Bioinformatics, 32(4), 2016, 599–601
doi: 10.1093/bioinformatics/btv691
Advance Access Publication Date: 20 November 2015
Applications Note

OXFORD

Genome analysis

ParTIES: a toolbox for *Paramecium* interspersed DNA elimination studies

Cyril Denby Wilkes, Olivier Arnaiz and Linda Sperling*

Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 15, 2015; revised on October 28, 2015; accepted on November 14, 2015

Abstract

Motivation: Developmental DNA elimination occurs in a wide variety of multicellular organisms, but ciliates are the only single-celled eukaryotes in which this phenomenon has been reported. Despite considerable interest in ciliates as models for DNA elimination, no standard methods for identification and characterization of the eliminated sequences are currently available.

Results: We present the *Paramecium* Toolbox for Interspersed DNA Elimination Studies (ParTIES), designed for *Paramecium* species, that (i) identifies eliminated sequences, (ii) measures their presence in a sequencing sample and (iii) detects rare elimination polymorphisms.

Availability and implementation: ParTIES is multi-threaded Perl software available at https://github.com/oarnaiz/ParTIES. ParTIES is distributed under the GNU General Public Licence v3.

Contact: linda.sperling@i2bc.paris-saclay.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Programmed DNA elimination during somatic development is widely distributed in animal species. First discovered by Boveri in ascaris worms in the 19th century, the phenomenon has to date been characterized in multiple species of nematodes, insects, arachnids, crustaceans, lampreys, fish, birds and mammals, and can involve the programmed reduction of up to 85% of the genome (review: Wang and Davis, 2014). Ciliates, the only unicells that undergo somatic DNA elimination, resemble animals in that they make a germ/soma distinction. One or more diploid germ line nuclei and a polyploid somatic nucleus coexist in a unique cytoplasm. Only the somatic nucleus is transcriptionally active during vegetative growth. As in metazoans, somatic DNA elimination in ciliates can silence transposable elements and cellular genes (Chen *et al.*, 2014), regulate gene dosage (Nowacki *et al.*, 2010) and determine mating type (Cervantes *et al.*, 2013; Singh *et al.*, 2014).

Among ciliates, *Paramecium* is an outstanding model to study DNA elimination. Sexual processes are readily controlled under laboratory conditions and a third of the germ line genome is lost through two types of reproducible, programmed deletions: (i) repeated sequences are heterogeneously eliminated leading to chromosome fragmentation; (ii) single copy elements, called Internal Eliminated Sequences (IESs), are precisely excised. Somatic DNA is also endoreplicated to reach ~800 haploid copies. As the 45 000 IESs in the *Paramecium tetraurelia* genome interrupt non-coding and coding sequences (Arnaiz *et al.*, 2012), their elimination is essential to reconstitute open reading frames (ORFs). Both types of DNA elimination depend on a piggyBac domesticated transposase (named PiggyMac) (Baudry *et al.*, 2009), which may be guided by short-RNA driven epigenetic signals (Lepère *et al.*, 2008, 2009).

Cost decrease in High Throughput Sequencing (HTS) is allowing researchers to produce massive genome-wide data to study DNA elimination, but specific bioinformatic methods are still lacking. Here we describe ParTIES: *Paramecium* Toolbox for IES Interspersed DNA Elimination studies. With Illumina DNA-Seq paired-end reads and a somatic reference genome as input, ParTIES performs IES identification, quantitates their presence in the sample and detects rare excision polymorphisms. Benchmarks are provided in Supplementary Materials.

2 Description

Given an input somatic reference genome and any Illumina DNA-Seq sample, be it from germ line or from somatic DNA, ParTIES provides 3 complementary methods which can be run consecutively or independently (Fig. 1).

2.1 Insertion identification

IES identification is important not only for mechanistic studies of DNA elimination, but also to investigate IES origin and evolution across *Paramecium* species. IES identification is a 2-step process. First, an exhaustive list of potential insertion sites is compiled by Method of Identification by Alignment Anomalies (MIRAA). Second, Method of Identification by Comparison of Assemblies (MICA) determines the insertion sequences and their positions. MIRAA uses read mapping to detect an excess of read ends at a given site. Partially mapped reads are presumed to contain additional sequence. Reads that perfectly match the somatic reference at these sites are discarded. The filtered reads are assembled with Velvet (Zerbino and Birney, 2008) to produce contigs potentially containing IESs. MICA carries out global comparison of the contigs and the somatic reference, followed by local realignment to define inserted segments precisely. Optionally for *Paramecium*, the local alignments can be further adjusted to ensure that the ends of the inserted segments conform to the PiggyMac cleavage requirements, namely that insertions be bounded by TA dinucleotides. The IES identification output is a standard GFF3 file.

The above procedure may be simplified if the user already has an assembly containing insertions (such as a germ line reference genome) or if Velvet is not suitable for the available sequencing data, which the user can optionally assemble with a preferred protocol.

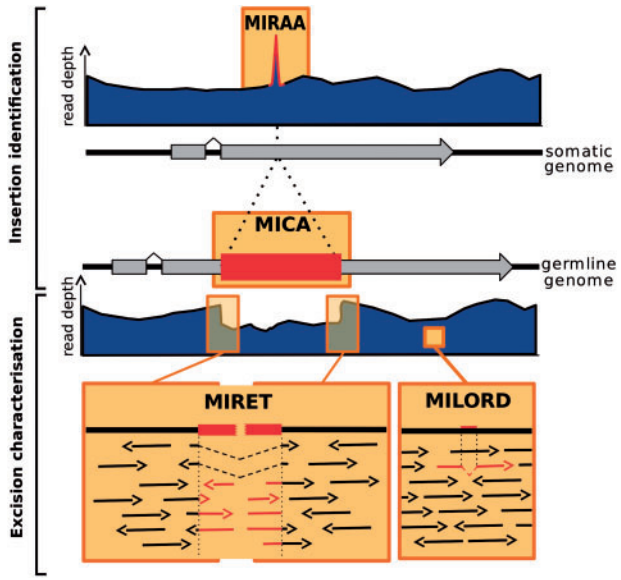

Fig. 1. ParTIES toolbox. Somatic and germ line genomes are evoked by solid black lines with exons (grey boxes) and IES (red box). The dark blue regions represent alignment of Illumina short reads on reference somatic (upper) and germ line (lower) genomes. Arrows in the insets represent mapped reads. MIRAA identifies breakpoints based on excess read coverage and MICA identifies insertions by comparing contigs, assembled from the input reads, to the reference somatic genome; together they output a list of IESs found in a sample. MIRET uses alignments of the short reads to measure IES retention in a sample while MILORD looks for rare deletions in the reads indicative of excision polymorphism

2.2 Excision characterization

2.2.1 IES retention

Many experiments are designed to see whether IESs are excised or retained in the somatic genome after experimental depletion of a factor potentially involved in DNA elimination. The **M**ethod of **IES** **Ret**ention (MIRET) module was designed to quantitate the presence of each IES in the genome given a DNA-Seq sample. MIRET uses alignment of the sample reads on the somatic reference to count reads that cross the IES excision junction, designated IES^- reads. MIRET uses alignment of the reads on an IES-containing reference to count reads crossing the junction between an IES and its flanking sequence, designated IES^+ reads. MIRET then calculates a 'boundary score', defined as the ratio of IES^+ reads over the sum of IES^- and IES^+ reads for that boundary. MIRET can also calculate an 'IES retention score' that uses the same counts as the boundary score, with the additional restriction that a read that crosses both ends of an IES is counted only once in the IES retention score calculation.

Determination of the statistical significance of a retention score requires a control sample, provided by sequencing the somatic DNA of untreated cells. Comparison of retention scores of the experimental and control samples is performed to test the null hypothesis that a given IES has the same retention score in both samples. The statistical tests are provided by the R environment and take into account read depth (cf. Sup Mat for details).

2.2.2 Rare deletion events

The IES excision machinery is error-prone and sometimes deletes a segment of somatic DNA or uses an alternative boundary during elimination of a *bona fide* IES (Duret *et al.*, 2008). In order to catalogue these events and evaluate their frequency, we developed the **M**ethod of **I**dentification and **L**ocalization **o**f **R**are **D**eletions (MILORD) module. MILORD looks for a deletion in a read compared to a reference genome. It identifies partially mapped reads and then tries to realign the unmapped part of the read. If a coherent unique alignment is found, a deletion segment is recorded.

3 Discussion

We benchmarked ParTIES using real and simulated data with the *P. tetraurelia* 72 Mb somatic reference genome (Aury *et al.*, 2006) and IES reference set (Arnaiz *et al.*, 2012). The results are presented in Supplementary Materials, and can be used to plan optimal, cost-effective sequencing experiments. We found the minimal requirement for high sensitivity and specificity IES identification and excision quantification is 35× sequencing of a short-insert library of 75 nt paired-end reads, provided the sequencing sample contains at least 25% germ line DNA.

The ParTIES package is expected to set the standard for quantitative analysis of *Paramecium* genomes and DNA elimination.

Acknowledgements

We are grateful to Laurent Duret and Franck Picard for help with statistical tests.

Funding

This work was funded by the CNRS and by the ANR-12-BSV6-0017 'INFERNO' and the ANR-14-CE10-0005-03 'PiggyPack' grants. CDW was supported by a PhD fellowship from the MENRT and by the ANR. This work was carried out in the context of the CNRS-supported European Research Group 'Parameciom Genome Dynamics and Evolution'.

Conflict of Interest: none declared.

References

Arnaiz,O. *et al*. (2012) The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet.*, **8**), e1002984.

Aury,J.-M. *et al*. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, **444**, 171–178.

Baudry,C. *et al*. (2009) PiggyMac: a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev.*, **23**, 2478–2483.

Cervantes,M.D. *et al*. (2013) Selecting one of several mating types through gene segment joining and deletion in *Tetrahymena thermophila*. *PLoS Biol.*, **11**, e1001518.

Chen,X. *et al*. (2014) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell*, **158**, 1187–1198.

Duret,L. *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.

Lepère,G. *et al*. (2008) Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in *Paramecium tetraurelia*. *Genes Dev.*, **22**, 1501–1512.

Lepère,G. *et al*. (2009) Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. *Nucleic Acids Res.*, **37**, 903–915.

Nowacki,M. *et al*. (2010) RNA-mediated epigenetic regulation of DNA copy number. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 22140–22144.

Singh,D.P. *et al*. (2014) Genome-defence small RNAs exapted for epigenetic mating-type inheritance. *Nature*, **509**, 447–452.

Wang,J. and Davis,R.E. (2014) Programmed DNA elimination in multicellular organisms. *Curr. Opin. Genet. Dev.*, **27**, 26–34.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.