Bioinformatics, 38(16), 2022, 4030–4032
https://doi.org/10.1093/bioinformatics/btac442
Advance Access Publication Date: 4 July 2022
Applications Note

Genome analysis

HiC-TE: a computational pipeline for Hi-C data analysis to study the role of repeat family interactions in the genome 3D organization

Matej Lexa [1,2,*], Monika Cechova[1], Son Hoang Nguyen[1], Pavel Jedlicka[2], Viktor Tokan[2], Zdenek Kubat[2], Roman Hobza[2] and Eduard Kejnovsky[2,*]

[1]Faculty of Informatics, Masaryk University, 60200 Brno, Czech Republic and [2]Department of Plant Developmental Genetics, Institute of Biophysics of the Czech Academy of Sciences, 61200 Brno, Czech Republic

*To whom correspondence should be addressed.
Associate Editor: Can Alkan

Received on February 15, 2022; revised on June 14, 2022; editorial decision on June 28, 2022; accepted on June 30, 2022

Abstract

Motivation: The role of repetitive DNA in the 3D organization of the interphase nucleus is a subject of intensive study. In studies of 3D nucleus organization, mutual contacts of various loci can be identified by Hi-C sequencing. Typical analyses use binning of read pairs by location to reduce noise. We use binning by repeat families instead to make similar conclusions about repeat regions.

Results: To achieve this, we combined Hi-C data, reference genome data and tools for repeat analysis into a Nextflow pipeline identifying and quantifying the contacts of specific repeat families. As an output, our pipeline produces heatmaps showing contact frequency and circular diagrams visualizing repeat contact localization. Using our pipeline with tomato data, we revealed the preferential homotypic interactions of ribosomal DNA, centromeric satellites and some LTR retrotransposon families and, as expected, little contact between organellar and nuclear DNA elements. While the pipeline can be applied to any eukaryotic genome, results in plants provide better coverage, since the built-in TE-greedy-nester software only detects tandems and LTR retrotransposons. Other repeats can be fed via GFF3 files. This pipeline represents a novel and reproducible way to analyze the role of repetitive elements in the 3D organization of genomes.

Availability and implementation: https://gitlab.fi.muni.cz/lexa/hic-te/.

Contact: lexa@fi.muni.cz

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

1 Introduction

The eukaryotic genome is hierarchically packed in the nucleus allowing DNA replication and gene transcription to take place in a spatially and temporally regulated fashion. A significant part of eukaryotic genomes is made up of transposable elements (TEs) and satellite DNA, where e.g. LTR retrotransposons constitute up to 90% of genomes in some species (Liehr, 2021; Schnable *et al.*, 2009; Wicker *et al.*, 2018). TEs are often embedded in cellular regulatory networks (Feschotte, 2008) where they rewire the gene expression programs (Slotkin and Martienssen, 2007). Many examples of the domestication of TEs for specific cellular functions have been observed (Jangam *et al.*, 2017; Sinzelle *et al.*, 2009). Moreover, the 3D organization of the interphase nucleus is recently a subject of intensive study.

Methods of high-throughput mapping of DNA–DNA interactions, such as chromosome conformation capture (Hi-C), now allow the study of long-distance interactions in eukaryotic nuclei. Because of technical issues, these have mostly avoided repetitive parts of the genome. A better understanding of the interaction of the main repeat classes can help uncover their genomic role. A recent study demonstrated the role of TEs in organizing the human and mouse genomes (Lu *et al.*, 2021) and similar analysis in other organisms is hitherto missing.

Here, we present a new sequence processing pipeline to identify and quantify interactions of TEs, satellite DNA and rDNA in nuclei, especially those that participate in long-distance ($\geq$1 Mb) or inter-chromosomal contacts with frequencies that differ from baseline expectations of randomness.

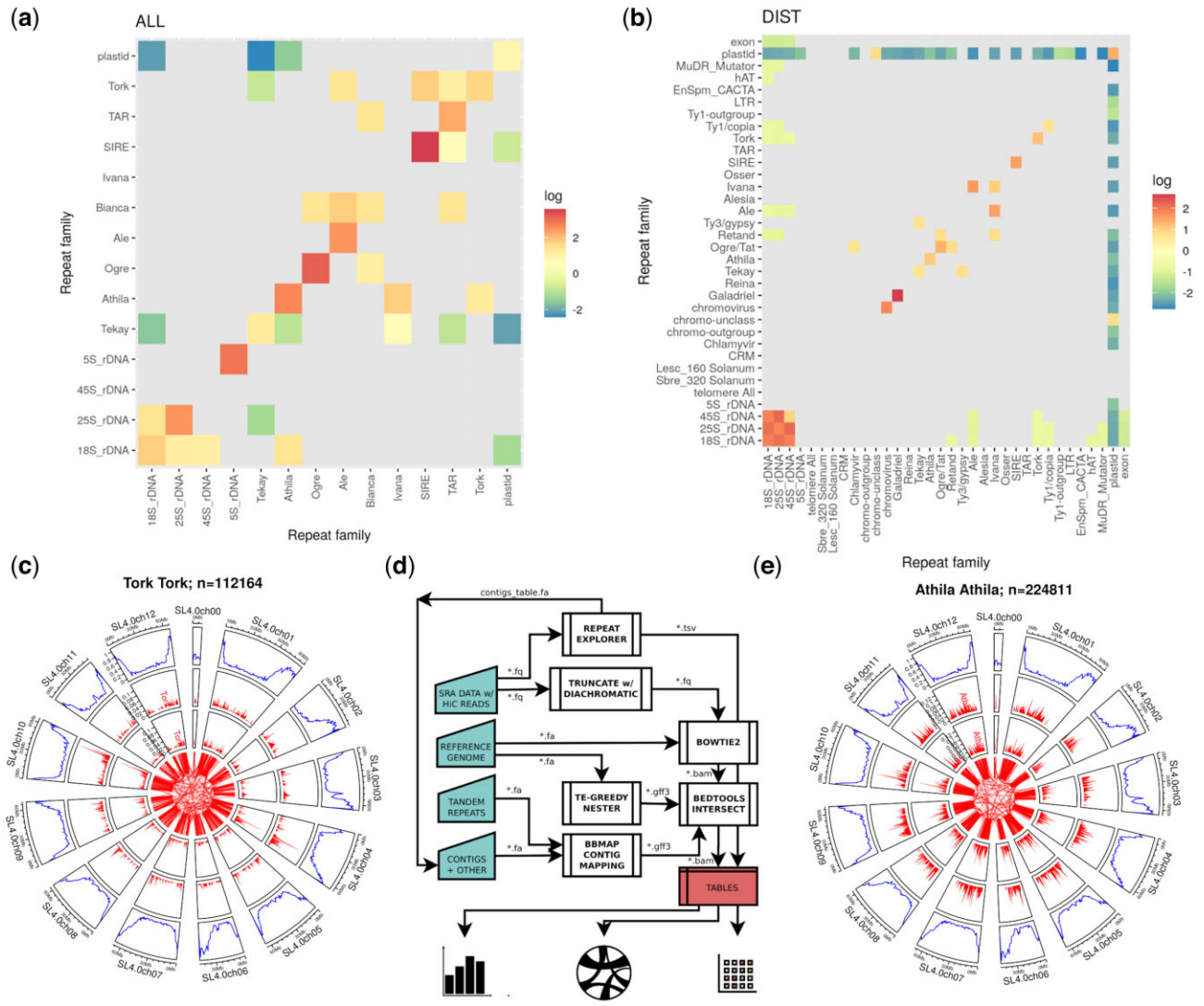

Fig. 1. HiC-TE pipeline. (**a**) Reference-free and (**b**) reference-based heatmaps generated by the pipeline for SRR5748725, showing repeat family pairs and their label-permutation normalized contact frequency; (**c**, **e**) Circos plots for the same showing chromosomal locations of representative contacts for the given family pair; (**d**) block diagram showing the overall data flow in the HiC-TE pipeline. Some details were omitted for clarity (the full graph produced by the pipeline is shown in Supplementary Fig. S1) [left—main data inputs; bottom—main data outputs; double edged rectangles—main processes running external tools; FASTA (*.fa), FASTQ (*.fq), BAM, GFF3—main sequence and annotation data formats passed between processes]

2 Nextflow pipeline description and testing

Our pipeline (Fig. 1d, Supplementary Fig. S1) integrates sequence analysis from several sources: assembled genome repeat annotation [TE-greedy-nester (Lexa *et al.*, 2020), PlantSat database (Macas *et al.*, 2002)], medium and long-distance contact information (Hi-C experiments) and repetitive NGS read clustering [Repeat Explorer (Novak *et al.*, 2013)]. TE-greedy-nester is our previously developed tool for structure-based detection of LTR retrotransposons that can even assign fragmented TEs to their families. However, a large number of tools in this area exist as recently reviewed by Rodriguez and Makałowski (2022). Currently, alternative annotations from such sources can be applied via a GFF3 file or consensus sequences that will be mapped to reference by the pipeline (see 'DATA' and 'OPTIONAL PARAMETERS' sections in source code README.md). Repeat Explorer is a popular and time-tested solution for graph-clustering repetitive sequencing reads without the need for a reference genome. We included these two tools to complement each other, as each method has different strengths. For example, many reference genomes lack repetitive regions that are hard to assemble, a situation in which reference-free approach might be advantageous, although the quality of reference genomes is gradually improving with T2T sequencing (Nurk *et al.*, 2022). The pipeline was implemented with Nextflow (Di Tommaso *et al.*, 2017) to allow for flexibility and scalability, using a recent installation of Ubuntu Linux with all dependencies included. In addition, we provide a tested containerized version allowing runs with Docker/Singularity deployment (see 'RUNNING THE PIPELINE' in source code README.md). As a result, all the figures and tables are fully reproducible and can be easily generated. We summarize the memory, disk and time requirements in Supplementary Tables S1 and S2 and Supplementary Figures S2–S4.

To test 'HiC-TE', we used a publicly available dataset on the tomato (*Solanum lycopersicum*) (Dong *et al.*, 2017) with two technical replicates for each of three plants. The minimal input dataset represents a HiC experiment (FASTQ) with the name of the restriction enzyme used in the protocol, a reference genome sequence (FASTA), exon annotations (GFF3) and a set of tandem repeats and satellite DNA to be mapped (PlantSat FASTA) with the telomeric repeat sequence array as a minimum. We verified that the pipeline produces consistent results and that the computational replicates are less variable than any other replicates. We analyzed Hi-C contacts from reads clustered with Repeat Explorer (Fig. 1a) and reference-mapped long-distance interactions

(spanning ≥1 Mb or between sequences located on different chromosomes) (Fig. 1b). The main output is a series of heatmaps showing high/low values of normalized contacts in diverging colors, while fields (repeat family pairs) with missing values are shown in gray. The pipeline also relies on initial trimming of raw reads with Diachromatic (Hansen *et al.*, 2019), read mapping via Bowtie2 (Langmead and Salzberg, 2012) and BBmap, overlap/intersection analysis with bedtools (Quinlan and Hall, 2010) and data manipulation and visualization in R/Bioconductor with extra packages (Supplementary Note S1) (Gel and Serra, 2017; Gu *et al.*, 2014; Indahl *et al.*, 2018; Pedersen and Shemanarev, 2020; Wickham, 2007). Before visualization in heatmaps, the data are normalized using three different normalization techniques (Supplementary Note S2) to account for background HiC signal and the fact that repeat families have varying frequencies. Circos plots allow to understand chromosomal localization of the contacts, such as the chromosome number or whether it is in a gene-rich or gene-poor area (Fig. 1c and e). Normalized values that are too close to 1, or based on samples with a low number of reads (configurable by the user) are shown as gray fields in these heatmaps.

3 Discussion

The pipeline contains two modes of repeat annotation, reference-based and reference-free. While reference-based data contain chromosomal positions and allow the calculation of distances, the reference-free mode avoids the necessity to discern real and apparent read mapping, which is especially problematic when dealing with repeats and short reads. Our tool contrasts traditional methods of binning HiC contacts (Golicz *et al.*, 2020; Sun *et al.*, 2020; Zheng *et al.*, 2019) (for the discussion see Supplementary Note S3). It has a potential, based on frequency of interactions of specific centromeric or telomeric repeats, to reveal distinct local organizations of chromosomes, such as Rabl, Rosette or Bouquet arrangement (Tiang *et al.*, 2012) (see Supplementary Note S4 for further discussion of biological relevancy).

The focus of traditional HiC data analysis pipelines on unique mapping made us realize that multiple-mapping reads could still be assigned to a family of repeats. To this end, we built a HiC read mapping pipeline that explores this possibility in tandem with Repeat Explorer software, producing a report of likely interaction partners among repeat families. Another branch of our computations produces similar output relying on reads that Bowtie2 maps to annotated repeats. This Nextflow pipeline can identify and quantify the contacts of specific repeats in the 3D nucleus. Using real biological data from public databases we have shown that with proper normalization techniques, known (and possibly also unknown) interaction partners can be revealed among annotated repeat families.

Acknowledgements

We thank Christopher Johnson for critical reading of the manuscript, Jan Hoidekr for help with deployment of the pipeline in the MetaCentrum environment. Computational resources were supplied by the project 'e-Infrastruktura CZ' (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic. M.C. is the holder of Martina Roeselova Memorial Fellowships 2020.

Funding

This work was supported by the Czech Science Foundation [21-00580S].

Conflict of Interest: none declared.

Data availability

The data underlying this article are available in Zenodo, at: https://dx.doi.org/10.5281/zenodo.6628770 (pipeline info and output) and https://dx.doi.org/10.5281/zenodo.6628543 (source code freeze).

References

Di Tommaso,P. *et al.* (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
Dong,P. *et al.* (2017) 3D chromatin architecture of large plant genomes determined by local a/B compartments. *Mol. Plant.*, **10**, 1497–1509.
Feschotte,C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, **9**, 397–405.
Gel,B. and Serra,E. (2017) KaryoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, **33**, 3088–3090.
Golicz,A. *et al.* (2020) Rice 3D chromatin structure correlates with sequence variation and meiotic recombination rate. *Commun. Biol.*, **3**, 235.
Gu,Z. *et al.* (2014) Circlize implements and enhances circular visualization in R. *Bioinformatics*, **30**, 2811–2812.
Hansen,P. *et al.* (2019) Computational processing and quality control of Hi-C, capture Hi-C and Capture-C data. *Genes*, **10**, 548.
Indahl,U. *et al.* (2018) A similarity index for comparing coupled matrices. *J. Chemom.*, **32**, e3049.
Jangam,D. *et al.* (2017) Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.*, **33**, 817–831.
Langmead,B. and Salzberg,S. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
Lexa,M. *et al.* (2020) TE-greedy-nester: structure-based detection of LTR retrotransposons and their nesting. *Bioinformatics*, **36**, 4991–4999.
Liehr,T. (2021) Repetitive elements in humans. *Int. J. Mol. Sci.*, **22**, 2072.
Lu,J. *et al.* (2021) Homotypic clustering of L1 and B1/Alu repeats compartmentalizes the 3D genome. *Cell Res.*, **31**, 613–630.
Macas,J. *et al.* (2002) PlantSat: a specialized database for plant satellite repeats. *Bioinformatics*, **18**, 28–35.
Novak,P. *et al.* (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.
Nurk,S. *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44–53.
Pedersen,T. and Shemanarev,M. (2020) *ragg: graphic devices based on AGG. R package version 1.1.3*. https://ragg.r-lib.org, https://github.com/r-lib/ragg.
Quinlan,A. and Hall,I. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
Rodriguez,M. and Makałowski,W. (2022) Software evaluation for de novo detection of transposons. *Mob. DNA*, **13**, 14.
Schnable,P. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
Sinzelle,L. *et al.* (2009) Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell. Mol. Life Sci.*, **66**, 1073–1093.
Slotkin,R. and Martienssen,R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, **8**, 272–285.
Sun,L. *et al.* (2020) Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. *Nat. Commun.*, **11**, 1886.
Tiang,C.-L. *et al.* (2012) Chromosome organization and dynamics during interphase, mitosis, and meiosis in plants. *Plant Physiol.*, **158**, 26–34.
Wicker,T. *et al.* (2018) Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.*, **19**, 103.
Wickham,H. (2007) Reshaping data with the reshape package. *J. Stat. Softw.*, **21**, 1–20.
Zheng,Y. *et al.* (2019) Generative modeling of multi-mapping reads with mhi-c advances analysis of hi-c studies. *eLife*, **8**, 1141–1156.