TRCompDB: A reference of human tandem repeat sequence and composition variation from long-read assemblies

Bida Gu (bidagu@usc.edu)[1] and Mark J.P. Chaisson (mchaisso@usc.edu)[1,2]

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, United States.
[2]To whom correspondence should be addressed

Abstract

Tandem repeats (TRs), including short tandem repeats (STRs) and variable-number tandem repeats (VNTRs), are hypermutable genetic elements consisting of tandem arrays of repeated motifs. TR variation can modify gene expression and has been implicated in over 50 diseases through repeat mutation and pathogenic expansion. Recent advances in long-read sequencing (LRS) enable the comprehensive profiling of TR variation in large cohorts. We previously developed `vamos`, a tool for annotating motif count and composition in LRS samples. Here, we expanded the functionality of `vamos` with new methods to construct motif databases that enhanced motif consistency, and a toolset `tryvamos` for rapid analysis using vamos output. We demonstrate that the `vamos` motif composition annotations more accurately reflect underlying genomes than other approaches for TR annotation. By applying `vamos` to 360 LRS assemblies of diverse ancestries, we constructed TRCompDB, a reference database of tandem repeat variation across 805,485 STR and 370,468 VNTR loci on the CHM13 reference genome. Using `tryvamos` for genome-wide testing, we identified 6,039 loci exhibiting strong signatures of population divergence in length or composition, yielding insight into stratification of TR loci.

Keywords: long-read sequencing, tandem repeat, motif composition

1 Introduction

Tandem repeats (TRs) refer to sequences of one or more DNA base pairs that are repeated in tandem along the chromosome. TRs with repeating units (motifs) less than 7 base pairs are categorized as short tandem repeats (STR) or microsatellites, while those with motifs of 7 base pairs or more fall into the variable number tandem repeat (VNTR) classification. As one of the most abundant and variable types of genetic variations, TRs are known to have high mutation rates that are typically several orders of magnitude higher than simple mutations [1, 2], as explained by slipped strand mispairing [3], unequal crossover [4, 5], and gene conversion [4, 6]. Their polymorphic nature makes TRs valuable markers for exploring the genetic diversity of human genome.

Over 50 TRs have been reported to be associated with diseases, highlighting the significance of TR profiling in current research [7]. Computational tools for TR genotyping can be categorized based on their output type, distinguishing between interval-based structural variant (SV) callers and motif copy number callers. Standard SV callers, such as pbsv [8] and Sniffles [9], detect variants from clustered breakpoints of mapped reads. While these methods can optionally use boundaries of tandem repeats to improve clustering of breakpoints, the calls are represented by deletion or insertion genomic intervals. A major challenge for interval-based callers is the rapid increase in the number of TR alleles as more genomes are sequenced, leading to challenges in interpreting multiallelic variation in populations. Consequently, downstream analyses often employ merging heuristics to group alleles [10, 11, 12], which results in a considerable loss of genetic diversity. Alternatively, motif copy number callers estimate the number of repeats by using read depth information from genomic alignments [13], hidden Markov models [14, 15, 16], or alignment to pangenomes [17]. However, these methods either provide only overall copy number estimates, lose accuracy with complex motifs, or are restricted to short-read data, thereby inadequately representing the sequence composition of many TRs.

Recent advancements in long-read sequencing technology have enabled the routine resolution of TRs on a genomic scale in large cohort studies [11, 18, 19, 20]. These efforts have produced long-read data and high-quality haplotype-resolved assemblies for hundreds of individuals, providing valuable resources for TR research. To overcome the limitations of existing TR profiling tools, which often simplify allele representation at the expense of losing variational information, we previously developed a tool `vamos` [21]. This tool generates accurate and informative TR motif annotations by decomposing query TR sequences at pre-configured loci on the GRCh38 reference genome into their efficient motif compositions. For each TR locus, the full motif set comprises all motifs collected from a population of genomes. The efficient motif set is a subset of this full set, optimized to minimize the number of motifs while minimizing information loss. This loss is quantified by the total edit distance incurred when replacing removed motifs with their most similar counterparts among the remaining efficient motifs. The level of condensation between the efficient motif set and the full motif set is controlled by a single parameter, q. The annotation of query TR sequences as concatenations of efficient motifs, which yield the smallest edit distance, is achieved using the StringDecomposer algorithm [22].

In this study, we present multiple advances of the `vamos` method and software ecosystem. First, we present an enhanced framework that extends the motif database for any reference genome (e.g., T2T-CHM13 [23]) for use with `vamos`. The updated database also provides more accurate TR boundaries and motif definitions than our preliminary analysis [21]. By annotating 360 publicly available high-quality long-read haplotype assemblies with the improved TR database from four sources (HGSVC [11], Human Pangenome Reference Consortium (HPRC) [18], Chinese Pangenome Consortium (CPC) [19], and Arab Pangenome Reference (APR) [20]), we establish a comprehensive TR allele reference TRCompDB (TR Composition DataBase) that incorporates diverse population ancestries. Additionally, we demonstrate the accuracy and richness of compositional information provided by `vamos` through a comparison with the similar tool TRGT [16]. Finally, we introduce a toolset called `tryvamos` to analyze combined population datasets from `vamos` output.

2 Methods

2.1 Construction of TR boundaries and motifs from HPRC genomes

We created our TR locus and motif database using 94 HPRC haplotypes [18]. Tandem repeats were first annotated using Tandem Repeat Finder [24] and RepeatMasker [25] with default parameters. Each TR call was characterized by an interval in the genome and a consensus motif that was repeated, potentially with variation, within this interval. The TR calls sharing overlaps on the same assembly were merged. In the majority of instances the shortest consensus motif among all merged calls was selected as the new consensus motif. In relatively few instances ($< 1\%$), two-base pair STR or homopolymer annotations were nested in higher-order tandem repeats, and the annotation of the larger instance more accurately reflected the tandem repeat structure of the locus. To address this we defined nested repeats as annotations with a one or two base motif and with a length less than half the length of the total spanning region of all overlapping loci. In these instances the motif for the longest overlapping annotation was used. Subsequently, the basis of our TR database was formed by mapping TR boundaries to the GRCh38 and CHM13 reference genomes.

Constructing a TR database from multiple genomes is challenging due to inconsistencies in TR boundaries and motif decomposition. A TR with a 3-base consensus in one genome might be misinterpreted as a 6-base consensus in another, leading to discrepancies and complicating allele comparisons. This is not explicitly solved by `vamos` efficient motif selection. To address this, we used k-means clustering to identify inconsistent TR boundaries and motifs before merging TR boundaries across samples. The clustering groups overlapping TR calls based on genomic proximity and consensus similarity, using boundary coordinates and consensus motif lengths as input features (Fig. 1). The optimal number of clusters was determined by the highest Silhouette score. Clusters supported by fewer than three genomes were excluded. Each remaining cluster was refined by averaging boundary coordinates and deriving a new motif consensus sequence from the multiple sequence alignment of all motif consensus sequences within the cluster. Overlapping clusters were merged into a single TR locus, with the motif consensus from the most-supported cluster used to re-decompose each genomic TR sequence into motifs via the StringDecomposer algorithm [22]. To ensure proper decomposition, boundary motifs differing by more than 20% in length from the

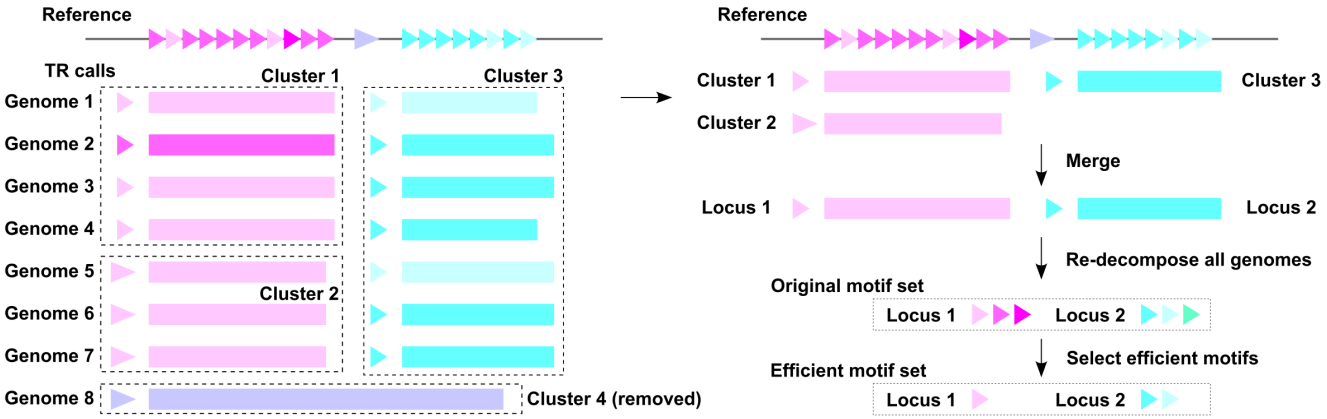

Figure 1: **Overview of the clustering algorithm for the construction of TR boundaries and motifs.** Preprocessed TR calls (TRF and RepeatMasker) are first clustered by boundary coordinates and length of consensus motifs before merging across clusters. Poorly supported clusters are removed and the consensus motif of the best supported cluster is selected for each locus after merging. Next, the motif set used for efficient motif selection is obtained by re-decomposing all input genomic TR sequences using the consensus motif, followed by selection of efficient motifs.

consensus, as well as single-base motifs for non-homopolymers, were excluded. The remaining motifs formed the complete motif set for each locus. For computational efficiency, TR loci longer than 10,000 bases or with more than 500 unique motifs were filtered out. Finally, efficient motifs were obtained with a compression level of $q = 0.1$ as described previously [21].

2.2 Tandem repeat analysis using `vamos` output (`tryvamos`)

The variation annotated by `vamos` differs from conventional sequence variants because they are highly multiallelic, and differ in both variant length and sequence composition. We developed a toolset `tryvamos` that has functions to analyze `vamos` TR annotations. This toolset is designed to accept `vamos` VCF files as input and provides functions for panel VCF combination (`--combineVCF`), feature matrix generation for high-dimensional analysis (`--quickFeature`), panel allele visualization (`--waterfallPlot`) (Fig. 2), allelic comparison of paired samples (`--pairwiseCompare`), as well as genome wide tests and visualization of divergent loci between two panels (`--testTwoPanels`). As the scope of LRS studies increases, the `tryvamos` toolset will undergo iterative enhancements to include new capabilities.

3 Results

3.1 TR composition effectively features major population groups in TRCompDB

We constructed motif databases for 1,243,954 loci in GRCh38 and 1,175,953 loci in CHM13 (Figure 3a). Due to the imprecise definition of centromeres on GRCh38, more centromeric TRs were retained, resulting in a higher total number of loci on GRCh38. This release contains 289,891 (23.3%) loci in GRCh38 and 252,975 (21.5%) loci in CHM13 that overlap or are overlapped by transposable elements by at least 50%. Unless otherwise specified, our analysis is based on the CHM13 reference. Homopolymer loci are known to have low accuracy with most sequencing technologies [26] and were excluded. The PCAs of the 360 haplotype genomes in TRCompDB using either TR annotation length or counts of the most frequent motif show clear clustering of major population groups (Fig. 3b). This clustering resembles what is observed in SNP data [27], indicating that TR loci contain population-specific information. We used `tryvamos --testTwoPanels` to search for significantly diverging TR loci comparing

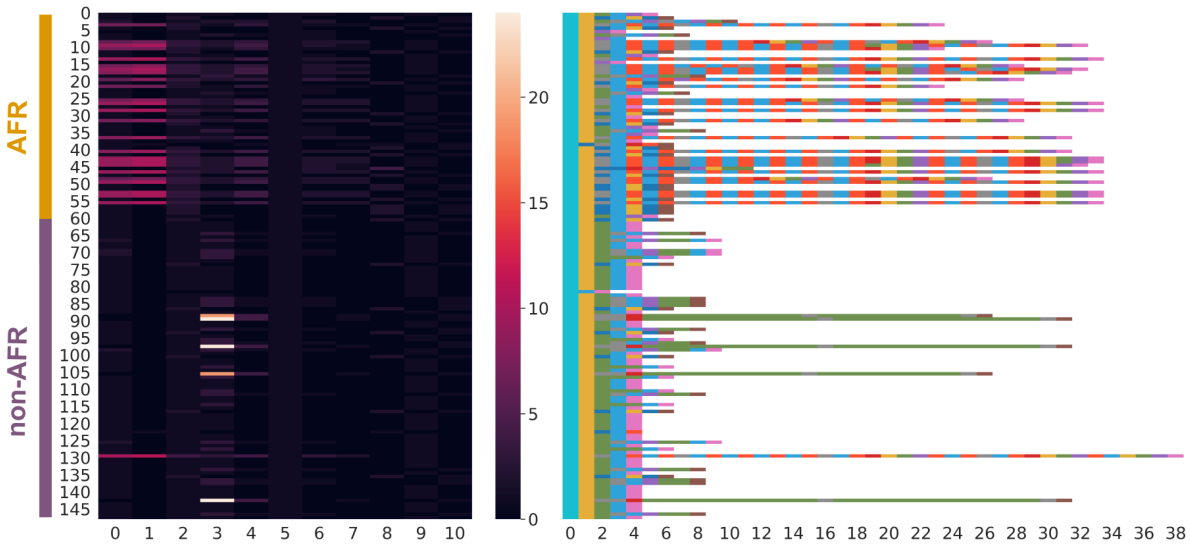

Figure 2: **Example population-specific VNTR locus at chr1:3511428-3511720 of 146 HGSVC and HPRC haplotypes.** For each haplotype, counts of each motifs (indexed from 0 to 10) are visualized in the left heatmap and allele landscape by motif compositions is visualized in the right waterfall plot.

62 African and 86 non-African genomes in the HGSVC and HPRC datasets. To avoid results from extremely small sample variance or low sample counts, we excluded loci with small variance (< 1) in the t-test group and loci with fewer than 10 samples in either population. This resulted in 117,737 and 999,327 tested loci for the t-test and KS-test groups, respectively, revealing 4,701 significantly diverging loci (corrected p-value < 0.05; $p < 10^{-8}$). Figure 2 shows the motif distribution and allele landscape of an example population-divergent TR.

3.2 The improved database represents more consistent definition of TR motifs

To compare with the previously published database for the GRCh38 reference, we aligned 795,133 GRCh38 TR loci between the two versions, permitting a 5-base offset on each boundary. We defined motif inconsistency as the average pairwise distance of motifs to the consensus motif. The average ratio of the motif inconsistency between the old and the new database is 1.5 for the full motif set and 1.6 for the efficient motif set ($q = 0.1$) indicating that the new motif database has more closely related individual motifs.

3.3 Efficient motifs ($q = 0.1$) represents majority of motif compositions

Our previous work demonstrated that the `vamos` approach preserves substantially more compositional diversity for TR alleles compared to variant merging approaches [21]. We computed the ratio of motifs remaining after motif selection to the total count of motifs for two measures, one considering the total cardinality of motifs, and the other only counting unique instances of motifs (Fig. 3c). The first measure provides information for how frequent the efficient motifs are while the second measure gives the compression of the motif set. On average 29.0% ($\pm$ 16.5%) of unique motifs are replaced in the efficient motif set, representing only 7.6% ($\pm$ 4.3%) of the total motifs. This indicates replaced motifs only constitutes a small proportion of the full motif set. The TR loci showing greater diversity measured by a higher count (> 10) of unique motifs undergo more condensation during the efficient motif selection procedure, demonstrating that elevated TR diversity in most loci primarily arises from the emergence of extra motifs that are rare and/or less representative in terms of sequence compositions. Moreover, our analysis reveals distinct trends regarding motif compression between VNTRs and STRs. Specifically, VNTRs with fewer than 10 unique motifs exhibit less compression compared to STRs with a similar count of unique motifs. However, a reversal of this trend is observed for TRs with more than 10 unique motifs. This implies that the

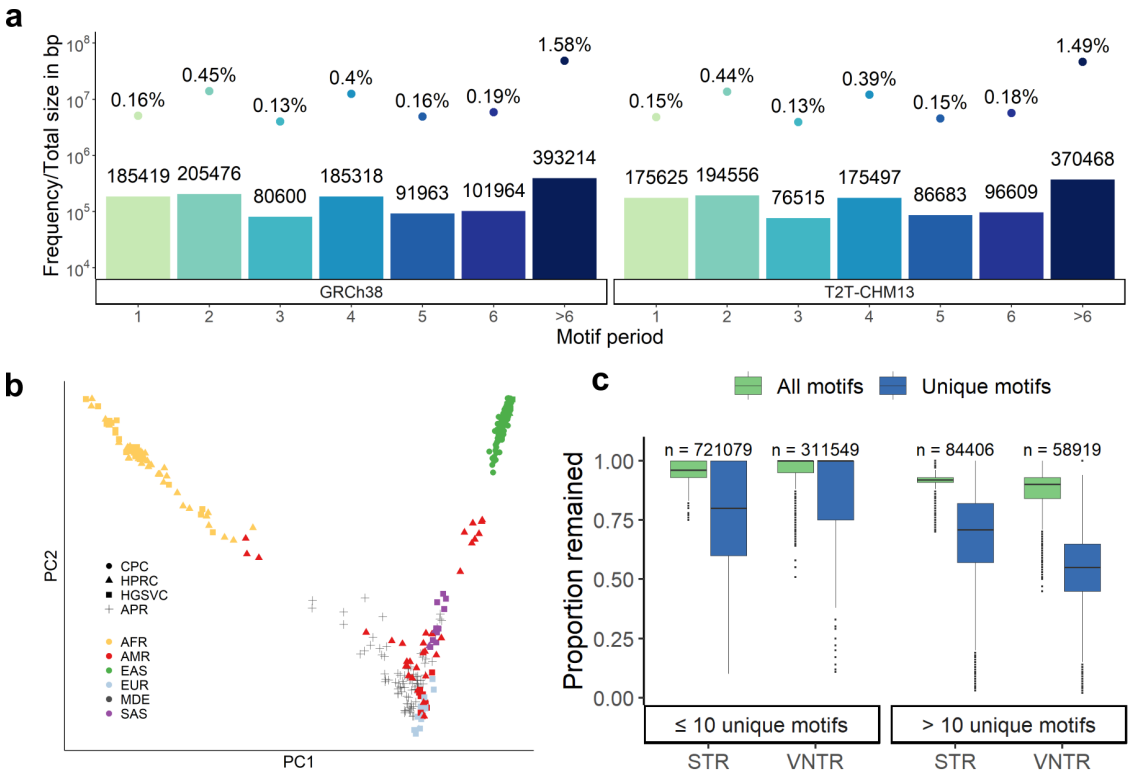

Figure 3: **a**. Frequency and size distribution of TRs in the `vamos` databases on GRCh38 and CHM13 reference genomes. Labeled numbers indicate counts and genomic percentages by size. **b**. PCA of 360 genomes used in TRCompDB by counts of the most frequent motif. **c**. Proportion of remained motifs in the efficient motif set ($q = 0.1$) by all motif (green) or unique motif (blue).

relative amplification of locus diversity is more pronounced for VNTRs than STRs as the number of unique motifs increases, largely attributable to the longer and more variable nature of VNTR motifs.

3.4 Comparison to TRGT on compositional accuracy of motif annotations

We compared `vamos` with TRGT [16] to evaluate their performance in annotating TR composition. While `vamos` is developed for both raw sequencing reads and assemblies, TRGT is not optimized to analyze assemblies. We thus compared both tools only on the 30X HG002 HiFi data from the NCBI-PRJNA586863 project. Moreover, both tools use predefined TR boundaries and motif definitions on GRCh38 and are provided with default TR databases. The TRGT default database encompasses 937,122 loci all defined by a single motif and added with 25 bp flanks on each boundary, although TRGT supports annotation by multiple motifs per locus. On the other hand, the `vamos` default database (efficient motif set by $q = 0.1$) contains 1,240,349 loci with no added flanks, where 986,975 loci are defined with more than one motif. To ensure fair comparison, we compared the two tools on three sets of TRs, namely the default TRGT loci with the default single motif (the "TRGT-single group"), the default `vamos` loci with the default single/multiple motifs (the "vamos-multiple" group), and the default vamos loci, but with only the most frequent motif counted from the input genomes for each locus (the "vamos-single" group). Since `vamos` annotates all sequences within the pre-defined TR boundaries as repeat motifs, the 25 bp flanks of the default TRGT loci were removed for both tools.

We assessed annotation accuracy by comparing the edit distances between the nucleotide sequences reconstructed from each tool's motif annotations and the corresponding HPRC HG002 assembly sequences. The `vamos` annotations

were converted reconstructions by concatenating motif sequences in the order of the annotation string. In contrast, TRGT annotates TRs as possibly discontinuous segments that are represented as repetition of a single motif. The total counts of each input motif in the entire annotation is reported, but not the number of repetitions of a motif in a specific segment. This makes sequence reconstruction challenging if the same motif is annotated into discontinuous segments within a TR. Consequently, loci with such TRGT annotations were excluded from the comparison. Finally, given the typical presence of approximately 5 bp non-repetitive flanking sequences on the default TRGT TRs after flank removal, we considered a tool superior if it showed at least a 10 edit distance advantage over the other tool when compared to the assembly sequence. After exclusion of loci with challenging sequence reconstruction, 883,103, 1,126,655, and 968,017 loci remained for the "TRGT-single", "vamos-single", and "vamos-multiple" groups, respectively. As shown in Table 1, the nucleotide sequences reconstructed from `vamos` show higher similarity to the assembly sequences in all three groups, suggesting that `vamos` produces more accurate annotations for TR motif compositions.

TR set	`vamos` better	comparable	TRGT better
TRGT-single	64,527 (7.3%)	817,037 (92.5%)	1,539 (0.2%)
vamos-single	55,779 (5.0%)	1,068,695 (94.8%)	2,181 (0.2%)
vamos-multiple	34,916 (3.6%)	931,753 (96.3%)	1,348 (0.1%)

Table 1: **Comparison of annotation accuracy of `vamos` and TRGT.**

4 Discussion

Recent efforts to generate long-read sequencing data and high-quality assemblies for large cohort studies has brought increased interest in the study of TRs [26]. Annotation of TR compositions using a representative subset of motifs overcomes the issue of oversimplifying allele diversity in existing TR profiling tools. Such approach relies on obtaining accurate and consistent definition of TR boundaries and motifs from a collection of genomes. By clustering TR calls sharing genomic proximity and consensus similarity, our proposed TR boundary and motif construction framework generates consistent motifs that result in TR annotations evidencing meaningful population diversity. This framework, together with the previously published tool `vamos`, is in principle applicable to study of TR motif compositions for any species on any reference genomes. While the exclusion of poorly supported TR clusters in the k-means procedure may discard rare TR loci in a population, this issue can be mitigated as more high-quality genomes become available.

Comparison to TRGT shows that `vamos` produces more accurate motif annotations. Unlike TRGT, which annotates only sequences closely matching the input motif(s), `vamos` annotates the entire TR sequence. Consequently, many instances where `vamos` appears more accurate result from TRGT's lack of annotations due to low similarity between the input motif(s) and the genomic TR sequence. This issue is prevalent in loci with low sequence purity and a single input motif. Identifying all such cases is challenging. However, we estimate that 16,588 (TRGT-single), 12,128 (vamos-single), and 1,652 (vamos-multiple) loci have few motif matches found by TRGT due to the motif(s) not reflecting the TR sequence composition and at least ten times as many motif annotations are "forced" at corresponding loci by the decomposition algorithm in `vamos`. This reflects the importance of carefully defined motif sets reflecting composition of the query TR sequences when applying both tools to these loci.

Although TR length has been frequently associated with biological function, variation in the motif composition of TRs have been underexplored in biological studies. The `tryvamos` toolset, with its infrastructure for genome-wide statistical analyses and visualization capabilities, enables detailed examination of TR compositions using motif annotations generated by `vamos` on long-read sequencing data. Given the expanding research community and growing interest in TR studies, ongoing efforts will be dedicated to enhancing `tryvamos` with additional functionalities to meet emerging research needs. Finally, the annotations for 360 high-quality genome assemblies

covering approximately 1.2 million TRs in the TRCompDB database, serves as a reference for global diversity of neutral variation in tandem repeats sequences.

5 Competing interests

No competing interest is declared.

6 Author contributions statement

B.G. wrote the manuscript and conducted the analysis. M.J.P.C. wrote the manuscript.

7 Code and Data Availability

- The TR construction pipeline and `tryvamos` is available at: `https://github.com/chaissonlab/vamos`
- Assemblies produced by the Human Genome Structural Variation Consortium are available through `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/assemblies/`.
- Assemblies produced by the Human Pangenome Reference Consortium are available through `https://github.com/human-pangenomics/HPP_Year1_Assemblies/blob/main/assembly_index/Year1_assemblies_v2_genbank.index`.
- Assemblies produced by the Chinese Pangenome Consortium are available through `https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA011422`.
- Assemblies produced by the Arab Pangenome Reference project are available through `https://mbruniversity.sharepoint.com/sites/apr/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2Fapr%2FShared%20Documents%2FArab%20Pangenome%20Reference%2FAPR%20Assemblies&p=true&ga=1`.
- The HG002 HiFi data is available from NCBI project PRJNA586863.
- Motif set files and combined vcf of TR annotations (TRCompDB) are available through zenodo at `https://zenodo.org/records/13263615`.

References

[1] Matthieu Legendre, Nathalie Pochet, Theodore Pak, and Kevin J Verstrepen. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome research*, 17(12):1787–1796, 2007.

[2] Kristin A Eckert and Suzanne E Hile. Every microsatellite is different: Intrinsic dna features dictate mutagenesis of common microsatellites present in the human genome. *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center*, 48(4):379–388, 2009.

[3] Gene Levinson and George A Gutman. Slipped-strand mispairing: a major mechanism for dna sequence evolution. *Molecular biology and evolution*, 4(3):203–221, 1987.

[4] Alec J Jeffreys, David L Neil, and Rita Neumann. Repeat instability at human minisatellites arising from meiotic recombination. *The EMBO Journal*, 17(14):4147–4157, 1998.

[5] Hélène Debrauwère, Jérôme Buard, Jacques Tessier, Dominique Aubert, Gilles Vergnaud, and Alain Nicolas. Meiotic instability of human minisatellite ceb1 in yeast requires dna double-strand breaks. *Nature genetics*, 23 (3):367–371, 1999.

[6] Frédéric Pâques, Guy-Franck Richard, and James E Haber. Expansions and contractions in 36-bp minisatellites by gene conversion in yeast. *Genetics*, 158(1):155–166, 2001.

[7] Mark JP Chaisson, Arvis Sulovari, Paul N Valdmanis, Danny E Miller, and Evan E Eichler. Advances in the discovery and analyses of human tandem repeats. *Emerging topics in life sciences*, 7(3):361–381, 2023.

[8] pbsv, 2023. URL `https://github.com/PacificBiosciences/pbsv`.

[9] Moritz Smolka, Luis F Paulin, Christopher M Grochowski, Dominic W Horner, Medhat Mahmoud, Sairam Behera, Ester Kalef-Ezra, Mira Gandhi, Karl Hong, Davut Pehlivan, et al. Detection of mosaic and population-level structural variants with sniffles2. *Nature biotechnology*, pages 1–10, 2024.

[10] Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications*, 10(1):1–16, 2019.

[11] Peter Ebert, Peter A Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537):eabf7117, 2021.

[12] Melanie Kirsche, Gautam Prabhu, Rachel Sherman, Bohan Ni, Alexis Battle, Sergey Aganezov, and Michael C Schatz. Jasmine and iris: population-scale structural variant comparison and analysis. *Nature Methods*, 20(3): 408–417, 2023.

[13] Paras Garg, Alejandro Martin-Trujillo, Oscar L Rodriguez, Scott J Gies, Elina Hadelia, Bharati Jadhav, Miten Jain, Benedict Paten, and Andrew J Sharp. Pervasive cis effects of variation in copy number of large tandem repeats on local dna methylation and gene expression. *The American Journal of Human Genetics*, 108(5): 809–824, 2021.

[14] Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, and Vineet Bafna. Targeted genotyping of variable number tandem repeats with advntr. *Genome research*, 28(11):1709–1719, 2018.

[15] Mehrdad Bakhtiari, Jonghun Park, Yuan-Chun Ding, Sharona Shleizer-Burko, Susan L Neuhausen, Bjarni V Halldórsson, Kári Stefánsson, Melissa Gymrek, and Vineet Bafna. Variable number tandem repeats mediate the expression of proximal genes. *Nature communications*, 12(1):1–12, 2021.

[16] Egor Dolzhenko, Adam English, Harriet Dashnow, Guilherme De Sena Brandine, Tom Mokveld, William J Rowell, Caitlin Karniski, Zev Kronenberg, Matt C Danzi, Warren A Cheung, et al. Characterization and visualization of tandem repeats at genome scale. *Nature Biotechnology*, pages 1–9, 2024.

[17] Tsung-Yu Lu, Human Genome Structural Variation Consortium Munson Katherine M. 2 Lewis Alexandra P. 2 Zhu Qihui 3 Tallon Luke J. 4 Devine Scott E. 4 Lee Charles 3 5 6 Eichler Evan E. 2 7, and Mark JP Chaisson. Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nature Communications*, 12(1):4250, 2021.

[18] Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K Lucas, Jean Monlong, Haley J Abel, et al. A draft human pangenome reference. *Nature*, 617(7960):312–324, 2023.

[19] Yang Gao, Xiaofei Yang, Hao Chen, Xinjiang Tan, Zhaoqing Yang, Lian Deng, Baonan Wang, Shuang Kong, Songyang Li, Yuhang Cui, et al. A pangenome reference of 36 chinese populations. *Nature*, pages 1–10, 2023.

[20] Nasna Nassir, Mohamed Almarri, Muhammad Kumail, Nesrin Mohamed, Bipin Balan, Shehzad Hanif, Maryam AlObathani, Bassam Jamlalail, Hanan Elsokary, Dasuki Kondaramage, et al. A draft arab pangenome reference. *bioRxiv*, pages 2024–07, 2024.

[21] Jingwen Ren, Bida Gu, and Mark JP Chaisson. Vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biology*, 24(1):175, 2023.

[22] Tatiana Dvorkina, Andrey V Bzikadze, and Pavel A Pevzner. The string decomposition problem and its applications to centromere analysis and assembly. *Bioinformatics*, 36(Supplement_1):i93–i101, 2020.

[23] Arang Rhie, Sergey Nurk, Monika Cechova, Savannah J Hoyt, Dylan J Taylor, Nicolas Altemose, Paul W Hook, Sergey Koren, Mikko Rautiainen, Ivan A Alexandrov, et al. The complete sequence of a human y chromosome. *Nature*, 621(7978):344–354, 2023.

[24] Gary Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research*, 27(2): 573–580, 1999.

[25] Smit et al. Repeatmasker, 2013. URL `http://www.repeatmasker.org`.

[26] Adam C English, Egor Dolzhenko, Helyaneh Ziaei Jam, Sean K McKenzie, Nathan D Olson, Wouter De Coster, Jonghun Park, Bida Gu, Justin Wagner, Michael A Eberle, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nature Biotechnology*, pages 1–12, 2024.

[27] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, et al. The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.