Bioinformatics Advances, 2025, **00**, vbaf066
https://doi.org/10.1093/bioadv/vbaf066
Advance Access Publication Date: 8 April 2025
Application Note

Genome analysis

TRACKing tandem repeats: a customizable pipeline for identification and cross-species comparison

Carolina L. Adam[1,*], Joana Rocha[2], Peter Sudmant[2], Rori Rohlfs[1,3]

[1]Institute of Ecology and Evolution, University of Oregon, Eugene, Oregon 97403, United States
[2]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, United States
[3]School of Computer and Data Sciences, University of Oregon, Eugene, OR 97403, United States

*Corresponding author. Institute of Ecology and Evolution, University of Oregon, Onyx Bridge 362, 1318 Franklin Blvd, Eugene, Oregon 97403, United States. E-mail: carolinaladam@gmail.com.

Associate Editor: Aida Ouangraoua

Abstract

Summary: TRACK is a user-friendly Snakemake workflow designed to streamline the discovery and comparison of tandem repeats (TRs) across species. TRACK facilitates the cataloging and filtering of TRs based on reference genomes or T2T transcripts, and applies reciprocal LiftOver and sequence alignment methods to identify putative homologous TRs between species. For further analyses, TRACK can be used to genotype TRs and subsequently estimate and plot basic population genetic statistics. By incorporating key functionalities within an integrated workflow, TRACK enhances TR analysis accessibility and reproducibility, while offering flexibility for the user.

Availability and implementation: The TRACK toolkit with step-by-step tutorial is freely available at https://github.com/caroladam/track.

1 Introduction

Tandem repeats (TRs) are repetitive genomic sequences characterized by their abundance in genomes, high mutation rates, and presence of multiple alleles within a population (Gymrek 2017), making them a major source of genetic variation (Kashi *et al.* 1997). By accumulating mutations faster than single nucleotide polymorphisms (SNPs) (Sun *et al.* 2012), TRs are an easy target for natural selection and act as central players in rapid evolution (Gemayel *et al.* 2010). Although ubiquitous in all eukaryotic genomes (Richard *et al.* 2008), TRs gained prominence in human and non-human primate studies due to their role as epigenetic and gene expression modulators and their association with many human diseases (Gymrek *et al.* 2016). TR expansions, for instance, are linked to the pathogenesis of multiple types of cancer (Erwin *et al.* 2023) and neurological-related conditions, such as Huntington's disease (MacDonald *et al.* 1993) and Friedreich's ataxia (Campuzano *et al.* 1996).

The cumulative evidence that TRs are associated with the evolution of complex traits (Press *et al.* 2014) and likely evolved under selective pressures (Liang *et al.* 2015) highlights the importance of understanding their variation across species through comparative analyses. In great apes, e.g. there is evidence for the role of TRs in chromosomal rearrangements (Farré *et al.* 2011) and gene expression divergence (Bilgin Sonay *et al.* 2015). Thus, building comparative TR frameworks can provide critical insights into evolutionary processes.

The propensity of TRs for homoplasy, coupled with sequencing technology limitations that hindered accurate sequencing of long TRs, has historically restricted their widespread use in cross-species comparative studies (Hodel *et al.* 2016). The advent of long-read sequencing technologies provides the means to overcome these restrictions. The Telomere-to-Telomere (T2T) Consortium delivered the first gapless human genome assembly, CHM13, correcting inaccuracies in GRCh38 (Nurk *et al.* 2022), and has since expanded to include six non-human primate genomes (Yoo *et al.* 2025). These high-resolution data and the availability of new analysis tools tailored for TRs (Mousavi *et al.* 2021, Dolzhenko *et al.* 2024) open avenues for comparative studies with unparalleled resolution, enabling the discovery of previously undetected TRs, which may help generate new evolutionary hypotheses.

A suite of tools has been developed to analyze intraspecific TR variation in long-read data (e.g. Mitsuhashi *et al.* 2019, Chiu *et al.* 2021, Mousavi *et al.* 2021). However, an integrated tool that streamlines the entire process—from generating species-specific TR catalogs to genotyping TRs and assessing homology for interspecific TR variation analysis—is still lacking. To unify the discovery and analysis of shared TR loci across species, we developed the Tandem Repeat Analysis and Comparison Kit (TRACK). This Snakemake workflow integrates TR identification, filtering, homology assessment, and genotyping into a single consolidated tool (Fig. 1).

We built upon well-established TR analysis tools, such as Tandem Repeat Finder (Benson 1999), to identify TRs in chromosome-level genome assemblies, and incorporate Liftover (Hinrichs *et al.* 2006), which converts genomic

Received: December 12, 2024; **Revised:** March 14, 2025; **Editorial Decision:** March 25, 2025; **Accepted:** April 7, 2025

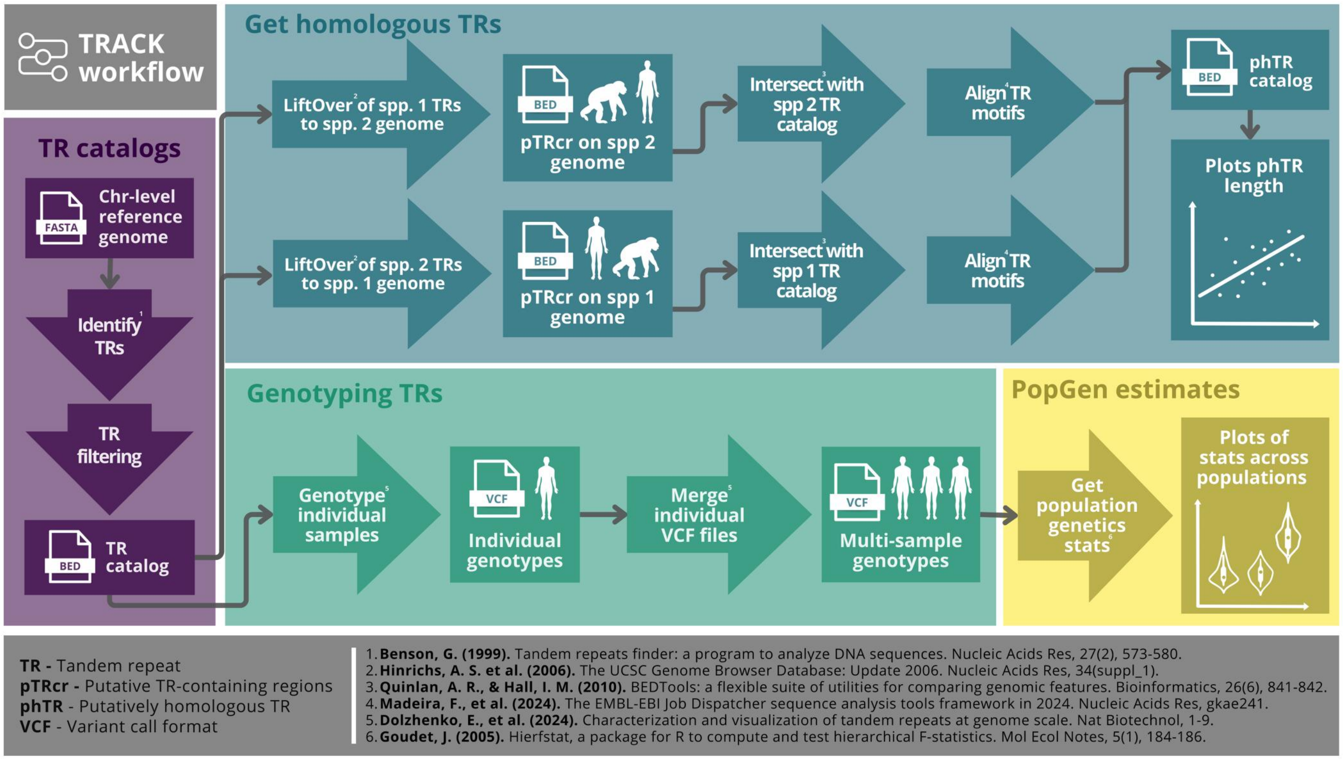
TRACK workflow
TR catalogs
FASTA
Chr-level reference genome
Identify[1] TRs
TR filtering
BED
TR catalog
Get homologous TRs
LiftOver[2] of spp. 1 TRs to spp. 2 genome
BED
pTRcr on spp 2 genome
Intersect[3] with spp 2 TR catalog
Align[4] TR motifs
LiftOver[2] of spp. 2 TRs to spp. 1 genome
BED
pTRcr on spp 1 genome
Intersect[3] with spp 1 TR catalog
Align[4] TR motifs
BED
phTR catalog
Plots phTR length
Genotyping TRs
Genotype[5] individual samples
VCF
Individual genotypes
Merge[5] individual VCF files
VCF
Multi-sample genotypes
PopGen estimates
Get population genetics stats[6]
Plots of stats across populations
TR - Tandem repeat
pTRcr - Putative TR-containing regions
phTR - Putatively homologous TR
VCF - Variant call format
1. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 27(2), 573-580.
2. Hinrichs, A. S. et al. (2006). The UCSC Genome Browser Database: Update 2006. Nucleic Acids Res, 34(suppl_1).
3. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842.
4. Madeira, F., et al. (2024). The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. Nucleic Acids Res, gkae241.
5. Dolzhenko, E., et al. (2024). Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol, 1-9.
6. Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. Mol Ecol Notes, 5(1), 184-186.

Figure 1. A simplified workflow of the Tandem Repeat Analysis and Comparison Kit (TRACK) features.

coordinates between different genome assemblies, to map homologous TR regions between species. Users can define custom thresholds for genomic region overlap and TR motif sequence similarity, allowing fine-tuned putative homologous TR catalog generation. Additionally, TRACK leverages the newly developed Tandem Repeat Genotyping Tool (Dolzhenko *et al.* 2024), tailored for HiFi long-read data, enabling high-resolution TR analysis in population-level datasets. Unlike other long-read genotyping tools (Mitsuhashi *et al.* 2019, Chiu *et al.* 2021), TRGT provides both allele length and sequence composition, allowing the analysis of non-constant variants, i.e. where the fraction of nucleotides within a given TR that matches the consensus motif is <100%. TRACK also enhances usability by providing processed genotype summaries and allele distributions in an accessible tabular format, enabling downstream analyses without requiring direct manipulation of VCF files. This integrated approach improves efficiency and reproducibility, reducing the need for extensive post-processing and custom scripting.

2 Pipeline description

2.1 Generating TR catalogs

TRACK uses Tandem Repeat Finder (TRF) version 4.09 (Benson 1999) to generate the TR reference catalogs from chromosome-level reference genomes. TRACK default parameters are in Table 1, which results in catalogs with TR minimum length >12 bp. To eliminate redundancy from multiple computations at the same genomic index position or variation in score values, overlapping TRs are initially merged. Subsequently, TRs exceeding 10 Kbp in total length or having a copy number <2.5 are filtered out.

Table 1. TRACK default parameters for Tandem Repeat Finder (TRF) run.

Parameter	TRACK default value
Match score	2
Mismatch score	5
Indel score	7
Matching probability	80
Indel probability	10
Minimum alignment score	24
Maximum period size	2000

2.2 Identification of putative homologous TRs

utative TR homology is assessed in a pairwise manner. TRACK begins by conducting a LiftOver (Hinrichs *et al.* 2006) analysis using a TR reference catalog from a target genome (tTRc) and a chain file that describes the conversion between genome positions from a target genome assembly to a query genome assembly. Chain files can be obtained directly from the UCSC Genome Browser or be custom-made by the user by performing whole-genome alignment of target and query genomes and subsequent conversion of the alignment file to chain file. The liftover step produces a file of putative TR-containing regions (pTRcr) in the query genome. To reduce bias in homology detection, TRACK performs the LiftOver analysis bidirectionally, with the two genomes of interest serving as both target and query. This yields two lifted bed files for pairwise comparison. The resulting pTRcr file is then intersected with the original TR reference catalog of the query genome used in the comparison, retaining only regions that meet a user-defined overlap threshold. This means that the lifted region and the corresponding TR in the reference catalog must overlap by at least the specified percentage to be kept for further analysis.

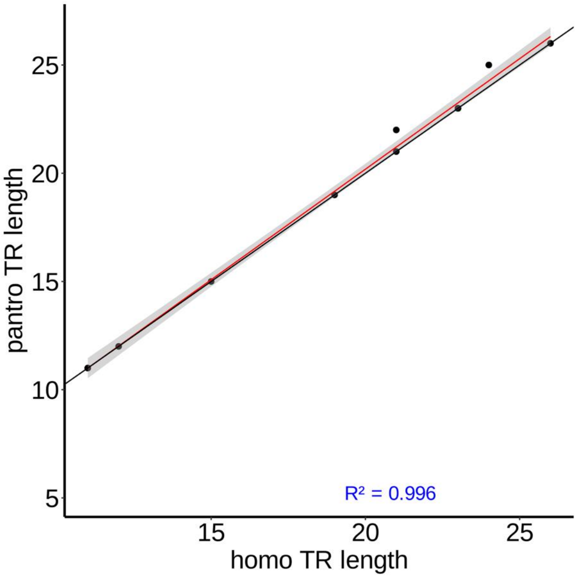

Figure 2. Scatterplot of the total length of 1000 randomly subsetted TRs from the shared TR catalog between human and chimpanzee T2T reference genomes using a threshold of 10% overlap and 95% sequence similarity.

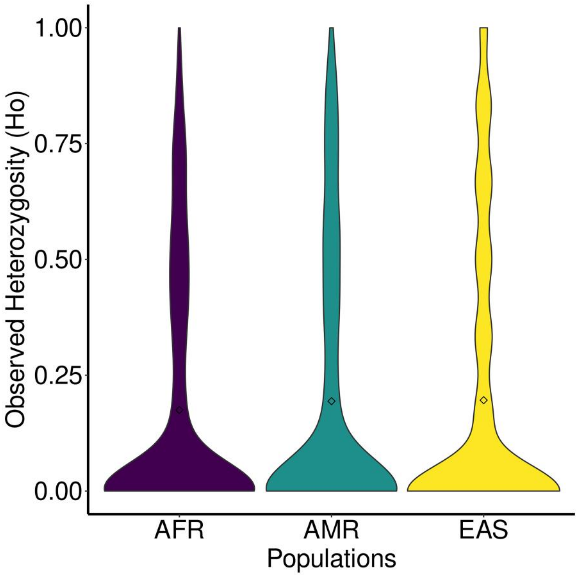

Figure 3. Observed heterozygosity estimates from a random subset of 1000 TR loci from TRACK's human T2T TR catalog genotyped in individuals of the Human Pangenome Reference Consortium (HPRC) (Wang *et al.* 2022). AFR—Africa; AMR—Americas; EAS—Asia.

To verify the similarity of these TRs based on sequence composition, TRACK conducts pairwise global alignments between the motifs of the putative homologous TRs from each species using the Needleman-Wunsch algorithm (Madeira *et al.* 2024). Then, TRACK keeps TRs that meet a user-specified threshold for motif sequence similarity. The two resulting files from each pairwise comparison are intersected to produce a final catalog of putatively homologous TRs (phTRs) in bed format. Each row contains the index position, motif and TR lengths, and sequence composition of the homologous pair. TRACK also provides a visualization feature to create a scatterplot comparing the total length of shared TRs between two individuals of different species (Fig. 2).

2.3 Genotyping TRs and basic population genetic statistics

One of the many uses of TR catalogs is genotyping variants in within-species population datasets. TRACK provides an integrated module for preparing and structuring long-read data to genotype TR variants using the Tandem Repeat Genotyping Tool (TRGT) (Dolzhenko *et al.* 2024). The output is a VCF file containing genotyped TR variants across multiple samples. Additionally, TRACK estimates basic population genetic statistics, such as observed heterozygosity and genetic diversity, and generates violin plots to visualize the results (Fig. 3). TRACK's genotype functionality also converts the merged VCF into a data frame containing genotypes per sample for each TR. Additionally, TRACK also provides a data frame containing estimates for the number of unique alleles in the VCF and the range of allele lengths per locus.

3 Benchmarking

The benchmarking runs were performed using the chimpanzee and human T2T TR catalogs available in the TRACK repository, which contain 3 076 618 (225MB) and 2 756 609 (186MB) repeats, respectively. Bidirectional liftover required ∼254 MB of RAM and 30.32 s of actual CPU time, completing 1m7s of total wall-clock time. Motif sequence alignment had a total runtime of 8 h 20 m, 1.96 h of cumulative CPU time, and 1.9 GB RAM, running with 18 processor threads in parallel. All analyses were performed on a Lenovo Legion personal computer with a 13th-generation Intel Core i7-13700H processor.

4 Conclusion

TRACK proposes a novel, streamlined approach to identifying shared TRs between species, simplifying some of the often complex and time-consuming steps of comparative TR analysis. By integrating multiple features from different established tools—from catalog generation and cross-species comparison to population-level genotyping and diversity estimates—TRACK improves the accessibility of TR analyses while increasing reproducibility across analyses. Its flexibility allows users to initiate analyses at different pipeline stages, accommodating diverse research needs. As genomic data continues to expand, tools like TRACK are indispensable for uncovering biologically meaningful insights into TRs, particularly from long-read sequencing data.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and constructive feedback, which greatly improved the quality of this manuscript.

Author contributions

Carolina de Lima Adam (Conceptualization, Data curation, Methodology, Formal analysis, Software, Writing – original draft, Writing – review & editing), Joana Rocha (Data curation, Writing – review & editing), Peter Sudmant (Resources, Writing – review and editing), Rori Rohlfs (Conceptualization, Funding acquisition, Resources, Writing – review & editing)

Conflict of interest

None declared.

Funding

This work was supported by the National Science Foundation (NSF) [2144878 to R.R.].

Data availability

The data underlying this article are available in *caroladam/track* at https://github.com/caroladam/track.

References

Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**:573–80. https://doi.org/10.1093/nar/27.2.573

Bilgin Sonay T, Carvalho T, Robinson MD *et al.* Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res* 2015;**25**:1591–9. https://doi.org/10.1101/gr.190868.115

Campuzano V, Montermini L, Moltò MD *et al.* Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 1996;**271**:1423–7. https://doi.org/10.1126/science.271.5254.1423

Chiu R, Rajan-Babu IS, Friedman JM *et al.* Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol* 2021;**22**:224.

Dolzhenko E, English A, Dashnow H *et al.* Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol* 2024;**42**:1606–14. https://doi.org/10.1038/s41587-023-02057-3

Erwin GS, Gürsoy G, Al-Abri R *et al.* Recurrent repeat expansions in human cancer genomes. *Nature* 2023;**613**:96–102. https://doi.org/10.1038/s41586-022-05515-1

Farré M, Bosch M, López-Giráldez F *et al.* Assessing the role of tandem repeats in shaping the genomic architecture of great apes. *PLoS One* 2011;**6**:e27239. https://doi.org/10.1371/journal.pone.0027239

Gemayel R, Vinces MD, Legendre M *et al.* Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* 2010;**44**:445–77. https://doi.org/10.1146/annurev-genet-072610-155046

Gymrek M. A genomic view of short tandem repeats. *Curr Opin Genet Dev* 2017;**44**:9–16. https://doi.org/10.1016/j.gde.2017.01.012

Gymrek M, Willems T, Guilmatre A *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* 2016;**48**:22–9. https://doi.org/10.1038/ng.3461

Hinrichs AS, Karolchik D, Baertsch R *et al.* The UCSC genome browser database: update 2006. *Nucleic Acids Res* 2006;**34**:D590–8. https://doi.org/10.1093/nar/gkj144

Hodel RGJ, Segovia-Salcedo MC, Landis JB *et al.* The report of My death was an exaggeration: a review for researchers using microsatellites in the 21st century. *Appl Plant Sci* 2016;**4**:1600025. https://doi.org/10.3732/apps.1600025

Kashi Y, King D, Soller M. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet* 1997;**13**:74–8. https://doi.org/10.1016/S0168-9525(97)01008-1

Liang K-C, Tseng JT, Tsai S-J *et al.* Characterization and distribution of repetitive elements in association with genes in the human genome. *Comput Biol Chem* 2015;**57**:29–38. https://doi.org/10.1016/j.compbiolchem.2015.02.007

MacDonald ME, Ambrose CM, Duyao MP *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 1993;**72**:971–83. https://doi.org/10.1016/0092-8674(93)90585-E

Madeira F, Madhusoodanan N, Lee J *et al.* The EMBL-EBI job dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res* 2024;**52**:W521–5. https://doi.org/10.1093/nar/gkae241

Mitsuhashi S, Frith MC, Mizuguchi T *et al.* Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol* 2019;**20**:58–17.

Mousavi N, Margoliash J, Pusarla N *et al.* TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* 2021;**37**:731–3. https://doi.org/10.1093/bioinformatics/btaa736

Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53.

Press MO, Carlson KD, Queitsch C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet* 2014;**30**:504–12. https://doi.org/10.1016/j.tig.2014.07.008

Richard G-F, Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev* 2008;**72**:686–727. https://doi.org/10.1128/MMBR.00011-08

Sun JX, Helgason A, Masson G *et al.* A direct characterization of human mutation based on microsatellites. *Nat Genet* 2012;**44**:1161–5. https://doi.org/10.1038/ng.2398

Wang T, Antonacci-Fulton L, Howe K *et al.*; Human Pangenome Reference Consortium. The human pangenome project: a global resource to map genomic diversity. *Nature* 2022;**604**:437–46. https://doi.org/10.1038/s41586-022-04601-8

Yoo D, Rhie A, Hebbar P *et al.* Complete sequencing of ape genomes. *Nature* 2025;1–18. https://doi.org/10.1038/s41586-025-08816-3

Bioinformatics Advances, 2025, 00, 1–4
https://doi.org/10.1093/bioadv/vbaf066
Application Note