The Journal of Open Source Software

reper: Genome-wide identification, classification and quantification of repetitive elements without an assembled genome

Niklas Terhoeven[1, 2], **Jörg Schultz**[1, 2], **and Thomas Hackl**[3]

1 Center for Computational and Thoeretical Biology, Universität Würzburg **2** Lehrstuhl für Bioinformatik, Universität Würzburg **3** Department of Civil and Environmental Engineering, Massachusetts Institute of Technology

DOI: 10.21105/joss.00527

Software

- Review
- Repository
- Archive

Submitted: 21 December 2017
Published: 08 February 2018

Summary

Repetitive elements constitute a substantial fraction of most eukaryotic genomes. Still, their actual amount differs strongly between species. For example, the genome of *Saccharomyces cervisiae* contains only about 3 % repeats (Kim et al. 1998), *Arabidopsis* harbours 14 % (The Arabidopsis Genome Initiative 2000), human 50 % (Lander et al. 2001) and wheat even 90 % (Clavijo et al. 2017).

Annotation and Classification of these elements is a pivotal step in the annotation of each genome. Furthermore, tracing their history can give ample insights into the evolution of a genome and thereby, of a species. Accordingly, different methods for repeat annotation have been developed (A. Smit, Hubley, and Green 2013; Benson 1999; Gymrek et al. 2012). Still, typically they rely on an assembled genome sequence – a prerequisite which can lead to erroneous results. As repetitive elements are highly similar assembly algorithms will collapse repeat variants into a single occurrence or not assemble the repetitive regions at all. Thus, the annotation of repeat regions and thereby the characterization of their content and diversity solely based on an assembled genome sequence can give misleading results.

To address this challenge, we developed reper, a kmer based method to detect, classify and quantify repeats in next generation sequencing (NGS) data without the need of a genome assembly. Our pipeline samples reads with high kmer coverage directly from the NGS dataset. The kmer counts are acquired using jellyfish (Marçais and Kingsford 2011). This subset is assembled using the transcriptome assembler Trinity (Grabherr et al. 2011), allowing reper to recover repeat variants at a high resolution. To create exemplar sequences of each repeat in the genome, the assembled repeats are clustered using cd-hit (W. Li and Godzik 2006; Fu et al. 2012). These are further classified based on homology to known repeats using multiple blast (Camacho et al. 2009) searches. Since reper was developed with a focus on plant data, the default classification libraries are REdat (Nussbaumer et al. 2012) for repeats, and refseq (O'Leary et al. 2016) for chloroplast and mitochondrial sequences. The reference database, however, can easily be customized to the user's needs. A configuration script for the popular, but proprietary database repbase is provided with the package as well. Next, the repeat content is quantified on sequence, cluster and class level based on read mappings using bowtie2 and samtools (Langmead and Salzberg 2012; H. Li et al. 2009). Finally, the repeat landscape can be analyzed and graphically represented with the R script provided with the pipeline. Currently, reper is specifically customized to work with paired-end Illumina data, but support of long-read technologies such as PacBio and Nanopore is in development.

To date, there is only a single software package with a similar functionality to reper,

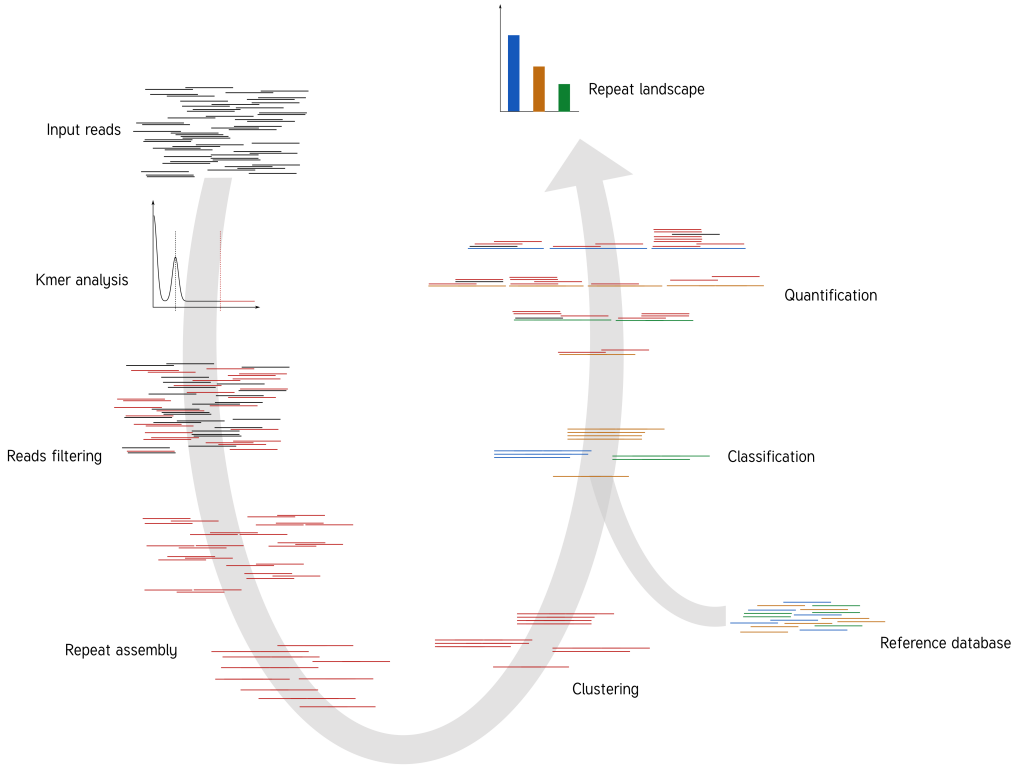

Figure 1: schematic overview of workflow

namely dnaPipeTE (Goubert et al. 2015). Still, it relies on dependencies like RepeatMasker, which has to be installed independently as well as the proprietary repeat database repbase by giri. Contrasting, The reper source code is available on github under the MIT license. To further ease installation and usage, a Docker container with a complete reper installation is also provided. Since reper is usually run in an HPC environment where users don't have root or Docker rights, we furthermore made a singularity image available which can be used with standard user permissions.

We are currently using reper to analyze the repeat content in different plant genome sequencing projects. An example using *Beta vulgaris* data is given in the tutorial section of the reper wiki.

Acknowledgements

We would like to thank Markus Ankenbrand and Frank Förster for valuable discussions and their support and advice on different topics like Docker and pipeline design.

References

Benson, G. 1999. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research* 27 (2):573–80. https://doi.org/10.1093/nar/27.2.573.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications."

BMC Bioinformatics 10 (1):421. https://doi.org/10.1186/1471-2105-10-421.

Clavijo, Bernardo J., Luca Venturini, Christian Schudoma, Gonzalo Garcia Accinelli, Gemy Kaithakottil, Jonathan Wright, Philippa Borrill, et al. 2017. "An Improved Assembly and Annotation of the Allohexaploid Wheat Genome Identifies Complete Families of Agronomic Genes and Provides Genomic Evidence for Chromosomal Translocations." *Genome Research* 27 (5):885–96. https://doi.org/10.1101/gr.217117.116.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data." *Bioinformatics* 28 (23):3150–2. https://doi.org/10.1093/bioinformatics/bts565.

Goubert, Clément, Laurent Modolo, Cristina Vieira, Claire ValienteMoro, Patrick Mavingui, and Matthieu Boulesteix. 2015. "De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes Albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes Aegypti)." *Genome Biology and Evolution* 7 (4):1192–1205. https://doi.org/10.1093/gbe/evv050.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data Without a Reference Genome." *Nature Biotechnology* 29 (7):644–52. https://doi.org/10.1038/nbt.1883.

Gymrek, Melissa, David Golan, Saharon Rosset, and Yaniv Erlich. 2012. "lobSTR: A Short Tandem Repeat Profiler for Personal Genomes." *Genome Research* 22 (6):1154–62. https://doi.org/10.1101/gr.135780.111.

Kim, Jin M., Swathi Vanguri, Jef D. Boeke, Abram Gabriel, and Daniel F. Voytas. 1998. "Transposable Elements and Genome Organization: A Comprehensive Survey of Retrotransposons Revealed by the Complete *Saccharomyces Cerevisiae* Genome Sequence." *Genome Research* 8 (5):464–78. https://doi.org/10.1101/gr.8.5.464.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822):860–921. https://doi.org/10.1038/35057062.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4):357–59. https://doi.org/10.1038/nmeth.1923.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.

Li, W., and A. Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22 (13):1658–9. https://doi.org/10.1093/bioinformatics/btl158.

Marçais, Guillaume, and Carl Kingsford. 2011. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers." *Bioinformatics* 27 (6):764–70. https://doi.org/10.1093/bioinformatics/btr011.

Nussbaumer, Thomas, Mihaela M. Martis, Stephan K. Roessner, Matthias Pfeifer, Kai C. Bader, Sapna Sharma, Heidrun Gundlach, and Manuel Spannagl. 2012. "MIPS PlantsDB: A Database Framework for Comparative Plant Genome Research." *Nucleic Acids Research* 41 (D1):D1144–D1151. https://doi.org/10.1093/nar/gks1153.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." *Nucleic Acids Research* 44 (D1):D733–D745. https://doi.org/10.1093/nar/gkv1189.

Smit, AFA, R Hubley, and P Green. 2013. "RepeatMasker Open-4.0." http://www.repeatmasker.org.

The Arabidopsis Genome Initiative. 2000. "Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana." *Nature* 408 (6814):796–815. https://doi.org/10.1038/35048692.