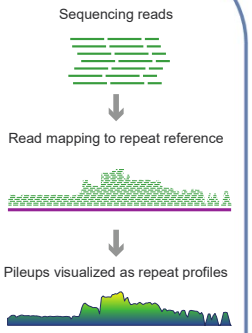
RepeatProfiler: a pipeline for visualization and comparative analysis of repetitive DNA profiles

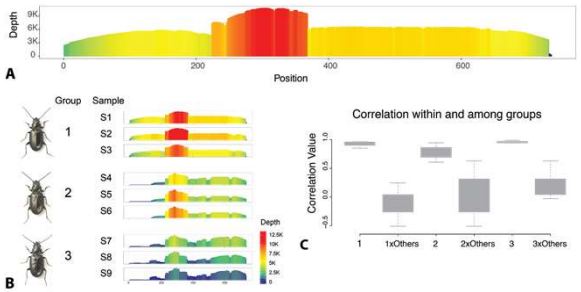
Sherif Negm, Anya Greenberg, Amanda Larracuente & John Sproul
University of Rochester, Department of Biology

Overview

Repetitive DNA can reveal signals of evolutionary history over short time scales that may not be evident in sequences from slower-evolving genomic regions [1]. Many tools for studying repeats are directed toward organisms with existing genomic resources, including genome assemblies and repeat libraries. However, signals in repeat variation may prove especially valuable in disentangling evolutionary histories in diverse non-model groups, for which genomic resources are limited. Here we introduce RepeatProfiler, a tool for generating, visualizing, and comparing repetitive DNA profiles from low-coverage, short-read sequence data. RepeatProfiler facilitates comparative study of repeats across samples and repeats to provide a high-resolution data source for studies on species delimitation, genome evolution, and repeat biology over short evolutionary time scales.

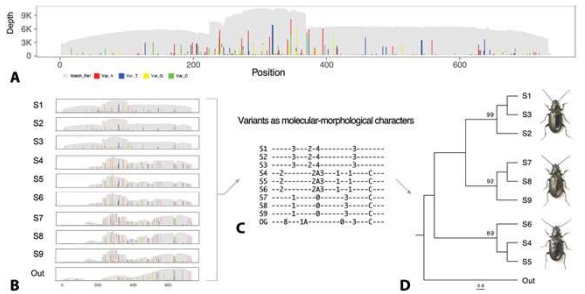


Standard Output



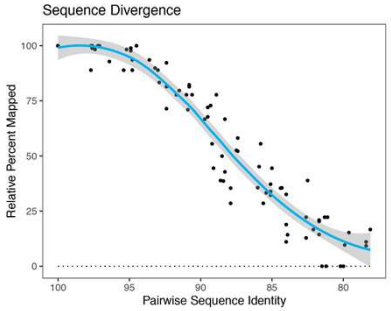
RepeatProfiler produces visually enhanced read-depth profiles (A) to simplify visual comparison of patterns across samples (B). The pipeline also enables statistical comparison of profile shape within and between user-defined sample groups. This correlation analysis measures the degree of similarity in the variable of coverage depth for each position across the reference sequence between two samples. We use Spearman's rank correlation coefficient to calculate correlation values for pairwise comparisons of all samples, and summarize within- and between-group comparisons with boxplots (C).

Phylogenetic analysis of variant signatures

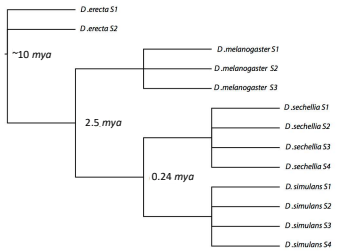


The pipeline also facilitates comparison of profiles across samples in a phylogenetic framework. The pipeline summarizes information contained in variant-enhanced profiles (A) by identifying abundant variants and encoding those variants in a phylip file as molecular-morphological characters (C) which can then be analyzed as morphological data using phylogenetic software. This example shows profiles of nine individual samples belonging to three *Bembidion* ground beetle species and one outgroup (B). The phylogenetic tree (D) correctly groups samples as species based on the variant in profiles (B, C).

Validating RepeatProfiler

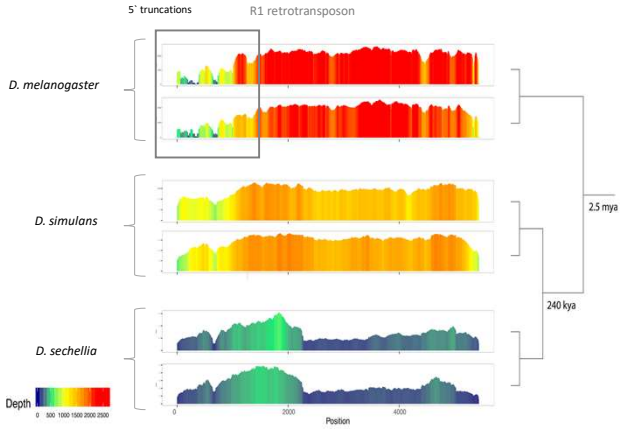


RepeatProfiler's default read mapping[3] settings allow for considerable divergence between the reference sequence and the reads being mapped, enabling comparative study across clades showing shallow to moderate sequence divergence among samples.



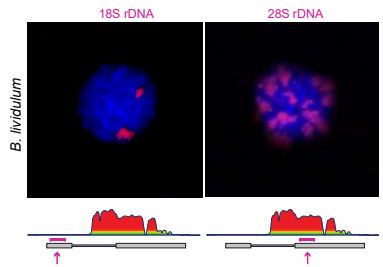
We validated our approach to comparing variant signatures through phylogenetic analysis by using RepeatProfiler to analyze 37 abundant transposable elements (TEs) in four closely related *Drosophila* species. We conducted phylogenetic analysis of phylip files summarizing variants for each TE as morphological data in IQ-Tree[4] and found that trees show good phylogenetic signal over short divergence times. This is a consensus tree of all 37 TE trees showing correct sample groupings and branching pattern among samples.[2]

Species Boundaries and Repeat Biology



Repetitive DNA profiles of two individuals for each of three *Drosophila* species showing strong species-specific signatures despite their recent evolutionary divergence (see tree on right with dates from [2]). In addition to providing evidence of species boundaries, repeat profiles can hold information related to repeat biology. This R1 element shows evidence of 5` truncations (i.e., reduced coverage in gray box) suggesting this element is recently active in this species, as active non-LTR elements are known to accumulate 5` truncations.

RepeatProfiler and Genome Architecture



Fluorescence in situ hybridization shows that repeat profiles can provide evidence of genome-scale repatterning of repeat architecture over short time-scales. These FISH images show the recent mobilization or rDNA-like sequences that has led to genome-wide repatterning of repeats across in ground beetle species.[1]

Features & Summary

- Produces visually enhanced repeat profiles across repeats, and samples
- Quantitative comparative analysis of profile features across samples
- High-resolution data source for species delimitation and genome evolution
- Provides insight's into of repeat biology and genome architecture
- Installation on Mac, Linux, and Windows using Homebrew and Docker
- Available from https://github.com/johnssproul/RepeatProfiler

Dependencies: R[5],python, Bowtie2[3], SAMtools[6], & R packages (ggplot2[7]: ggpubr, scales, reshape2)

References: 1. Sproul, J.S., *Barton, L.M., Maddison, D.R. 2020. Repetitive DNA profiles reveal evidence of rapid genome evolution and reflect species boundaries in ground beetles. Systematic Biology; syaa030, https://doi.org/10.1093/sysbio/syaa030. 2. Garrigan D, Kingan SB, Geneva AJ, et al. Genome sequencing reveals complex speciation in the Drosophila simulans clade. Genome Res. 2012;22(8):1499–1511. doi:10.1101/gr.130922.111. 3. Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012). https://doi.org/10.1038/nmeth.1923. 4. Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, Bui Quang Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies, Molecular Biology and Evolution, Volume 32, Issue 1, January 2015, Pages 268–274, https://doi.org/10.1093/molbev/msu300. 5. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 6. Wickham H (2016). & Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), pp.2078-2079. 7. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.