated the nature of the putative g111288 somatic MAC nanochromosome. Next generation sequencing reads from a pool of F1 progeny cells were mapped to the germline MIC genome. This allowed assembly of an entire g111288 nanochromosome with telomeres at both ends and indicated that it derives from seven MDSs on a MIC chromosome with the g111288 open reading frame entirely contained within the first MDS (Figure 3B). RNA sequencing from developmental time-points confirmed that g111288 is transcribed from 40 to 60 hr after mixing of both parental strains (Figure 3C). In addition, alignment of RNA-seq reads to the other six MDSs on the g111288 nanochromosome suggested the possibility that the other six MDSs of the g111288 nanochromosome could have coding potential. To assess the nanochromosome's relative copy number in different cell lines, we performed qPCR to target different amplicons across the g111288 nanochromosome using template genomic DNA from parental cells and F1 progeny lines. A two order of magnitude copy number increase was consistently observed in the JRB510 parent line relative to the reference JRB310 strain (Figure 3D). Moreover, three F1 lines displayed copy levels somewhat higher than the JRB510 parental strain, and the other two F1 lines appeared to have few to no copies of the nanochromosome, like strain JRB310. Southern hybridization with a probe targeting a MAC-specific MDS-MDS junction region confirmed the presence of the nanochromosome in MAC DNA from parental strain JRB510 as well as two F1 cell lines (SLC89 and SLC92; Seegmiller et al. 1996) (Figure 3E).

Since g111288 is present in the somatic genome of several F1 lines and at a low level in one parent, we assessed whether the coding sequence is transcribed during asexual (vegetative) growth. However, we did not detect any transcripts from this locus outside the middle and late stages of developmental, corresponding to approximately 48 hours after mixing of mating-compatible cells (Figure 3F). Swart et al. (2013) previously reported that many other MAC nanochromosomes have developmental-specific expression (Swart et al. 2013), suggesting that g111288 is a strain-specific nanochromosome, retained only in the MAC genome of JRB510 and passed on to its F1 progeny.

Few ncRNAs map to TGLO loci

Oxytricha's genome rearrangements and DNA deletion are regulated by noncoding RNAs (ncRNAs). For example, Otiwi-1-bound piRNAs map to retained MDSs but not germline-limited regions or IESs (Fang et al. 2012), and long template RNAs map to nanochromosomes in the MAC genome (Lindblad et al. 2017). Hence, we mapped template RNAs and Otiwi-1-associated piRNAs to the MIC genome and assessed their coverage in TGLO loci and the g111288 locus. We found that Otiwi-1 piRNAs map to MDSs more heavily than TGLOs (Figure 4A). Otiwi-1 piRNAs aligned to g111288, which is retained at a low somatic copy level in one parent (Figure 4B), but piRNAs are present at a reduced level compared to neighboring MDSs. Template RNA coverage was also significantly higher in MDSs compared to TGLOs (Figure 4C), although the strain-specific TGLO g111288 lacked any template RNAs despite being encoded by the JRB510 MAC (Figure 4D).

Synthetic RNA injection can protect TGLO loci from genomic deletion

g111288 presents an example of a potential coding sequence that is present in the somatic MAC of one strain while eliminated as a TGLO in another strain. We decided to test whether exposure to artificial RNAs during development could reprogram the germline-limited status of TGLOs, thereby retaining them on MAC nanochromosomes. Given our previous observations that exposure to non-coding RNAs can reprogram IES retention in the MAC (Fang et al. 2012; Khurana et al. 2018 RNA; Beh et al. 2019) we used RNA injection to test whether exposure to targeting RNA could reprogram the retention of two TGLO loci during development (Figure 5A). We targeted two TGLO loci that are encoded in the IESs of other MAC loci. One of the two candidates, g67186, was previously predicted to encode a histone 2B gene (Chen et al. 2014), while the other, g104149, did not contain any predicted conserved domains. The two candidates are also among the highest expressed TGLOs that mapped within IESs, facilitating our strategy (Figure 5A). Importantly, we also observed that our candidate TGLOs lacked Otiwi-1 piRNAs and template RNAs during the sexual life cycle (Figure 5B), suggesting that the cell does not endogenously encode their somatic retention during the sexual life cycle.

PCR from cell cultures derived from single injected cells, followed by Sanger sequencing indicated that RNA injection did reprogram IES+TGLO retention in some progeny, with varying levels of retention based on differences in PCR band sizes. Some products contained small deletions in the retained sequence relative to the reference MIC locus, but no deletions affected the ORF (Figure 5C and 5D). No F1 lines from uninjected WT parental cells contained the TGLO sequences, suggesting that RNA injection specifically programs the somatic DNA retention (Figure 5E right and Figure 5F right).

RNA programmed IES retention was previously shown to be heritable after subsequent sexual cycles, so we also tested whether the IES+TGLO insertions were retained after backcrossing to a parental strain. PCR amplification from genomic DNA of backcrossed pools of cells indicated that the retained TGLO g104149 was partially heritable for at least two more generations (Figure 5E, left). The other retained TGLO, g67186, was partially heritable for one backcrossed generation (Figure 5F, left). A second band corresponding to the wild-type product was present in both backcrosses, consistent with the presence of WT nanochromosomes in the backcrosses to the wild-type parental strain.

Retained TGLOs are transcribed outside usual developmental program

Our engineered strains that retain TGLO loci are unique in their ability to encode previously eliminated germline sequences in their macronucleus. Genome-wide transcription start site profiling in asexually growing *O. trifallax* cells showed that transcription initiation typically occurs in the subtelomeric sequence of somatic nanochromosomes that encode a single gene, and this is usually within approximately one hundred bases of the transcribed coding sequence (Beh et al. 2019). Since the retained TGLO reading frames are nested within the protein coding sequences of a flanking gene, but also retain their own putative upstream and downstream regulatory sequences, we assessed the expression of retained TGLOs. We collected total RNA from asexually growing cells with the retained TGLO (line 4 for g104149 and lines 3 and 5 for g67186), as well as WT parental lines, and from a WT developmental time-course when TGLOs are normally transcribed, and amplified cDNA ends using 5' RACE (Figure 6A). We found that retained TGLO loci were now transcribed during both the asexual life cycle as well as at their normal developmental pattern (Figure 6B bottom and Figure 6C bottom). The size of the RACE products were similar for the retained lines as well as during normal developmental expression, suggesting transcription initiation in the vicinity of the endogenous TSS , albeit with more variation for g104149.

Given the structural differences between the somatic MAC nanochromosome in asexually growing cells vs. the differentiating MAC during the sexual life cycle, the transcriptional environment of the two nuclei could differ greatly. We used qRT-PCR to compare the transcription levels of retained TGLO loci during the asexual life cycle vs. WT TGLO expression during development, and found that the transcription level of retained TGLOs is approximately an order of magnitude higher during the WT developmental timepoint compared to artificial expression during the asexual life cycle in retained lines (Figures 6D and 6E).

Discussion

Here, we introduce the definition of TGLO as a transcribed germline-limited DNA sequence with the ability to encode a putative protein. We show that the *O. trifallax* germline MIC genome contains abundant TGLOs that are transcribed to varying degrees in WT cells during development, and are then eliminated from the somatic MAC. This suggests that TGLO gene expression may be regulated by DNA elimination. The presence of conserved domains and predicted functions found in TGLO datasets also support this hypothesis. The TGLOs that lack conserved domains could be enriched in lineage-specific or orphan functions, because eggNOG-mapper is a strict annotation tool that recognizes orthologous domains if they are conserved across many species. It is also possible that the TGLO set may contain some ncRNAs that survived poly(A) selection due to the AT-richness of the genome. Moreover, as ciliates have heterochromatic MIC genomes that are not amenable to transcription (Gorovsky and Woodard 1969) and may lack RNA polymerase II (Khurana et al. 2014), it is possible that they evolved mechanisms to shut down developmental gene expression via programmed DNA elimination.

The earlier report of 810 germline-limited genes in *O. trifallax* assumed that germline-limited coding sequences would be deleted before the cell returned to the asexual life cycle (Chen et al. 2014). Here we present evidence instead that the timing of DNA elimination of TGLOs is heterogeneous during the sexual life cycle. Furthermore, we note the transient addition of *de novo* telomeres in unexpected locations accompanying TGLO transcription, a step that might activate them for transcription. Conceptually similar, in a related ciliate *Euplotes crassus*, DNA processing during the sexual life cycle is responsible for modulating the transcription of one of three telomerase catalytic subunit genes (Karamysheva et al. 2003). Finally, our DNA sequencing results suggest that most TGLOs are indeed eliminated from the somatic MAC by the end of the sexual life cycle. However, we cannot exclude the possibility that a subset of TGLOs persist longer, as further research into later developmental time-points could reveal.

We also observed that at least one germline-encoded ORF, g111288, is actually present at a low somatic copy level in one parental cell line. Unlike TGLOs, g111288 is variably retained as a high copy nanochromosome in some F1 progeny. Presumably, the presence of ncRNAs derived from one parent can program retention of the chromosome in F1 cells, as well as copy number variation in the progeny (Nowacki et al. 2010), but the incomplete penetrance of somatic g111288 heritability correlates with its low somatic copy number in the JRB510 cell line. Curiously, g111288 does not appear to be transcribed from the somatic MAC in either the parent nor F1 progeny. This is unexpected because the entire coding sequence is present on its own nanochromsome along with its putative upstream and downstream regulatory sequences. However, it is possible that its gene expression requires other *trans*-acting regulatory factors specific to the developmental life cycle.

The case of g111288 is also noteworthy because it appears capable of being either germline-restricted or somatic-encoded. At the level of smaller MDS or IES regions, flexibility between being retained vs. deleted has been observed before but on the timescale of either generations in the laboratory (Fang et al. 2012) or intra-species variation in the wild (Mollenbeck et al. 2006) rather than an intraspecies difference (Vitali et al. 2019). This feature itself could contribute to the birth of new genes, since new coding sequences can sometimes arise from retained noncoding sequences if transcribed and functional (Neme and Tautz 2016; Neme et al. 2017). A previous study in *Tetrahymena* reported that a set of developmentally transcribed somatic minichromosomes are gradually eliminated from the MAC after genome rearrangement (Lin et al. 2016). Moreover, a specific minichromosome in one *Tetrahymena* species might be germline-limited in another species. This *Tetrahymena* example and our functional experiments that reprogram somatic TGLO retention in *O. trifallax* suggest that TGLOs might be a reservoir of sequences with somatic coding potential. We can envision an evolutionary model by which germline-encoded sequences can gain access to the somatic genome where they would be expressed. A deeper intraspecies survey of MAC and MIC genomes, together with developmental RNAseq to survey expression, would be needed to test this hypothesis.

Our ability to program the somatic retention of specific TGLOs via ncRNA injection is a unique feature of the present study. This had the ability to unmask gene expression of targeted TGLOs outside their normal developmental program. *Tetrahymena thermophila* also has non-maintained chromosomes that are lost soon after expression during development and can be fused to adjacent regions to program their retention in the somatic MAC (Feng et al. 2017). Here we have extended this general phenomenon to *Oxytricha* and showed that somatic retention subverts the cell's endogenous transcription of the gene locus. This supports the hypothesis that TGLO elimination represses their gene expression. In our example, the misexpression of a single TGLO locus did not affect cell viability, but the ensemble of loci may need to be silenced during asexual growth.

Data availability

All cell stocks are available upon request. Illumina sequencing datasets were uploaded to the NCBI Short Read Archive under the BioProject PRJNA665991. The authors affirm that all data necessary for confirming the conclusions of the article are present within the manuscript and figures.

Acknowledgements

The authors thank Virginia A. Zakian, Gertrud M. Schüpbach, Samuel Sternberg, and the current and past members of the Landweber lab for their helpful feedback throughout the study. We are grateful to Xiao Chen for bioinformatic help with selecting TGLO candidates. We would also like to thank Wei Wang and Jessica Buckles at the Princeton University Genomics Core Facility for their high-throughput sequencing expertise.

Funding

This work was funded by NIH R35GM122555 and NSF DMS1764366 to LFL.

References

Allen, S. E., Hug, I., Pabian, S., Rzeszutek, I., Hoehener, C., and Nowacki, M. (2017). Circular concatemers of ultra-short DNA segments produce regulatory RNAs. 168: 990-999.

Allen, S. E. and Nowacki, M. (2020). Roles of noncoding RNAs in ciliate genome architecture. J. Mol. Biol. S0022-2836: 30026-30027.

Beh, L. Y., Debelouchina, G. T., Clay, D. M., Thompson, R. E., Lindblad, K. A., Hutton, E. R., Bracht, J. R., Sebra, R. P., Muir, T. W., and Landweber, L. F. (2019). Identification of a DNA N6-adenine methyltransferase complex and its impact on chromatin organization. Cell 177: 1781-1796.

Bracht, J. R., Wang, X., Shetty, K., Chen, X., Uttarotai, G. J., Callihan, E. C., McCloud S. S., Clay, D. M., Wang, J., Nowacki, M., and Landweber, L. F. (2017). Chromosome fusions triggered by noncoding RNA. RNA Biol. 14: 620-631.

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. and Evol. 33: 2337-2344.

Burns, J., Kukushkin, D., Lindblad, K., Chen, X., Jonoska, N., and Landweber, L. F. (2016). : a database of ciliate genome rearrangements. Nucleic Acids Res. 44: D703-D709.

Chang, W. J., Stover, N. A., Addis, V. M., and Landweber, L. F. (2004). A micronuclear locus containing three protein-coding genes remains linked during macronuclear development in the spirotrichous ciliate Holosticha. Protist 155: 245-255.

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., Sebra, R. P., and Landweber, L. F. (2014).

The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell 158: 1187-1198.

Chen, X. and Landweber, L. F. (2016). Phylogenomic analysis reveals genome-wide purifying selection on TBE transposons in the ciliate *Oxytricha*. Mob. DNA 7: 2.

Clay, D. M., Yerlici, V. T., Villano, D. J., and Landweber, L. F. (2019). Programmed Chromosome Deletion in the Ciliate *Oxytricha trifallax*. G3 9: 3105-3118.

Curtis, E. A. and Landweber, L. F. (2006). Evolution of gene scrambling in ciliate micronuclear genes. Ann. N. Y. Acad. Sci. 870: 349-350.

Fang, W., Wang, X., Bracht, J. R., Nowacki, M., and Landweber, L. F. (2012). Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement. Cell 151: 1243–1255.

Feng, L., Wang, G., Hamilton, E. P., Xiong, J., Yan, G., Chen, K., Chen, X., Dui, W., Plemens, A., Khadr, L., et al. (2017). A germline-limited piggyBac transposase gene is required for precise excision in *Tetrahymena* genome rearrangement. Nucleic Acids Res. 45: 9481–9502.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28: 3150-3152.

Gorovsky, M. A. and Woodard, J. (1969). Studies on nuclear structure and function in *Tetrahymena pyriformis*: I. RNA synthesis in macro-and micronuclei. The Journal of cell biology, 42: 673-682.

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 *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. eLife 5: e19090.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol. Biol. Evol. 34: 2115–2122.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47: D309–D314.

Itoh, Y., Kampf, K., Pigozzi, M. I., and Arnold, A. P. (2009). Molecular cloning and characterization of the germline-restricted chromosome sequence in the zebra finch. Chromosoma 118: 527-536.

Karamysheva, Z., Wang, L., Shrode, T., Bednenko, J., Hurley, L. A., and Shippen, D. E. (2003). Developmentally programmed gene elimination in *Euplotes crassus* facilitates a switch in the telomerase catalytic subunit. Cell, 113: 565-576.

Khurana, J. S., Clay, D. M., Moreira, S., Wang, X., and Landweber, L. F. (2018). Small RNA-mediated regulation of DNA dosage in the ciliate *Oxytricha*. RNA 24: 18-29.

Khurana, J. S., Wang, X., Chen, X, Perlman, D. H., and Landweber, L. F. (2014). Transcription-independent functions of an RNA polymerase II subunit, Rpb2, during genome rearrangement in the ciliate, *Oxytricha trifallax*. Genetics 197: 839-849.

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37: 907–915.

Kinsella, C. M., Ruiz-Ruano, F. J., Dion-Côté, A. M., Charles, A. J., Gossmann, T. I., Cabrero, J., Kappei, D., Hemmings, N., Simons, M. J. P., Camacho, J. P. M., et al. (2019). Programmed DNA elimination of germline development genes in songbirds. Nat. Commun. 10: 5468.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25: 2078-2079.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30: 923-930.

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.

Lindblad, K. A., Bracht, J. R., Williams, A. E., and Landweber, L. F. (2017). Thousands of RNA-cached copies of whole chromosomes are present in the ciliate *Oxytricha* during development. RNA 23: 1200-1208.

Lindblad, K. A., Pathmanathan, J. S., Moreira, S., Bracht, J. R., Sebra, R. P., Hutton, E. R., and Landweber, L. F. (2019). Capture of complete ciliate chromosomes in single sequencing reads reveals widespread chromosome isoforms. BMC Genom. 20: 1037.

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.

Meyer, G. F. and Lipps, H. J. (1981). The formation of polytene chromosomes during macronuclear development of the hypotrichous ciliate *Stylonychia mytilus*. Chromosoma, 82: 309-314.

Möllenbeck, M., Cavalcanti, A. R., Jönsson, F., Lipps, H. J., and Landweber, L. F. (2006). Interconversion of germline-limited and somatic DNA in a scrambled gene. J. Mol. Evol., 63: 69-73.

Neme, R., Amador, C., Yildirim, B., McConnell, E., and Tautz D. (2017). Random sequences are an abundant source of bioactive RNAs or peptides. Nat. Ecol. Evol. 1: 0217.

Neme, R. and Tautz, D. (2016). Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. eLife 5: e09977.

Nowacki, M., Haye, J. E., Fang, W., Vijayan, V., and Landweber, L. F. (2010). RNA-mediated epigenetic regulation of DNA copy number. Proc. Natl. Acad. Sci. U.S.A. 107: 22140-22144.

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: 925-928.

Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T. G., and Landweber, L. F. (2008). RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature 451: 153–158.

Pigozzi, M. I. and Solari, A. J. (1998). Germ cell restriction and regular transmission of an accessory chromosome that mimics a sex body in the zebra finch, *Taeniopygia guttata*. Chromosome Res. 6: 105-113.

Pigozzi, M. I. and Solari, A. J. (2005). The germ-line-restricted chromosome in the zebra finch: recombination in females and elimination in males. Chromosoma 114: 403-409.

Prescott, D. M. (1994). The DNA of ciliated protozoa. Microbiol. Rev. 58: 233–267.

Scotto-Lavino, E., Du, G., and Frohman, M. A. (2006). 5' end cDNA amplification using classic RACE. Nat. Protoc. 1: 2555-2562.

Seegmiller, A., Williams, K. R., Hammersmith, R. L., Doak, T. G., Witherspoon, D., Messick, T., Storjohann, L. L., and Herrick, G. (1996). Internal eliminated sequences interrupting the *Oxytricha* 81 locus: allelic divergence, conservation, conversions, and possible transposon origins. Mol. Biol. Evol. 13: 1351-1362.

Smit, A. F. A., Hubley, R., and Green, P. RepeatMasker Open-4.0. (2013) <http://www.repeatmasker.org>.

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.

Smith, J. J., Baker, C., Eichler, E. E., and Amemiya, C. T. (2012). Genetic consequences of programmed genome rearrangement. Curr. Biol. 22: 1524–1529.

Spear, B. B. and Lauth, M. R. (1976). Polytene chromosomes of *Oxytricha*: biochemical and morphological changes during macronuclear development in a ciliated protozoan. Chromosoma 54: 1-13.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34: W435-W439.

Swart, E. C., Bracht, J. R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J. S., Goldman, A. D., Nowacki, M., Schotanus, K., et al. (2013). The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol. 11: e1001473.

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.

Timoshevskiy, V. A., Lampman, R. T., Hess, J. E., Porter, L. L., Smith, J. J. (2017). Deep ancestry of programmed genome rearrangement in lampreys. Dev. Biol. 429: 31-34.

Torgasheva, A. A., Malinovskaya, L. P., Zadesenets, K. S., Karamysheva, T. V., Kizilova, E. A., Akberdina, E. A., Pristyazhnyuk, I. E., Shnaider, E. P., Volodkina, V. A., Saifitdinova, A. F. et al. (2019). Germline-restricted chromosome (GRC) is widespread among songbirds. Proc. Natl. Acad. Sci. U.S.A. 116: 11845-11850.

Vitali, V., Hagen, R., and Catania, F. (2019). Environmentally induced plasticity of programmed DNA elimination boosts somatic variability in *Paramecium tetraurelia*. Genome Res., 29: 1693-1704.

Wang, J., Gao, S., Mostovoy, Y., Kang, Y., Zagoskin, M., Sun, Y., Zhang, B., White, L. K., Easton, A., Nutman, T. B. et al. (2017). Comparative genome analysis of programmed DNA elimination in nematodes. Genome Res. 27: 2001-2014.

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.

Yerlici, V. T., Lu, M. W., Hoge, C. R., Miller, R. V., Neme, R., Khurana, J. S., Bracht, J. R., and Landweber, L. F. (2019). Programmed genome rearrangements in *Oxytricha* produce transcriptionally active extrachromosomal circular DNA. Nucleic Acids Res. 47: 9741–9760.

Figure legends

Figure 1: Germline-limited ORFs are expressed during *Oxytricha trifallax* genome rearrangement.

(A) Left: Steps for predicting transcribed germline-limited open reading frames (TGLOs) in the *Oxytricha trifallax* germline MIC genome. Center: Total number of computationally predicted candidates remaining after each pipeline step. Right: Total number of previously reported MIC-limited genes (Chen et al. 2014) remaining after each pipeline step. Yellow numbers (also leftmost numbers for the center and right columns) indicate high transcription TGLOs. Red numbers (also rightmost numbers for the center and right columns) indicate low transcription TGLOs. **(B)** EggNOG-mapper-predicted functions and conserved domains in high transcription TGLOs (upper yellow box) and low transcription TGLOs (lower red box). Blue text indicates that the associated TGLOs were validated by detection of peptides at a timepoint during rearrangement (Chen et al. 2014). **(C)** RNA sequencing reads from throughout the *Oxytricha* developmental life cycle were aligned to the MIC genome. Log_2-normalized RNA-seq read counts were calculated for high and low transcription TGLOs, as well as one thousand (randomly selected) somatic MAC-encoded genes across the *Oxytricha trifallax* developmental life cycle (hours labeled post mixing of compatible mating types: 0 hr, JRB310 and JRB510 cells mixed together; 12 hr, MIC meiosis; 18 hr, zygotic MAC formation; 24 hr, early rearrangement; 36 hr, mid-rearrangement; 48 hr, late rearrangement/disappearance of parental MAC; 60 hr, late rearrangement; 72 hr, end of rearrangement). Color scale refers to the log_2-normalized RNA expression. Rows are hierarchically clustered to clearly group TGLOs by their relative RNA expression at each time-point.

Figure 2: DNA from TGLO loci is eliminated from the developing MAC.

(A) Log_2-normalized DNA copy number of high and low transcription TGLOs across the *Oxytricha trifallax* developmental life cycle based on mapping of whole cell DNA sequencing reads to the MIC genome. Color scale refers to the $\log_2$-normalized DNA copy number across each time course. Rows are hierarchically clustered to clearly group TGLOs by their relative DNA copy number at each time-point. **(B)** Nested telomere suppression PCR targeting the upstream telomere addition site of selected TGLOs in genomic DNA samples collected throughout the *Oxytricha trifallax* developmental life cycle. Detected PCR products correspond to a specific TGLO locus with an upstream telomere addition site that is not observed in asexually growing cells, and no assessed telomere-capped TGLOs are detectable after 74 hours.

Figure 3: Parental cells can carry a strain-specific germline-limited ORF.

(A) Top: PCR targeting g111288 or Actin II using genomic DNA from F1 lines, parent lines, and other mutant F1 lines used in this study. Bottom: Genome track showing the approximate location of g111288 PCR primers. Yellow: g111288, light blue: flanking MDSs. **(B)** The germline genome locus containing g111288 with mapped F1 reads from a pool of asexually growing F1 cells. Yellow: g111288, light blue: MDSs, dark blue: assembled g111288 MDSs from pooled F1 reads, gray triangles: observed telomere addition sites. **(C)** The germline genome locus (bottom) containing g111288 (yellow) and strain-specific MDSs (dark blue) with mapped RNA-seq coverage (black) from several time-points during asexual growth (starved or encysted cells) and hours post mixing of mating types during the sexual life cycle. **(D)** Top: Copy number relative to mitochondrial rDNA based on qPCR targeting several amplicons on the g111288 nanochromosome, an IES within the corresponding germline locus, and two unrelated somatic loci. Bottom: The germline genome locus containing g111288 with qPCR primer locations indicated. Yellow: g111288, light blue: MDSs, dark blue: assembled g111288 MDSs from pooled F1 reads, black arrows: qPCR primers. **(E)** Top: Southern blot of parental and F1 MAC DNA detected using an MDS-MDS junction spanning DNA probe. Bottom: MIC genome track showing the portions of MDSs 1 and 2 detected. **(F)** Top: RT-PCR targeting g111288 or Actin II using RNA from the same cell lines as in (A). Bottom: Genome track showing the approximate location of g111288 RT-PCR primers. Yellow: g111288, light blue: MDSs.

Figure 4: TGLO loci have few Otiwi-1 piRNAs and template RNAs.

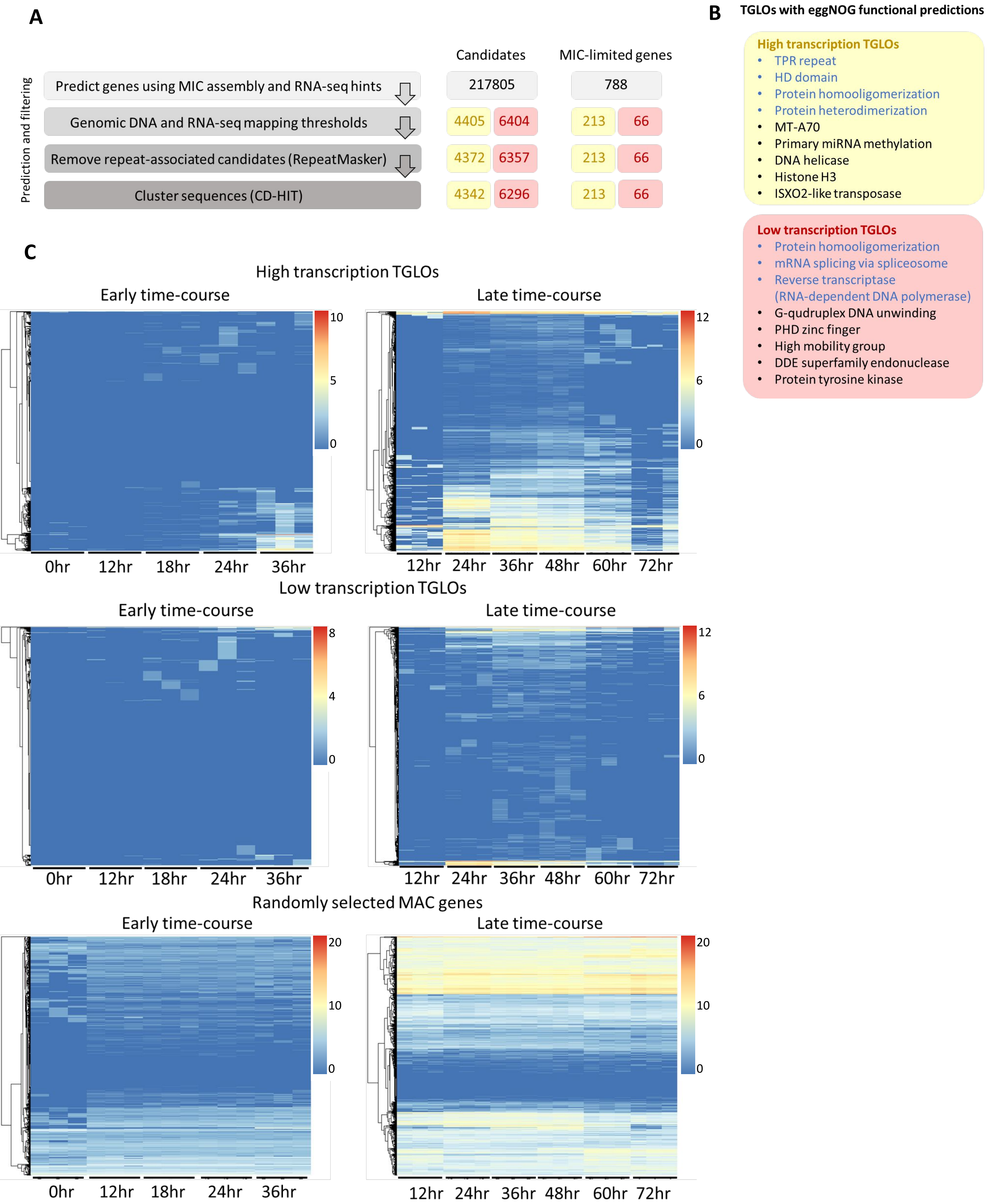
(A) Distribution of normalized mapping quality-filtered Otiwi-1 piRNA read counts (Fang et al. 2012) mapped to High and Low transcription TGLOs compared to MDSs. Read counts were normalized to reads per kilobase million (RPKM). Brackets and asterisks indicate statistically significant differences between distributions. Statistical significance was assessed using the non-parametric Kolmogorov–Smirnov (KS) test, and P<0.05 was considered statistically significant. **(B)** The germline genome locus containing the strain-specific TGLO g111288 (yellow), MDSs (blue), and mapping of Otiwi-1-associated piRNAs (gray) from several time-points during rearrangement onto MDS 1-7 of g111288 and the flanking MDSs for other loci. **(C)** Distribution of normalized mapping quality-filtered template RNA read counts (Lindblad et al. 2017) mapped to High and Low transcription TGLOs compared to MDSs. Read counts were normalized to RPKM. Brackets and asterisks indicate statistically significant differences between distributions. Statistical significance was assessed using the non-parametric KS test, and P<0.05 was considered statistically significant. **(D)** The germline genome locus containing the strain-specific TGLO g111288 (yellow), MDSs (blue), and mapped template RNA coverage (gray) from several time-points during rearrangement.

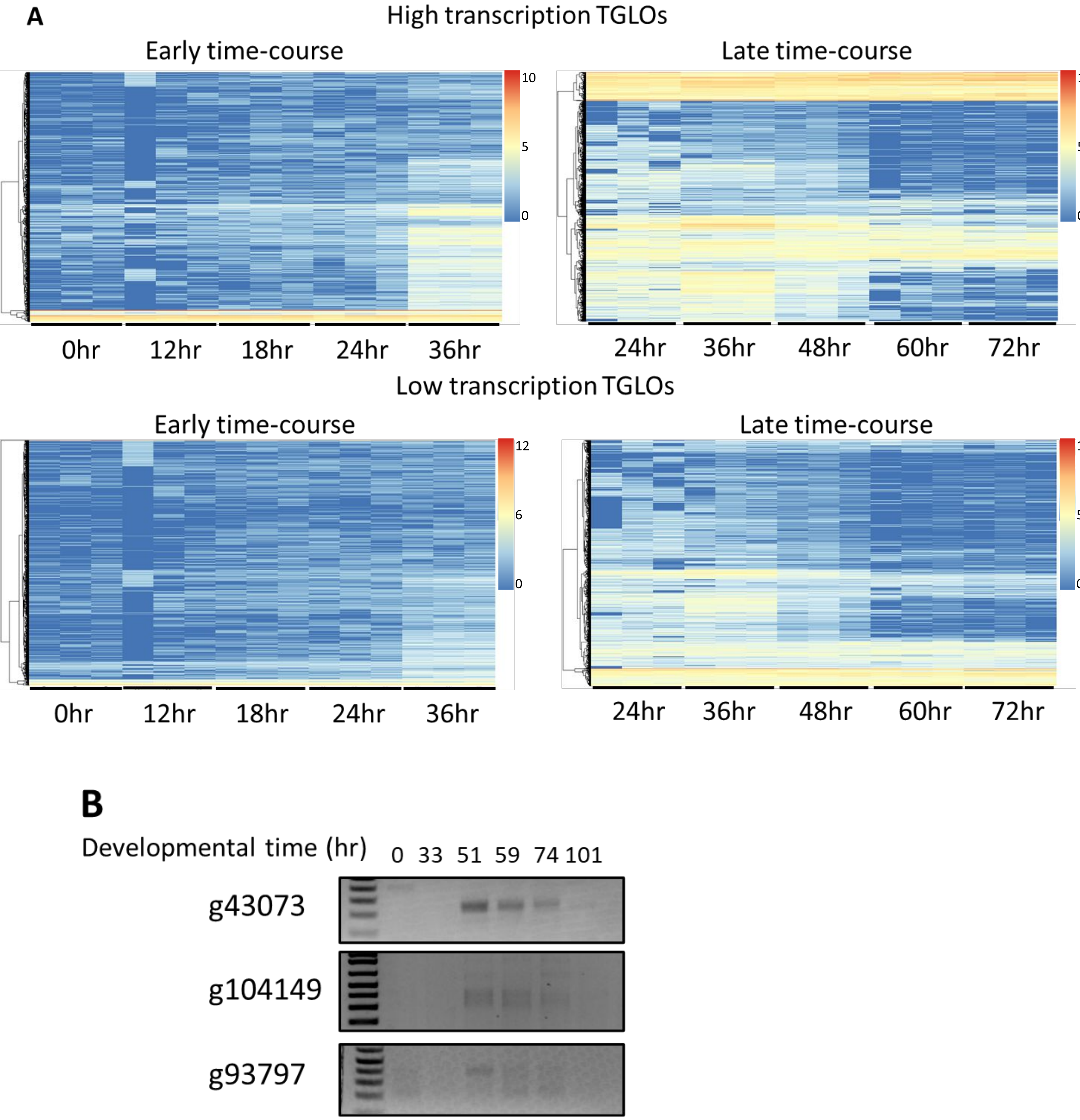
Figure 5: RNA injection programs heritable TGLO retention.

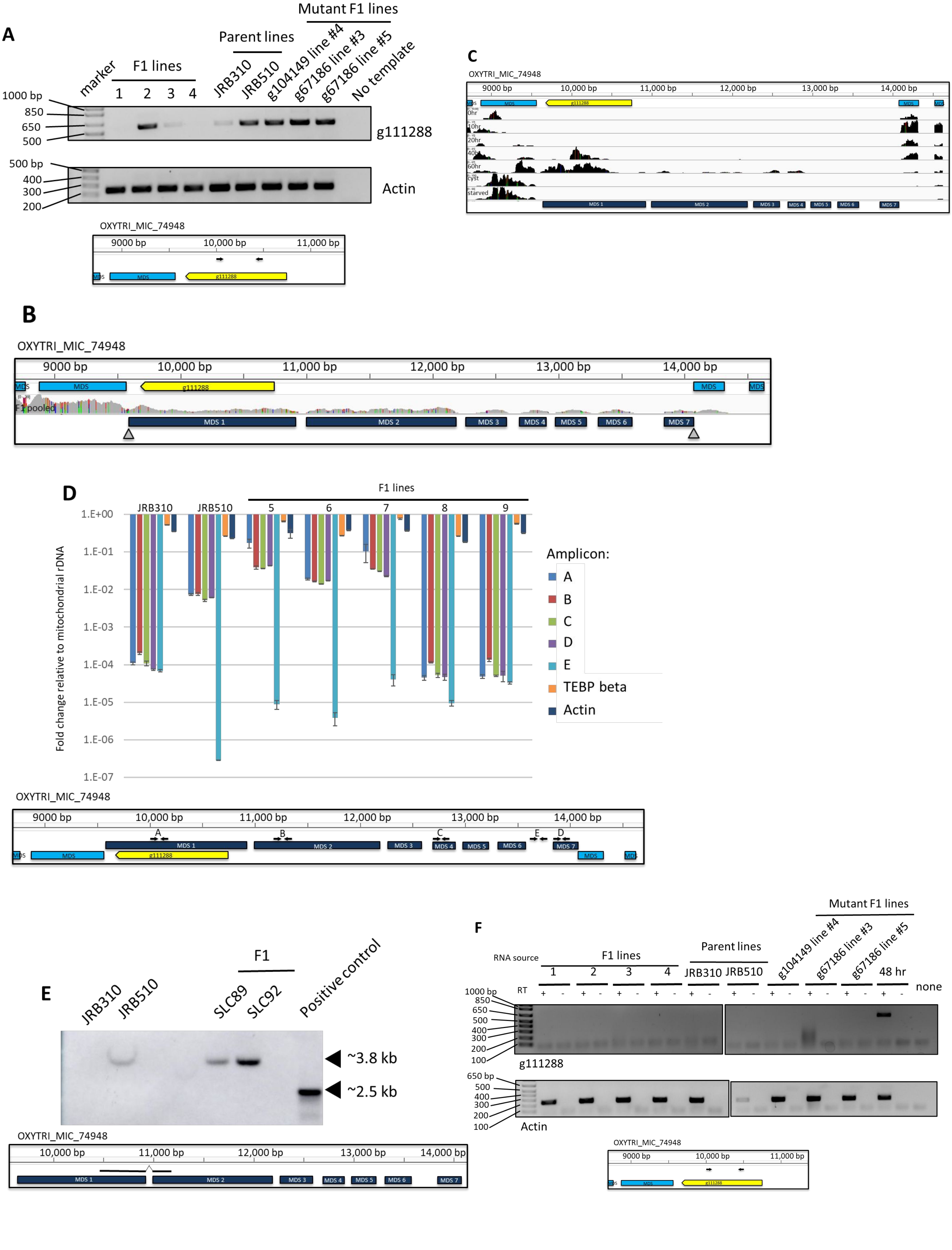
(A) Synthetic RNA injection scheme to program the retention of a TGLO (yellow) in an IES between two MDSs (blue blocks). Possible products can include telomere-capped (black blocks) nanochromosomes with the entire IES plus TGLO flanked by the MDSs of the wild-type flanking locus. **(B)** The germline genome loci containing the programmed retention candidate TGLOs g104149 and g67186 (yellow), MDSs (blue), and mapped piRNA or template RNA coverage (gray) from several time-points during rearrangement. **(C)** Top: Cell culture PCR targeting the IES containing g104149 from cell lines derived from single RNA injected mating pairs. Middle: The expected retention product containing g104149 with PCR primer locations. Yellow: g104149, light blue: MDSs, black arrows: PCR primers. Bottom: Sanger sequencing chromatograms from PCR reactions in (B) aligned to the expected retention product containing g104149 (yellow). **(D)** Top: Cell culture PCR targeting the IES containing the predicted histone 2B TGLO g67186 from cell lines derived from single RNA injected mating pairs. Middle: The expected retention product containing g67186 with PCR primer locations. Yellow: g67186, light blue: MDSs, black arrows: PCR primers. Bottom: Sanger sequencing chromatograms from PCR reactions aligned to the expected retention product containing g67186 (yellow). **(E)** Top: PCR targeting the IES containing g104149 using genomic DNA from parental cells, F1 retention cells, F1 retention cells backcrossed to parental cells, and unmanipulated F1 lines. Bottom: The expected retention product containing g104149 with PCR primers. Yellow: g104149, light blue: MDSs, black arrows: PCR primers. **(F)** Top: PCR targeting the IES containing the predicted histone 2B TGLO g67186 using genomic DNA from parental cells, F1 retention cells, F1 retention cells backcrossed to parental cells, and unmanipulated F1 lines. Bottom: The expected retention product containing g67186 with PCR primers. Yellow: g67186, light blue: MDSs, black arrows: PCR primers.

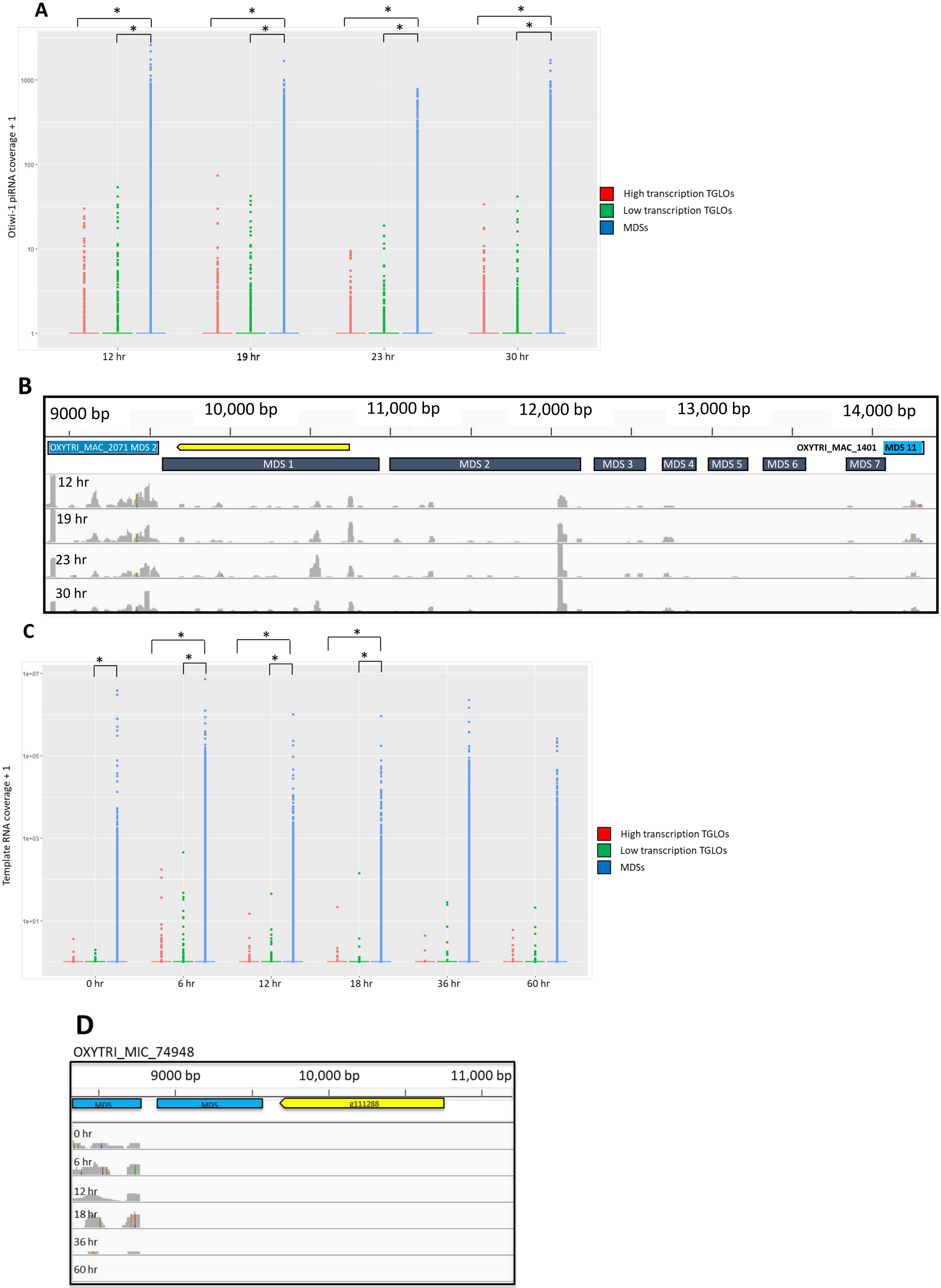
Figure 6: Retained TGLOs are misexpressed during asexual life cycle.

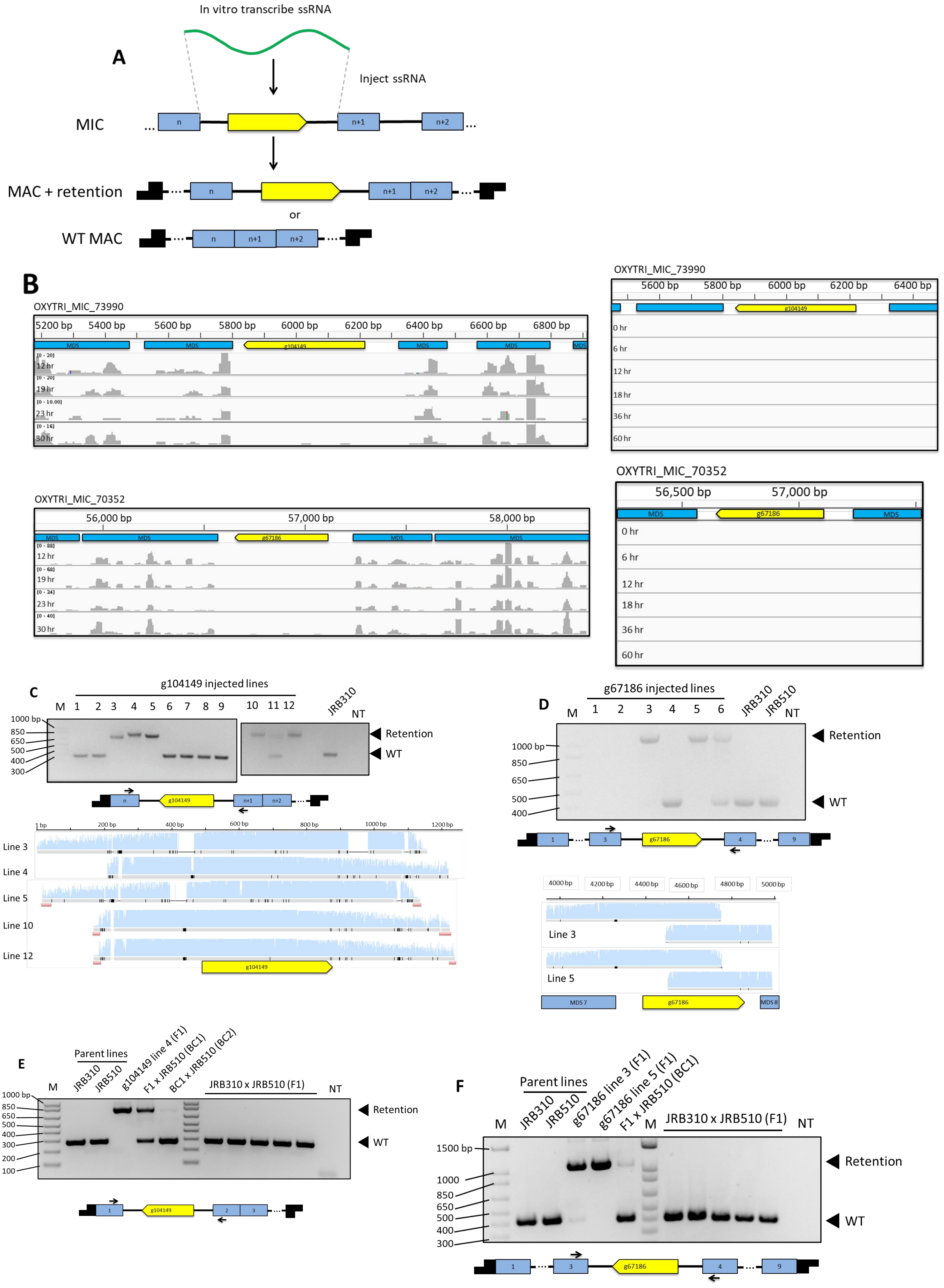
(A) Possible transcription start sites (black arrows) on a hypothetical rearranged somatic nanochromosome after RNA injection to somatically retain TGLOs (yellow). Green: target transcript deriving from TGLO's putative upstream regulatory sequence. Blue blocks: MDSs. Black blocks: Telomeres. **(B)** Top: Germline genome locus containing g104149 (yellow) and gene-specific 5' RACE primers used to amplify transcription start site. Bottom: 5' RACE products targeting the g104149 or Actin II transcription start site in RNA from F1 cells from TGLO retention engineered line # 4, parental cells (JRB310 and JRB510), and late-rearrangement mated cells (WT 48hr). TdT: terminal transferase. **(C)** Top: Germline genome locus containing g67186 (yellow) and gene-specific 5' RACE primers used to amplify transcription start site. 5' RACE products targeting the g67186 or Actin II transcription start site in RNA from F1 TGLO retention engineered line # 3 and 5, parental cells (JRB310 and JRB510), and late-rearrangement mated cells (WT 48hr). TdT: terminal transferase. **(D)** g104149 or Actin II RNA transcript levels based on qRT-PCR relative to mitochondrial rRNA. Error bars: standard deviation of three biological replicates. **(E)** g67186 or Actin II RNA transcript levels based on qRT-PCR relative to mitochondrial rRNA. Error bars: standard deviation of three biological replicates.

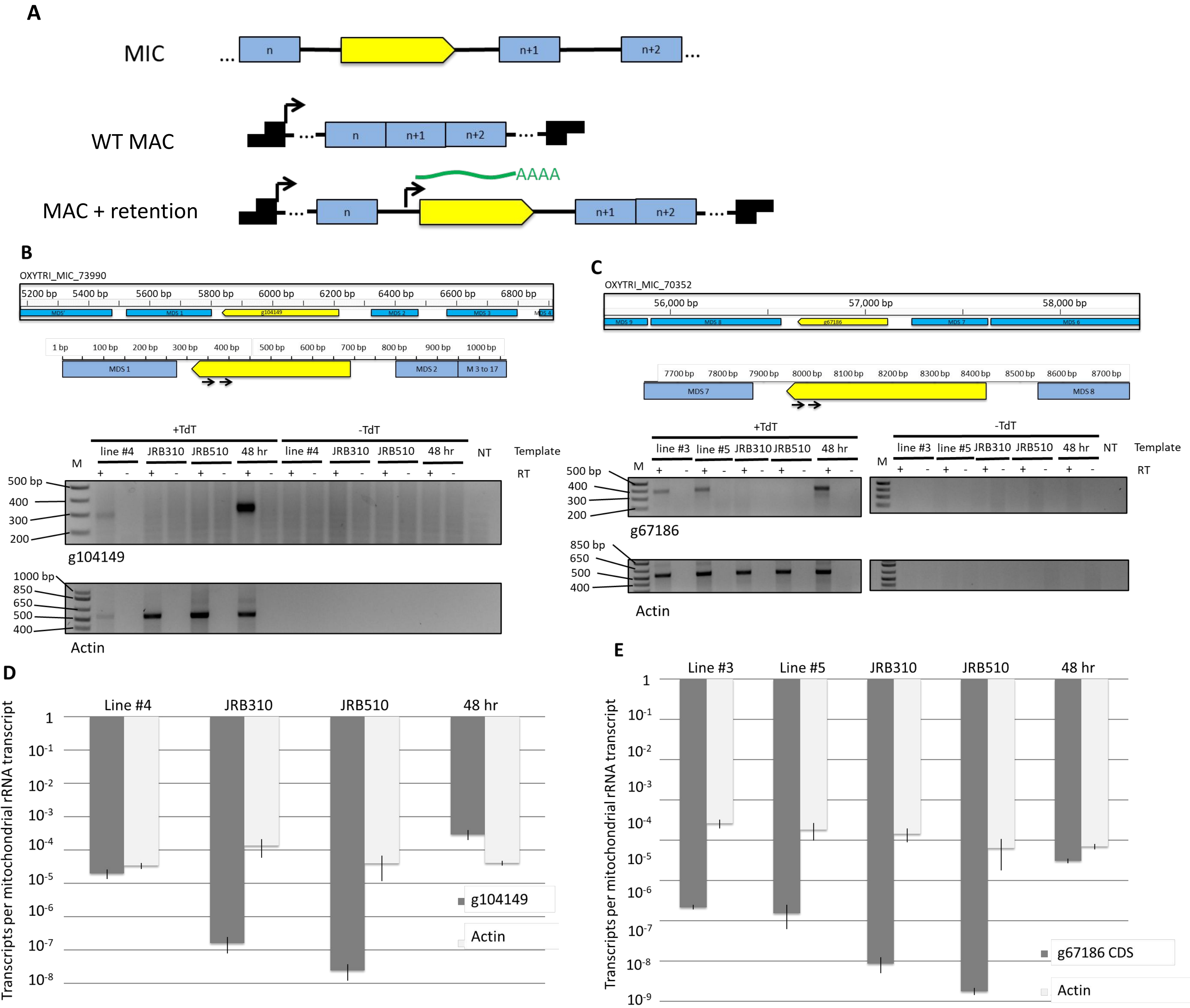

A
Prediction and filtering
Candidates
MIC-limited genes
Predict genes using MIC assembly and RNA-seq hints
217805
788
Genomic DNA and RNA-seq mapping thresholds
4405
6404
213
66
Remove repeat-associated candidates (RepeatMasker)
4372
6357
213
66
Cluster sequences (CD-HIT)
4342
6296
213
66
B
TGLOs with eggNOG functional predictions
High transcription TGLOs
TPR repeat
HD domain
Protein homooligomerization
Protein heterodimerization
MT-A70
Primary miRNA methylation
DNA helicase
Histone H3
ISXO2-like transposase
Low transcription TGLOs
Protein homooligomerization
mRNA splicing via spliceosome
Reverse transcriptase (RNA-dependent DNA polymerase)
G-qudruplex DNA unwinding
PHD zinc finger
High mobility group
DDE superfamily endonuclease
Protein tyrosine kinase
C
High transcription TGLOs
Early time-course
10
5
0
0hr
12hr
18hr
24hr
36hr
Late time-course
12
6
0
12hr
24hr
36hr
48hr
60hr
72hr
Low transcription TGLOs
Early time-course
8
4
0
0hr
12hr
18hr
24hr
36hr
Late time-course
12
6
0
12hr
24hr
36hr
48hr
60hr
72hr
Randomly selected MAC genes
Early time-course
20
10
0
0hr
12hr
18hr
24hr
36hr
Late time-course
20
10
0
12hr
24hr
36hr
48hr
60hr
72hr


A
High transcription TGLOs
Early time-course
10
5
0
0hr
12hr
18hr
24hr
36hr
Late time-course
10
5
0
24hr
36hr
48hr
60hr
72hr
Low transcription TGLOs
Early time-course
12
6
0
0hr
12hr
18hr
24hr
36hr
Late time-course
10
5
0
24hr
36hr
48hr
60hr
72hr
B
Developmental time (hr)
0
33
51
59
74
101
g43073
g104149
g93797


A
marker
F1 lines
1
2
3
4
Parent lines
JRB310
JRB510
Mutant F1 lines
g104149 line #4
g67186 line #3
g67186 line #5
No template
1000 bp
850
650
500
g111288
500 bp
400
300
200
Actin
OXYTRI_MIC_74948
9000 bp
10,000 bp
11,000 bp
MDS
g111288
B
OXYTRI_MIC_74948
9000 bp
10,000 bp
11,000 bp
12,000 bp
13,000 bp
14,000 bp
MDS
g111288
F1 pooled
MDS 1
MDS 2
MDS 3
MDS 4
MDS 5
MDS 6
MDS 7
C
OXYTRI_MIC_74948
9000 bp
10,000 bp
11,000 bp
12,000 bp
13,000 bp
14,000 bp
MDS
g111288
0hr
10hr
20hr
40hr
60hr
cyst
starved
MDS 1
MDS 2
MDS 3
MDS 4
MDS 5
MDS 6
MDS 7
D
F1 lines
JRB310
JRB510
5
6
7
8
9
1.E+00
1.E-01
1.E-02
1.E-03
1.E-04
1.E-05
1.E-06
1.E-07
Fold change relative to mitochondrial rDNA
Amplicon:
A
B
C
D
E
TEBP beta
Actin
OXYTRI_MIC_74948
9000 bp
10,000 bp
11,000 bp
12,000 bp
13,000 bp
14,000 bp
A
B
C
E
D
MDS 1
MDS 2
MDS 3
MDS 4
MDS 5
MDS 6
MDS 7
MDS
g111288
E
JRB310
JRB510
F1
SLC89
SLC92
Positive control
~3.8 kb
~2.5 kb
OXYTRI_MIC_74948
10,000 bp
11,000 bp
12,000 bp
13,000 bp
14,000 bp
MDS 1
MDS 2
MDS 3
MDS 4
MDS 5
MDS 6
MDS 7
F
Mutant F1 lines
Parent lines
RNA source
F1 lines
1
2
3
4
JRB310
JRB510
g104149 line #4
g67186 line #3
g67186 line #5
48 hr
none
RT
+
-
1000 bp
850
650
500
400
300
200
100
g111288
650 bp
500
400
300
200
100
Actin
OXYTRI_MIC_74948
9000 bp
10,000 bp
11,000 bp
MDS
g111288


A
*
*
Otiwi-1 piRNA coverage + 1
1000
100
10
1
12 hr
19 hr
23 hr
30 hr
🟥 High transcription TGLOs
🟩 Low transcription TGLOs
🟦 MDSs
B
9000 bp
10,000 bp
11,000 bp
12,000 bp
13,000 bp
14,000 bp
OXYTRI_MAC_2071 MDS 2
OXYTRI_MAC_1401
MDS 11
MDS 1
MDS 2
MDS 3
MDS 4
MDS 5
MDS 6
MDS 7
12 hr
19 hr
23 hr
30 hr
C
*
*
Template RNA coverage + 1
1e+07
1e+05
1e+03
1e+01
0 hr
6 hr
12 hr
18 hr
36 hr
60 hr
🟥 High transcription TGLOs
🟩 Low transcription TGLOs
🟦 MDSs
D
OXYTRI_MIC_74948
9000 bp
10,000 bp
11,000 bp
MDS
MDS
g111288
0 hr
6 hr
12 hr
18 hr
36 hr
60 hr


A
In vitro transcribe ssRNA
Inject ssRNA
MIC
n
n+1
n+2
MAC + retention
or
WT MAC
B
OXYTRI_MIC_73990
5200 bp
5400 bp
5600 bp
5800 bp
6000 bp
6200 bp
6400 bp
6600 bp
6800 bp
MDS
g104149
[0 - 20]
12 hr
[0 - 20]
19 hr
[0 - 10.00]
23 hr
[0 - 16]
30 hr
OXYTRI_MIC_73990
5600 bp
5800 bp
6000 bp
6200 bp
6400 bp
g104149
0 hr
6 hr
12 hr
18 hr
36 hr
60 hr
OXYTRI_MIC_70352
56,000 bp
57,000 bp
58,000 bp
MDS
g67186
[0 - 88]
12 hr
[0 - 68]
19 hr
[0 - 34]
23 hr
[0 - 40]
30 hr
OXYTRI_MIC_70352
56,500 bp
57,000 bp
MDS
g67186
0 hr
6 hr
12 hr
18 hr
36 hr
60 hr
C
g104149 injected lines
M
1
2
3
4
5
6
7
8
9
10
11
12
JRB310
NT
1000 bp
850
650
500
400
300
Retention
WT
n
g104149
n+1
n+2
1 bp
200 bp
400 bp
600 bp
800 bp
1000 bp
1200 bp
Line 3
Line 4
Line 5
Line 10
Line 12
g104149
D
g67186 injected lines
M
1
2
3
4
5
6
JRB310
JRB510
NT
1000 bp
850
650
500
400
Retention
WT
1
3
g67186
4
9
4000 bp
4200 bp
4400 bp
4600 bp
4800 bp
5000 bp
Line 3
Line 5
MDS 7
g67186
MDS 8
E
Parent lines
M
JRB310
JRB510
g104149 line 4 (F1)
F1 x JRB510 (BC1)
BC1 x JRB510 (BC2)
JRB310 x JRB510 (F1)
NT
1000 bp
850
650
500
400
300
200
100
Retention
WT
1
g104149
2
3
F
Parent lines
M
JRB310
JRB510
g67186 line 3 (F1)
g67186 line 5 (F1)
F1 x JRB510 (BC1)
M
JRB310 x JRB510 (F1)
NT
1500 bp
1000
850
650
500
400
300
Retention
WT
1
3
g67186
4
9


A
MIC
n
n+1
n+2
WT MAC
n
n+1
n+2
MAC + retention
AAAA
n
n+1
n+2
B
OXYTRI_MIC_73990
5200 bp
5400 bp
5600 bp
5800 bp
6000 bp
6200 bp
6400 bp
6600 bp
6800 bp
g104149
1 bp
100 bp
200 bp
300 bp
400 bp
500 bp
600 bp
700 bp
800 bp
900 bp
1000 bp
MDS 1
MDS 2
M 3 to 17
+TdT
-TdT
line #4
JRB310
JRB510
48 hr
NT
Template
M
RT
500 bp
400
300
200
g104149
1000 bp
850
650
500
400
Actin
C
OXYTRI_MIC_70352
56,000 bp
57,000 bp
58,000 bp
g67186
7700 bp
7800 bp
7900 bp
8000 bp
8100 bp
8200 bp
8300 bp
8400 bp
8500 bp
8600 bp
8700 bp
MDS 7
MDS 8
+TdT
-TdT
line #3
line #5
JRB310
JRB510
48 hr
NT
Template
M
RT
500 bp
400
300
200
g67186
850 bp
650
500
400
Actin
D
Transcripts per mitochondrial rRNA transcript
Line #4
JRB310
JRB510
48 hr
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
g104149
Actin
E
Transcripts per mitochondrial rRNA transcript
Line #3
Line #5
JRB310
JRB510
48 hr
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
g67186 CDS
Actin