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

[1] *Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic*
[2] *Department of Plant Developmental Genetics, Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 61200 Brno, Czech Republic*

* **lexa@fi.muni.cz**

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.

Keywords: nextflow pipeline, Hi-C, repetitive DNA, transposabele elements, interphase nucleus, genome architecture

Background

A significant part of eukaryotic genomes, namely in plants, comprises transposable elements (TEs) and satellite DNA, where e.g. LTR retrotransposons constitute up to 90% of genomes [1–3]. Despite their initial consideration as junk DNA, many functions of repeats have been revealed during the last decades demonstrating their role in the structure and function of genomes and cells. TEs are often embedded in cellular regulatory networks [4] where they re-wire the gene expression programs [5]. Many examples of the domestication of TEs for specific cellular functions have been observed [6, 7]. 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. Distinct organizations of chromosomes in interphase nuclei were revealed exhibiting interactions of centromeric or telomeric repetitive DNA.

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 nuclei. 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 [8] but similar analysis in plants is hitherto missing. Additionally, since centromeres and telomeres are mostly composed of repetitive DNA, such analyses have the potential to verify the Rabl and Rosette organizations at the molecular level.

Here we present a new sequence processing pipeline to identify and quantify interactions of transposable elements, satellite DNA and rDNA in nuclei, especially those that participate in long-distance (>1Mbp) or interchromosomal contacts with frequencies that differ from the baseline expectations.

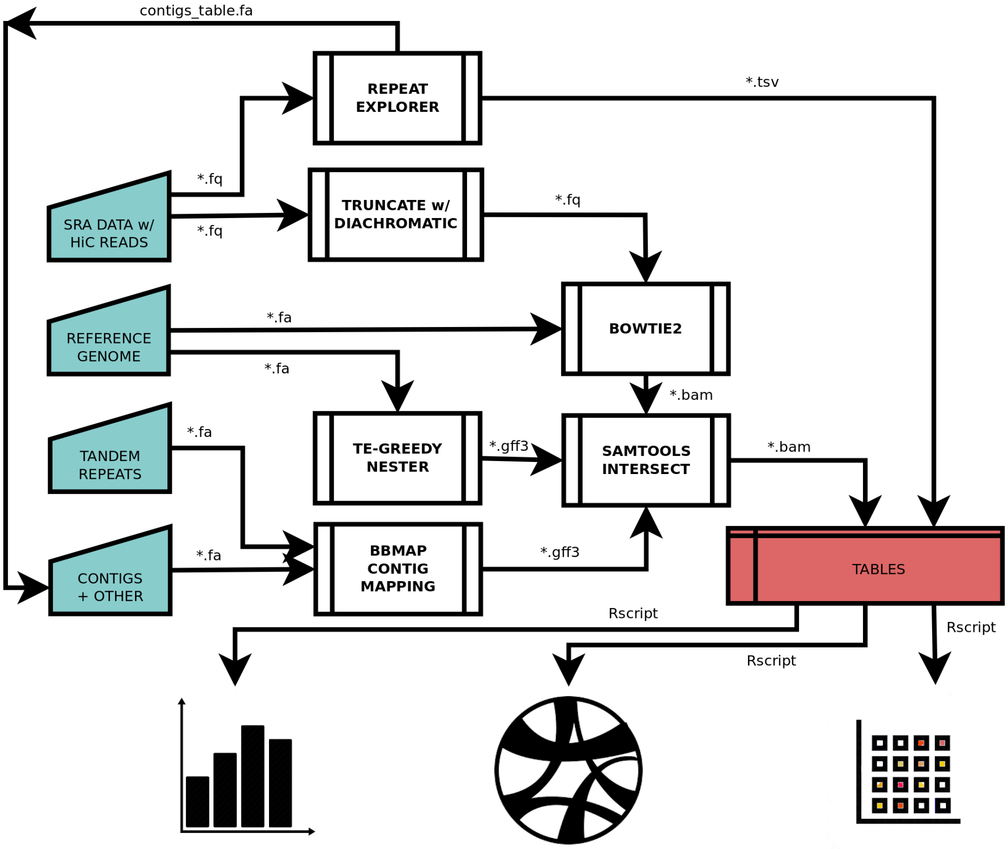

Figure 1. Block diagram showing the overall data flow in the HiC-TE pipeline. Some details were omitted for greater clarity (the full graph as produced by the pipeline is shown in Additional File 1)(blue - main data inputs; red - main data outputs; double edged rectangles - main processes running external bioinformatics tools; FASTA (*.fa), FASTQ (*.fq), BAM, GFF3 - main sequence and annotation data formats passed between processes; Rscript - R visualization scripts).

Results

Our pipeline (Fig.1, Additional File 1) integrates genome sequence analysis from several sources: assembled genome annotation for the transposable elements and satellite DNA (TE-greedy-nester, PlantSat database), medium and long-distance contact information (Hi-C sequencing experiments) and repetitive sequencing read clustering (Repeat Explorer 2). The pipeline was implemented with Nextflow 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 (manual and alternative test config file provided, see Additional File 2 and pipeline repository). 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 Additional File 3 and 4.

To test the pipeline, we used a publicly available dataset with six independent Hi-C experiments on the tomato (Solanum lycopersicum) with two technical replicates for each of three plants (from [9]. We verified that the pipeline produces consistent results and that the computational replicates are less variable than any other replicates present. We analyzed Hi-C contacts from reads clustered with Repeat Explorer (Fig.2a) and reference-mapped long-distance interactions (spanning more than 1 Mbp or between sequences located on different chromosomes)(Fig.2b). The main output is a series of heatmaps showing high and low values of normalized contacts in diverging colors, while fields (repeat family pairs) with missing values are shown in grey. This is typically caused by extremely low copy number of one of the families (either in data or after normalization randomization).

We found that in tomato, interactions among genes are limited to distal arm regions (euchromatin) while repetitive DNA interactions show higher presence in closer proximity to the centromere (heterochromatin) (Fig.2c,d). When two interacting regions belong to the same repeat family we talk about homotypic interactions (diagonals in Fig.2a,b). More than one third of repeat families (rDNA, DNA TEs, Ale, Reina, Galadriel) displayed such a pattern. The homotypic interactions were even more common when short-distance interactions were included, here they formed about 80% of all interactions (Fig.2a, Additional File 5). We showed that a number of individual interacting repeat families displayed a preference for another repeat family when analyzing pairwise long-distance interactions. This was most pronounced in ribosomal DNA, LTR TEs (e.g chromoviruses, Retand, Ogre/Tat, Alesia, Ivana, Chlamyvir, Ty1-outgroup), some tandem repeats (Lpen_370, Lesc_160) and DNA transposons (MuDR_Mutator)(Fig.2b).

Discussion

Here we presented a novel pipeline combining Hi-C data, plant reference genome data and tools for the characterization of genomic repeats. This pipeline can quickly identify and quantify the contacts of specific repeats in the 3D nucleus. Nextflow allowed us to formulate the pipeline in a modular manner and conveniently publish ready-to-run code, be it on individual computers or HPC environments, with docker and singularity containers to support all necessary dependencies and their required configuration. The modules (separate Nextflow processes) communicate via standard file formats. Therefore it should be straightforward to modify the computation to use a different piece of software or different input data. Compared to the non-repetitive fraction of the genome, for which a plethora of tools and pipelines exist, the repetitive sequences, such as those used in this pipeline, are challenging in terms of reliable mapping and the number that can be successfully clustered into families.

The pipeline contains two modes of repeat annotation, reference-free and reference-based. Being able to compare the results from both increases the robustness of the results. 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. In the case of tomato analyzed here, we included the unassembled "chromosome 00" in the reference-based mode which should have resulted in a more precise mapping and slightly distorted distance-based calculations, such as binning contacts into TAD and DIST categories.

While we took extra care to provide several modes of normalization to be able to pinpoint statistically and biologically significant contacts among the transposable elements or satellite DNA families, individual evaluation may still be needed in some

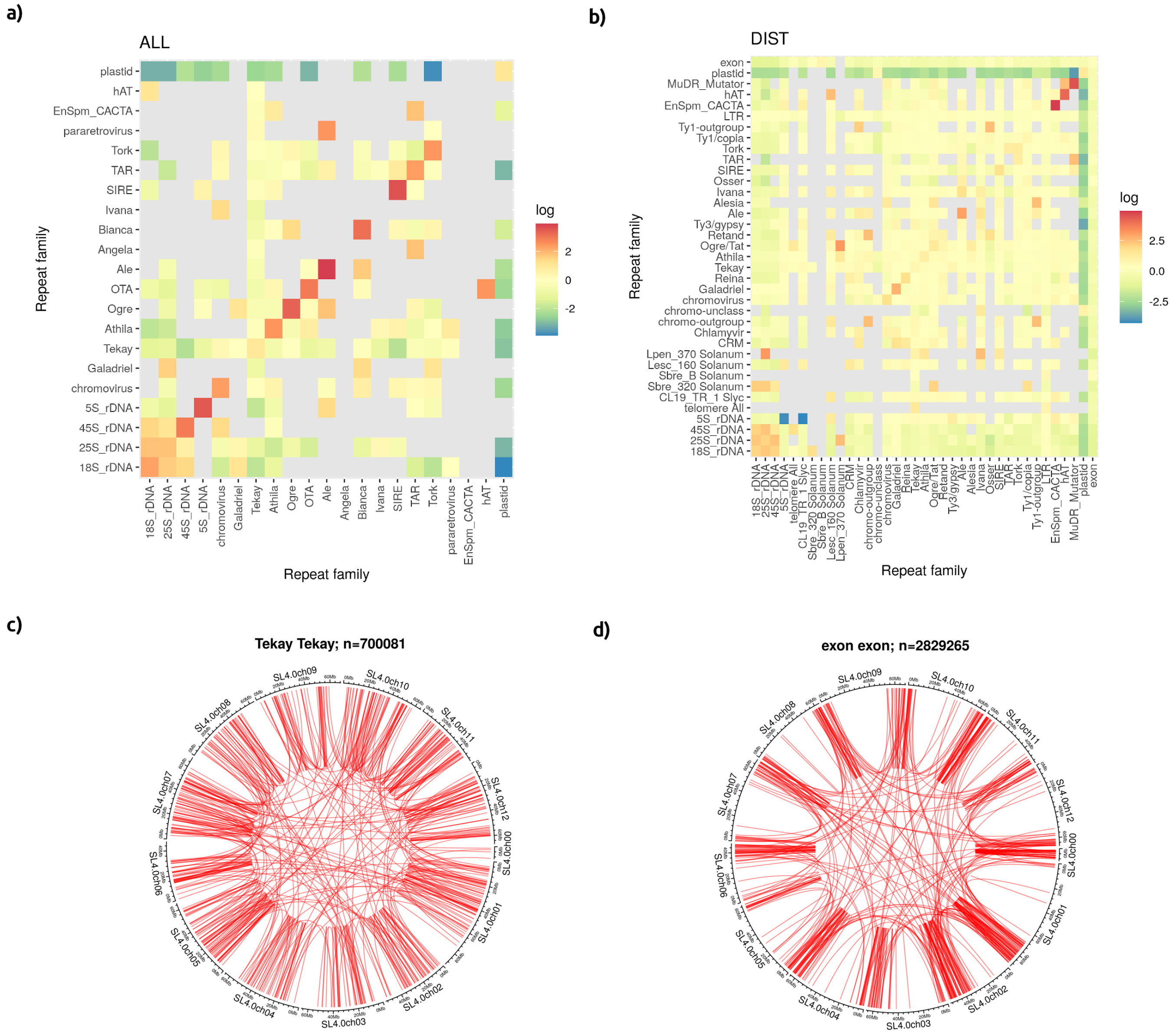

Figure 2. Results of testing the HiC-TE pipeline against 6 sequencing datasets from Dong et al. (2017). Examples from run SRR5748729. a) heatmap of all repeat family contacts based on HiC reads clustered with Repeat Explorer; grey fields are shown for pairs where missing values prevented normalization b) heatmap of long-distance repeat family contacts based on mapping HiC reads to annotated reference genome; grey fields are shown for pairs where missing values prevented normalization c) circular plot of Hi-C-supported interactions between regions annotated as TE family "Tekay"; d) circular plot of Hi-C-supported interactions between regions of the genome annotated as "exon".

cases (for more details on normalisation see Methods and Additional File 2). For example, in Fig.2a which shows all contacts that can be attributed to a repeat, results may partly reflect the length of the elements and the ability of Hi-C reads to bridge genomic regions a few hundreds of bases apart. However, the presence of the same result in long-distance contacts (Fig.2b) would suggest that individuals of the respective families tend to cluster and perhaps form exclusive domains in the nucleus. The primary DNA sequences, namely abundant repetitive elements embedded in the genome, may in this way instruct genome folding and aid genome compartmentalization. It is known that different types of chromatin regions tend to fold in different ways, with heterochromatic chromatin displaying a different average Hi-C interaction frequencies compared to euchromatin regions (Homer) [9, 10]. Individual repeats or entire repeat families can play a role in 3D organization by e.g. demarcating TAD boundaries [11] or harboring binding sites for architectural proteins [12]. Our pipeline has a potential, based on frequency of interactions of specific centromeric or telomeric repeats, to reveal these distinct local organizations of chromosomes in interphase nuclei, or even more global ones, such as Rabl, Rosette or Bouquet arrangement [13].

3-D contacts between repeats can possibly participate in processes of gene conversion or ectopic recombination [14]. Gene conversion contributes to LTR retrotransposon homogenization, while ectopic recombination helps to delete genomic regions. Since gene conversion is strongest in ribosomal genes [15], rDNA loci served us as an internal positive control. Indeed, rDNA clusters showed strong interactions with each other (Fig.2a,b). While the high homogeneity of rDNA has a functional consequence (the need for a large amount of the same functional rDNA molecules), the homogenization of TEs by gene conversion could be beneficial in ectopic recombination (and subsequent genome downsizing) and thus represents a tool for the regulation of genome size.

Methods

Data from six sequencing runs from a Hi-C experiment on the tomato [9], Additional File 5) were fed into our Nextflow [16] computational pipeline “HiC-TE” (Fig.1, Additional File 1). The pipeline combines read trimming with Diachromatic [17], TE annotation with TE-greedy-nester [18] and Repeat Explorer [19], satellite/tandem-repeat annotation with TAREAN [20], TRF [21] and PlantSat [22]. Read mapping is done via Bowtie2 [23] and BBmap with a subsequent overlap/intersection analysis with bedtools [24] and visualization in R/Bioconductor [25, 26] using the following packages: circlize [27], dplyr [28], GenomicRanges [29], ggplot2 [30], gplots [31], karyoplotteR [32], MatrixCorrelation [33], ragg [34], reshape2 [35], Rsamtools [36], rtracklayer [37], stringr [38]. This pipeline generated tables, heatmaps and circular plots showing frequency of Hi-C interaction between repeats and other annotated features in the genome (for code see GitLab repository). Before visualization in heatmaps, the data is normalized to account for the fact that repeat families have varying frequencies. As there are several ways to carry out such normalizations, each with its own biases, we therefore generated heatmaps using a range of normalization techniques (details in Additional File 2 and 5). Joint probability normalization assumes Hi-C contacts occur between independent positions and normalizes contact counts against a product of frequencies of the interacting families. Label permutation uses a sample set with family labels subjected to permutation. Annotation interval reshuffling uses a shuffled version of annotation files to normalize contacts.

Availability of data and materials

The source code and documentation for the Nextflow pipeline is available at http://gitlab.fi.muni.cz/hic-te/.

Supporting Information

Additional Files

Additional file 1

Suplementary Figure — Flow chart of the Nextflow HiC-TE pipeline (output from running Nextflow with the -graph switch)

Additional file 2

HiC-TE manual

Additional file 3

Suplementary Tables — 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)

Additional file 4

Suplementary Tables — 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)

Additional file 5

PDF files with 6 complete sets of outputs, numbered by SRR ID and the name of output

Acknowledgements

We thank Christopher Johnson for critical reading of this manuscript. We thank Jan Hoidekr for the help with the 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. Monika Cechova is the holder of Martina Roeselova Memorial Fellowships 2020.

Funding

This research was supported by the Czech Science Foundation (grant 21-00580S).

Competing interests

The authors declare no conflict of interest.

Authors' contributions 166

Conceptualization ML, EK, PJ; workflow implementation and data analysis ML, MC, SHN; methodology ML, MC, PJ and VT; docker, singularity container preparation SHN; validation PJ; writing - original draft preparation EK, ML, MC, PJ; writing - review and editing ZK and RH; supervision EK; funding acquisition EK and ML. All authors have read and agreed to the published version of the manuscript.

References

1. Charles, M., Belcram, H., Just, J., Huneau, C., Viollet, A., Couloux, A., et al.: Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics **180**, 1071–1086 (2008)

2. Schnable, P., Ware, D., Fulton, R., Stein, J., Wei, F., Pasternak, S., Liang, C., et al.: The B73 maize genome: complexity, diversity, and dynamics. Science **326**, 1112–1115 (2009)

3. Wicker, T., Grundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-Gonzales, R., et al: Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biology **19**, 103 (2018)

4. Feschotte, C.: Transposable elements and the evolution of regulatory networks. Nature Rev Genet **9**, 397–405 (2008)

5. Slotkin, R., Martienssen, R.: Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet **8**, 272–285 (2007)

6. Sinzelle, L., Izsvak, Z., Ivics, Z.: Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell Mol Life Sci **66**, 1073–1093 (2009)

7. Jangam, D., Feschotte, C., Betran, E.: Transposable element domestication as an adaptation to evolutionary conflicts. Trends Genet **33**, 817–831 (2017)

8. Lu, J., Chang, L., Li, T., Wang, Y., Yin, Y., et al: Homotypic clustering of L1 and B1/Alu repeats compartmentalizes the 3D genome. Cell Research **31**, 613–630 (2021)

9. Dong, P., Tu, X., Chu, P.-Y., Lu, P., Zhu, N., Grierson, D., Du, B., Li, P., Zhong, S.: 3D chromatin architecture of large plant genomes determined by local A/B compartments. Molecular Plant **10**(12), 1497–1509 (2017)

10. The Tomato Genome Consortium: The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**, 635–641 (2012)

11. Zhang, Y., Preissl, S., Amaral, M., et al: Transcriptional active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat Gen **51**, 1380–1388 (2019)

12. Choudhary, M., Friedman, R., Wang, J., Jang, H., Zhuo, X., Wang, T.: Co-opted transposons help perpetuate conserved higher-order chromosomal structures. bioarXiv **314**, 485342 (2018)

13. Tiang, C.-L., He, Y., Pawlowski, W.: Chromosome organization and dynamics during interphase, mitosis, and meiosis in plants. Plant Physiology **158**, 26–34 (2012)

14. Sun, X.-Q., Li, D.-H., Xue, J.-Y., Yang, S.-H., Zhang, Y.-M., Li, M.-M., Hang, Y.-Y.: Insertion DNA accelerates meiotic interchromosomal recombination in Arabidopsis thaliana. Mol Biol Evol **33**(8), 2044–2053 (2016)

15. Matyasek, R., Lim, K.Y., Kovarik, A., Leitch, A.R.: Ribosomal DNA evolution and gene conversion in Nicotiana rustica. Heredity (Edinb) **91**(3), 268–275 (2003)

16. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., Notredame, C.: Nextflow enables reproducible computational workflows. Nature Biotechnology **35**(4), 316–319 (2017)

17. Hansen, P., Gargano, M., Hecht, J., Ibn-Salem, J., Karlebach, G., Roehr, J.T., Robinson, P.: Computational processing and quality control of Hi-C, Capture Hi-C and Capture-C data. Genes **10**(19), 548 (2019)

18. Lexa, M., Jedlicka, P., Vanat, I., Cervenansky, M., Kejnovsky, E.: TE-greedy-nester: structure-based detection of LTR retrotransposons and their nesting. Bioinformatics **36**(20), 4991–4999 (2020)

19. Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J.: RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics **29**(6), 792–793 (2013)

20. Novak, P., Robledillo, L.A., Koblizkova, A., Vrbova, I., Neumann, P., Macas, J.: TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Research **45**(12), 111 (2017)

21. Benson, G.: Tandem repeats finder: a program to analyze DNA sequences. Nucl Acids Res **27**(2), 573–580 (1999)

22. Macas, J., T., M., Nouzova, M.: PlantSat: a specialized database for plant satellite repeats. Bioinformatics **18**, 28–35 (2002)

23. Langmead, B., Salzberg, S.: Fast gapped-read alignment with Bowtie 2. Nature Methods **9**, 357–359 (2012)

24. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)

25. R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2020). `https://www.R-project.org/`

26. Huber, W., Carey, V., Gentleman, R., Anders, S., Carlson, M., Carvalho, B., Bravo, H., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K., Irizarry, R., Lawrence, M., Love, M., MacDonald, J., Obenchain, V., Oles, A., Pages, H., Reyes, A., Shannon, P., Smyth, G., Tenenbaum, D., Waldron, L., Morgan, M.: Orchestrating high-throughput genomic analysis with Bioconductor. Annals Bot **12**, 115 (2015)

27. Gu, Z., Gu, L., Eils, R., Schlesner, M., Brors, B.: Circlize implements and enhances circular visualization in R. Bioinformatics **30**(19), 2811–2812 (2014)

28. Wickham, H., François, R., Henry, L., Müller, K.: dplyr: A grammar of data Manipulation. R package version 1.0.5 (2021). `https://CRAN.R-project.org/package=dplyr`

29. Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., et al.: Software for computing and annotating genomic ranges. PLoS Comput Biol **9**(8), 1003118 (2013)

30. Wickham, H.: Ggplot2: Elegant Graphics for Data Analysis. Springer, New York (2003)

31. Warnes, G., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B.: gplots: various R programming tools for plotting data. R package version 3.1.1 (2020). `https://CRAN.R-project.org/package=gplots`

32. Gel, B., Serra, E.: KaryoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics **33**(19), 3088–3090 (2017)

33. Indahl, U., Næs, T., Liland, K.: A similarity index for comparing coupled matrices. Journal of Chemometrics, 3049 (2018)

34. Pedersen, T., Shemanarev, M.: ragg: graphic devices based on AGG. R package version 1.1.3 (2020). `https://CRAN.R-project.org/package=ragg`

35. Wickham, H.: Reshaping data with the reshape package. Journal of Statistical Software **21**(12), 1–20 (2007)

36. Morgan, M., Pages, H., Obenchain, V., Hayden, N.: Rsamtools: binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 2.2.3 (2020). `http://bioconductor.org/packages/Rsamtools`

37. Lawrence, M., Gentleman, R., Carey, V.: Rtracklayer: an R package for interfacing with genome browsers. Bioinformatics **25**, 1841–1842 (2009)

38. Wickham, H.: stringr: simple, consistent wrappers for common string operations. R package version 1.4.0 (2019). `https://CRAN.R-project.org/package=stringr`