Briefings in Bioinformatics, 2026, 27, bbag031
https://doi.org/10.1093/bib/bbag031
Published: 5 February 2026
Review

OXFORD

Computational tools for tandem repeat detection using long-read sequencing

Qian Liu [1,2,*] and Jincheng Li [1]

[1] Nevada Institute of Personalized Medicine, College of Sciences, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, United States
[2] School of Life Sciences, College of Sciences, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, United States

*Corresponding author. Nevada Institute of Personalized Medicine, College of Sciences, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, United States.
E-mail: qian.liu@unlv.edu.

Abstract

Tandem repeats (TRs) play essential roles in a variety of biological functions, and their abnormal expansions are significantly implicated in phenotypic variation and cause >60 human diseases. However, long TR regions cannot be reliably detected using short-read sequencing, and long-read sequencing enables accurate genome-wide detection of TRs. In recent years, various computational tools have been developed to detect and genotype TRs from long-read data. In this survey, we systematically categorize and review 39 computational tools designed for TR detection, visualization and functional interpretation. We discuss their strengths and limitations for TR detection from long-read sequencing data, highlighting current challenges and future directions to advance long-read TR detection methodologies.

Keywords tandem repeats, computational tools, long-read sequencing

Introduction

Tandem repeats (TRs) are directly adjacent repetitions of specific nucleotide motifs in genomes. For example, the trinucleotide motif 'CAG' repeated consecutively forms a TR (Fig. 1A). Based on the sizes of TR motifs, TRs could be grouped into three categories [1], microsatellites (or short tandem repeats or STRs with motif lengths of 1–9 bp), minisatellites with motif lengths of 10–99 bp and satellites with motif lengths ≥100 bp. Microsatellites and minisatellites are collectively known as variable number of tandem repeats (VNTRs) [1]—please note that in some published works [2, 3], VNTRs solely refer to minisatellites. TRs constitute ∼6% of the human genome [4], and are among the most dynamic and mutable regions, accounting for >70% of structural variants [5, 6] longer than 50 bp. These TRs regulate gene expression [7–12], influence chromatin structures [13–16] and affect genome stability [17–19]. For example, increasing STR length in promoter regions is associated with altered gene expression variability in Saccharomyces cerevisiae [8], and STRs contribute to gene expression variation [7, 10, 12] at a genome-wide scale in humans [11]. Disease-associated STRs have also been found to co-localize with chromatin domain boundaries [13], and satellite TRs can influence chromatin architecture, inducing paramutation [14] and promoting heterochromatin establishment and spreading [16]. In addition, AT-dinucleotide repeats, as well as CAGAGG and CACAG motifs, have been shown to drive fragile site formation [17] and serve as structure-forming sites of fork collapse [18, 19].

TRs exhibit high polymorphism across individuals and populations [20] and are promising markers to study complex traits [21], genetic diversity, population structure, evolutionary dynamics as well as human disorders. For example, the expansions of specific TRs are known to cause over 60 human disorders [22–39], including Huntington's disease [1], fragile X syndrome [1], various spinocerebellar ataxias [1] and cancer [40]. The list of repeat expansion disorders (REDs) continues to grow [41]. In many neurological REDs, the number of motif units in TR regions inversely correlates with disease onset age and positively with disease severity: larger expansions usually lead to earlier and more severe disease manifestations [42, 43]. Expanded TRs in coding or regulatory regions can influence the stability, splicing and translation of RNA molecules, and lead to toxic RNAs or aberrant proteins when expanded beyond normal thresholds. In summary, TRs have been recognized as biologically and clinically important genomic variants, serving as promising therapeutic targets and diagnostic/prognostic biomarkers in clinical diagnostics [44], forensic analysis and paternity testing [45]. Thus, there is an urgent need to determine TR expansion locations and quantify TR repeat counts for individual genomes.

TRs can be detected by several approaches, including Southern blotting [46], Sanger sequencing [47], polymerase chain reaction

Received: September 23, 2025. **Revised:** December 1, 2025. **Accepted:** January 12, 2026

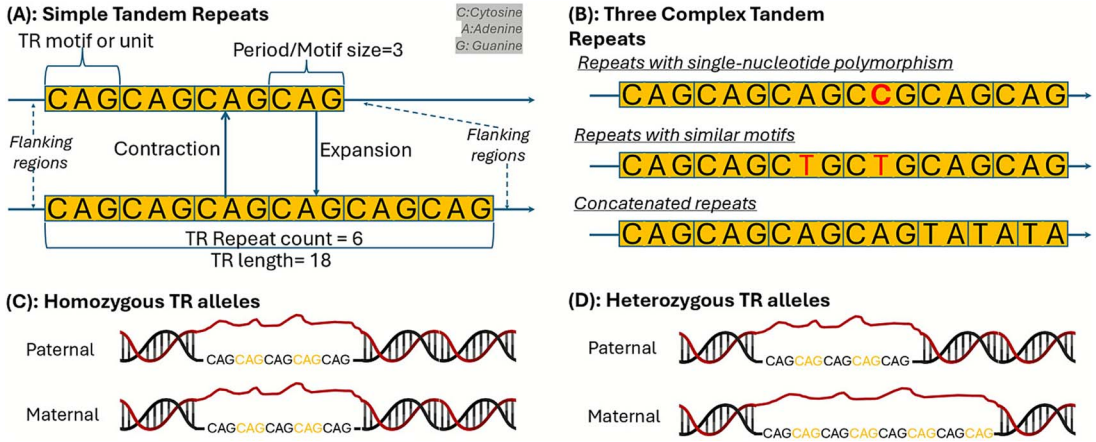

Figure 1 Tandem repeats (TRs) and their haplotypes. (A) Schematic illustration of a trinucleotide tandem repeat and its expansion or contraction. (B) Representative examples of three types of complex TRs. (C) Example of a homozygous TR with five repeat units on both alleles. (D) Example of a heterozygous TR with five repeat units on the paternal allele and eight repeat units on the maternal allele.

(PCR) [48], optical genome mapping [49] and various short-read sequencing techniques [50–57]. However, these approaches have their own inherent limitations. The first three traditional experimental approaches are useful to validate TRs at a known locus, but they are labor-intensive, gene-specific, and cannot be used for genome-wide applications. Short-read sequencing could overcome this limitation and enable genome-wide TR profiling. Thus, various short-read based TR detection tools are developed [50–57] including *de novo* detection tools like ExpansionHunter Denovo [51], STRling [52] and superSTR [53]. But short-read sequencing is limited by read lengths (typically 100–200 bp), which are often insufficient to span long or complex TRs; As a result, short-read based TR tools cannot reliably detect large TR expansions that often have the greatest biological impact. Optical genome mapping [49] is an alternative strategy to detect longer TRs. But optical genome mapping does not provide nucleotide-level resolution and may yield imprecise estimates of repeat lengths.

In contrast, long-read sequencing techniques, i.e. single-molecule real-time sequencing (SMRT) from Pacific Biosciences [58, 59] and Oxford nanopore technologies [60, 61] (ONT), overcome these limitations. These long-read platforms generate reads spanning hundreds or tens of thousands of bp, allowing direct measurements of complete TR regions and their flanking regions. Long reads thus enable accurate TR sizing, motif detection as well as improved alignment quality in low-complexity TR regions. As long-read sequencing has become more prevalent in the last several years, the importance of TR detection has gained increasing attention particularly in the context of long-read sequencing techniques. For example, in 2024, TRGTdb was built on 100 Pcbio HiFi sequencing data [5] to study genomic and population-wide variability of 937 122 TRs. Similarly, hundreds of ONT genomes have been investigated to comprehensively catalog human genetic variations including TR variants [62]. In 2025, a large-scale study characterized 3.6 billion TR alleles from 1027 long-read genomes to explore genome-wide patterns of TR pathogenicity [63].

This long-read advancement has catalyzed the development of 24 computational tools for TR detection and quantification using long-read data. Several publications that discussed the recent progress of TR characterization [64, 65] included a summary of long-read TR

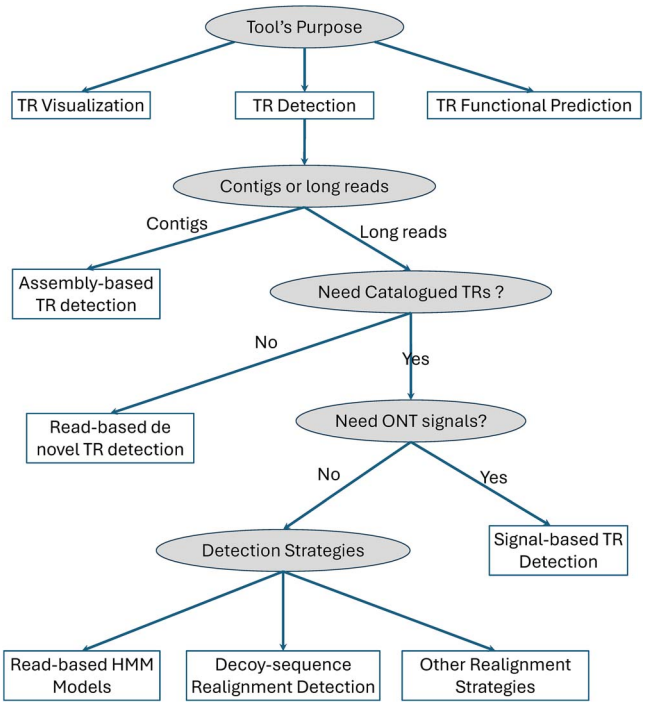

Figure 2 Categorization logic for different TR-associated tools especially for TR detection using long reads. Oval circles: categorization criteria; rectangles: tool categories. TR: tandem repeats.

detection tools but only with at most 10 tools reviewed. A comprehensive survey of long-read TR detection tools is urgently needed to guide future tool development, maximize effective utility of expanding long-read datasets, and facilitate large-scale TR studies. In this review, we systematically summarize computational tools for TR detection using long-read sequencing as well as tools for TR visualization and functional annotation according to the categorization criteria in Fig. 2. The detailed comparison (Supplementary Table S1) of these tools will help researchers and developers to select appropriate strategies and drive the development of next-generation TR detection tools optimized for long-read data.

Categories of tandem repeat detection tools

TR detection tools can be grouped according to their input requirements and algorithm design, as illustrated in Fig. 2. Based on whether the tools require predefined TR loci, these tools can be categorized into catalogue-based tools that focus on characterizing predefined known TR sites (typically derived from reference genomes), and *de novo* detection tools that do not require prior knowledge of TR locations and are designed to identify novel or uncatalogued TRs directly from the input data. Based on input data type, these tools can be categorized into read-based tools that operate directly on long reads, signal-based tools that work on ONT sequencing signals, and assembly-based tools that analyze assembled sequences (e.g. contigs or reference genomes). Assembly-based tools typically analyze reference genomes, and while they are not considered as a primary focus of this review, we still briefly discuss them because they are incorporated as a core component into several read-based tools.

Most detection tools are read-based that can be further categorized by their core detection strategies into two groups: (i) hidden Markov models (HMM)-based tools, which use HMM to infer TR boundaries, motifs, or repeat counts in a probabilistic framework, and (ii) realignment-based tools, which rely on customized alignment strategies (e.g. to decoy references, flanking regions, or profile models) to improve TR boundary resolution and repeat quantification.

Based on this classification, we will discuss and highlight detection strategies, different capabilities, limitations, and intended use cases of existing TR detection tools, as shown in Table 1.

Assembly-based detection of *de novo* tandem repeats

Assembly-based detection tools require input sequences derived from assembled genomes or reference genomes. These tools identify the location and motifs of TRs directly from the input sequences. Most are designed for *de novo* TR discovery, particularly to catalogue TRs in reference genomes. They can also be applied to contigs or scaffolds generated by assembly tools. These tools can be broadly categorized into three common algorithmic types, although other specialized approaches also exist.

One type of detection strategies is pattern recognition, such as Phobos [67], Tandem Repeats Finder (TRF) [70], mreps [66], vamos [73], TRStalker [72], and XSTREAM [75]. These tools treat DNA sequences as a string and detect TRs and their motifs using combinatorial pattern matching. They scan the entire input sequences for adjacent TR units of varying motif size and use a certain degree of mismatches between TR units to detect adjacent TR units that either match exactly or approximately. The difference is that TRF [70] uses a probabilistic scoring system to penalize mismatches including indels, while mreps [66] use Hamming distance to tolerate k-mismatches. The second type of tools is to use self-alignment of input sequences (like Ribbit [68] and Dot2dot [69]) or between local adjacent regions (tandemSWAN [71]). These tools conduct alignment between an input sequence to itself or its compressed sequences to identify self-similar regions. When visualizing the alignment as dot-plot, TR regions demonstrate diagonal stripes parallel to the main diagonal. These two types of tools usually perform well on clean and low-error sequences from reference genomes or assembly sequences, but they are not suitable for noisy long-read data. The third type of detection algorithms is wavelet analysis [74]. These tools convert DNA sequences into numerical vectors based on frequency profiles of each nucleotide type in the input sequences, and apply wavelet transforms and image segmentation for TR detection. This method can detect TR regions with diverse or imperfect motifs, including low-complexity or degenerate regions that may not be captured by string-based methods. However, it requires manual parameter tuning, such as fixed energy thresholds and color channel selection, and thus may not generalize well across different genomic contexts. It also relies on energy thresholds and visual segmentation rather than nucleotide-level alignment, leading to poor base-level precision in defining repeat boundaries.

In summary, assembly-based tools are effective for curated or assembled sequences, but they cannot directly analyze diploid or polyploid genomes with multiple haplotypes unless haplotypes are resolved prior to analysis. They are often used as components within more complex long-read TR detection pipelines discussed below.

Long-read tools for catalogued tandem repeat detection

Most TR detection tools for long-read sequencing data are designed to detect TR regions at known or catalogued TR loci. These tools typically require two main inputs: an alignment file of long reads and a predefined TR locus including genomic coordinates and motif information. Alignment files in BAM format are usually generated by using long-read specific aligners, e.g. minimap2 [105], to map long reads for an individual sample to a reference genome.

Then, these tools detect TRs following a three-step pipeline as shown in Fig. 3. First, for an input TR locus, candidate reads aligning to the TR region usually including flanking sequences are extracted. Second, the extracted reads are processed using different algorithms to identify TR boundaries and TR repeat counts for each sequenced read. Third, TR counts from multiple reads are aggregated for the TR locus to genotype TR repeat counts for each haplotype. Gaussian Mixture Model (GMM) or hierarchical clustering is used to infer the most likely TR repeat counts for the input samples. The output typically consists of two identical TR counts for a homozygous TR locus (Fig. 1C) and two distinct TR counts for a heterozygous TR locus (Fig. 1D). If the input contains a list of predefined TR loci, this detection process is repeated for each TR locus. In this process, the key challenge is the accurate identification of TR boundaries within each read, particularly distinguishing the TR region from its flanking sequences. This step strongly influences repeat count estimation and downstream genotyping accuracy. To address this challenge, four main strategies are adopted in different tools for TR boundary identification, which will be described in detail below.

Read-based hidden Markov model strategies

HMMs are a powerful probabilistic framework to model sequential data with underlying hidden structures. In an HMM, a sequence of hidden (unobserved) states emits a sequence of observable outputs, with both state transitions and emissions controlled by probability distributions. HMMs are particularly effective for noisy or imperfect data, making them well-suited for TR detection in long-read sequencing where error rates are high and motif boundaries may be ambiguous. In long-read TR detection, long reads serve as the observed data, while the underlying TR structures, such as TR motif content, repeat units and TR boundaries, are modeled as hidden states. Then, HMMs can be used to infer the start and end positions of TRs in long reads, TR motifs, and/or TR repeat counts per read.

Table 1 The summary of TR-associated tools. Tools are categorized according to the criteria in Fig. 2. TR: tandem repeats.

Category	Tools	Strengths	Limitations	Applications
Assembly-based *de novo* TR detection	mreps [66], Phobos [67], Ribbit [68]/Dot2dot [69], TRF [70], tandemSWAN [71], TRStalker [72], vamos [73], wavelet analysis [74], XSTREAM [75]	✓Robust genome-wide TR annotation ✓Capable of discovering novel TR loci	✓Cannot handle sequencing errors ✓Limited by assembly availability and quality	✓Annotation of reference genome or contigs ✓Allele information of TRs is not required ✓Accurate TR estimation
Read-based HMM detection for catalogued TRs	adVNTR [76], code-adVNTR [77], LongTR [78], PacmonsTR [79], RepeatHMM [80], TRGT [81], TRGT-denovo [82]	✓Robust with sequencing errors ✓Can handle complex TRs with known motifs ✓High genotyping accuracy	✓Difficult to scale to longer TR expansions ✓Model initialization is sensitive ✓Need known TR loci	✓Know TR loci ✓Sequencing data is available ✓Need precise genotyping
Decoy-sequence realignment detection for catalogued TRs	NanoMnT [83], NanoRepeat [84], NanoSTR [85], STRdust (in pathSTR) [86]	✓Better tolerance for extensive insertions/deletions (indels) within TRs ✓Improve resolution and accuracy of TR boundary identification	✓Computationally expensive and time-consuming for ultra-long TR regions ✓Only work with perfect repeat ✓Need known TR loci	✓Know TR loci ✓Sequencing data is available ✓TR repeat regions are not very long
Other read realignment strategy for catalogued TRs	ATaRVa [87], Repeat Detector (RD) [88], Tandem-genotypes [89], TRcaller [90], TREAT [91]	✓Higher computational efficiency ✓Work better with longer TR regions	✓Less robust for complex, highly degenerate, or high-error long-read data ✓Low accuracy for motif resolution ✓Need known TR loci	✓Know TR loci ✓Sequencing data is available ✓Rapid screening for TR variations ✓TR length estimation when extreme motif precision is not required
Signal-based detection for catalogued TRs	DeepRepeat [92], NanoSatellite [93], STRique [94], WarpSTR [95]	✓Work better with higher sequencing errors	✓Extremely high computational resource requirements ✓Susceptible to signal drift ✓Relies on ONT technology	✓Know TR loci ✓Sequencing data and ONT signal data are available ✓Sequencing data has higher sequencing errors
Read-based *de novo* TR detection	MotifScope [96], RepLong [97], Straglr [98], TRiCoLOR [99]	✓Capable of identifying novel, individual-specific, or uncatalogued TR loci in genomes	✓High time and space complexity ✓Higher false positive rate, requiring complex post-processing and validation	✓TR loci are unknown ✓Sequencing data is available ✓Discovery of novel TRs
TR visualization	ModDotPlot [100], pathSTR [86], StainedGlass [101]	✓Visualize TR structures ✓Investigate TR distribution.	✓Do not perform TR detection or genotyping ✓Serve only as an ancillary tool	✓Result validation after TR detection or genotyping ✓Population or individual specific TR investigation
Functional prediction of TRs	deepSTR [102], STRAS [103], RExPRT [104]	✓Assesses the pathogenicity potential and regulatory impact of TR variation ✓Filters for biologically relevant candidates	✓Model accuracy and interpretability depend on the quality of training data ✓Prediction results still require experimental validation	✓Prioritization and functional annotation of large volumes of genotyped TR variants ✓Assessing clinical relevance

Several TR detection tools have leveraged HMMs to perform accurate, flexible, and error-tolerant TR identification, including adVNTR [76], code-adVNTR [77], LongTR [78], RepeatHMM [80], PacmonsTR [79], TRGT [81] and TRGT-denovo [82]. These tools apply HMM on each extracted long read that is aligned to the TR locus input and then genotype TRs for each allele. The difference is how they train or use HMM for TR detection. RepeatHMM [80] uses HMM to infer TR boundaries and then calculates TR repeat counts in each long read, while PacmonsTR [79] applies global alignment to localize TR boundaries and then uses pair-HMM to estimate TR repeat counts in the regions defined by TR boundaries. In contrast, TRGT [81] and TRGT-denovo [82] apply HMM to interpret complex TR structures, while adVNTR [76] and code-adVNTR [77] train locus-specific HMMs for a known VNTR and align long reads with the trained HMM to infer TR repeat counts with high accuracy for targeted VNTR loci. Different from other tools which detect TR counts and then perform genotyping, LongTR [78] first identifies haplotype sequences to represent possible haplotype-sensitive TR structures, and then uses HMM to align reads to these candidate haplotypes to determine haplotype-specific TR repeat counts.

By incorporating HMMs, these tools can handle sequencing errors, model motif variability, and improve TR repeat count estimation. However, HMM models have several limitations. First, HMM might not capture long-range dependencies [106] and cannot scale well for

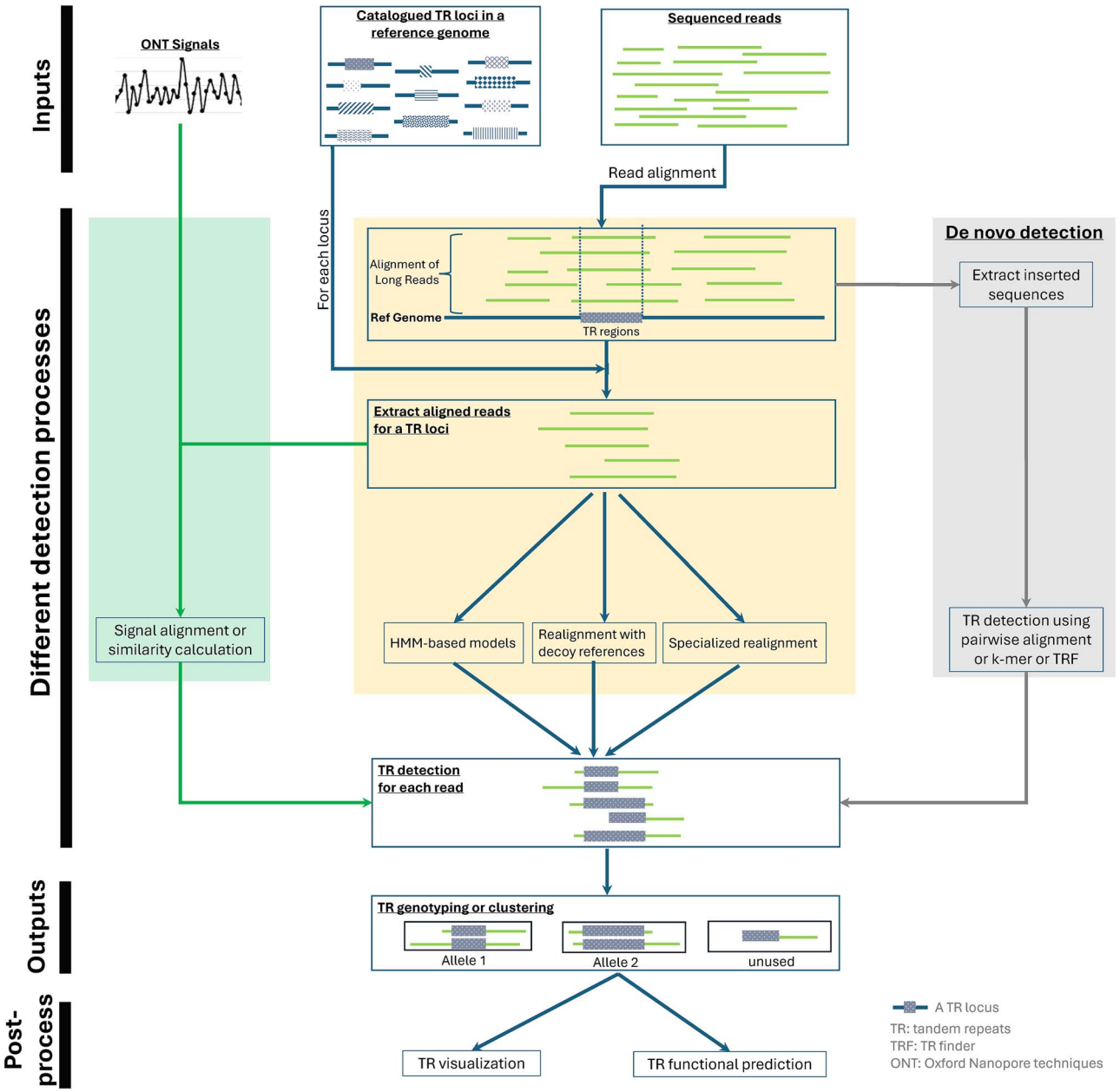

Figure 3 The detection framework of catalogue-based TR detection and de novo detection (in gray, right). Catalogued-based detection included read-based algorithms (in yellow, middle) and signal-based algorithms (in green, left). Left vertical bars suggest the four stages: inputs, detection process, outputs as well as post-processing of TRs. We use a TR locus for demonstration purposes and all TRs from catalogued loci will undergo similar detection processes. After detecting TRs, TR visualization tools could be used to visualize TR regions, while functional prediction tools enable further investigation of TRs in biological functions. TR: tandem repeats.

longer TRs or TR motifs. Second, HMM is sensitive to initialization and parameter tuning, leading to suboptimal detection performance for generalized detection of diverse TR motifs [107], unless specific HMMs are designed for each targeted TR loci [76, 77] or sequencing technique-specific HMMs are trained [76]. Third, standard HMM struggles to model complex TR regions [80], while heuristic strategies are designed [81].

Read-based decoy-sequence realignment strategies

TR regions in sequenced long reads exhibit two notable characteristics: greater variability in TR repeat counts and high-error rates compared to their flanking and confidently mapped genomic regions. Higher variation in TR repeat counts usually leads to large insertions or deletions during long-read alignment, while elevated sequencing errors can cause alignment issues to accurately map long reads across boundaries. Both compromise alignment quality and the estimation of TR repeat counts. To address these issues, several tools conduct realignment of extracted TR-candidate reads against custom reference sequences (decoy sequences).

Based on the observation of repeat count variability, tools like NanoRepeat [84] and NanoSTR [85] construct decoy reference sequences that contain the flanking regions as well as simulated TR segments with predefined repeat counts. Then, they realign extracted

reads of a TR locus to a library of decoy sequences and use the best-matching decoy to determine TR boundaries and repeat counts. These tools can tolerate indels and errors within TRs better than standard genomic alignment tools. But a major limitation of these tools is the computational cost to detect longer TR regions with thousands of TR repeat counts: a large number of decoy sequences need to be generated, and the realignment process is time consuming.

Alternatively, based on the observations that TR regions tend to have higher error rates than flanking regions especially on ONT data, tools such as STRdust in pathSTR [86] and NanoMnT [83] generate decoy reference sequences consisting of the left and right flank regions of a TR locus without TR regions. They realign extracted reads to the decoy sequences to determine TR boundaries in a long read and then TR repeat counts are calculated accordingly for genotyping. While these tools avoid generating extensive decoy libraries and are more computationally efficient, they do not account for the content of the TR region itself, and therefore cannot resolve complex TR motifs and the estimation of TR repeat counts may be imprecise especially for noisy long reads and interrupted TRs. These tools are also sensitive to the alignment quality of flanking sequences, while reads with longer TR regions often have short flanking sequences which lead to low-quality alignment.

Read-based other realignment strategies

Instead of generating decoy reference sequences with varying repeat lengths, several tools employ customized realignment strategies to more efficiently determine TR boundaries in long-read sequencing data. These tools endeavor to use specialized tools to realign extracted reads for a TR locus so that original alignment issues caused by TR expansion/contraction are reduced. For example, Repeat Detector (RD) [88] uses pfsearch to align long reads to predefined motif-based profiles. The profiles model TR structures and sequencing errors, motif variability, or partial motif units. Tandem-genotypes [89] uses last-split, a split-read aligner from the LAST aligner suite, to support split-read alignment. This method is particularly effective at detecting large insertions or deletions caused by TR expansions or contractions. By aligning reads in segments before and after TR regions, last-split can preserve high-confidence alignment across TR boundaries and estimate the size of intervening TR segments, even when the full repeat content cannot be directly aligned due to its length or complexity.

In contrast, tools like ATaRVa [87] and TREAT [91] conduct realignment against only flanking regions of a known TR locus to determine TR boundaries in long reads. They identify TR boundaries based on where long reads cease to align to upstream flank regions and begin to align to downstream flank regions, and consider intervening unaligned regions as TR regions. These tools are computationally efficient, but they share key limitations with the second type of decoy-based realignment strategy: they cannot robustly handle complex or highly error-prone TRs, nor precisely estimate TR repeat counts, especially for long or imperfect TRs. By comparison, TRcaller [90] takes a different strategy: it determines TR boundaries according to the alignment in a BAM file, extracts corresponding TR segments and compares extracted segments to repeat motif sequences for TR genotyping. TRcaller substantially reduces analysis time but relies on alignment quality in input BAM files.

Signal-based detection strategies

In addition to basecalled reads, ONT technique generates the changes of ionic current signals (referred to as ONT signals) which capture electrical disruption when DNA or RNA molecules pass through nanopores. ONT signals contain rich and fine-grained information that might be lost after basecalling, and basecalling accuracy tends to be lower in low-complexity regions, such as TR regions, limiting precise TR detection of read-based tools that rely solely on basecalled reads. To overcome this limitation, several specialized tools are designed to detect TRs via direct analysis of ONT signals. These tools usually first perform ONT basecalling to generate long reads, and then conduct anchoring of ONT signals to basecalled sequences so that each base (precisely, a 5-mer in the R9.4 chemistry) is associated with a corresponding segment of ONT signals. After that, they conduct signal-based comparisons for those reads aligned to a TR locus to determine TR regions and estimate TR repeat counts within individual reads.

These signal-based detection tools often differ in how they compare ONT signals for TR detection. STRique [94] employs a profile HMM to align real ONT signals with predefined expected signals of TR sequences, while NanoSatellite [93] and WarpSTR [95] use dynamic time warping (DTW) to compare ONT signals against reference signals. These three tools rely on precomputed reference signal templates rather than deriving TR signals from input files, which may introduce detection bias particularly when actual TR structures in input files deviate from reference signals. To address this issue, DeepRepeat [92] adopts a reference-signal-free approach by exploiting the self-similarity of ONT signals across adjacent repeat units within the same read. It applies a deep convolutional neural network model to learn and recognize TR signal patterns directly from the input data, potentially improving sensitivity and robustness of TR detection.

These signal-based tools enhance TR detection, especially in those regions where basecalling is unreliable. However, they come with three trade-offs. First, ONT signals are usually 10 times larger than basecalled read data, leading to significantly higher computational demands in both storage and data processing. Second, signal data is affected by signal drift, scaling differences and batch effects across sequencing runs and chemistries, requiring careful calibration to avoid the detection of non-TR patterns. Third, ONT basecalling accuracy continues to improve with advances in deep learning models and chemistry updates (e.g. R10.4.1), gradually reducing the need for ONT signal analysis for many, though not all, TR loci. In summary, signal-based TR detection tools offer unique advantages in quantifying TR regions particularly in regions where basecalling fails. However, due to their computational intensity and improving quality of basecalled reads, these tools are best used selectively, such as in the analysis of difficult-to-resolve loci.

Read-based detection of *de novo* tandem repeats

Most long-read TR detection methods described above require predefined TR loci as input. As a result, they are unable to detect novel, rare, individual-specific or population-specific TRs, especially in understudied species or non-model organisms where comprehensive TR catalogs are lacking. To address this limitation, several *de novo* TR detection tools have been developed, including RepLong [97], MotifScope [96], Straglr [98] and TRiCoLOR [99]. These tools investigate long reads without requiring known TR loci. Instead, they identify TRs on long reads or within unmapped or inserted regions of alignments in a reference-free or partially reference-guided strategy. After identifying TR regions and estimating read-level TR repeat counts, some tools proceed to genotyping using reference alignment, while others output per-read TR counts without further genotyping.

The major difference in these tools is TR detection methods. RepLong [97] performs pairwise alignments between long reads to identify overlapping regions enriched in repeated sequences, constructs a read-overlap network, and applies community detection algorithms to cluster reads sharing similar TR structures. This approach relies on the redundancy of TRs across reads, and thus is sensitive to high error rates in long reads. MotifScope [96] uses k-mer frequency analysis to detect local enrichments of repeated k-mers, which can indicate TR presence in a long read. However, its performance suffers from long-read noise. These two methods leverage overlapping TR content across long reads, and tend to have low boundary resolution, particularly in the presence of motif interruptions or non-uniform repeat patterns. Instead, Straglr [98] assumes that TR expansions often appear as insertions, and identifies large insertions from the alignment of long reads. It then applies TRF [70] to identify TR regions in the inserted sequences. However, Straglr is highly dependent on accurate insertion detection and on TRF's capability of tolerating long-read sequencing errors. Thus, Straglr is not robust across all repeat types or error profiles.

Different from the tools above, TRiCoLOR [99] takes a hybrid approach for *de novo* TR discovery. It first calculates Shannon entropy across reference genome regions to identify low-complexity, TR-rich segments. It then constructs consensus sequences from aligned reads over each detected TR-rich region, and applies an approximate string-matching approach to identify repeat patterns. While this method can detect novel TRs, it has several limitations. (i) It still relies on TR regions from reference genomes, reducing its utility in unassembled regions or nonreference alleles in diverse populations as well as in non-model organisms. (ii) Shannon entropy is a coarse measure of sequence complexity, leading to false positive and false negative TR detection from reference genomes. (iii) Pattern-matching approach cannot handle complex TR structures. (iv) It relies on accurate consensus sequence generation, which is often challenging in repetitive regions due to alignment ambiguity, high error rates, and heterogeneity of repeat length.

In summary, these *de novo* methods are useful to explore TR diversity in long-read datasets. But their performance varies depending on TR structures, sequencing error profiles, and the availability or quality of reference information. Each tool faces specific trade-offs in terms of detection accuracy, boundary resolution, and robustness to noisy or complex regions.

Tandem repeat visualization

TR regions can be visualized at the single-read, consensus sequence or reference sequence level to better understand their structure, periodicity, and variability. One effective method of visualization is through self-similarity plots (like StainedGlass [101] and ModDotPlot [100]) where each sequence is compared against itself to reveal internal repetitive patterns. In these plots, TR regions appear as off-diagonal parallel lines that shift consistently from the main diagonal. These visual signatures are useful for identifying TRs with consistent or imperfect motif structures, including compound, nested, or interrupted repeats. These visualization signatures are similar to what the assembly-based TR detection tools like Ribbit [68] and Dot2dot [69] do. However, these tools often operate on assembled or reference sequences or polished consensus sequences, since high error rates in raw long-read data obscure TR patterns in self-similarity plots.

Alternatively, pathSTR [86] is designed to operate at the individual genome level, with a focus on both TR visualization and comparative analysis. It generates multiple plot types to illustrate TR length distributions, individual genotypes, and deviations from population distribution. pathSTR [86] is designed to work with TR genotyping outputs from existing TR detection tools and provides insightful visualizations to assess TR variability and detect outliers.

Tandem repeat functional prediction

Given that TRs play crucial roles in various biological processes and are known to cause dozens of human disorders, several computational methods have been developed to predict the functional impact of TR variants. They take TRs and/or their expansion as input to machine learning algorithms or statistical approaches to predict how TRs are associated with biological functions or human diseases. For example, STRAS [103] and RExPRT [104] apply machine learning approaches to assess the pathogenic potential of TR expansions at specific loci. They trained classifiers to predict whether a given TR expansion is likely to be pathogenic or benign, after incorporating features such as genomic context, conservation, and known disease associations. deepSTR [102] introduces a deep learning model that specifically predicts TRs that are likely to influence transcriptional regulation, particularly those located near transcription start sites. It learns sequence patterns associated with TR-mediated transcription initiation in the Cap Analysis of Gene Expression (CAGE) data from the FANTOM5 consortium [108], providing insight into how certain TR variations may influence transcriptional regulation.

Instead of using machine learning algorithms, pathSTR [86] focuses on population-level variation and comparative analysis. It leverages a database of TR allele distributions derived from a large control cohort to compare an individual TR profile against population-level distributions. By identifying abnormal TR expansions and outlier TRs, pathSTR can flag TR candidates that may be biologically relevant or associated with disease phenotypes. Thus, pathSTR is powerful to screen pathogenic TR expansions and understand population-level TR variation.

In summary, these machine learning and population-informed tools reflect a growing recognition that TR variation is widespread and functionally important. They can significantly enhance TR interpretation in the context of health and disease studies.

Discussion and conclusion

In this review, we discussed 24 computational tools for TR detection using long reads. We systematically categorized these tools according to their input requirements (e.g. basecalled reads, ONT signals and assembled contigs as well as predefined TR loci) and core detection strategies (e.g. HMM, pattern recognition and realignment). For each group, we discussed the core algorithms, tool-specific variations, detection strengths and limitations. We found that most existing tools are designed for catalogue-based detection, relying on predefined TR loci and being incapable of discovering truly *de novo* TRs. They thus have limited applications in personalized genomics, rare variant detection, and the study of population-specific or species-specific repeats. HMM-based frameworks and realignment strategies have been the dominant approaches for read-based tools over the past decade, while string-based pattern recognition remains more common in assembly-based analyses.

Applications of different tools

These tools can be applied depending on the domain requirements. First, to detect novel TRs when allele-specific resolution is not needed and assembly contigs are available, assembly-based tools are appropriate. If assembly data are unavailable or allele-specific detection is required, read-based *de novo* TR tools can instead identify previously uncatalogued TRs with allele-level detail. Second, when ONT raw signal data are available and higher precision is desired, signal-based TR tools offer improved accuracy. Third, for simple and relatively short TR regions, decoy-sequence realignment tools can provide more accurate TR calls. When TR lengths or repeat structures are unknown, HMM-based TR tools serve as a more general and flexible solution. More details are discussed in Table 1 as well as in Supplementary Table S1.

Evaluation of tandem repeat detection

To assess the accuracy of TR detection, computational tools typically estimate repeat counts for a set of known TR loci and compare these estimates with benchmark values. Accuracy is usually quantified using the absolute difference between the estimated and benchmark repeat counts, and smaller differences suggest better performance. Benchmark repeat counts can be established using several strategies: (i) Simulated TR sequences with predefined repeat counts [76, 77, 80, 85, 88–90, 92, 94, 98, 99]; (ii) Experimental determination using Sanger sequencing or high-quality Illumina short reads or PacBio HIFI reads, which have low error rates [80, 82, 84, 87]; (iii) Repeat counts inferred from assembly contigs after error correction using both short- and long-read sequencing data [83, 84, 91, 92, 97]; (iv) Phenotype-linked validation, in which predicted repeat counts are evaluated based on known TR-expansion disease loci [76, 77, 80, 86, 88, 90–92, 96, 98, 103, 104, 109]. Each of these strategies provides only a partial assessment of TR detection tools. Simulated datasets (strategy 1) fail to fully reproduce the complex and platform-specific error profiles present in real long-read data. Experimental validation (strategy 2) is generally limited to shorter TR loci, whereas many long-read tools are designed to detect much larger repeat expansions. Assembly-based validation (strategy 3) typically yields overall TR lengths but may lack allele-specific resolution. Phenotype-based evaluation (strategy 4) is restricted to a small number of disease-associated loci and does not permit broad quantitative benchmarking. As a result, comprehensive and standardized evaluation of TR detection tools remain lacking, and current assessments capture only portions of their strengths and limitations.

Comparison of long-read sequencing platform

Two long-read sequencing platforms are both useful to detect TRs, but they differ markedly in read accuracy, error structure and read length. On one hand, existing studies found that PacBio HiFi reads generated more accurate TR genotypes for shorter TRs [91], while ONT outperforms PacBio HiFi to capture longer TRs [91, 110]. This is reasonable, because ONT could generate more longer reads than PacBio HiFi, while the base accuracy of PacBio HiFi reads is much better than ONT reads. On the other hand, ONT reads and PacBio HiFi reads have unique TR calls [78, 110] and their combination might be a better solution for a comprehensive TR calling.

Further improvements

Future TR detection tools could be significantly improved in the following four areas. First, enabling efficient and effective discovery of *de novo* TRs remains a central challenge, particularly for identifying individual- and population-specific TRs or somatic TR expansions in tumor samples. Although several TR tools [96–99] have been proposed to detect *de novo* TR loci, they are often computationally intensive or overly sensitive to high sequencing error rates. As a result, they cannot provide a comprehensive, genome-wide investigation of *de novo* TRs, especially when applied to large-scale or population-level long-read datasets. Second, enhancing detection accuracy requires addressing the intrinsic constraints in HMM-based and pattern-matching approaches which are two widely used approaches in TR detection. These existing methods struggle with noisy long-read data, interrupted or nested repeats, and large motif complexity of TRs, because they assume relatively clean periodicity and stable motif structures. Third, synergistic combinations of different approaches, such as integrating long- and short-read based detection tools as well as combining assembly-based and mapping-based strategies, may substantially improve robustness. But these hybrid strategies frequently produce inconsistent TR calls across data types or algorithms. Resolving these discrepancies still depends on heuristic or ad hoc filtering rules, which limit the reliability and scalability of genome-wide TR detection. Fourth, enhancing computational efficiency and scalability is becoming increasingly urgent given the rapid growth of long-read sequencing datasets. Modern computing infrastructures, such as GPU-accelerated architectures and parallelized frameworks, offer excellent opportunities for this enhancement. However, most existing tools, particularly those supporting *de novo* TR discovery, were not designed with these advanced infrastructures. This creates a critical bottleneck for analyzing TR variation in large cohorts or population-scale sequencing initiatives.

Across all these areas, advancements in artificial intelligence (AI), particularly deep learning and large AI models, could be transformative. On one hand, AI models, such as AlphaGenome [111], demonstrate the ability to learn sequence features directly from raw genomic sequences, enabling recognition of complex, degenerate, or noncanonical TR structures that rule-based or alignment-dependent methods fail to detect. Such models can also leverage long-range genomic context and generalize across species or sequencing platforms. On the other hand, AI methods are inherently optimized for GPU-based execution, offering substantial improvements in speed and scalability. However, applying AI models for TR detection remains challenging. Training high-performing AI systems requires large, accurately annotated TR datasets that are currently limited, and model interpretability remains a major concern: it is difficult to determine how predictions are made or whether the inferred TR structures reflect true biological variation. Despite these obstacles, AI-based approaches hold considerable promise. By integrating large AI models with existing algorithmic and computational techniques, future TR detection frameworks may achieve far greater robustness, scalability, and accuracy in genome-wide TR discovery.

Overall, TR detection remains a rapidly evolving field. As long-read sequencing technologies continue to advance, and more personalized genomic data become available, next-generation TR detection tools will play an increasingly critical role in investigating the impact of TRs on health and disease.

Key Points

- 24 long-read tools were proposed to detect tandem repeats on PacBio and ONT sequencing data.
- This survey categorizes 39 TR-associated tools into several groups and summarizes the core algorithms of each group.
- This work discusses the limitations of existing TR-associated tools.

Author contributions

Qian Liu (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing), Jincheng Li (Data curation, Formal analysis, Methodology, Validation, Writing—review & editing)

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Funding

This research was funded by the National Institute of General Medical Sciences grant number P20GM121325 and a UNLV startup grant. The publication fees for this article were partially supported by the UNLV University Libraries Open Article Fund. The funders had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Data availability

There is no relevant data or code.

References

1. Depienne C, Mandel J-L. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am J Hum Genet* 2021;**108**:764–85. https://doi.org/10.1016/j.ajhg.2021.03.011
2. Vergnaud G, Denoeud F. Minisatellites: mutability and genome architecture. *Genome Res* 2000;**10**:899–907.
3. Zhang S, Song Q, Zhang P. *et al.* Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: implications for biological regulation and human traits. *Cell Genomics* 2024;**4**:100699. https://doi.org/10.1016/j.xgen.2024.100699
4. Cui Y, Ye W, Li JS. *et al.* A genome-wide spectrum of tandem repeat expansions in 338,963 humans. *Cell* 2024;**187**:2336–41. https://doi.org/10.1016/j.cell.2024.03.004
5. 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
6. English AC, Dolzhenko E, Ziaei Jam H. *et al.* Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat Biotechnol* 2025;**43**:431–42. https://doi.org/10.1038/s41587-024-02225-z
7. 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
8. Vinces MD, Legendre M, Caldara M. *et al.* Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 2009;**324**:1213–6. https://doi.org/10.1126/science.1170097
9. Voynov V, Verstrepen KJ, Jansen A. *et al.* Genes with internal repeats require the THO complex for transcription. *Proc Natl Acad Sci U S A* 2006;**103**:14423–8. https://doi.org/10.1073/pnas.0606546103
10. Verstrepen KJ, Jansen A, Lewitter F. *et al.* Intragenic tandem repeats generate functional variability. *Nat Genet* 2005;**37**:986–90.
11. Fotsing SF, Margoliash J, Wang C. *et al.* The impact of short tandem repeat variation on gene expression. *Nat Genet* 2019;**51**:1652–9. https://doi.org/10.1038/s41588-019-0521-9
12. Jakubosky D, D'Antonio M, Bonder MJ. *et al.* Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat Commun* 2020;**11**:2927. https://doi.org/10.1038/s41467-020-16482-4
13. Sun JH, Zhou L, Emerson DJ. *et al.* Disease-associated short tandem repeats Co-localize with chromatin domain boundaries. *Cell* 2018;**175**:224–238.e15. https://doi.org/10.1016/j.cell.2018.08.005
14. Stam M, Belele C, Dorweiler JE. *et al.* Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. *Genes Dev* 2002;**16**:1906–18.
15. Ebersole T, Okamoto Y, Noskov VN. *et al.* Rapid generation of long synthetic tandem repeats and its application for analysis in human artificial chromosome formation. *Nucleic Acids Res* 2005;**33**:e130–0.
16. Liu Z-W, Liu J, Liu F. *et al.* Depositing centromere repeats induces heritable intragenic heterochromatin establishment and spreading in Arabidopsis. *Nucleic Acids Res* 2023;**51**:6039–54. https://doi.org/10.1093/nar/gkad306
17. Irony-Tur Sinai M, Salamon A, Stanleigh N. *et al.* AT-dinucleotide rich sequences drive fragile site formation. *Nucleic Acids Res* 2019;**47**:9685–95. https://doi.org/10.1093/nar/gkz689
18. Shastri N, Tsai Y-C, Hile S. *et al.* Genome-wide identification of structure-forming repeats as principal sites of fork collapse upon ATR inhibition. *Mol Cell* 2018;**72**:222–238.e11. https://doi.org/10.1016/j.molcel.2018.08.047
19. Brown RE, Freudenreich CH. Structure-forming repeats and their impact on genome stability. *Curr Opin Genet Dev* 2021;**67**:41–51. https://doi.org/10.1016/j.gde.2020.10.006
20. Shi Y, Niu Y, Zhang P. *et al.* Characterization of genome-wide STR variation in 6487 human genomes. *Nat Commun* 2023;**14**:2092. https://doi.org/10.1038/s41467-023-37690-8
21. Lamkin M, Gymrek M. The emerging role of tandem repeats in complex traits. *Nat Rev Genet* 2024;**25**:452–3. https://doi.org/10.1038/s41576-024-00736-8
22. 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
23. Orr HT, Chung MY, Banfi S. *et al.* Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat Genet* 1993;**4**:221–6.
24. Kremer EJ, Pritchard M, Lynch M. *et al.* Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence P(CCG)n. *Science* 1991;**252**:1711–4.
25. Cossée M, Schmitt M, Campuzano V. *et al.* Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations. *Proc Natl Acad Sci USA* 1997;**94**:7452–7. https://doi.org/10.1073/pnas.94.14.7452

26. Chen Z, Morris HR, Polke J. *et al.* Repeat expansion disorders. *Pract Neurol* 2025;**25**:204–16. https://doi.org/10.1136/pn-2023-003938
27. Zhao X-N, Usdin K. The repeat expansion diseases: the dark side of DNA repair. *DNA Repair (Amst)* 2015;**32**:96–105.
28. La Spada AR, Taylor JP. Repeat expansion disease: Progress and puzzles in disease pathogenesis. *Nat Rev Genet* 2010;**11**:247–58. https://doi.org/10.1038/nrg2748
29. Sone J, Mitsuhashi S, Fujita A. *et al.* Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet* 2019;**51**:1215–21. https://doi.org/10.1038/s41588-019-0459-y
30. Zeng S, Zhang MY, Wang XJ. *et al.* Long-read sequencing identified intronic repeat expansions in *SAMD12* from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy. *J Med Genet* 2019;**56**:265–70. https://doi.org/10.1136/jmedgenet-2018-105484
31. Tian Y, Wang JL, Huang W. *et al.* Expansion of human-specific GGC repeat in neuronal intranuclear inclusion disease-related disorders. *Am J Hum Genet* 2019;**105**:166–76. https://doi.org/10.1016/j.ajhg.2019.05.013
32. Ebbert MTW, Farrugia SL, Sens JP. *et al.* Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol Neurodegener* 2018;**13**:46. https://doi.org/10.1186/s13024-018-0274-4
33. Seixas AI, Loureiro JR, Costa C. *et al.* A pentanucleotide ATTTC repeat insertion in the non-coding region of DAB1, mapping to SCA37, causes spinocerebellar ataxia. *Am J Hum Genet* 2017;**101**:87–103. https://doi.org/10.1016/j.ajhg.2017.06.007
34. Song JHT, Lowe CB, Kingsley DM. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. *Am J Hum Genet* 2018;**103**:421–30. https://doi.org/10.1016/j.ajhg.2018.07.011
35. Ishiura H, Doi K, Mitsui J. *et al.* Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet* 2018;**50**:581–90. https://doi.org/10.1038/s41588-018-0067-2
36. Cen Z, Jiang Z, Chen Y. *et al.* Intronic pentanucleotide TTTCA repeat insertion in the SAMD12 gene causes familial cortical myoclonic tremor with epilepsy type 1. *Brain* 2018;**141**:2280–8. https://doi.org/10.1093/brain/awy160
37. Mizuguchi T, Toyota T, Adachi H. *et al.* Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. *J Hum Genet* 2019;**64**:191–7. https://doi.org/10.1038/s10038-018-0551-7
38. Cortese A, Simone R, Sullivan R. *et al.* Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet* 2019;**51**:649–58. https://doi.org/10.1038/s41588-019-0372-4
39. Rafehi H, Szmulewicz DJ, Bennett MF. *et al.* Bioinformatics-based identification of expanded repeats: a non-reference intronic pentamer expansion in RFC1 causes CANVAS. *Am J Hum Genet* 2019;**105**:151–65. https://doi.org/10.1016/j.ajhg.2019.05.016
40. 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
41. Halman A, Lonsdale A, Oshlack A. Analysis of tandem repeats in short-read sequencing data: from genotyping known pathogenic repeats to discovering novel expansions. *Curr Protoc* 2024;**4**:e70010.
42. Gijselinck I, Van Mossevelde S, van der Zee J. *et al.* The C9orf72 repeat size correlates with onset age of disease, DNA methylation and transcriptional downregulation of the promoter. *Mol Psychiatry* 2016;**21**:1112–24. https://doi.org/10.1038/mp.2015.159
43. Langbehn D, Brinkman RR, Falush D. *et al.* A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin Genet* 2004;**65**:267–77.
44. Rajan-Babu I-S, Dolzhenko E, Eberle MA. *et al.* Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. *Nat Rev Genet* 2024;**25**:476–99. https://doi.org/10.1038/s41576-024-00696-z
45. Budowle B, Sajantila A. Short tandem repeats — how microsatellites became the currency of forensic genetics. *Nat Rev Genet* 2024;**25**:450–1. https://doi.org/10.1038/s41576-024-00721-1
46. Southern EM. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 1975;**98**:503–17.
47. Rocca MS, Ferrarini M, Msaki A. *et al.* Comparison of NGS panel and sanger sequencing for genotyping CAG repeats in the AR gene. *Mol Genet Genomic Med* 2020;**8**:e1207. https://doi.org/10.1002/mgg3.1207
48. Daunay A, Duval A, Baudrin LG. *et al.* Low temperature isothermal amplification of microsatellites drastically reduces stutter artifact formation and improves microsatellite instability detection in cancer. *Nucleic Acids Res* 2019;**47**:e141–1.
49. van der Sanden B, Neveling K, Pang AWC. *et al.* Optical genome mapping for applications in repeat expansion disorders. *Curr Protoc* 2024;**4**:e1094. https://doi.org/10.1002/cpz1.1094
50. Dashnow H, Lek M, Phipson B. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol* 2018;**19**:121. https://doi.org/10.1186/s13059-018-1505-2
51. Dolzhenko E, Bennett MF, Richmond PA. *et al.* ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol* 2020;**21**:102. https://doi.org/10.1186/s13059-020-02017-z
52. Dashnow H, Pedersen BS, Hiatt L. *et al.* STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. *Genome Biol* 2022;**23**:257. https://doi.org/10.1186/s13059-022-02826-4
53. Fearnley LG, Bennett MF, Bahlo M. Detection of repeat expansions in large next generation DNA and RNA sequencing data without alignment. *Sci Rep* 2022;**12**:13124. https://doi.org/10.1038/s41598-022-17267-z
54. Ziaei Jam H, Li Y, DeVito R. *et al.* A deep population reference panel of tandem repeat variation. *Nat Commun* 2023;**14**:6711. https://doi.org/10.1038/s41467-023-42278-3
55. Dolzhenko E, Deshpande V, Schlesinger F. *et al.* ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* 2019;**35**:4754–6. https://doi.org/10.1093/bioinformatics/btz431
56. Willems T, Zielinski D, Yuan J. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat Methods* 2017;**14**:590–2. https://doi.org/10.1038/nmeth.4267
57. Mousavi N, Shleizer-Burko S, Yanicky R. *et al.* Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res* 2019;**47**:e90–0.
58. Ardui S, Ameur A, Vermeesch JR. *et al.* Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* 2018;**46**:2159–68. https://doi.org/10.1093/nar/gky066
59. Eid J, Fehr A, Gray J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**23**:133–8. https://doi.org/10.1063/pt.5.023562

60. Sereika M, Kirkegaard RH, Karst SM. *et al.* Oxford nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods* 2022;**19**:823–6. https://doi.org/10.1038/s41592-022-01539-7
61. Lin B, Hui J, Mao H. Nanopore technology and its applications in gene sequencing. *Biosensors* 2021;**11**:214. https://doi.org/10.3390/bios11070214
62. Gustafson JA, Gibson SB, Damaraju N. *et al.* High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res.* 2024;**34**:2061–2073. https://doi.org/10.1101/gr.279273.124
63. Danzi MC, Xu IRL, Fazal S. *et al.* Detailed tandem repeat allele profiling in 1,027 long-read genomes reveals genome-wide patterns of pathogenicity. *bioRxiv* 2025. https://doi.org/10.1101/2025.01.06.631535
64. Tanudisastro HA, Deveson IW, Dashnow H. *et al.* Sequencing and characterizing short tandem repeats in the human genome. *Nat Rev Genet* 2024;**25**:460–75. https://doi.org/10.1038/s41576-024-00692-3
65. Chaushevska M, Alapont-Celaya K, Schack AK. *et al.* Get ready for short tandem repeats analysis using long reads-the challenges and the state of the art. *Front Genet* 2025;**16**. https://doi.org/10.3389/fgene.2025.1610026
66. Kolpakov R, Bana G, Kucherov G. Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* 2003;**31**:3672–8.
67. Mayer, C. Phobos 3.3.12. 2006.
68. Avvaru AK, Sharma A, Sowpati DT. Ribbit: accurate identification and annotation of complex tandem repeat sequences in genomes. bioRxiv. 2025.
69. Genovese LM, Mosca MM, Pellegrini M. *et al.* Dot2dot: accurate whole-genome tandem repeats discovery. *Bioinformatics* 2019;**35**:914–22. https://doi.org/10.1093/bioinformatics/bty747
70. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**:573–80.
71. Boeva V, Regnier M, Papatsenko D. *et al.* Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics* 2006;**22**:676–84.
72. Pellegrini M, Renda ME, Vecchio A. TRStalker: An efficient heuristic for finding fuzzy tandem repeats. *Bioinformatics* 2010;**26**:i358–66. https://doi.org/10.1093/bioinformatics/btq209
73. Ren J, Gu B, Chaisson MJP. Vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol* 2023;**24**:175. https://doi.org/10.1186/s13059-023-03010-y
74. Touati R, Tajouri A, Mesaoudi I. *et al.* New methodology for repetitive sequences identification in human X and Y chromosomes. *Biomed Signal Process Control* 2021;**64**:102207. https://doi.org/10.1016/j.bspc.2020.102207
75. Newman AM, Cooper JB. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* 2007;**8**:382. https://doi.org/10.1186/1471-2105-8-382
76. Bakhtiari M, Shleizer-Burko S, Gymrek M. *et al.* Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res* 2018;**28**:1709–19. https://doi.org/10.1101/gr.235119.118
77. Park J, Bakhtiari M, Popp B. *et al.* Detecting tandem repeat variants in coding regions using code-adVNTR. *iScience* 2022;**25**:104785. https://doi.org/10.1016/j.isci.2022.104785
78. Ziaei Jam H, Zook JM, Javadzadeh S. *et al.* LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. *Genome Biol* 2024;**25**:176. https://doi.org/10.1186/s13059-024-03319-2
79. Ummat A, Bashir A. Resolving complex tandem repeats with long reads. *Bioinformatics* 2014;**30**:3491–8. https://doi.org/10.1093/bioinformatics/btu437
80. Liu Q, Zhang P, Wang D. *et al.* Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med* 2017;**9**:65. https://doi.org/10.1186/s13073-017-0456-7
81. TRGT-ing the dark genome to accurately characterize tandem repeats at scale. *Nat Biotechnol* 2024;**42**:1504–5. https://doi.org/10.1038/s41587-023-02073-3
82. Mokveld T, Dolzhenko E, Dashnow H. *et al.* TRGT-denovo: accurate detection of *de novo* tandem repeat mutations. *bioRxiv* 2024. https://doi.org/10.1101/2024.07.16.600745
83. Park G, An H, Luo H. *et al.* NanoMnT: An STR analysis tool for Oxford nanopore sequencing data driven by a comprehensive analysis of error profile in STR regions. *GigaScience* 2025;**14**:giaf013. https://doi.org/10.1093/gigascience/giaf013
84. Fang L, Monteys AM, Dürr A. *et al.* Haplotyping SNPs for allele-specific gene editing of the expanded huntingtin allele using long-read sequencing. *Hum Genet Genomics Adv* 2023;**4**:100146. https://doi.org/10.1016/j.xhgg.2022.100146
85. Lang J, Xu Z, Wang Y. *et al.* NanoSTR: a method for detection of target short tandem repeats based on nanopore sequencing data. *Front Mol Biosci* 2023;**10**:1093519. https://doi.org/10.3389/fmolb.2023.1093519
86. De Coster W, Höijer I, Bruggeman I. *et al.* Visualization and analysis of medically relevant tandem repeats in nanopore sequencing of control cohorts with pathSTR. *Genome Res.* 2024;**34**:2074–2080. https://doi.org/10.1101/gr.279265.124
87. Sivakumar AK, Sudarsanam S, Sharma A. *et al.* ATaRVa: analysis of tandem repeat variation from long read sequencing data. bioRxiv. 2025.
88. Taylor AS, Barros D, Gobet N. *et al.* Repeat detector: versatile sizing of expanded tandem repeats and identification of interrupted alleles from targeted DNA sequencing. *NAR Genomics Bioinformatics* 2022;**4**:lqac089. https://doi.org/10.1093/nargab/lqac089
89. 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. https://doi.org/10.1186/s13059-019-1667-6
90. Wang X, Huang M, Budowle B. *et al.* TRcaller: a novel tool for precise and ultrafast tandem repeat variant genotyping in massively parallel sequencing reads. *Front Genet* 2023;**14**. https://doi.org/10.3389/fgene.2023.1227176
91. Tesi N, Salazar A, Zhang Y. *et al.* Characterizing tandem repeat complexities across long-read sequencing platforms with TREAT and otter. *Genome Res* 2024;**34**:1942–53. https://doi.org/10.1101/gr.279351.124
92. Fang L, Liu Q, Monteys AM. *et al.* DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biol* 2022;**23**:108. https://doi.org/10.1186/s13059-022-02670-6
93. De Roeck A, De Coster W, Bossaerts L. *et al.* NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biol* 2019;**20**:239. https://doi.org/10.1186/s13059-019-1856-3

94. Giesselmann P, Brändl B, Raimondeau E. *et al.* Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol* 2019;**37**:1478–81. https://doi.org/10.1038/s41587-019-0293-x
95. Sitarčík J, Vinař T, Brejová B. *et al.* WarpSTR: determining tandem repeat lengths using raw nanopore signals. *Bioinformatics* 2023;**39**:btad388. https://doi.org/10.1093/bioinformatics/btad388
96. Zhang Y, Hulsman M, Salazar A. *et al.* Multisample motif discovery and visualization for tandem repeats. *Genome Res* 2025;**35**:850–62. https://doi.org/10.1101/gr.279278.124
97. Guo R, Li YR, He S. *et al.* RepLong: de novo repeat identification using long read sequencing data. *Bioinformatics* 2018;**34**:1099–107. https://doi.org/10.1093/bioinformatics/btx717
98. Chiu R, Rajan-Babu I-S, Friedman JM. *et al.* Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol* 2021;**22**:224. https://doi.org/10.1186/s13059-021-02447-3
99. Bolognini D, Magi A, Benes V. *et al.* TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. *Gigascience* 2020;**9**:giaa101. https://doi.org/10.1093/gigascience/giaa101
100. Sweeten AP, Schatz MC, Phillippy AM. ModDotPlot—rapid and interactive visualization of tandem repeats. *Bioinformatics* 2024;**40**:btae493. https://doi.org/10.1093/bioinformatics/btae493
101. Vollger MR, Kerpedjiev P, Phillippy AM. *et al.* StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* 2022;**38**:2049–51. https://doi.org/10.1093/bioinformatics/btac018
102. Grapotte M, Saraswat M, Bessière C. *et al.* Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network. *Nat Commun* 2021;**12**:3297.https://doi.org/10.1038/s41467-021-23143-7
103. Zhang M. STRAS: a snakemake pipeline for genome-wide short tandem repeats annotation and score. *Hum Genet* 2024;**143**:735–8. https://doi.org/10.1007/s00439-024-02662-5
104. Fazal S, Danzi MC, Xu I. *et al.* RExPRT: a machine learning tool to predict pathogenicity of tandem repeat loci. *Genome Biol* 2024;**25**:39. https://doi.org/10.1186/s13059-024-03171-4
105. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191
106. Ma J, Wang S, Wang Z. *et al.* MRFalign: protein homology detection through alignment of Markov random fields. *PLoS Comput Biol* 2014;**10**:e1003500. https://doi.org/10.1371/journal.pcbi.1003500
107. Olson D, Wheeler T. ULTRA: a model based tool to detect tandem repeats. *ACM BCB* 2018;37–46.
108. Nobusada T, Yip CW, Agrawal S. *et al.* Update of the FANTOM web resource: enhancement for studying noncoding genomes. *Nucleic Acids Res* 2025;**53**:D419–24. https://doi.org/10.1093/nar/gkae1047
109. Liu Q, Tong Y, Wang K. Genome-wide detection of short tandem repeat expansions by long-read sequencing. *BMC Bioinformatics* 2020;**21**:542. https://doi.org/10.1186/s12859-020-03876-w
110. Chiu R, Rajan-Babu I-S, Friedman JM. *et al.* A comprehensive tandem repeat catalog of the human genome. *medRxiv* 2024. https://doi.org/10.1101/2024.06.19.24309173
111. Avsec Ž, Latysheva N, Cheng PJ. *et al.* AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. *bioRxiv* 2025. https://doi.org/10.1101/2025.06.25.661532

Briefings in Bioinformatics, 2026, **27**, bbag031
https://doi.org/10.1093/bib/bbag031
Review