HiC-TE: a Nextflow pipeline to study repeat interactions in the 3D genome

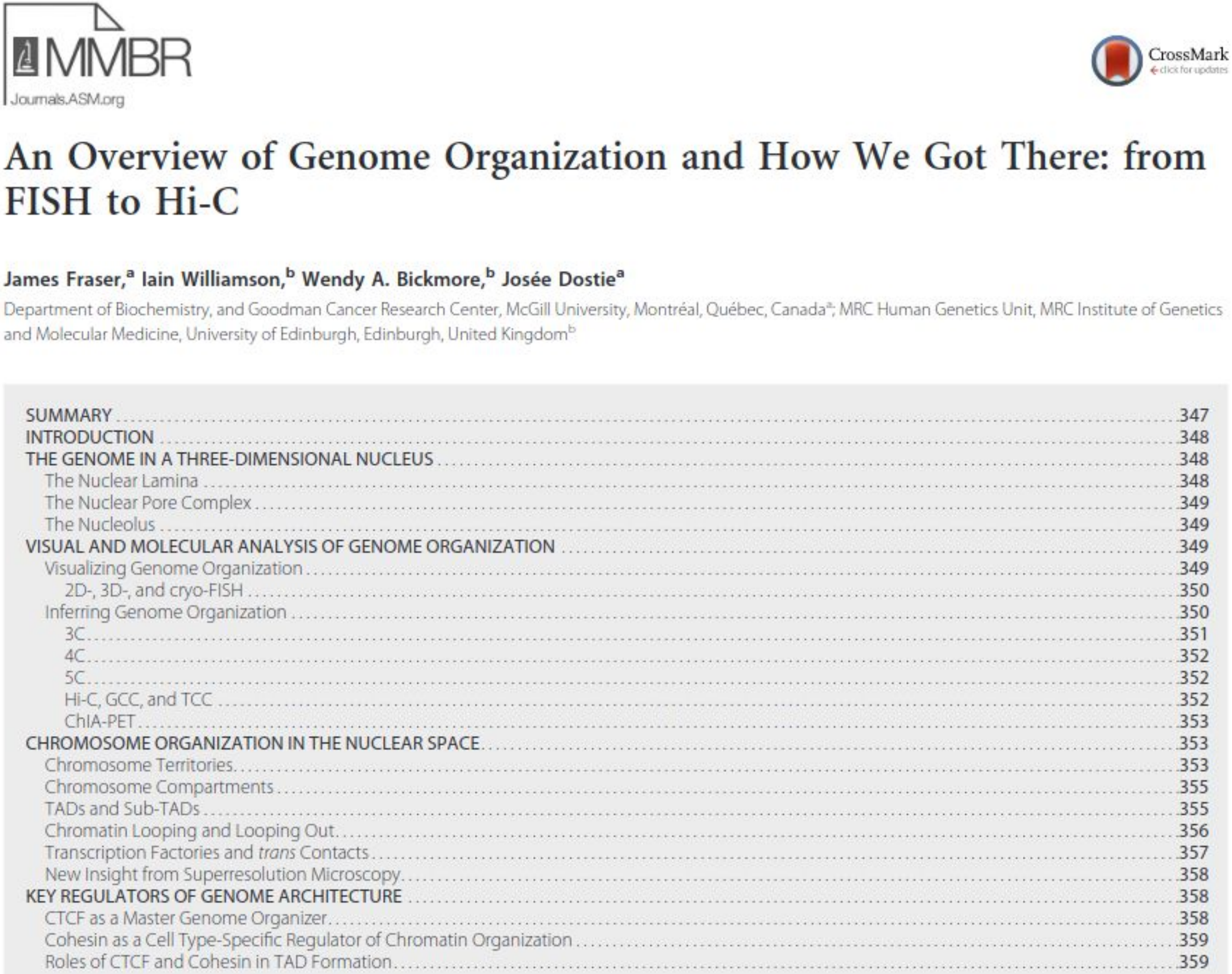
Matej Lexa

ENBIK 2022

Němčice u Kolína

13.-15.6.2022

Genome 3D organization



MMBR
Journals.ASM.org

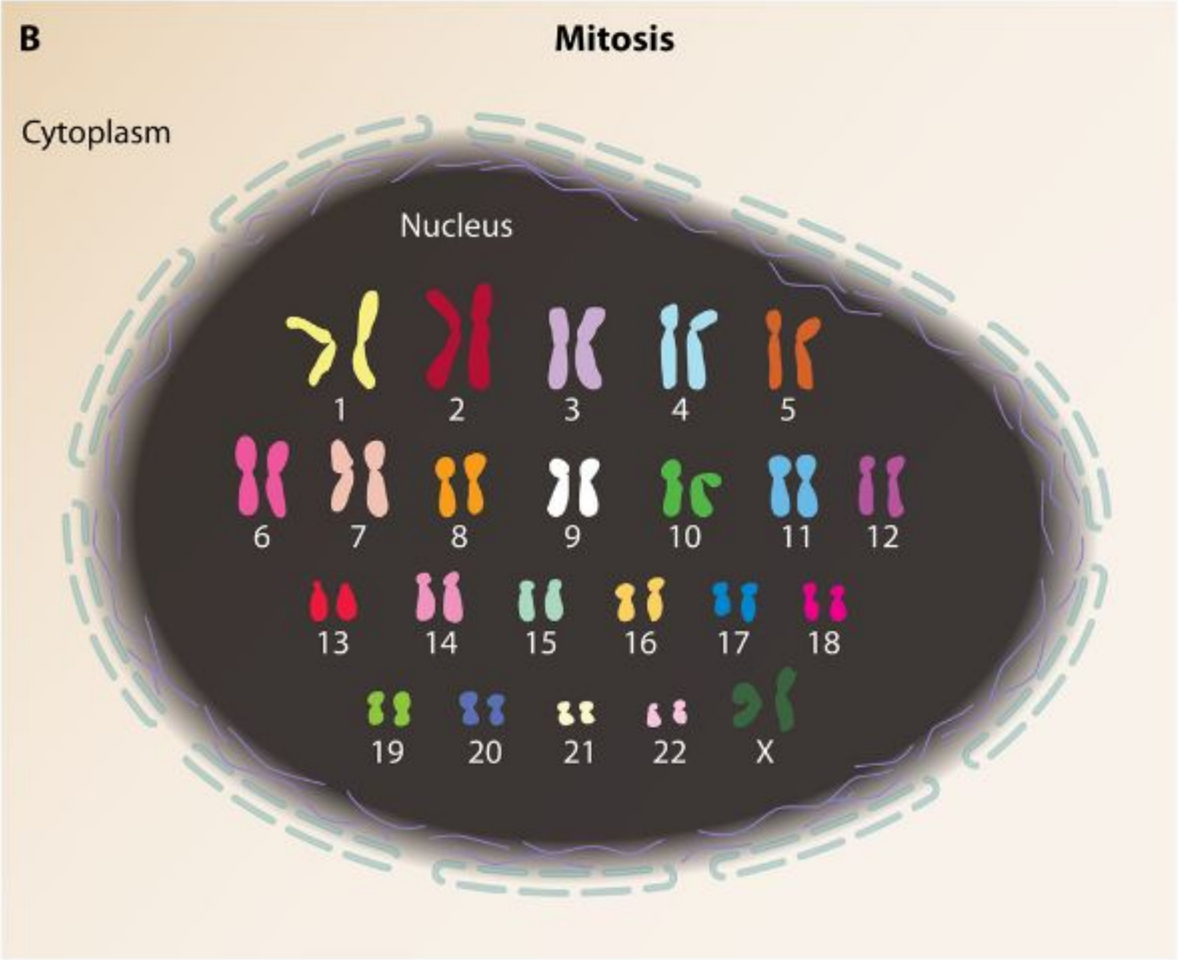
An Overview of Genome Organization and How We Got There: from FISH to Hi-C

James Fraser,[a] Iain Williamson,[b] Wendy A. Bickmore,[b] Josée Dostie[a]

Department of Biochemistry, and Goodman Cancer Research Center, McGill University, Montréal, Québec, Canada[a]; MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom[b]

SUMMARY 347
INTRODUCTION 348
THE GENOME IN A THREE-DIMENSIONAL NUCLEUS 348
The Nuclear Lamina 348
The Nuclear Pore Complex 349
The Nucleolus 349
VISUAL AND MOLECULAR ANALYSIS OF GENOME ORGANIZATION 349
Visualizing Genome Organization 349
2D-, 3D-, and cryo-FISH 350
Inferring Genome Organization 350
3C 351
4C 352
5C 352
Hi-C, GCC, and TCC 352
ChIA-PET 353
CHROMOSOME ORGANIZATION IN THE NUCLEAR SPACE 353
Chromosome Territories 353
Chromosome Compartments 355
TADs and Sub-TADs 355
Chromatin Looping and Looping Out 356
Transcription Factories and *trans* Contacts 357
New Insight from Superresolution Microscopy 358
KEY REGULATORS OF GENOME ARCHITECTURE 358
CTCF as a Master Genome Organizer 358
Cohesin as a Cell Type-Specific Regulator of Chromatin Organization 359
Roles of CTCF and Cohesin in TAD Formation 359

Genome 3D organization

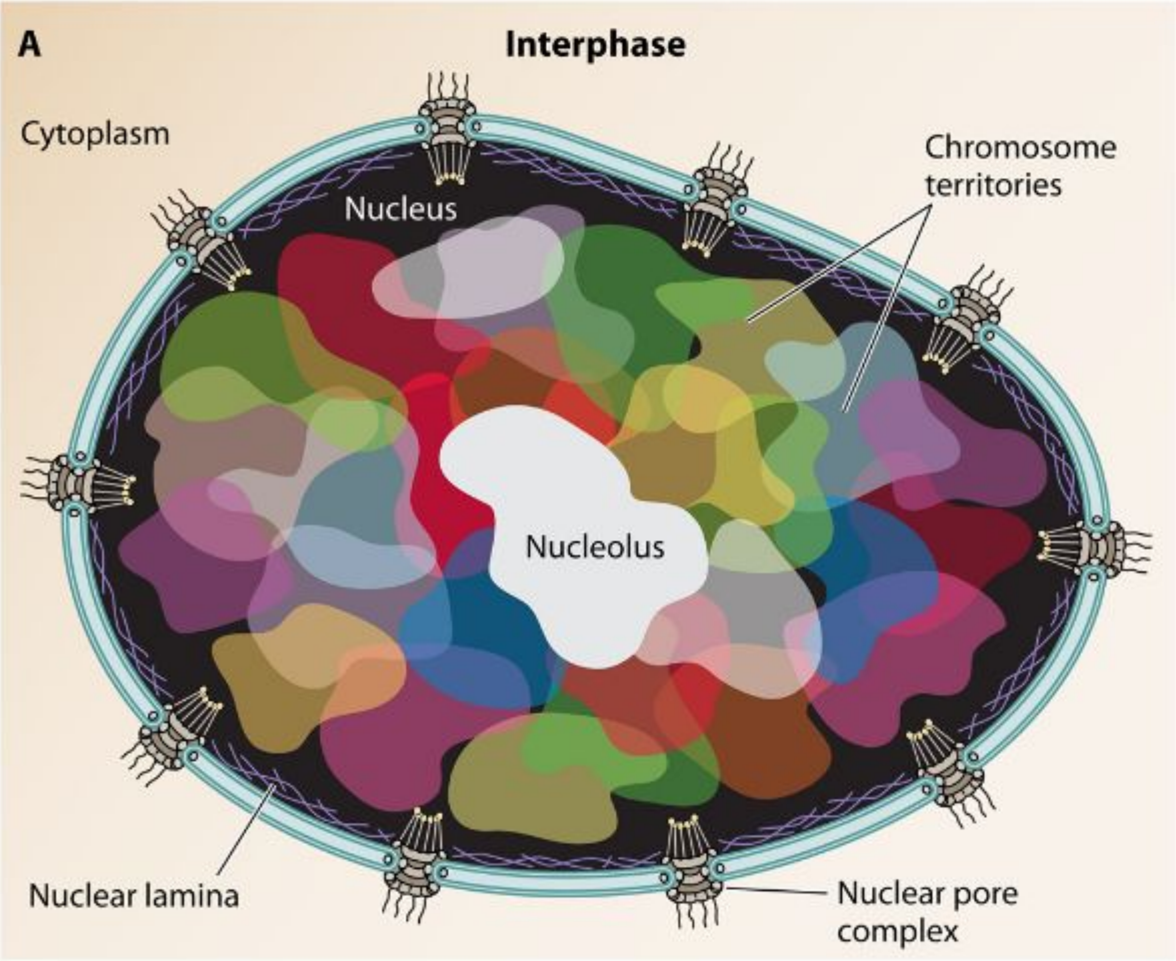


An Overview of Genome Organization and How We Got There: from FISH to Hi-C

James Fraser,[a] Iain Williamson,[b] Wendy A. Bickmore,[b] Josée Dostie[a]

Department of Biochemistry, and Goodman Cancer Research Center, McGill University, Montréal, Québec, Canada[a]; MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom[b]

Genome 3D organization

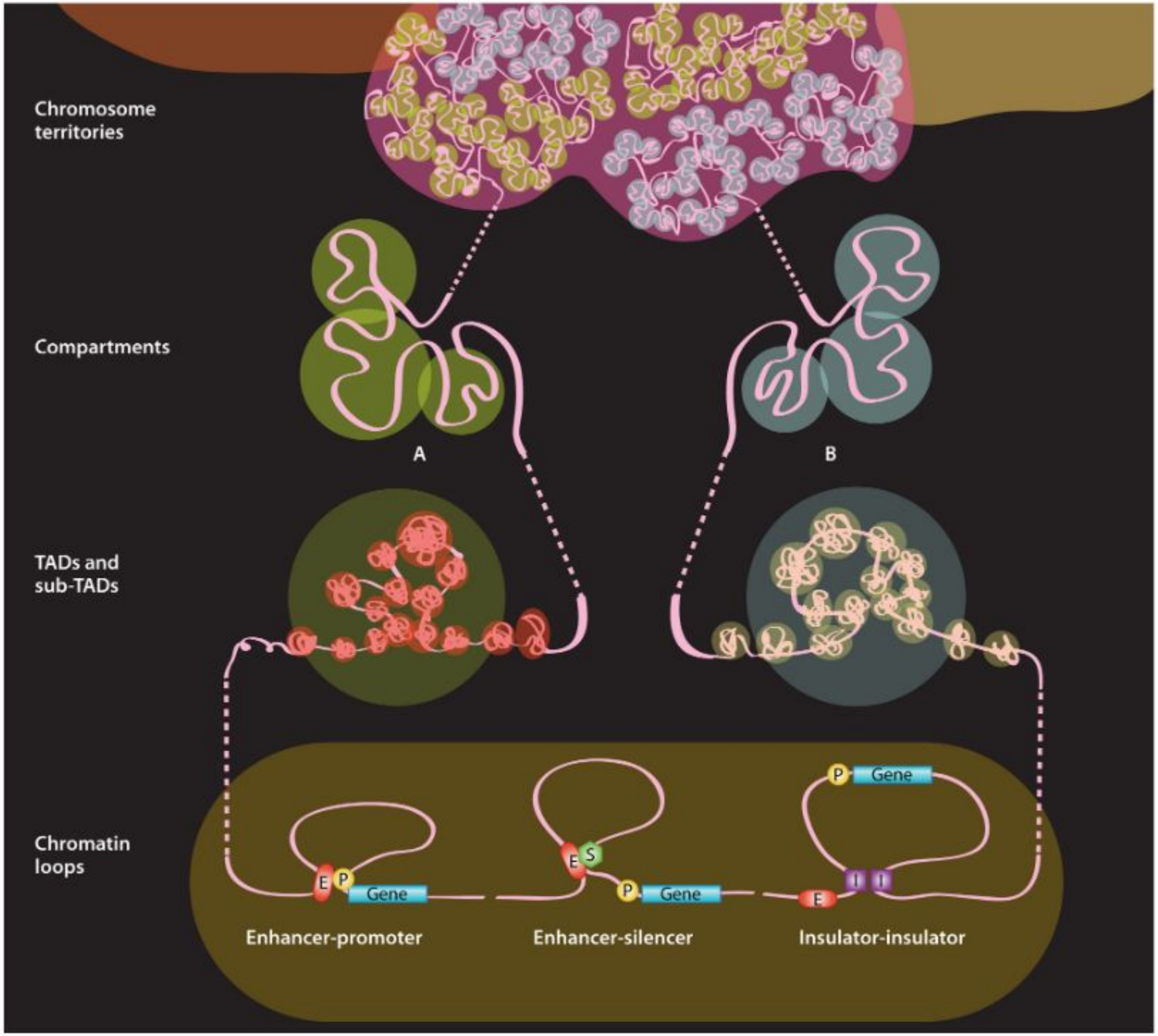


An Overview of Genome Organization and How We Got There: from FISH to Hi-C

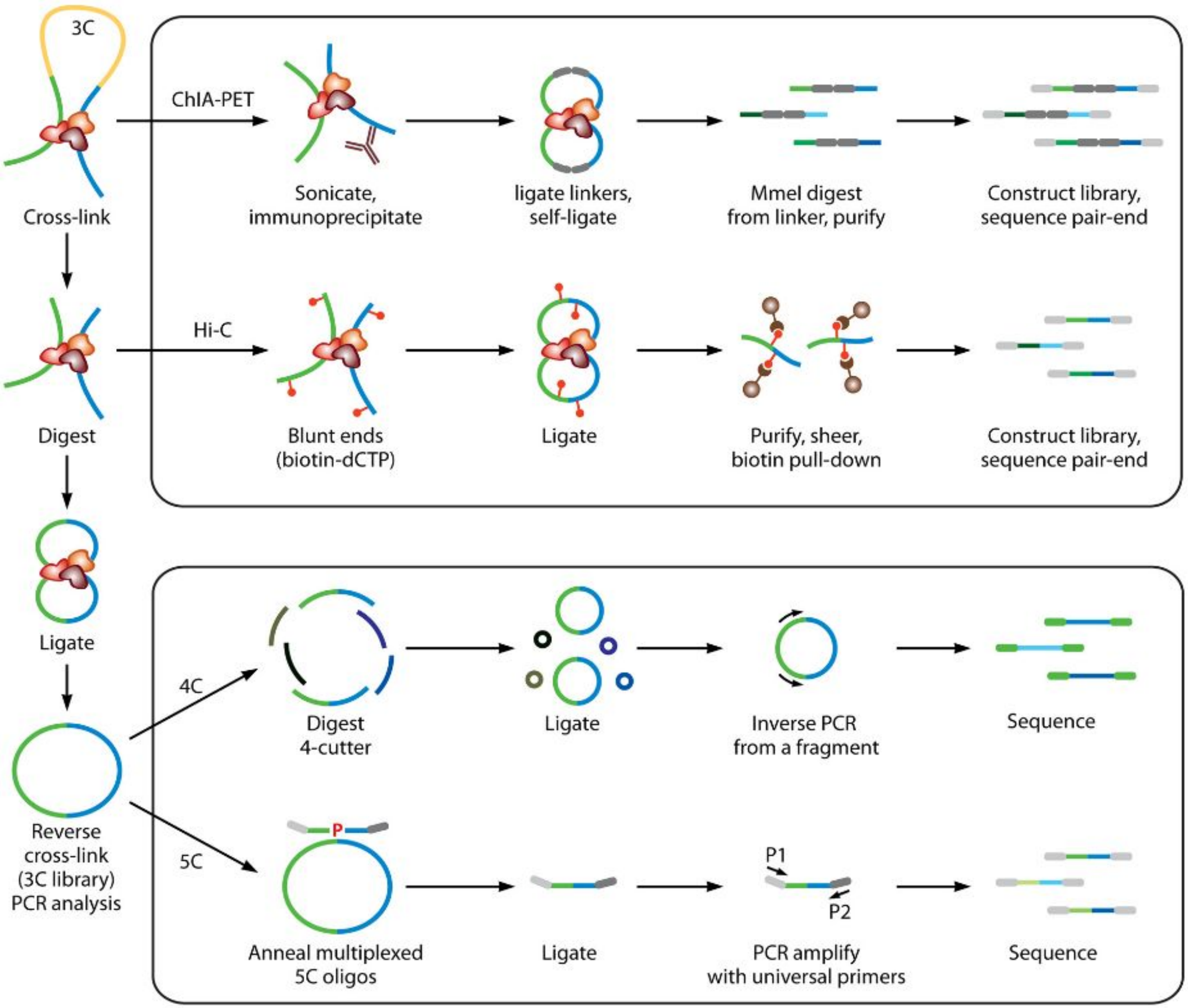
James Fraser,[a] Iain Williamson,[b] Wendy A. Bickmore,[b] Josée Dostie[a]

Department of Biochemistry, and Goodman Cancer Research Center, McGill University, Montréal, Québec, Canada[a]; MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom[b]

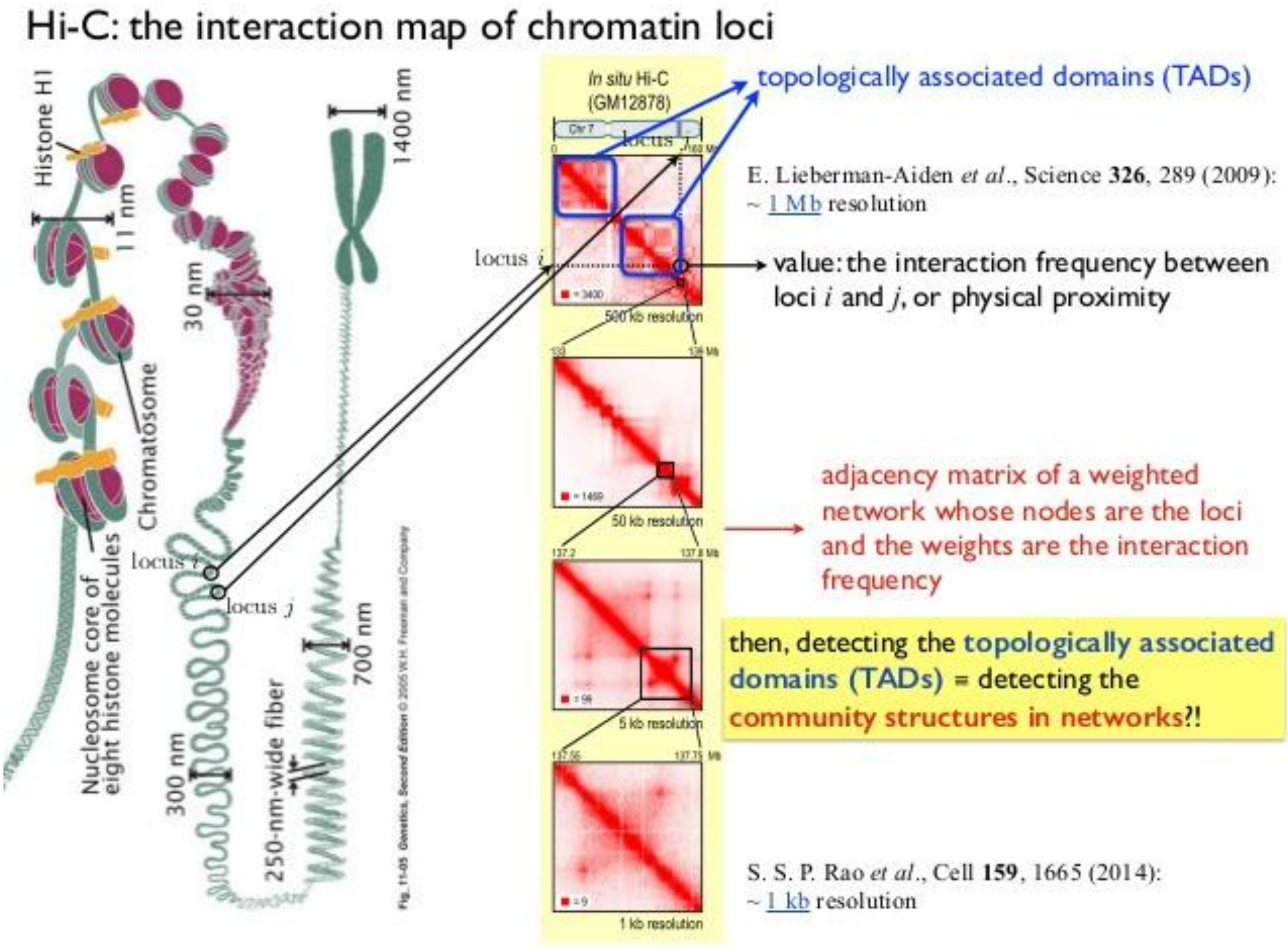
Genome 3D organization



3D seq methods - 3C, HiC


3C
Cross-link
ChIA-PET
Sonicate, immunoprecipitate
ligate linkers, self-ligate
MmeI digest from linker, purify
Construct library, sequence pair-end
Digest
Hi-C
Blunt ends (biotin-dCTP)
Ligate
Purify, sheer, biotin pull-down
Construct library, sequence pair-end
Ligate
Reverse cross-link (3C library) PCR analysis
4C
Digest 4-cutter
Ligate
Inverse PCR from a fragment
Sequence
5C
P
Anneal multiplexed 5C oligos
Ligate
P1
P2
PCR amplify with universal primers
Sequence

3D seq methods - 3C, HiC



Hi-C: the interaction map of chromatin loci

topologically associated domains (TADs)

E. Lieberman-Aiden *et al*., Science **326**, 289 (2009): ~ 1 Mb resolution

value: the interaction frequency between loci i and j, or physical proximity

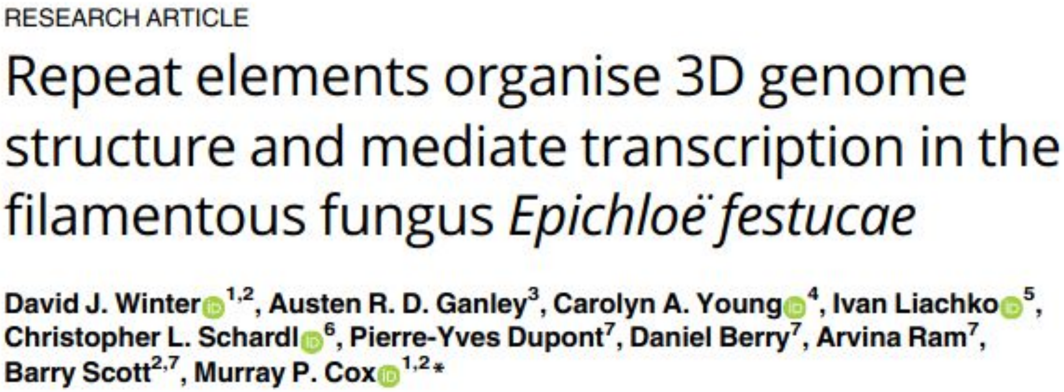
adjacency matrix of a weighted network whose nodes are the loci and the weights are the interaction frequency

then, detecting the topologically associated domains (TADs) = detecting the community structures in networks?!

S. S. P. Rao *et al*., Cell **159**, 1665 (2014): ~ 1 kb resolution

3D Genome: do repeats matter?

3D Genome: do repeats matter?



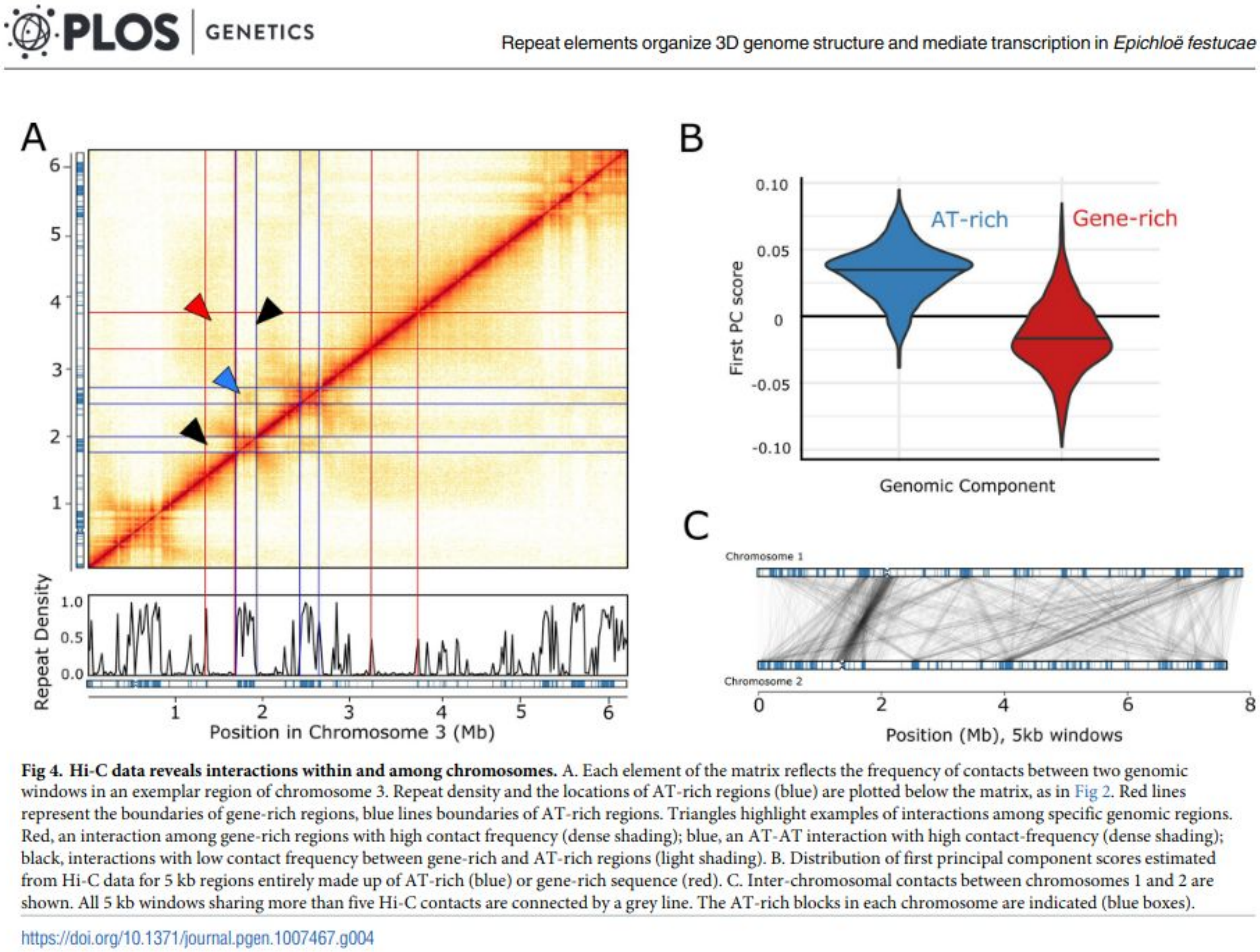
RESEARCH ARTICLE

Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus *Epichloë festucae*

David J. Winter[1,2], Austen R. D. Ganley[3], Carolyn A. Young[4], Ivan Liachko[5], Christopher L. Schardl[6], Pierre-Yves Dupont[7], Daniel Berry[7], Arvina Ram[7], Barry Scott[2,7], Murray P. Cox[1,2]*

"Our results reveal a genome in which very repeat-rich blocks of DNA with discrete boundaries are interspersed by gene-rich sequences that are almost repeat-free. In contrast to other species reported to date, the three-dimensional structure of the genome is anchored by these repeat blocks, which act to isolate transcription in neighbouring gene-rich regions. Genes that are differentially expressed in planta are enriched near the boundaries of these repeat-rich blocks, suggesting that their three-dimensional orientation partly encodes and regulates the symbiotic relationship formed by this organism."

3D seq methods - 3C, HiC



PLOS GENETICS

Repeat elements organize 3D genome structure and mediate transcription in *Epichloë festucae*

Fig 4. Hi-C data reveals interactions within and among chromosomes. A. Each element of the matrix reflects the frequency of contacts between two genomic windows in an exemplar region of chromosome 3. Repeat density and the locations of AT-rich regions (blue) are plotted below the matrix, as in Fig 2. Red lines represent the boundaries of gene-rich regions, blue lines boundaries of AT-rich regions. Triangles highlight examples of interactions among specific genomic regions. Red, an interaction among gene-rich regions with high contact frequency (dense shading); blue, an AT-AT interaction with high contact-frequency (dense shading); black, interactions with low contact frequency between gene-rich and AT-rich regions (light shading). B. Distribution of first principal component scores estimated from Hi-C data for 5 kb regions entirely made up of AT-rich (blue) or gene-rich sequence (red). C. Inter-chromosomal contacts between chromosomes 1 and 2 are shown. All 5 kb windows sharing more than five Hi-C contacts are connected by a grey line. The AT-rich blocks in each chromosome are indicated (blue boxes).

https://doi.org/10.1371/journal.pgen.1007467.g004

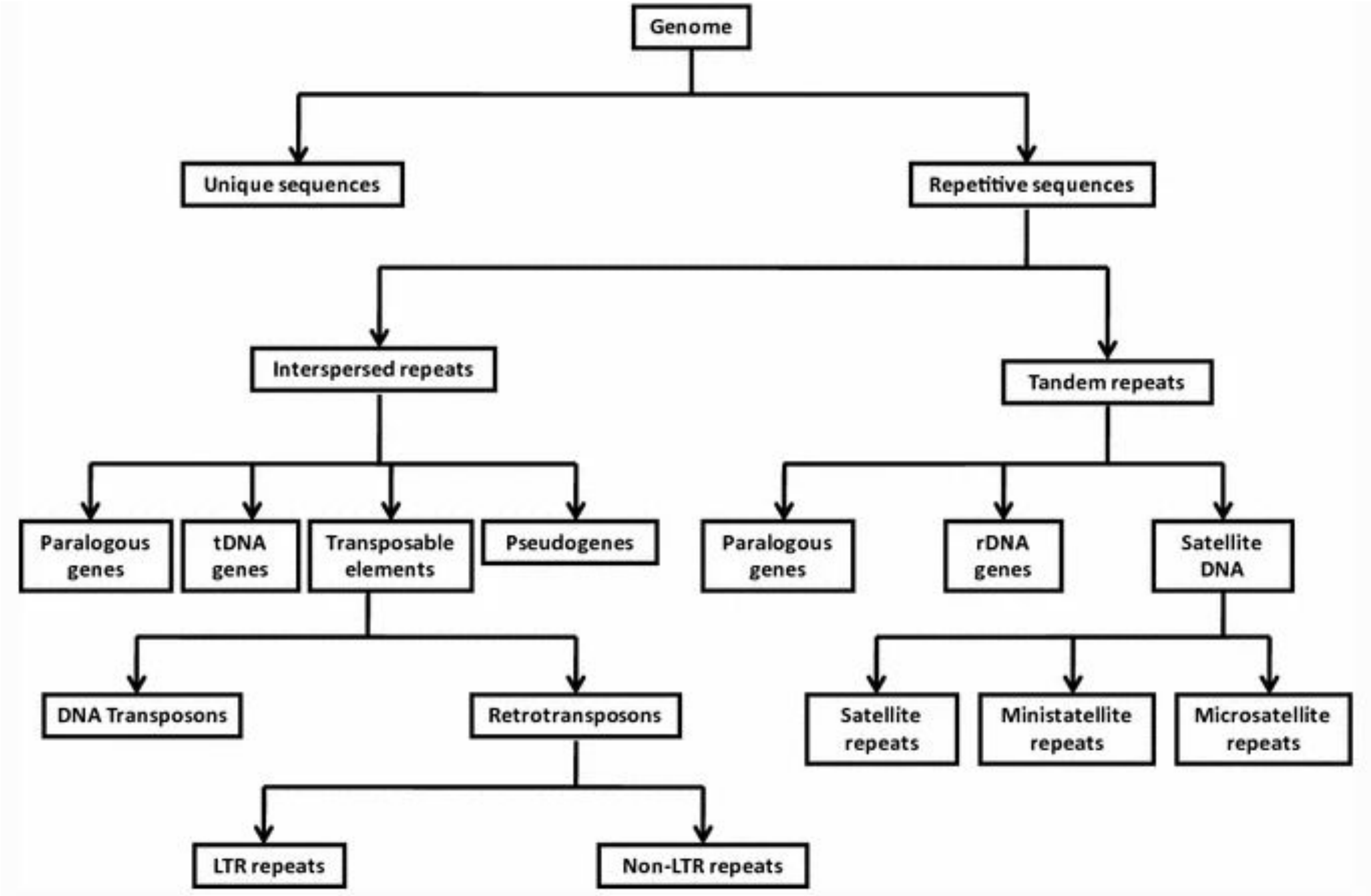
Questions/hypotheses

Can repeat contribution to interphase genome 3D organization be seen in HiC data from plants?

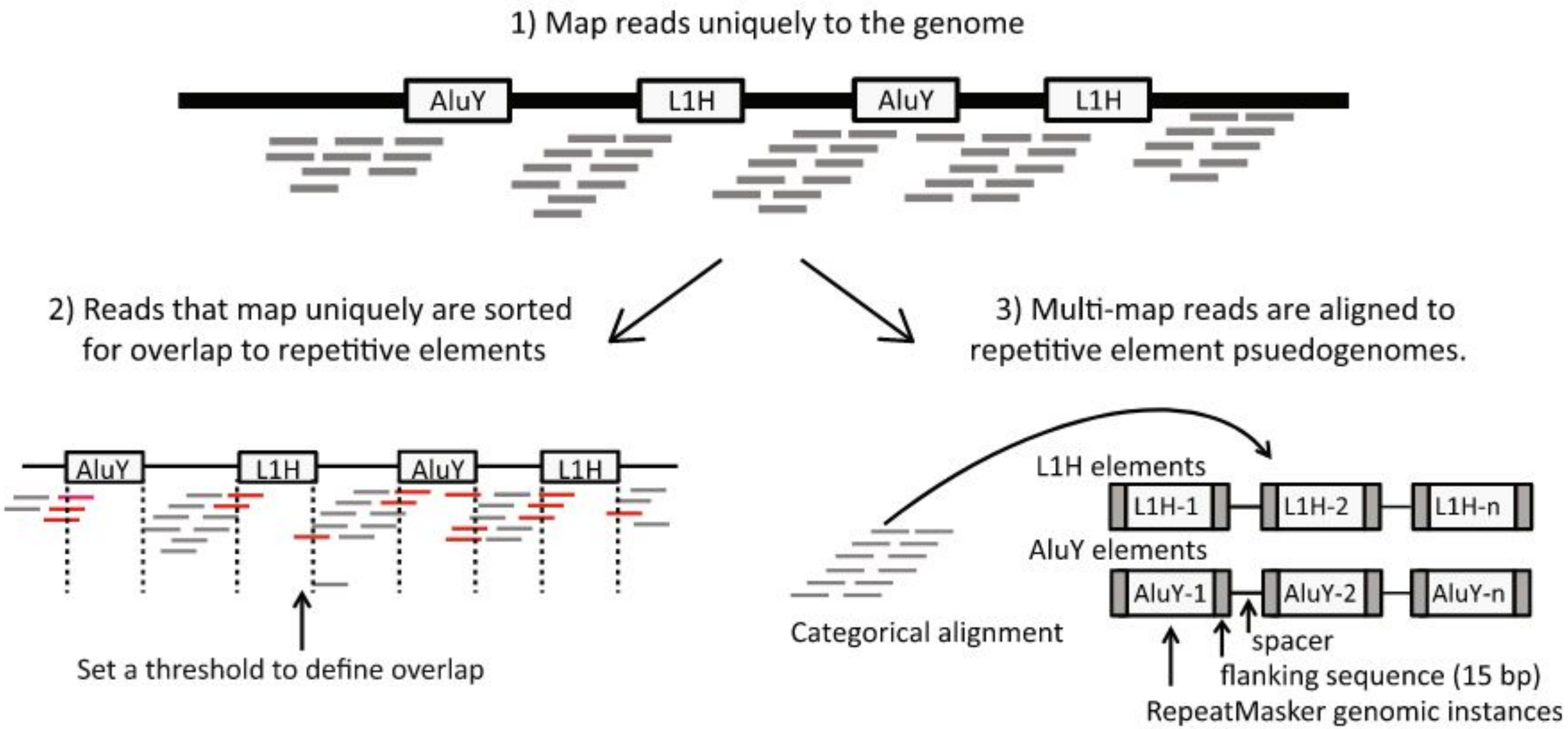
Which families of repeats may have such function?

Where in the genome are these effects concentrated?

Unique/multimapping reads



Unique/multimapping reads



1) Map reads uniquely to the genome

2) Reads that map uniquely are sorted for overlap to repetitive elements

3) Multi-map reads are aligned to repetitive element psuedogenomes.

Figure 1 ***RepEnrich*** **read mapping strategy.** Reads are mapped to the genome using the *Bowtie1* aligner. Reads mapping uniquely to the genome are assigned to subfamilies of repetitive elements based on their degree of overlap to *RepeatMasker* annotated genomic instances of each repetitive element subfamily. Reads mapping to multiple locations are separately mapped to repetitive element assemblies – referred to as repetitive element psuedogenomes – built from *RepeatMasker* annotated genomic instances of repetitive element subfamilies.

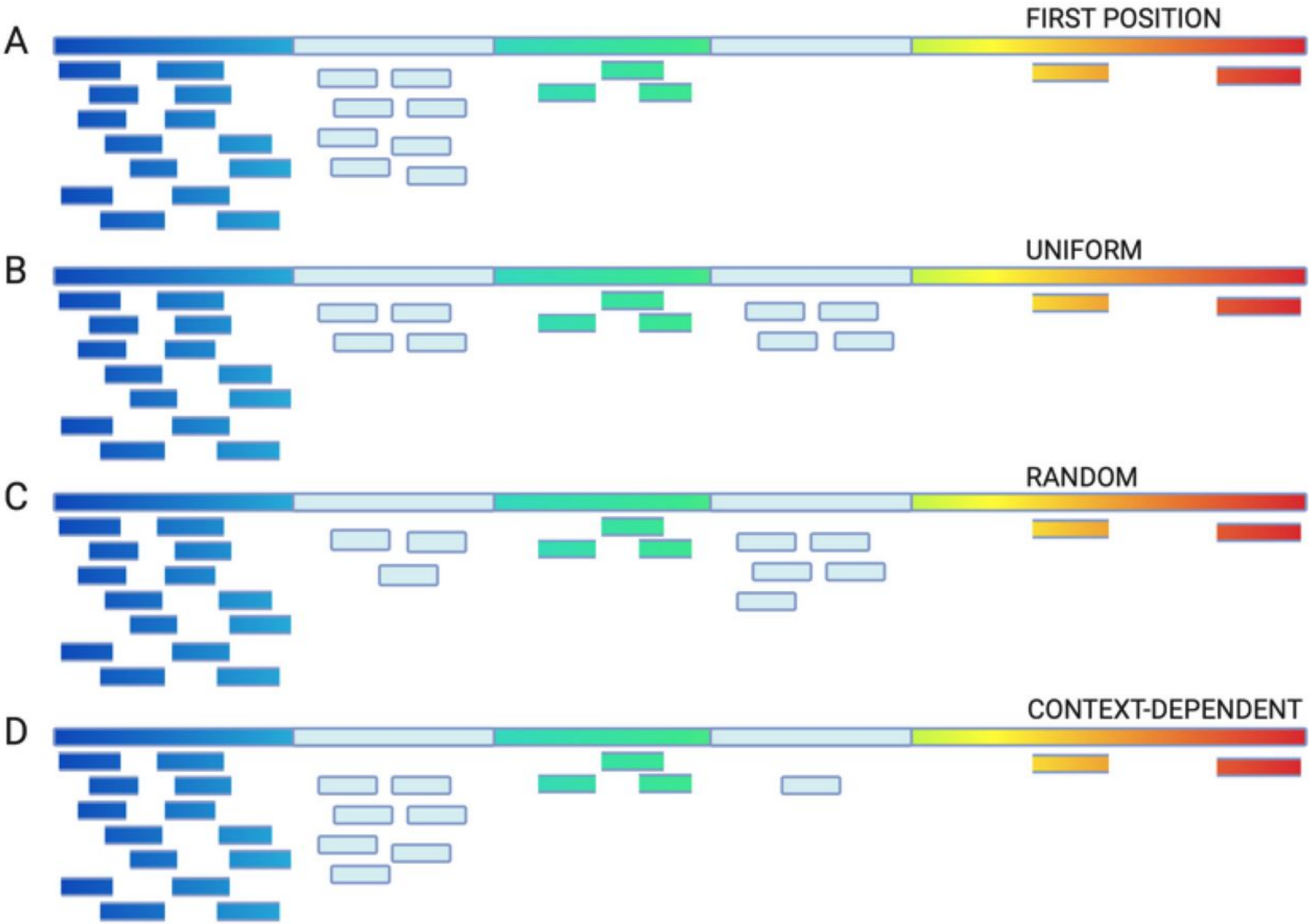
BMC Genomics. 2014; 15: 583.
Published online 2014 Jul 11. doi: 10.1186/1471-2164-15-583

PMCID: PMC4122776
PMID: 25012247

Transcriptional landscape of repetitive elements in normal and cancer human cells

Steven W Criscione,[1] Yue Zhang,[1] William Thompson,[2,3] John M Sedivy,[1] and Nicola Neretti[1,3]

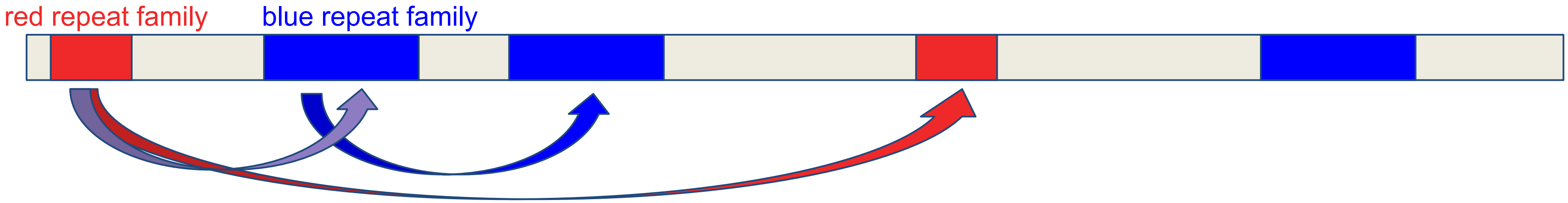
Unique/multimapping reads



Cechova, 2020

How to find contacts between repeat families?

- Mapping to the reference



For red and blue repeat families, the Hi-C read pairs can be formed between:

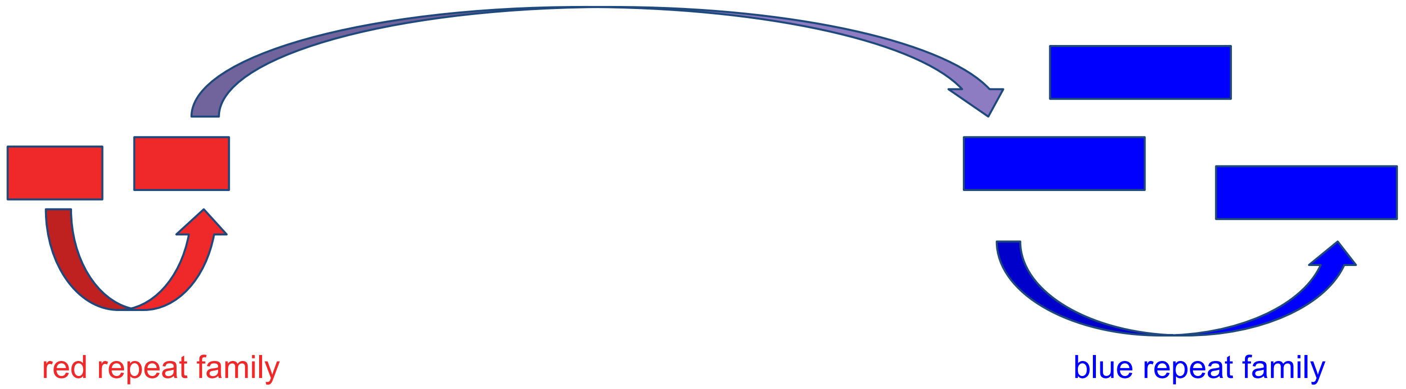
- ➢ red and red family
- ➢ red and blue family
- ➢ blue and blue family

Mapping to the reference brings challenges:

- ❖ how to deal with multi-mapping reads?
- ❖ imperfect representation of repeats in reference genomes

RepeatExplorer

- Find read pairs within and between clusters
- Label each cluster
- No reference necessary



=> complementary method to the reference-based approach

Do we see more contacts than expected?

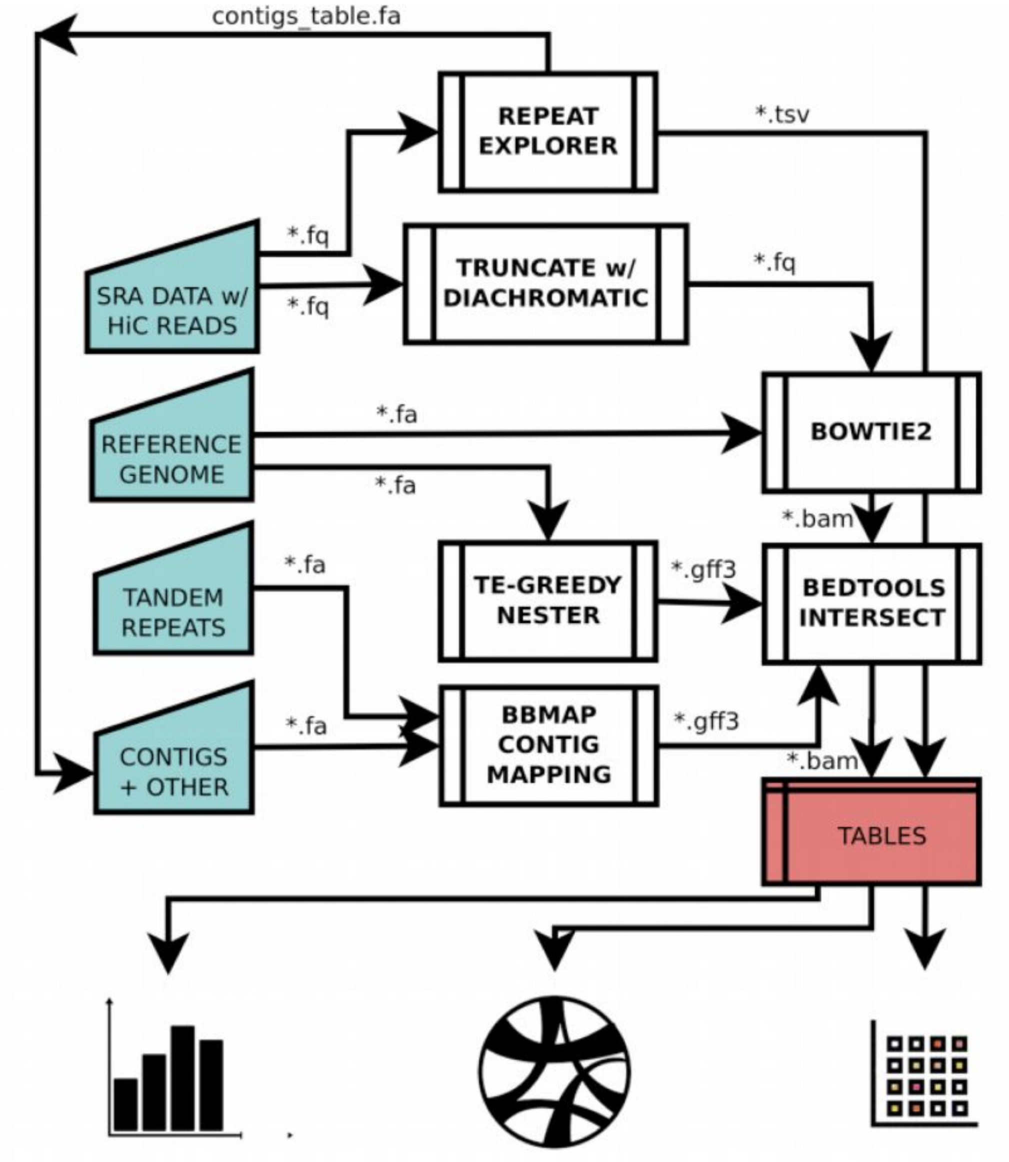
Three methods: **observed versus expected counts of read pairs**

- joint probability
  - expected counts are p(A,B) = p(A).p(B)

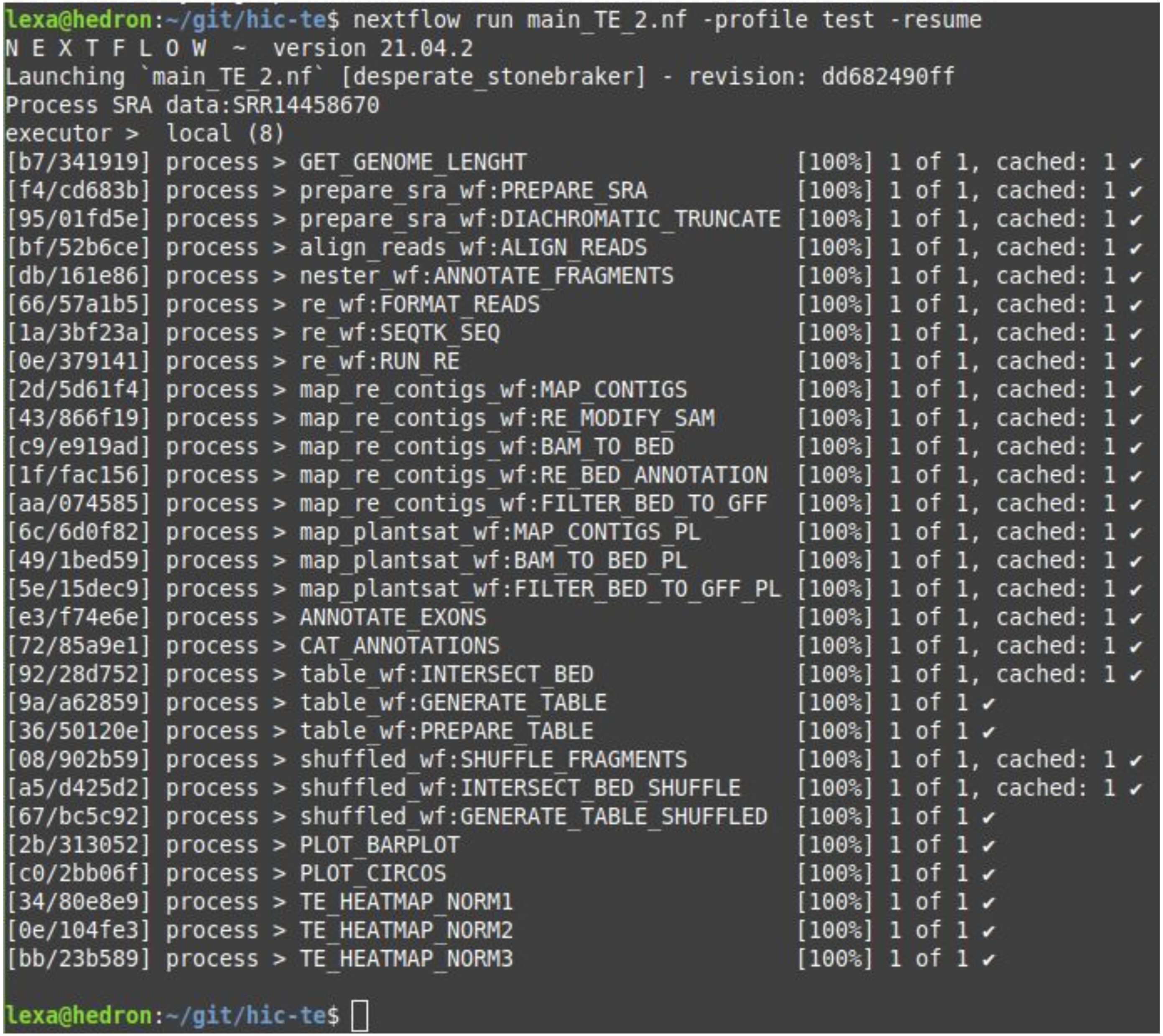
- label permutation
  - each row represents a read pair with annotations of family1 and family2, family names are assigned to random rows of the table

- annotation shuffling
  - annotation of repeat families are shuffled along the reference genome

HiC-TE pipeline



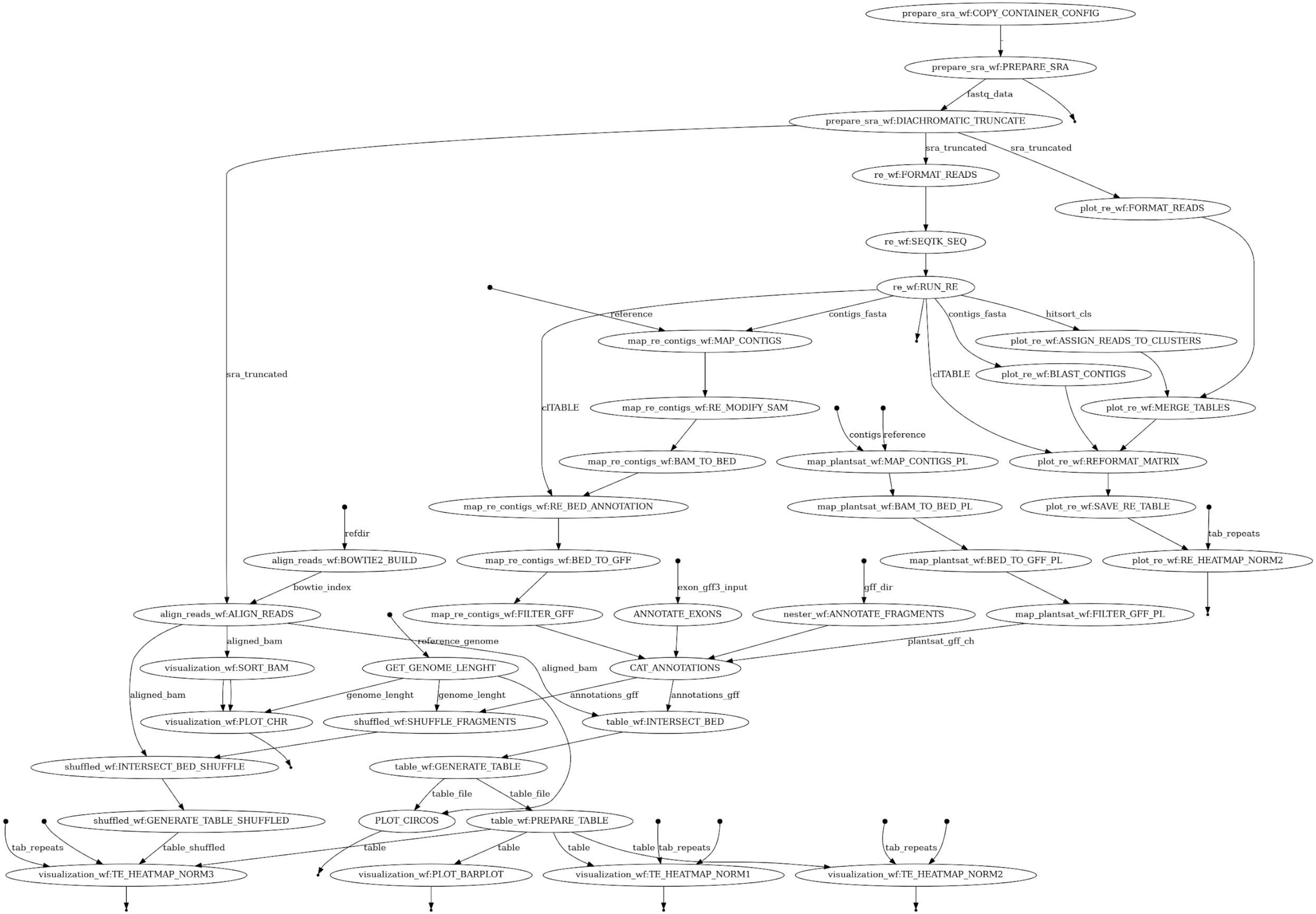
HiC-TE pipeline



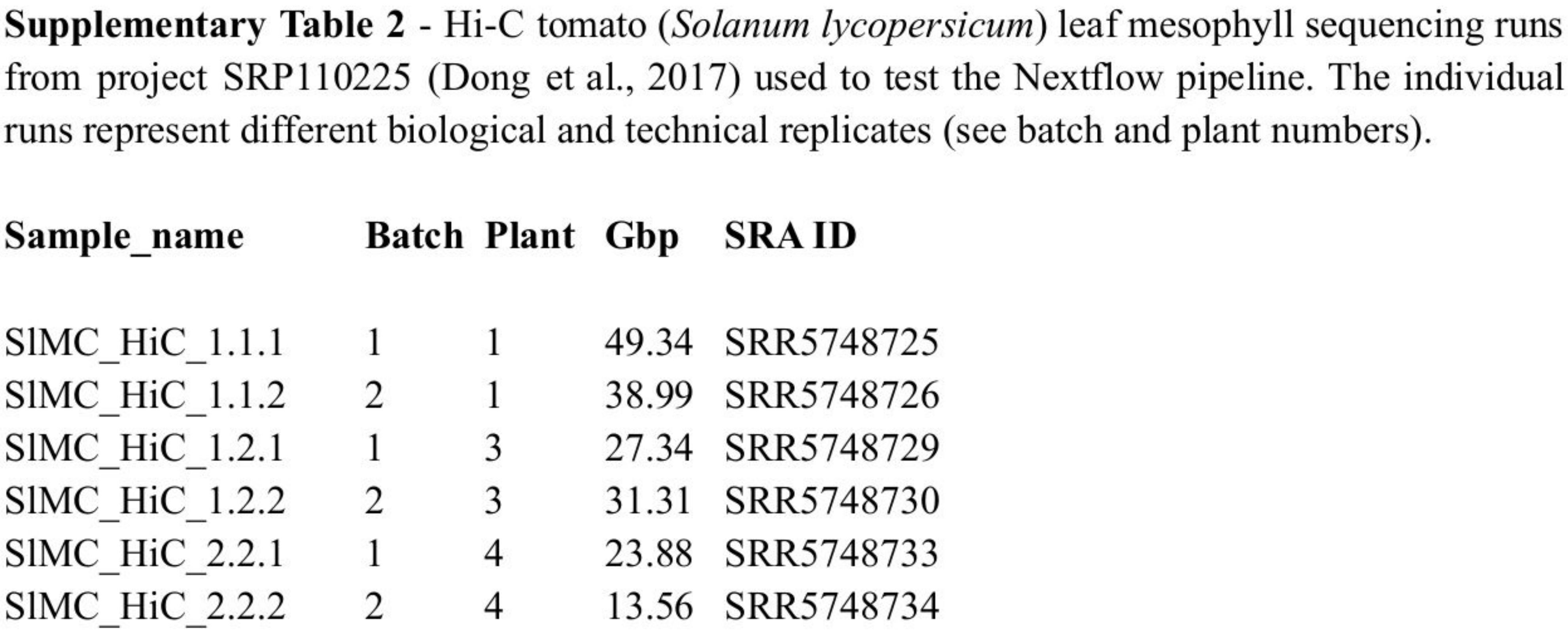
```
lexa@hedron:~/git/hic-te$ nextflow run main_TE_2.nf -profile test -resume
N E X T F L O W  ~  version 21.04.2
Launching `main_TE_2.nf` [desperate_stonebraker] - revision: dd682490ff
Process SRA data:SRR14458670
executor >  local (8)
[b7/341919] process > GET_GENOME_LENGHT                    [100%] 1 of 1, cached: 1 ✔
[f4/cd683b] process > prepare_sra_wf:PREPARE_SRA           [100%] 1 of 1, cached: 1 ✔
[95/01fd5e] process > prepare_sra_wf:DIACHROMATIC_TRUNCATE [100%] 1 of 1, cached: 1 ✔
[bf/52b6ce] process > align_reads_wf:ALIGN_READS           [100%] 1 of 1, cached: 1 ✔
[db/161e86] process > nester_wf:ANNOTATE_FRAGMENTS         [100%] 1 of 1, cached: 1 ✔
[66/57a1b5] process > re_wf:FORMAT_READS                   [100%] 1 of 1, cached: 1 ✔
[1a/3bf23a] process > re_wf:SEQTK_SEQ                      [100%] 1 of 1, cached: 1 ✔
[0e/379141] process > re_wf:RUN_RE                         [100%] 1 of 1, cached: 1 ✔
[2d/5d61f4] process > map_re_contigs_wf:MAP_CONTIGS        [100%] 1 of 1, cached: 1 ✔
[43/866f19] process > map_re_contigs_wf:RE_MODIFY_SAM      [100%] 1 of 1, cached: 1 ✔
[c9/e919ad] process > map_re_contigs_wf:BAM_TO_BED         [100%] 1 of 1, cached: 1 ✔
[1f/fac156] process > map_re_contigs_wf:RE_BED_ANNOTATION  [100%] 1 of 1, cached: 1 ✔
[aa/074585] process > map_re_contigs_wf:FILTER_BED_TO_GFF  [100%] 1 of 1, cached: 1 ✔
[6c/6d0f82] process > map_plantsat_wf:MAP_CONTIGS_PL       [100%] 1 of 1, cached: 1 ✔
[49/1bed59] process > map_plantsat_wf:BAM_TO_BED_PL        [100%] 1 of 1, cached: 1 ✔
[5e/15dec9] process > map_plantsat_wf:FILTER_BED_TO_GFF_PL [100%] 1 of 1, cached: 1 ✔
[e3/f74e6e] process > ANNOTATE_EXONS                       [100%] 1 of 1, cached: 1 ✔
[72/85a9e1] process > CAT_ANNOTATIONS                      [100%] 1 of 1, cached: 1 ✔
[92/28d752] process > table_wf:INTERSECT_BED               [100%] 1 of 1, cached: 1 ✔
[9a/a62859] process > table_wf:GENERATE_TABLE              [100%] 1 of 1 ✔
[36/50120e] process > table_wf:PREPARE_TABLE               [100%] 1 of 1 ✔
[08/902b59] process > shuffled_wf:SHUFFLE_FRAGMENTS        [100%] 1 of 1, cached: 1 ✔
[a5/d425d2] process > shuffled_wf:INTERSECT_BED_SHUFFLE    [100%] 1 of 1, cached: 1 ✔
[67/bc5c92] process > shuffled_wf:GENERATE_TABLE_SHUFFLED  [100%] 1 of 1 ✔
[2b/313052] process > PLOT_BARPLOT                         [100%] 1 of 1 ✔
[c0/2bb06f] process > PLOT_CIRCOS                          [100%] 1 of 1 ✔
[34/80e8e9] process > TE_HEATMAP_NORM1                     [100%] 1 of 1 ✔
[0e/104fe3] process > TE_HEATMAP_NORM2                     [100%] 1 of 1 ✔
[bb/23b589] process > TE_HEATMAP_NORM3                     [100%] 1 of 1 ✔

lexa@hedron:~/git/hic-te$ 
```

HiC-TE pipeline


prepare_sra_wf:COPY_CONTAINER_CONFIG
prepare_sra_wf:PREPARE_SRA
fastq_data
prepare_sra_wf:DIACHROMATIC_TRUNCATE
sra_truncated
sra_truncated
re_wf:FORMAT_READS
plot_re_wf:FORMAT_READS
re_wf:SEQTK_SEQ
re_wf:RUN_RE
reference
contigs_fasta
contigs_fasta
hitsort_cls
map_re_contigs_wf:MAP_CONTIGS
plot_re_wf:ASSIGN_READS_TO_CLUSTERS
sra_truncated
cITABLE
plot_re_wf:BLAST_CONTIGS
cITABLE
map_re_contigs_wf:RE_MODIFY_SAM
plot_re_wf:MERGE_TABLES
contigs
reference
map_re_contigs_wf:BAM_TO_BED
map_plantsat_wf:MAP_CONTIGS_PL
plot_re_wf:REFORMAT_MATRIX
map_re_contigs_wf:RE_BED_ANNOTATION
map_plantsat_wf:BAM_TO_BED_PL
plot_re_wf:SAVE_RE_TABLE
refdir
tab_repeats
align_reads_wf:BOWTIE2_BUILD
map_re_contigs_wf:BED_TO_GFF
map_plantsat_wf:BED_TO_GFF_PL
plot_re_wf:RE_HEATMAP_NORM2
bowtie_index
exon_gff3_input
gff_dir
align_reads_wf:ALIGN_READS
map_re_contigs_wf:FILTER_GFF
ANNOTATE_EXONS
nester_wf:ANNOTATE_FRAGMENTS
map_plantsat_wf:FILTER_GFF_PL
aligned_bam
reference_genome
plantsat_gff_ch
visualization_wf:SORT_BAM
GET_GENOME_LENGHT
aligned_bam
CAT_ANNOTATIONS
aligned_bam
genome_lenght
genome_lenght
annotations_gff
annotations_gff
visualization_wf:PLOT_CHR
shuffled_wf:SHUFFLE_FRAGMENTS
table_wf:INTERSECT_BED
shuffled_wf:INTERSECT_BED_SHUFFLE
table_wf:GENERATE_TABLE
table_file
table_file
shuffled_wf:GENERATE_TABLE_SHUFFLED
PLOT_CIRCOS
table_wf:PREPARE_TABLE
tab_repeats
table_shuffled
table
table
table
table
tab_repeats
tab_repeats
visualization_wf:TE_HEATMAP_NORM3
visualization_wf:PLOT_BARPLOT
visualization_wf:TE_HEATMAP_NORM1
visualization_wf:TE_HEATMAP_NORM2

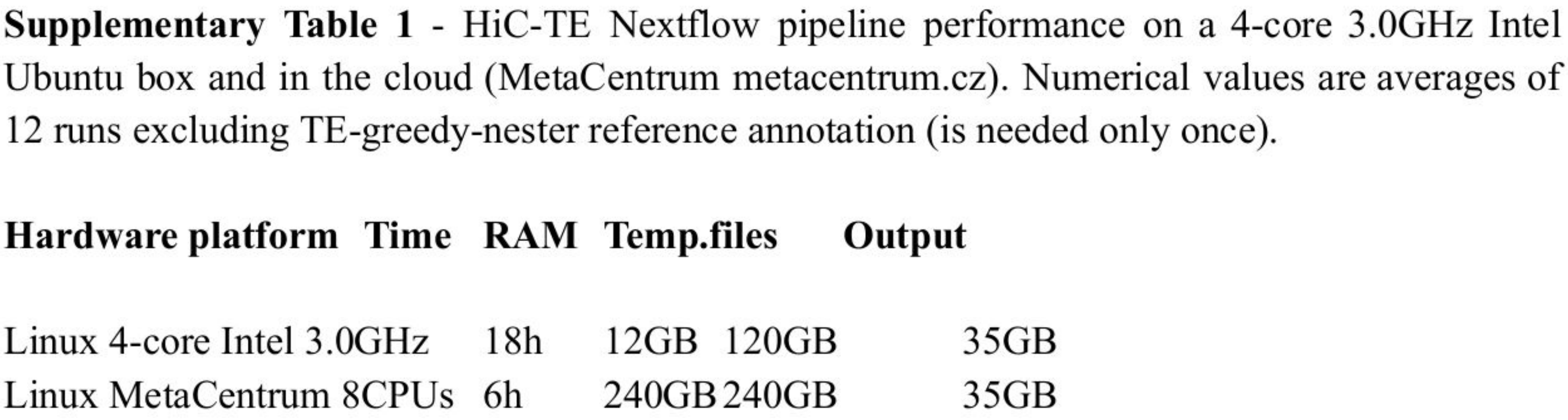
HiC-TE pipeline



Supplementary Table 2 - Hi-C tomato (*Solanum lycopersicum*) leaf mesophyll sequencing runs from project SRP110225 (Dong et al., 2017) used to test the Nextflow pipeline. The individual runs represent different biological and technical replicates (see batch and plant numbers).

Sample_name	**Batch**	**Plant**	**Gbp**	**SRA ID**
SlMC_HiC_1.1.1	1	1	49.34	SRR5748725
SlMC_HiC_1.1.2	2	1	38.99	SRR5748726
SlMC_HiC_1.2.1	1	3	27.34	SRR5748729
SlMC_HiC_1.2.2	2	3	31.31	SRR5748730
SlMC_HiC_2.2.1	1	4	23.88	SRR5748733
SlMC_HiC_2.2.2	2	4	13.56	SRR5748734

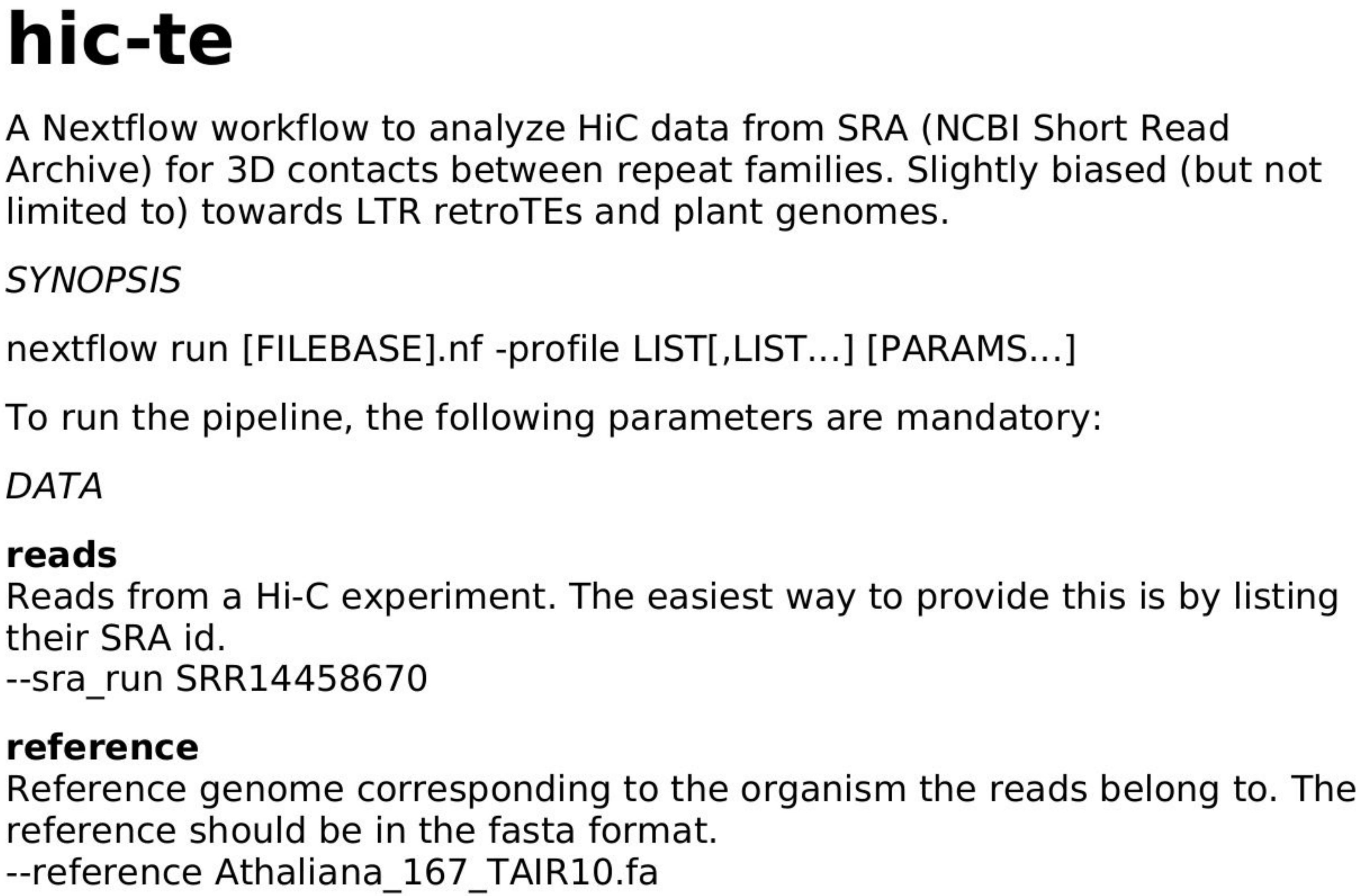
HiC-TE pipeline



Supplementary Table 1 - HiC-TE Nextflow pipeline performance on a 4-core 3.0GHz Intel Ubuntu box and in the cloud (MetaCentrum metacentrum.cz). Numerical values are averages of 12 runs excluding TE-greedy-nester reference annotation (is needed only once).

Hardware platform	Time	RAM	Temp.files	Output
Linux 4-core Intel 3.0GHz	18h	12GB	120GB	35GB
Linux MetaCentrum 8CPUs	6h	240GB	240GB	35GB

HiC-TE pipeline output



hic-te

A Nextflow workflow to analyze HiC data from SRA (NCBI Short Read Archive) for 3D contacts between repeat families. Slightly biased (but not limited to) towards LTR retroTEs and plant genomes.

SYNOPSIS

nextflow run [FILEBASE].nf -profile LIST[,LIST...] [PARAMS...]

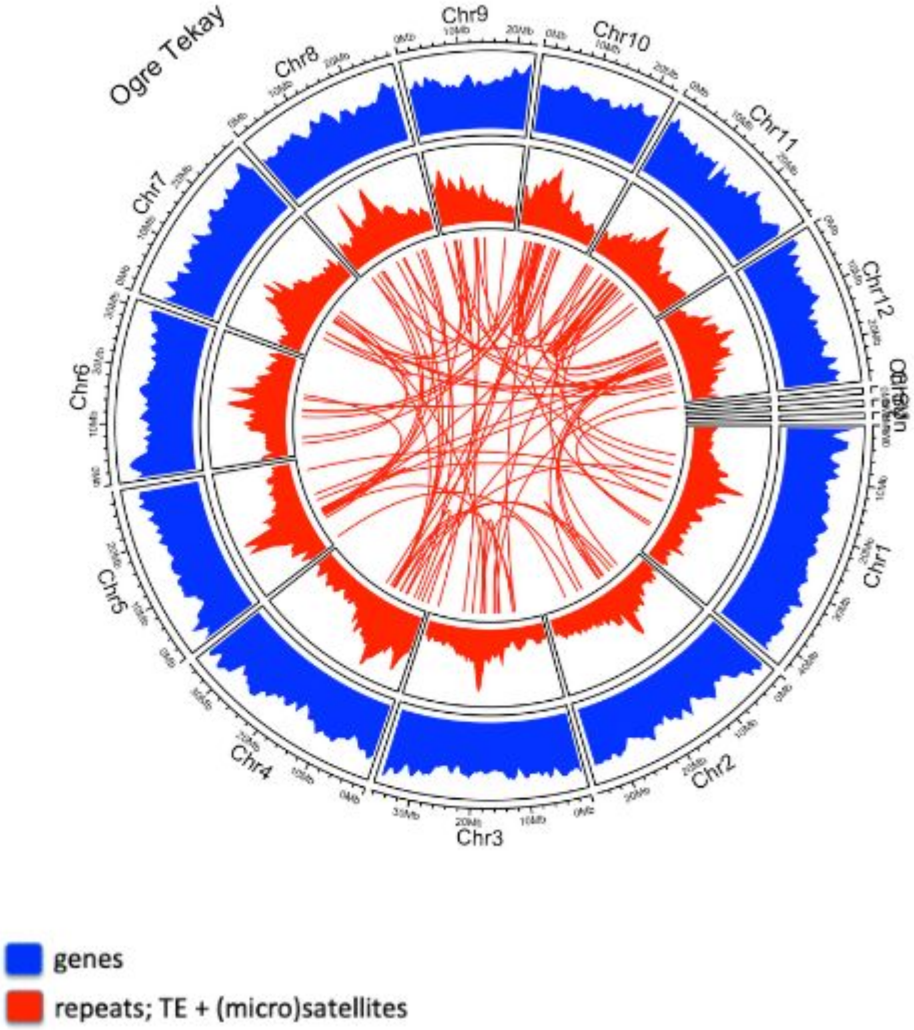
To run the pipeline, the following parameters are mandatory:

DATA

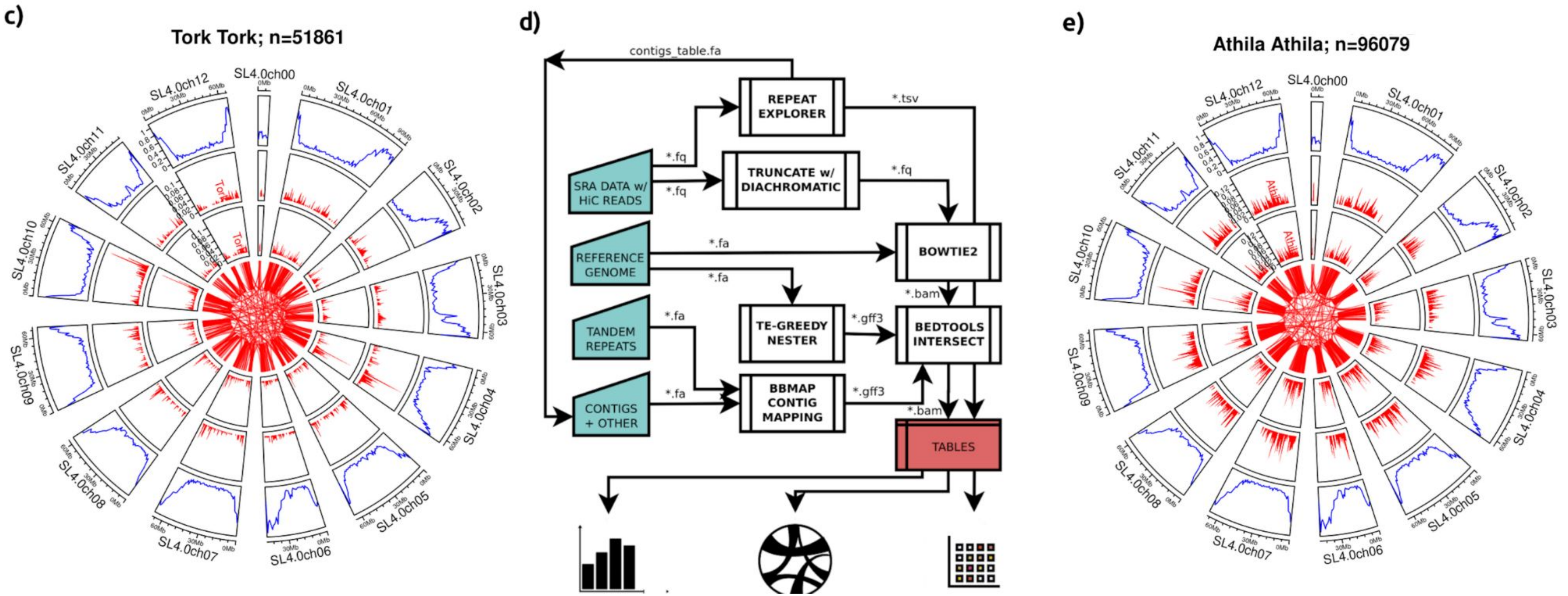
reads
Reads from a Hi-C experiment. The easiest way to provide this is by listing their SRA id.
--sra_run SRR14458670

reference
Reference genome corresponding to the organism the reads belong to. The reference should be in the fasta format.
--reference Athaliana_167_TAIR10.fa

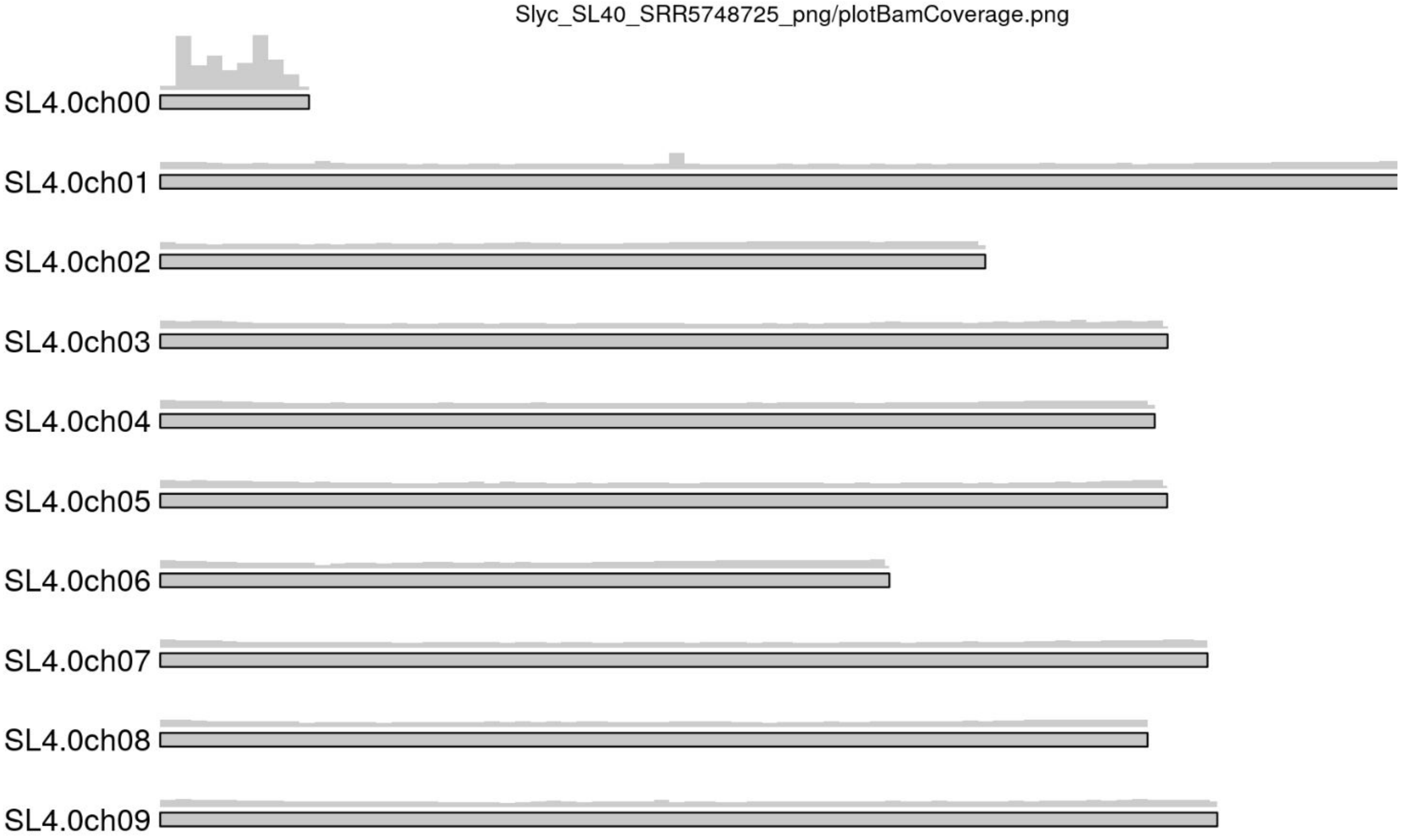
HiC-TE pipeline output



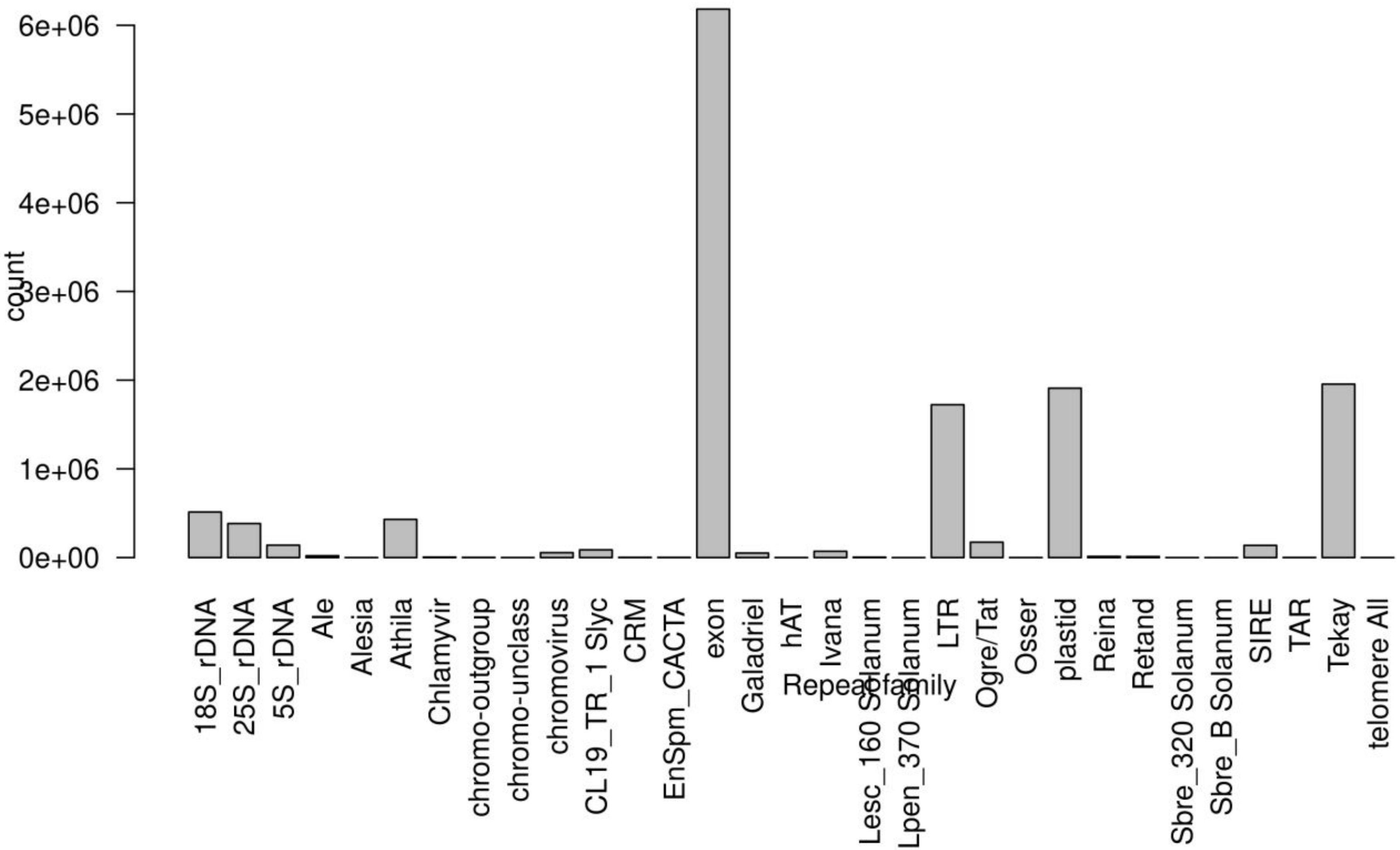
HiC-TE pipeline output



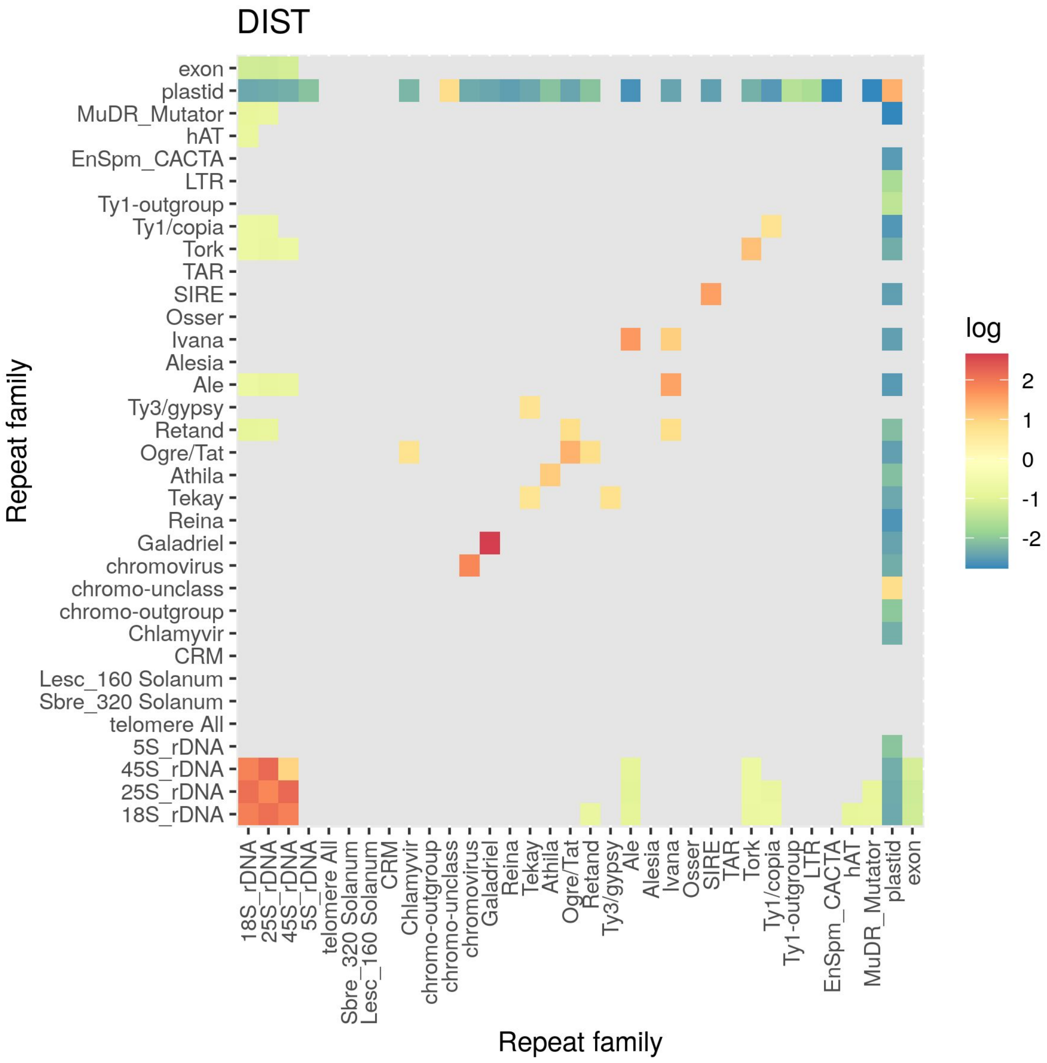
HiC-TE pipeline output



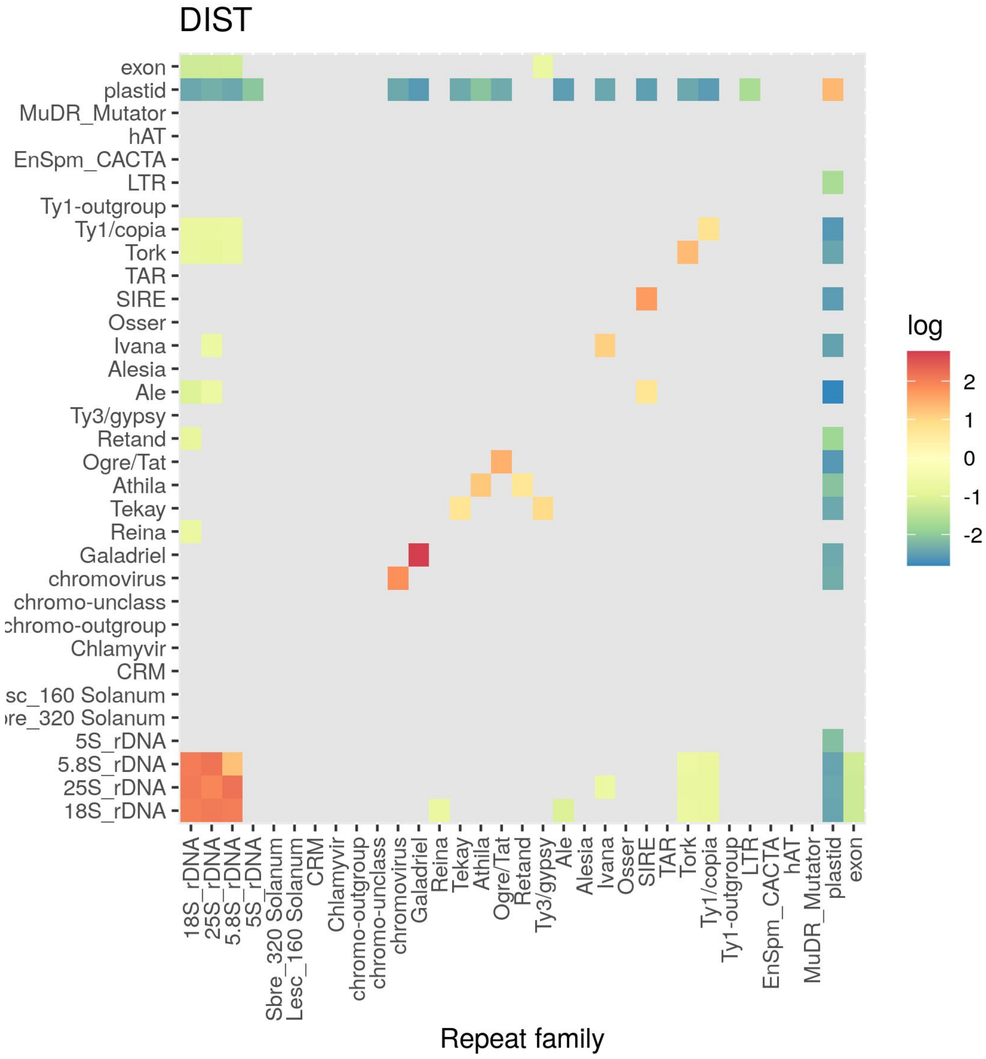
HiC-TE pipeline output



Different plant individual

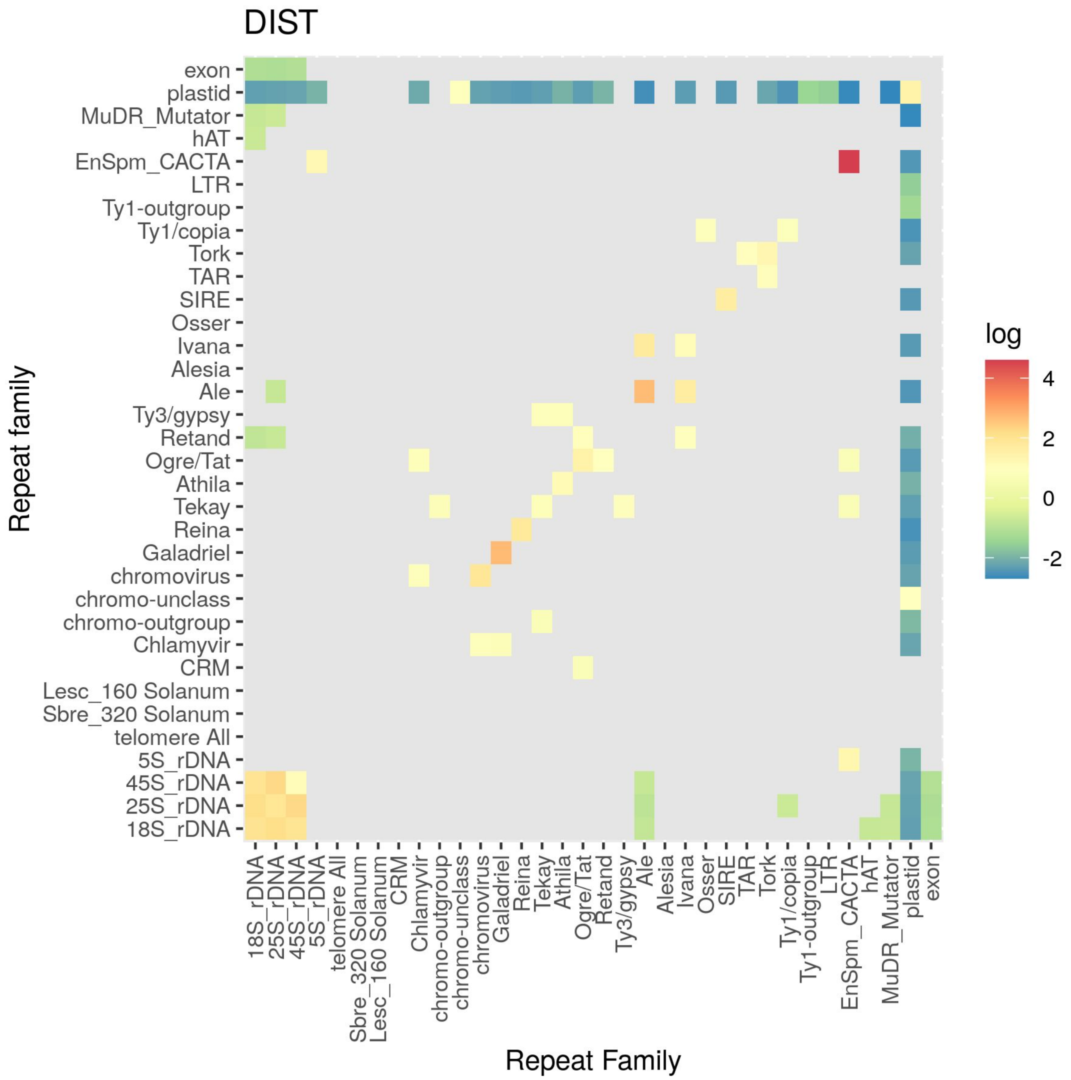


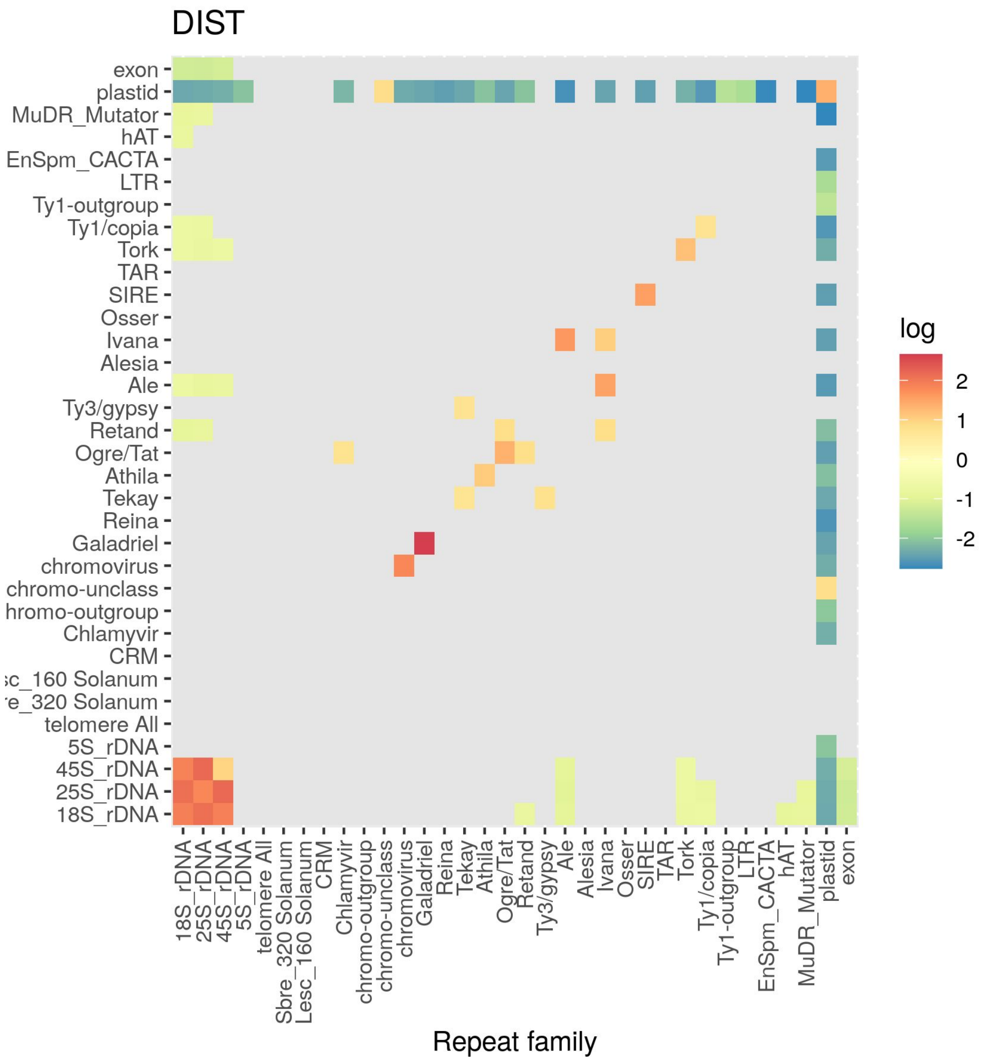
DIST



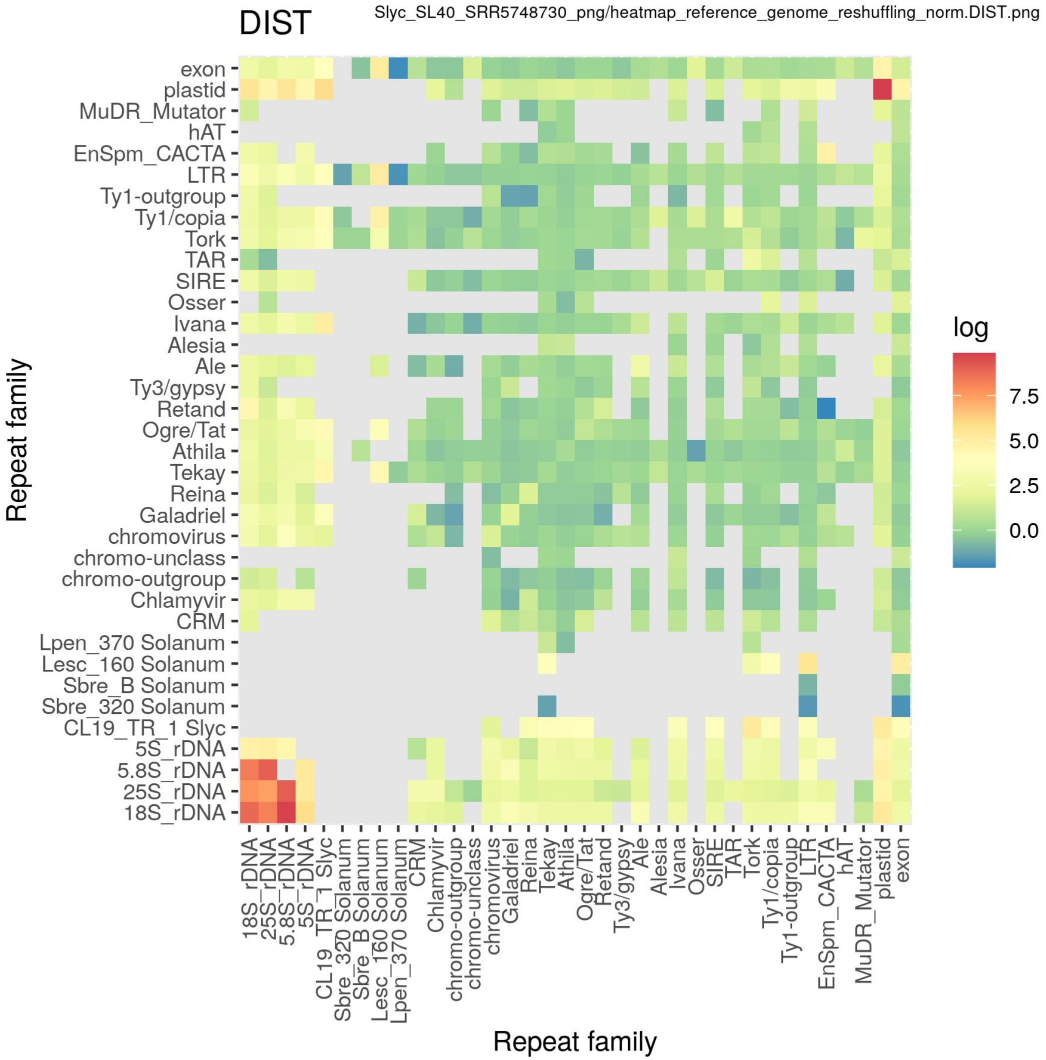
DIST

Different normalization method



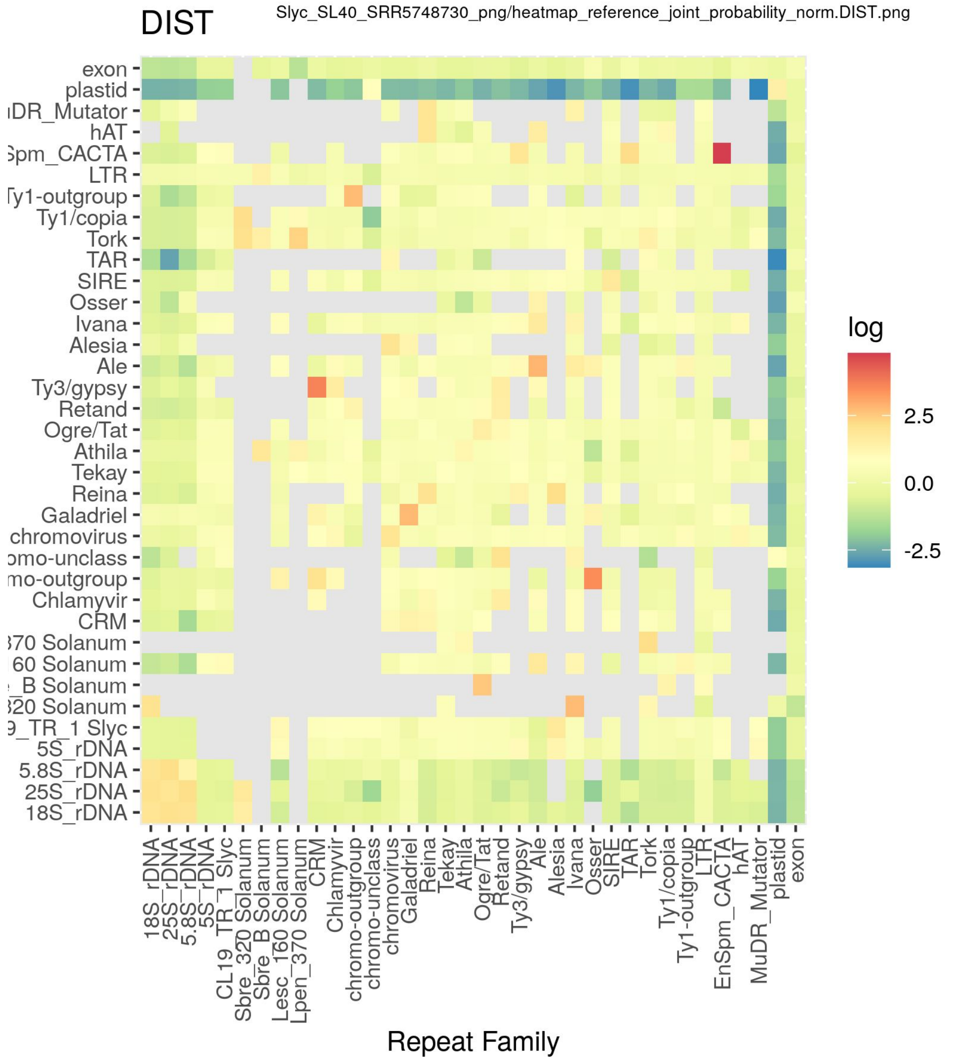


Different normalization method



DIST

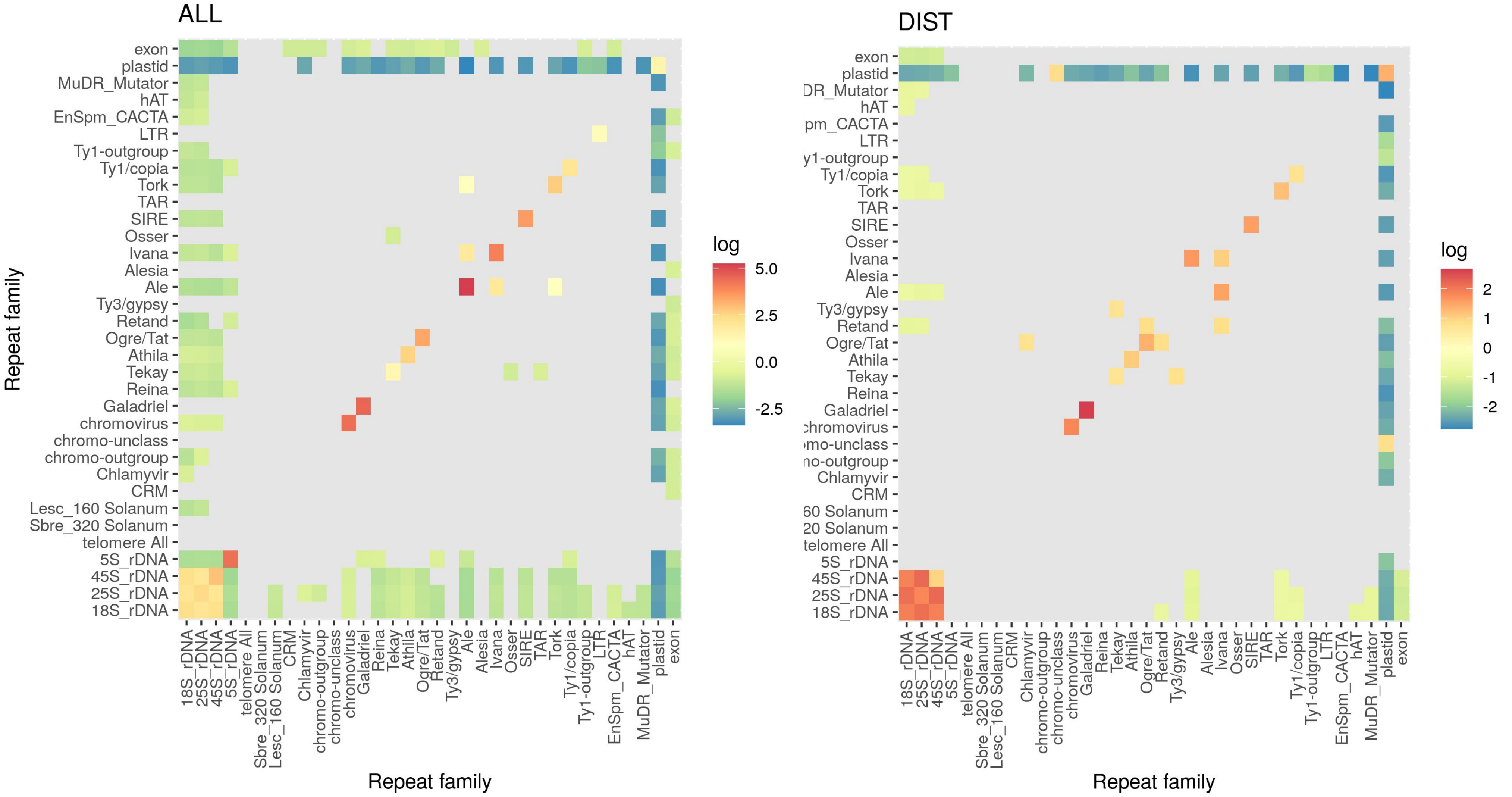
Slyc_SL40_SRR5748730_png/heatmap_reference_genome_reshuffling_norm.DIST.png



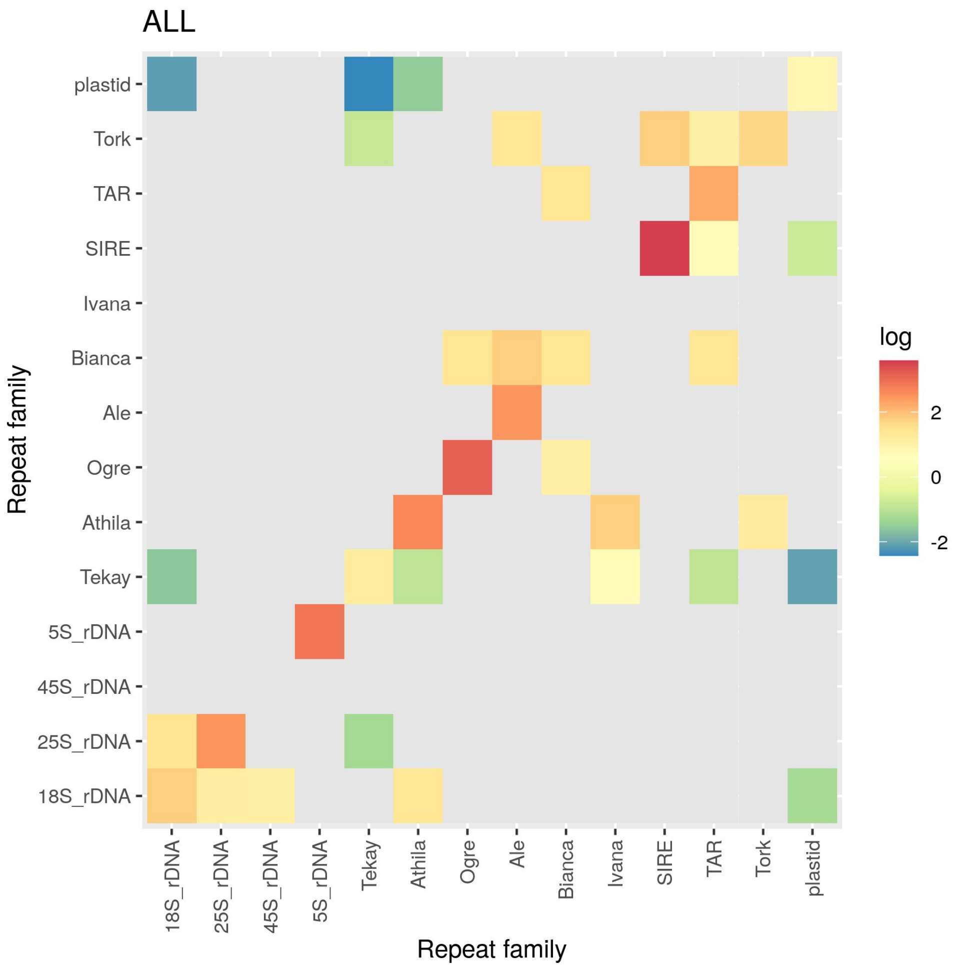
DIST

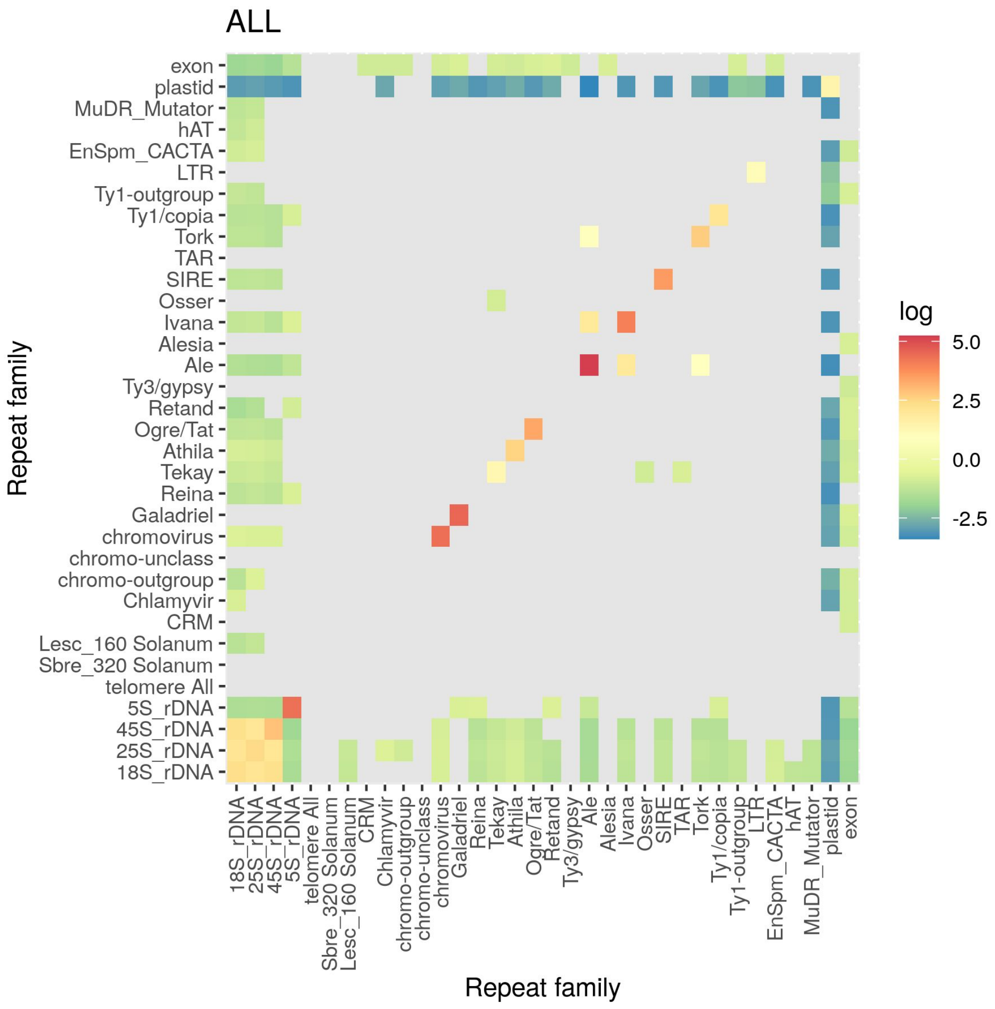
Slyc_SL40_SRR5748730_png/heatmap_reference_joint_probability_norm.DIST.png

All vs long-distance interactions



RepeatExplorer versus reference





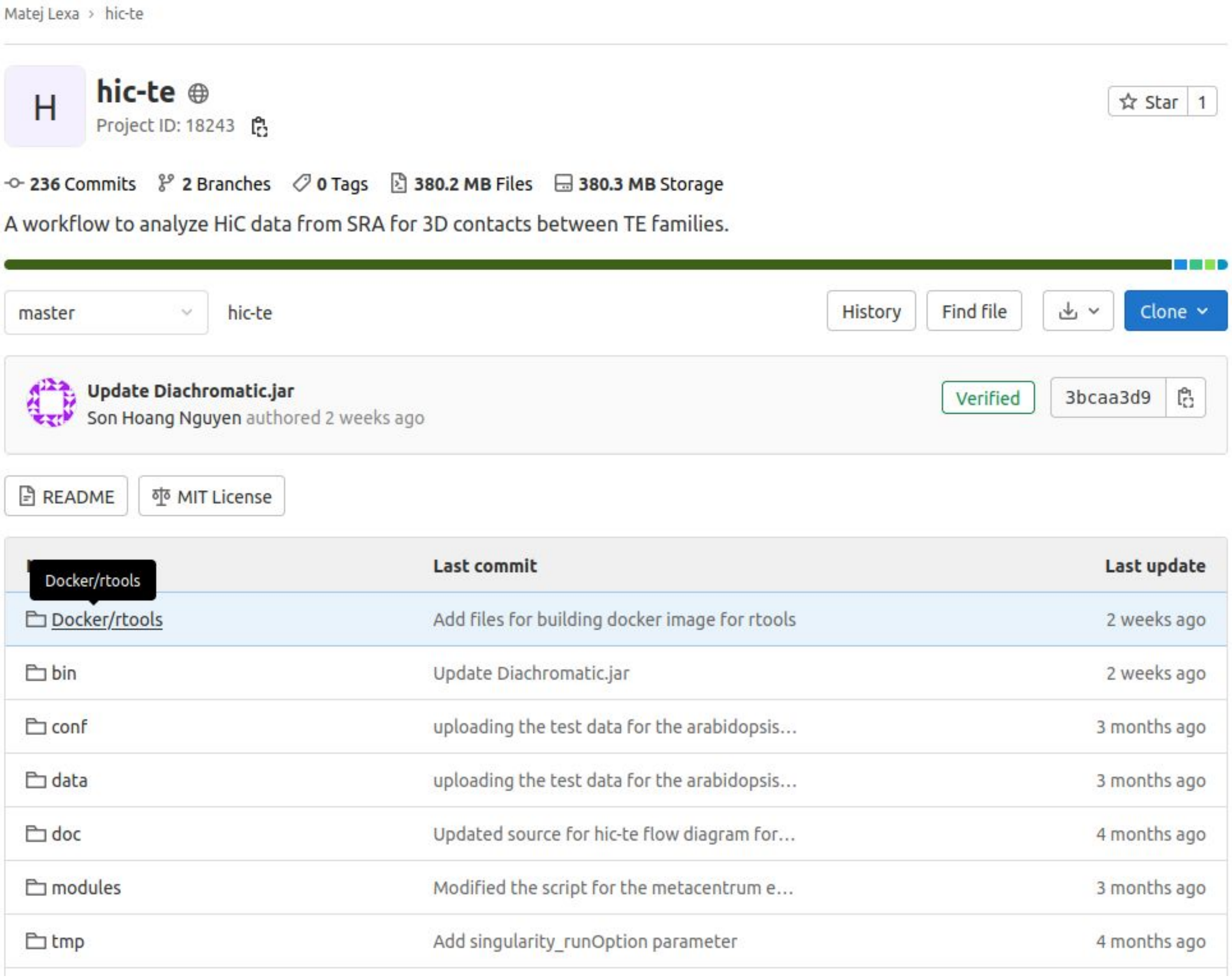
Method matter (when dealing with repeats)

- General patterns stay the same but the individual variation is present depending on the individual, normalization method, repeat dispersion, density etc.

- RepeatExplorer clusters offer unique insights into repeat contacts not available with reference-based methods

- Hi-C TE pipeline can be used to uncover chromatin contacts between repetitive annotations in plant genomes

HiC-TE pipeline contributors



Matej Lexa › hic-te

H **hic-te**

Project ID: 18243

☆ Star 1

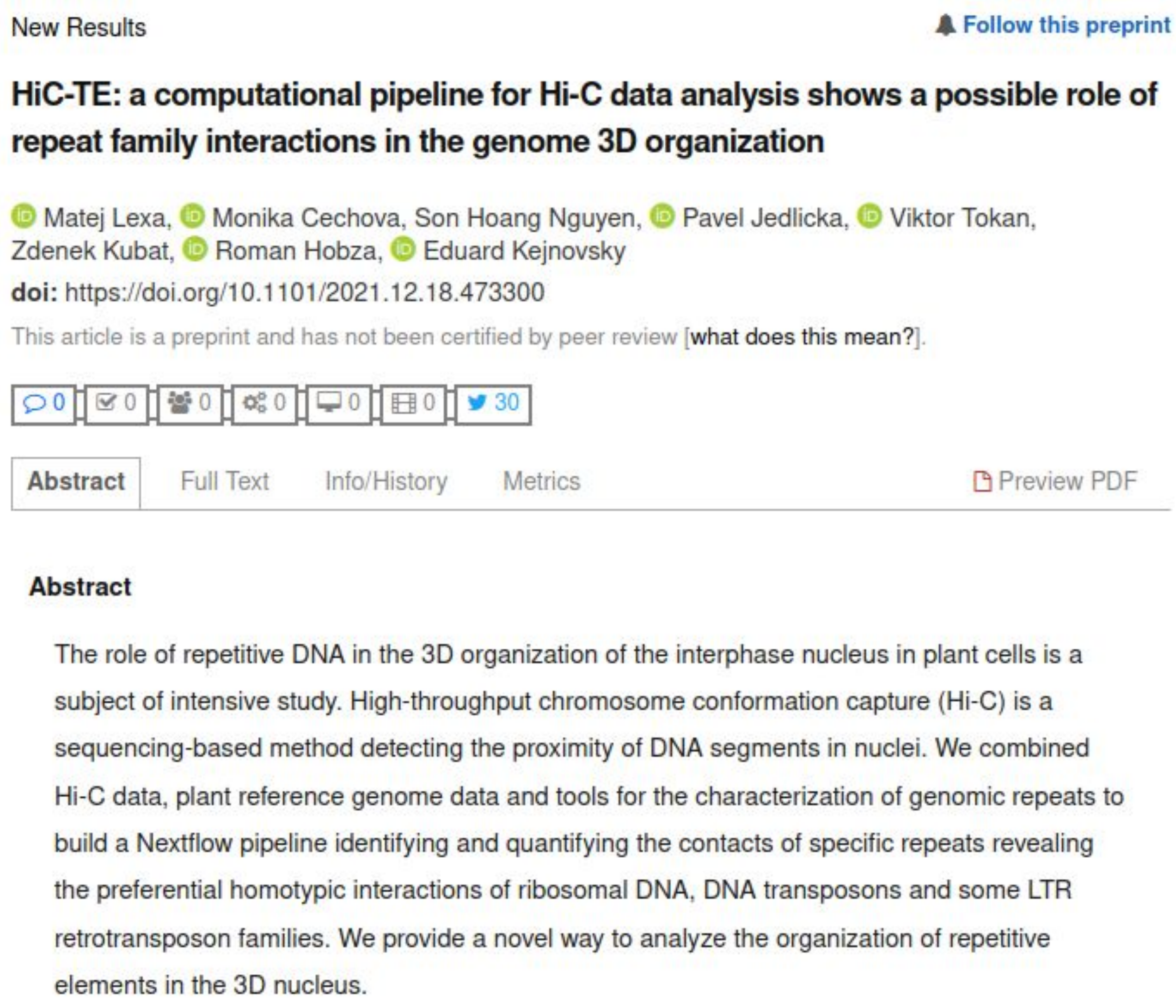
236 Commits · 2 Branches · 0 Tags · 380.2 MB Files · 380.3 MB Storage

A workflow to analyze HiC data from SRA for 3D contacts between TE families.

Update Diachromatic.jar
Son Hoang Nguyen authored 2 weeks ago

	Last commit	Last update
Docker/rtools	Add files for building docker image for rtools	2 weeks ago
bin	Update Diachromatic.jar	2 weeks ago
conf	uploading the test data for the arabidopsis…	3 months ago
data	uploading the test data for the arabidopsis…	3 months ago
doc	Updated source for hic-te flow diagram for…	4 months ago
modules	Modified the script for the metacentrum e…	3 months ago
tmp	Add singularity_runOption parameter	4 months ago

HiC-TE pipeline contributors



New Results

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

Matej Lexa, Monika Cechova, Son Hoang Nguyen, Pavel Jedlicka, Viktor Tokan, Zdenek Kubat, Roman Hobza, Eduard Kejnovsky

doi: https://doi.org/10.1101/2021.12.18.473300

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

The role of repetitive DNA in the 3D organization of the interphase nucleus in plant cells is a subject of intensive study. High-throughput chromosome conformation capture (Hi-C) is a sequencing-based method detecting the proximity of DNA segments in nuclei. We combined Hi-C data, plant reference genome data and tools for the characterization of genomic repeats to build a Nextflow pipeline identifying and quantifying the contacts of specific repeats revealing the preferential homotypic interactions of ribosomal DNA, DNA transposons and some LTR retrotransposon families. We provide a novel way to analyze the organization of repetitive elements in the 3D nucleus.

Thank you for your attention

INSTITUTE OF BIOPHYSICS

Pavel Jedlička (@pj_naruto)
Viktor Tokan
Zdeněk Kubát
Roman Hobza
Eduard Kejnovský

MASARYK UNIVERSITY

Matej Lexa (@matej_lexa)
Monika Čechová (@biomonika)
Son Hoang Nguyen