



TECHNICAL NOTE

TRiCoLoR: tandem repeat profiling using whole-genome long-read sequencing data

Davide Bolognini ^{1,2,*}, Alberto Magi³, Vladimir Benes², Jan O. Korbel⁴ and Tobias Rausch ^{2,4}

¹Department of Experimental and Clinical Medicine, University of Florence, Viale Pieraccini 6, Florence 50134, Italy; ²European Molecular Biology Laboratory (EMBL), GeneCore, Meyerhofstraße 1, Heidelberg 69117, Germany; ³Department of Information Engineering, University of Florence, Via di S. Marta 3, Florence 50134, Italy and ⁴European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstraße 1, Heidelberg 69117, Germany

*Correspondence address. Davide Bolognini, Via della Pineta 30, Quarrata 51039, Italy. E-mail: davidebolognini7@gmail.com  <http://orcid.org/0000-0002-8735-8093>

Abstract

Background: Tandem repeat sequences are widespread in the human genome, and their expansions cause multiple repeat-mediated disorders. Genome-wide discovery approaches are needed to fully elucidate their roles in health and disease, but resolving tandem repeat variation accurately remains a challenging task. While traditional mapping-based approaches using short-read data have severe limitations in the size and type of tandem repeats they can resolve, recent third-generation sequencing technologies exhibit substantially higher sequencing error rates, which complicates repeat resolution. **Results:** We developed TRiCoLoR, a freely available tool for tandem repeat profiling using error-prone long reads from third-generation sequencing technologies. The method can identify repetitive regions in sequencing data without a prior knowledge of their motifs or locations and resolve repeat multiplicity and period size in a haplotype-specific manner. The tool includes methods to interactively visualize the identified repeats and to trace their Mendelian consistency in pedigrees. **Conclusions:** TRiCoLoR demonstrates excellent performance and improved sensitivity and specificity compared with alternative tools on synthetic data. For real human whole-genome sequencing data, TRiCoLoR achieves high validation rates, suggesting its suitability to identify tandem repeat variation in personal genomes.

Keywords: long-read sequencing; tandem repeat variation; bioinformatics software

Background

Almost half of the human genome is estimated to be covered by repetitive sequences [1]. Among these, tandem repeats (TRs) have been found to be involved in a range of functions such as DNA repair, chromatin organization, telomere maintenance, and regulation of gene expression [2]. Most importantly, >40 diseases, primarily neurological, are known to be related to TR expansions [3]. Despite their clinical importance, accurately resolving TRs remains challenging in sequencing data sets mainly

because of insufficient read lengths failing to encompass entire expanded repeats or technological limitations, such as high sequencing error rates.

Prior methods for TR profiling in short-read sequencing data sets can be broadly classified as reference-based [4,5] or *de novo* [6,7] approaches. While the former investigates only reads spanning known TRs, the latter can identify TRs regardless of whether or not their repeat motif is annotated in the reference. Short-read methods are often inadequate to accurately resolve

Received: 4 June 2020; Revised: 7 August 2020; Accepted: 7 September 2020

© The Author(s) 2020. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

expanded TRs if the total repeat length is greater than the read length.

Long reads from third-generation sequencing technologies, namely, Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PB), have already proved invaluable for the discovery of large structural variants [8] and are obvious candidates for broadening the scope of detectable TRs. However, long reads exhibit high sequencing error rates that make it difficult to accurately decipher TRs, especially in low-complexity regions.

Few TR detection methods for long-read sequencing data have been developed so far. Examples include PacmonSTR [9], NCRF [10], TideHunter [11], NanoSatellite [12], and Tandem-genotypes [13]. However, these tools have some limitations, either because they are technology-specific (PacmonSTR and NanoSatellite), because they are not intended to be used genome-wide (NCRF, TideHunter, and NanoSatellite), or because they require substantial preprocessing steps preventing their large-scale use (Tandem-genotypes). Some tools also lack genotyping capabilities (NCRF and TideHunter), and none of the aforementioned methods is capable of profiling TRs *de novo* in regions that have previously not been annotated as harboring a TR.

TRiCoLoR addresses these shortcomings of existing, alignment-based tools by allowing users to rapidly identify and genotype TRs from haplotype-resolved long-read alignments. Once low-entropy repetitive regions have been identified in sequenced long reads, TRiCoLoR exploits partial order alignment (POA) [14] to compute haplotype-specific low-error consensus sequences [15] that are further processed by means of a fast regular expression (RegEx)-based approximate string-matching algorithm to resolve repeat motif and multiplicity of the discovered TRs. Detected TRs can be interactively visualized within their haplotype-specific sequence context for manual exploration of expanded or contracted repeats. For trio sequencing studies, TRiCoLoR additionally allows Mendelian inheritance patterns to be traced across TR genotypes.

Methods

TRiCoLoR (Tandem Repeats Caller for Long Reads) requires haplotype-resolved long-read alignments as input (Supplementary Note S1). It then runs a series of modules to identify and genotype TRs as described in detail below. A manual containing an in-depth explanation of how to install TRiCoLoR and run its various modules is available at Github [16], including use case examples.

Identifying repetitive regions *de novo*

TRiCoLoR can identify repetitive regions in haplotype-resolved BAM files *de novo*. This is achieved using the SENSOR module, which uses the Shannon entropy of DNA sequences to identify candidate repetitive segments in genomic sequences [17]. TRiCoLoR SENSOR scans in parallel the haplotype-specific BAM files and computes, for each sequencing read, its Shannon entropy content in non-overlapping, sliding windows of a pre-trained size (20 bp, by default). Genomic coordinates of windows in which multiple reads (≥ 5 , by default) support an entropy drop (≤ 1.23 , by default) are stored and those nearby are merged (those falling within 100 bp intervals, by default). The default entropy threshold of 1.23 efficiently discriminates between repetitive and non-repetitive DNA sequences using synthetic ONT and PB reads as shown in Supplementary Fig. S1 (see Supplementary Note S2). All candidate repetitive regions identi-

fied with this approach are eventually outputted in BED format. This pre-filtering of repetitive regions is fairly fast even in deep-coverage whole-genome data (see also Findings) and drastically reduces the computational time required for the subsequent TR profiling.

Profiling repetitive regions

TRiCoLoR can profile TRs in haplotype-resolved BAM files through the REFER (REpeats FindER) module. The input of REFER is a BED file generated by TRiCoLoR SENSOR. Alternatively, the BED file can be provided by the user on the basis of prior knowledge of clinically relevant TRs, for instance.

For each region in the BED file, REFER first fetches from the haplotype-specific BAM files the sequencing reads spanning the selected region and trims them, so that the length of each read is approximately the size of the region. Let $R = [S, E]$ be a region from the BED file, ranging from a start coordinate S to an end coordinate E for a given chromosome. Each sequencing read entirely spanning R is fetched and trimmed so that the actual sequence REFER stores is that included between S and E , which significantly improves the runtime of the subsequent POA algorithm to generate a consensus sequence.

Once the sequencing reads of interest have been collected and trimmed, TRiCoLoR uses SPOA [18], a single-instruction multiple-data accelerated version of the robust POA framework, to compute highly accurate consensus sequences with an approximate error reduction of $\sim 77\%$ and $\sim 88\%$ for ONT and PB, respectively (Supplementary Note S3 and Supplementary Fig. S2).

With the haplotype-specific consensus sequences at hand, REFER aligns these to the reference genome using minimap2 [19], which compared favorably to alternative aligners on synthetic data, in terms of both speed and mapping accuracy (Supplementary Note S4 and Supplementary Fig. S3). The reference-aligned low-error consensus sequences are then screened by a RegEx-based approximate string-matching algorithm, which has 3 processing steps: (i) identifying motifs (motifs of length ≤ 6 , by default) that are perfectly repeated a minimum number of times (5, by default); (ii) looking for approximate repetitions of the identified motifs to account for remaining consensus errors, i.e., imperfect repeated motifs up to a user-defined edit distance (≤ 1 , by default); and (iii) in case of multiple overlapping approximate repetitions, resolving these competing TR predictions using an N-gram model that favors the most frequently occurring perfect repeat motif.

Together with the haplotype-specific consensus sequences, the corresponding reference is screened in a similar manner, with few differences being noteworthy: (i) the algorithm assumes the reference does not contain errors and does not look for approximate repetitions of the motifs identified; and (ii) among overlapping repetitions, the longest repeat is taken.

TRs (those ≥ 50 bp, by default) varying between the haplotypes or the reference are eventually stored in BCF-compliant format. TRiCoLoR REFER also stores in the output folder several BED files describing the TRs identified (for both the reference and each haplotype) and haplotype-specific BAM files containing the aligned consensus sequences.

Visualizing identified repeats

The TRs profiled using TRiCoLoR REFER can be interactively visualized through the ApP (Alignment Plotter) module. This module takes as inputs the BED and the BAM files generated by TRi-

CoLoR REFER, together with an additional BED file describing 1 or more regions to plot.

TRiCoLoR ApP produces a static HTML file illustrating the alignment between the reference and the individual's haplotypes at single-base resolution, highlighting the TRs detected (Supplementary Note S5 and Supplementary Fig. S4A–C).

Tracing Mendelian inheritance patterns of identified repeats

In pedigree studies, assigned genotypes can be either Mendelian consistent or inconsistent. TRiCoLoR enables genotype consistency checks for TRs identified in the index child when haplotype-resolved long-read alignments for both parents are available. This is achieved through the SAGE (Sample GENotyper) module with special emphasis on the common situation that parents have been sequenced at low depth.

Using the same aforementioned TRiCoLoR REFER approach, SAGE computes haplotype-specific consensus alignments for each child TR in each parent. Next the module checks whether the parental TRs are more similar (i.e., have a lower edit distance) to the reference or to the TR identified in the child and assigns them the most likely genotype. Knowing the genotype of both parents, the module eventually flags each TR as Mendelian consistent or inconsistent with the “–mendel” parameter enabled. The output of TRiCoLoR SAGE is a multi-sample BCF file that contains the genotypes for the index child and both parents.

Findings

We benchmarked TRiCoLoR using both synthetic data generated with VISOR [20] and real, publicly available data from the Human Genome Structural Variation Consortium (HGSVC) [8].

Benchmarking TRiCoLoR on synthetic data

We used the TR simulator VISOR to generate synthetic ONT and PB alignments containing TR contractions and expansions. First, we simulated haplotype-resolved ONT and PB BAM files (the average length of simulated reads was set to 8,000 bp on the basis of statistics derived from recent ONT sequencing runs [21]; the substitution:insertion:deletion ratio was set to ~45:25:30 for the synthetic ONT reads and to ~15:50:35 for the synthetic PB reads, in accordance with findings described in Supplementary Note S3) exhibiting variable error rates (accuracy of reads ~0.85, ~0.90, and ~0.95) and depth of coverage (haplotype-specific depth of coverage 5–10 \times and 10–20 \times), with each BAM file harboring a heterozygous contraction or expansion of a known, randomly chosen, TR. At this stage, we simulated small TR contractions/expansions (contractions/expansions of 7 motifs on average) to evaluate the capability of our method to spot even minor changes in the TR multiplicity of the 2 haplotypes. For each group, we simulated 200 haplotype-resolved BAM files. Then, we evaluated the performance of TRiCoLoR in terms of precision (P), recall (R), and F1 score (F1) (Supplementary Note S6). In particular, P, R, and F1 values were calculated allowing no discrepancies, 1 discrepancy, or 2 discrepancies between the number of repeated motifs in the ground truth and the number of repeated motifs predicted by TRiCoLoR. Fig. 1 shows these findings for synthetic TR contractions (panel A) and expansions (panel B). TRiCoLoR demonstrated high P and R in all the simulated groups: our method always achieved an F1 close to 1 when allowing a single-motif discrepancy between simulated and predicted TRs and hit $P \sim 1$ and $R \sim 1$ when allowing

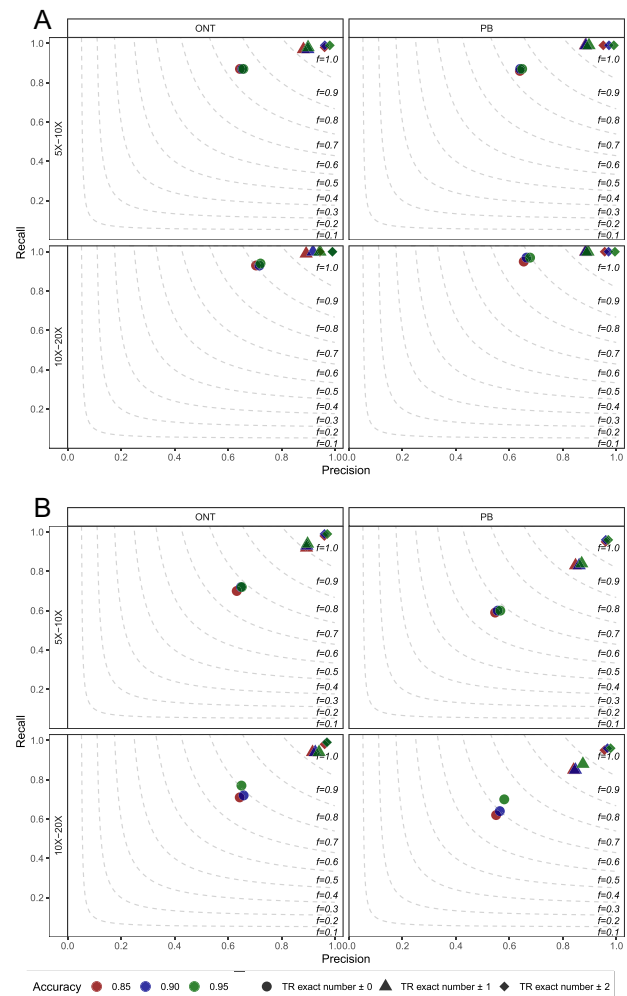


Figure 1: TRiCoLoR's P (x-axis), R (y-axis), and F1 (dashed lines) on synthetic TR contractions (A) and expansions (B). ONT and PB reads exhibit variable error rates (accuracy ~0.85, red; accuracy ~0.90, blue; accuracy ~0.95, green) and were simulated using variable haplotype-specific depth of coverage. P, R, and F1 were calculated allowing no motif discrepancies (circles), 1 motif discrepancy (triangles), or 2 motif discrepancies (rhombi) between TRiCoLoR's predictions and the number of repeated motifs in the ground truth.

up to 2 motif discrepancies. For both contractions and expansions the F1 depends on the coverage and input read accuracy as expected. In all the simulated TR contractions and expansions, TRiCoLoR was also able to properly identify the correct repeated motif, a few times shifted (e.g., a repeated TG instead of a repeated GT). Supplementary Fig. S5 illustrates these findings for the same simulated groups of Fig. 1, averaged over the different accuracy levels. Furthermore, as a proof of concept, we compared TRiCoLoR to a TR caller for long reads recently published, namely, NCRF. Using the same approach described above, we simulated 100 ONT and 100 PB BAM files (accuracy of reads ~0.90, depth of coverage for each haplotype 5–10 \times), each harboring a small TR contraction/expansion, and we ran both TRiCoLoR and NCRF on these data. Because NCRF cannot deal with BAM input, we slightly modified TRiCoLoR to store in FASTA format the sequences used for the consensus computation step, which could be processed through NCRF (Supplementary Note S7). Fig. 2 shows the correlation results between the number of repeated motifs in the ground truth and the number of re-

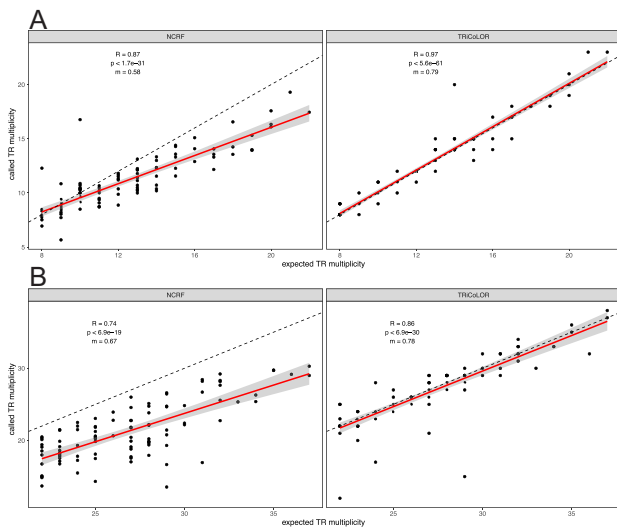


Figure 2: Correlation results between the number of repeated motifs in the ground truth (x-axis) and the number of repeated motifs predicted by TRiCoLoR and NCRF (y-axis) for synthetic TR contractions (A) and expansions (B). Each dot represents the synthetic contraction/expansion of a single TR. R is the Pearson correlation coefficient, p is the P-value of the linear regression analysis, m is the slope of the regression line, and the dashed line is the bisector of the first quadrant angle that marks the perfect correspondence between expected and predicted number of TRs.

peated motifs predicted by TRiCoLoR and NCRF for the simulated TR contractions (panel A) and expansions (panel B). For both TR contractions and expansions, TRiCoLoR got excellent R scores ($R = 0.97$ for contractions and $R = 0.86$ for expansions), outperforming NCRF ($R = 0.87$ for contractions and $R = 0.74$ for expansions). We next evaluated exceptionally long TR expansions because these have been implicated in several neurological disorders. For instance, the common fragile-X syndrome is related to a CGG-repeat usually consisting of ≤ 55 repeated motifs that expands to ≥ 200 repeated motifs. Following the simulation schema described above, we generated 100 ONT and 100 PB synthetic BAM files harboring TRs expanded by 200 motifs and we ran both TRiCoLoR and NCRF on these data. Supplementary Fig. S6 shows the correlation results between the number of repeated motifs in the ground truth and the number of repeated motifs predicted by TRiCoLoR and NCRF for the simulated long TR expansions. As above, TRiCoLoR achieved the best R score ($R = 0.73$), outperforming NCRF ($R = 0.53$).

Benchmarking TRiCoLoR on real data

We applied TRiCoLoR to call TRs *de novo* on publicly available ONT and PB human whole-genome sequencing data from the HGSVC project. In particular, we used the ONT sequencing data for HG00514 (Han Chinese), HG00733 (Puerto Rican), and NA19240 (Yoruban Nigerian) and the PB sequencing data for HG00731 (Puerto Rican, father), HG00732 (Puerto Rican, mother), and HG00733 (son).

We aligned the ONT FASTQ files to the human GRCh38 reference genome using minimap2, and we merged the chromosome-specific PB alignments using samtools [22]. We then split the ONT and PB alignments by haplotype with Alfred [23] using phased single-nucleotide variants from the HGSVC project. We calculated the coverage of the initial and the haplotype-resolved BAM files using mosdepth [24]. For all the

ONT samples, we identified an initial $\sim 20\times$ coverage (HG00733 $\sim 21\times$, HG00514 $\sim 23\times$, and NA19240 $\sim 24\times$), slightly reduced after splitting by haplotype owing to some unassigned reads (HG00733 $\sim 8\times$, HG00514 $\sim 9\times$, and NA19240 $\sim 10\times$ for each haplotype). For the PB samples, we identified a $\sim 42\times$ coverage for HG00733 and $\sim 21\times$ coverage for HG00731 and HG00732, reduced after splitting the data by haplotype (HG00733 $\sim 14\times$, HG00731 and HG00732 $\sim 8\times$ for each haplotype).

We then ran TRiCoLoR SENSOR using the default parameter settings on the HG00733 (ONT and PB), HG00514, and NA19240 individuals. Using an Ubuntu 16.04.6 LTS desktop with Intel® Xeon® processors X5460 (clock rate 2.93 GHz), the module took ~ 4 hours to scan the ONT samples and ~ 8 hours to scan the PB sample, which reflects the higher coverage available for PB. For the HG00733, HG00514, and NA19240 ONT individuals the module identified $\sim 160,000$, $\sim 190,000$, and $\sim 260,000$ low-entropy regions (mean length of the regions ~ 900 bp), which were reduced to $\sim 70,000$, $\sim 100,000$, and $\sim 160,000$, respectively, after filtering for regions with mean coverage > 8 . For the HG00733 PB individual the module identified $\sim 380,000$ low-entropy regions (mean length of the regions ~ 850 bp), which were reduced to $\sim 150,000$ after filtering for regions with mean coverage $> 10\times$. For HG00733, $\sim 97\%$ of the low-entropy regions originally identified in the ONT individual overlapped those in the PB one; owing to the different coverage distributions, this percentage was reduced to $\sim 31\%$ after filtering.

We ran TRiCoLoR REFER on the samples processed by TRiCoLoR SENSOR using the default parameter settings. With 7 processors on our Ubuntu desktop, the module took ~ 10 – 12 hours to profile TRs on the ONT individuals and ~ 14 hours to profile TRs on the PB individual.

We calculated the number of TRs properly called by TRiCoLoR using an alignment-free validation approach. Current benchmarks for TR calling in human genomes are mainly based on short-read sequencing and are biased towards regions of the genome that are easy to call with such a technology [25]. It has been shown that it is often impossible to accurately map or even assemble short reads originating from repetitive regions [26], and as a consequence, some TRs are missing from the available TR callsets. Following the idea from Dolle et al. [27] we first built full-text searchable FM indexes [28] for both the GRCh38 human reference FASTA and the high-quality Illumina FASTQ files of the HG00733, HG00514, and NA19240 individuals. Then, for each individual's variant identified by TRiCoLoR REFER, the validation algorithm (i) checks whether the variant sequence appears ≥ 1 time in the reference FM index: if so, using the consensus BAM files stored by TRiCoLoR REFER, the variant sequence is extended by 1 bp to the left and 1 bp to the right and step 1 is repeated; if not, the algorithm proceeds to the next step; and (ii) checks whether the variant appears ≥ 1 time in the corresponding Illumina FM index: if so, the variant is considered a valid call; if not, the variant is considered an invalid call. Taking into account possible errors both in the consensus sequences generated by TRiCoLoR and in the Illumina sequences, we counted as valid calls also variants that are found in the Illumina FM indexes with ≤ 2 bp discrepancies (i.e., their edit distance is ≤ 2). Limited by the length of the available Illumina sequences, using this approach we could not validate variant TRs longer than 124 bp. Overall, we got high validation ratios (ratios between the valid calls and the number of calls that could be assessed using short reads): $\sim 82\%$ for HG00733 (ONT and PB), $\sim 85\%$ for HG00514, and $\sim 86\%$ for NA19240 (Supplementary Fig. S7).

We eventually ran TRiCoLoR SAGE on the Puerto Rican PB trio HG00731, HG00732, and HG00733, with the default parame-

ter settings and the “-mendel” parameter enabled to check the Mendelian consistency of the TRs identified in HG00733. With 7 processors on our Ubuntu desktop, the module took ~2 hours to complete the analysis. Filtering for variants differing from the reference for ≥ 10 bp and for multi-allelic variants differing from each other by the same distance, we identified ~80% of Mendelian consistent TRs, which is low compared to trio-based single-nucleotide variant and InDel Mendelian consistency rates but above reported genotype agreement rates for structural variants in repetitive regions [29].

Among the Mendelian consistent TRs called by TRiCoLoR on the HG00733 PB individual, we identified 32 long TRs (≥ 150 bp) that were absent in the HGSVC ground truth for the same individual. To identify the cause of these apparent discrepancies, we aligned the HG00733 phased contigs from HGSVC to the GRCh38 human reference genome with minimap2, using the assembly-to-reference alignment mode and the parameters suggested by QUAST-LG [30], and we manually inspected the discordant TRs in the aligned contigs using IGV [31]. As reported in Table 1, out of 58 non-reference TR alleles identified by TRiCoLoR, we could visually confirm 42 (~75%) of them in the HGSVC assembly, which means that both TRiCoLoR and the HGSVC predicted the same variant type (deletion or insertion) and the predicted variant size is roughly similar (i.e., the difference does not exceed 50 bp). However, for the other 16 variants (~25%), the HGSVC assembly either did not contain the allele predicted by TRiCoLoR or did not cover the investigated region, which suggests that mapping-based and assembly-based approaches can be complementary for TR detection using long reads.

Discussion

TRiCoLoR is a comprehensive TR caller for long reads that supports the *de novo* identification of TRs in whole-genome sequencing data. TRiCoLoR profiles TRs through an efficient POA algorithm combined with a RegEx-based string-matching search, facilitating a robust and accurate discovery of the full spectrum of expanded and contracted TRs in personal genomes.

In comparison with previous tools, TRiCoLoR works with ONT and PB data seamlessly. TRiCoLoR also identifies TRs *de novo* and does not require *a priori* knowledge of annotated TR regions. The unique combination of features for genome-wide, *de novo* discovery and genotyping of TRs in ONT and PB data is to the best of our knowledge unmet by any other TR caller for long-read data. Besides the detection of TRs, TRiCoLoR visualizes TRs in their haplotype context and it can infer parental genotypes using low-coverage parental sequencing data.

TRiCoLoR has been designed for diploid organisms (Supplementary Note S8), and future work includes extending its feature set to polyploid species and haploid chromosomes (human Y chromosome). As a mapping-based approach, TRiCoLoR cannot identify repeats in unassembled regions of the genome (e.g., human centromeres and telomeres). Furthermore, the entropy threshold and window size for the *de novo* identification of repetitive stretches that we empirically estimated is well suited for short repeated motifs (2–3 bp) but may need adjustments for long motifs of higher nucleotide complexity. Last, by default TRiCoLoR profiles TRs with motif lengths ≤ 6 bp (also known as micro-satellites), excluding those with motif lengths ≥ 7 bp (also known as mini-satellites), which are less abundant in diploid organisms [32]. The RegEx algorithm can also be tuned to profile mini-satellites (i.e., by extending the “-size” parameter), but

TRiCoLoR has been extensively applied so far only to micro-satellites.

Given these limitations, future work will focus on extending TRiCoLoR to other ploidies, broadening the size spectrum of detectable repeat motif lengths and taking advantage of improved sequencing read accuracy (e.g., high-fidelity long reads from PB). The latter directly improves the RegEx-based identification of repeats used by TRiCoLoR, and we thus believe that TRiCoLoR is well suited to characterize the TR landscape in present and future long-read data sets, making it an instrumental tool for robustly deciphering the multiplicity of TRs in repeat-mediated clinical disorders.

Availability of Source Code and Requirements

Project name: TRiCoLoR

Project home page: <https://github.com/davidebolo1993/TRiCoLoR>. A dockerized version of TRiCoLoR is available at <https://hub.docker.com/r/davidebolo1993/tricolor>. On-line documentation is available at <https://davidebolo1993.github.io/tricolordoc>.

Operating system: Unix

Programming languages: Python, Bash, C++

Other requirements: Python 3.6 or higher, GCC 4.8 or higher, and CMake 3.2 or higher

License: GNU Lesser General Public License 3.0

RRID:SCR_018801

biotools ID: tricolor

Availability of Supporting Data and Materials

HGSVC whole-genome long-read sequencing data are available on the HGSVC website (<https://www.internationalgenome.org/human-genome-structural-variation-consortium>). Specifically:

ONT FASTQ files: http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20181210.ONT_rebasecalled

PB alignments: http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20180102_pacbio_blasr_reheader

Phased single-nucleotide variants: http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20170323.Str_and-seq_phased_FB%2BGATK.VCFs

Illumina FASTQ files: http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/illumina_wgs.sequence.index

HG00733 phased contigs: http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20180227_PhasedSVGenomes

HG00733 ground truth of structural variant calls: http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20180627_PanTechnologyIntegrationSet/HG00733_merged_nonredundant.vcf

The GRCh38 human reference genome used for alignments is available at http://ftp.ebi.ac.uk/1000g/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa. The corresponding annotated TRs can be accessed through the UCSC Table Browser tool (<http://genome.ucsc.edu>).

A whole-genome ONT FASTQ file of the *Arabidopsis thaliana* KBS-Mac-74 is available at <ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR217/003/ERR2173373/ERR2173373.fastq.gz>. The TAIR10 reference genome for *A. thaliana* can be downloaded through the Arabidopsis Information Resource database (<https://www.arabidopsis.org/index.jsp>). Several scripts used to perform the analyses described in this article and the TR calls generated by TRiCoLoR for the HGSVC individuals and the *A. thaliana* KBS-Mac-74 are

Table 1: Comparison between TRiCoLoR's mapping-based and HGSVC's assembly-based approaches for Mendelian consistent long TRs identified by TRiCoLoR on the HG0733 PB individual

Chromosome	Start	End	HGSVC assembly	TRiCoLoR call
chr1	23703657	23703893	DEL;INS	DEL;INS
chr1	223672571	223672681	INS;INS	INS;INS
chr10	69539376	69539572	INS;INS	INS;INS
chr11	79190887	79191145	REF;REF	DEL;INS
chr11	128436913	128437081	INS;INS	INS;INS
chr14	84276747	84276903	REF;DEL	INS;DEL
chr15	70364402	70364587	INS;NA	INS;INS
chr16	3529535	3529854	REF;DEL	LC;INS
chr17	27525992	27526118	INS;INS	INS;INS
chr18	44544809	44545037	INS;INS	INS;INS
chr18	59081301	59081379	INS;INS	INS;INS
chr18	71198388	71198450	REF;NA	REF;INS
chr2	160426201	160426342	INS;INS	INS;INS
chr2	211860947	211861156	DEL;NA	DEL;INS
chr21	35063465	35063588	INS;INS	INS;INS
chr22	46174187	46174274	REF;INS	REF;INS
chr3	13856835	13857013	DEL;INS	DEL;INS
chr4	13807826	13807982	REF;REF	REF;INS
chr4	18837113	18837320	INS;DEL	INS;DEL
chr4	81637241	81637408	DEL;DEL	DEL;DEL
chr5	54513584	54513735	REF;INS	INS;INS
chr6	25450910	25450975	REF;INS	REF;INS
chr6	55543085	55543393	INS;INS	INS;INS
chr6	106945844	106946002	DEL;DEL	DEL;INS
chr7	38610247	38610412	NA;DEL	INS;DEL
chr7	71847696	71847865	INS;INS	INS;INS
chr7	109663557	109663744	INS;DEL	INS;DEL
chr7	131933466	131933651	INS;INS	INS;INS
chr9	82850174	82850347	DEL;DEL	DEL;DEL
chr9	91622218	91622365	NA;NA	INS;REF
chr9	91634814	91634973	NA;NA	DEL;INS
chr9	116632126	116632280	INS;INS	INS;INS

DEL: deletion; INS: insertion; REF: reference allele; NA: region is not covered by the assembly or mis-assembled; LC: TRiCoLoR could not generate a consensus sequence for the allele owing to the low coverage in the region. The 2 alleles are separated by a semicolon.

available through the GitHub code repository of TRiCoLoR (<https://github.com/davidebolo1993/TRiCoLoR>). More in detail:

The <https://github.com/davidebolo1993/TRiCoLoR/tree/master/paper/data> folder contains the BED file with annotated TRs from the GRCh38 human reference genome (GRCh38.TRs.bed), a bash script that illustrates how to haplotype-resolve a long-read alignment using phased single-nucleotide variants (prepare.sh), a python script used for the Shannon entropy simulations (entropy.py), a python script used to calculate precision, recall, and F1 scores of TRiCoLoR on synthetic data (pr.py), and a couple of C++ source code files (fmindex.cpp and validate.cpp) for validating TRiCoLoR calls on real human data.

The <https://github.com/davidebolo1993/TRiCoLoR/tree/master/paper/samples> folder contains TRiCoLoR calls for the HGSVC individuals and the *A. thaliana* KBS-Mac-74 in standard BCF format.

A snapshot of the archival code is available in the GigaScience GigaDB database [33].

Additional Files

Supplementary Figure S1.
Supplementary Figure S2.
Supplementary Figure S3.
Supplementary Figure S4.
Supplementary Figure S5.
Supplementary Figure S6.
Supplementary Figure S7.

Supplementary Note S1.

Supplementary Note S2.

Supplementary Note S3.

Supplementary Note S4.

Supplementary Note S5.

Supplementary Note S6.

Supplementary Note S7.

Supplementary Note S8.

Abbreviations

bp: base pairs; F1: F1 score; HGSVC: Human Genome Structural Variation Consortium; ONT: Oxford Nanopore Technologies; P: precision; PB: Pacific Biosciences; POA: partial order alignment; R: recall; RegEx: regular expression; SENSOR: Shannon Entropy Scanner; TR: tandem repeat; UCSC: University of California Santa Cruz.

Competing Interests

The authors declare that they have no competing interests.

Funding

J.O.K. is supported by GraphGenomes grant 031L0184C. A.M. is supported by AIRC grant 20307. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

D.B. and T.R. designed and benchmarked the software. D.B. wrote the code. T.R. supervised the work. D.B. and T.R. co-wrote the manuscript draft. A.M., V.B., and J.O.K. contributed to the interpretation of the results, provided critical feedback, and helped to write the manuscript. All the authors read and approved the manuscript.

Acknowledgments

The authors thank HGSVC for data access, and EMBL GeneCore and IT for technical support.

References

- de Koning APJ, Gu W, Castoe TA, et al. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 2011;**7**(12):e1002384.
- Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* 2018;**19**(5):286–98.
- Paulson H. Repeat expansion diseases. *Hanb Clin Neurol* 2018;**147**:105–23.
- Gymrek M, Golan D, Rosset S, et al. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* 2012;**22**(6):1154–62.
- Highnam G, Franck C, Martin A, et al. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* 2013;**41**:e32.
- Doi K, Monjo T, Hoang PH, et al. Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics* 2014;**30**(6):815–22.
- Velasco A, James BT, Wells VD, et al. Look4TRs: A de novo tool for detecting simple tandem repeats using self-supervised hidden Markov models. *Bioinformatics* 2020;**36**(2):380–7.
- Chaisson MJP, Sanders AD, Zhao X, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* 2019;**10**:1784.
- Ummat A, Bashir A. Resolving complex tandem repeats with long reads. *Bioinformatics* 2014;**30**(24):3491–8.
- Harris RS, Cechova M, Makova KD, et al. Noise-cancelling repeat finder: Uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics* 2019;**35**(22):4809–11.
- Gao Y, Liu B, Wang Y, et al. TideHunter: Efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. *Bioinformatics* 2019;**35**(14):i200–7.
- 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**(1):239.
- Mitsuhashi S, Frith MC, Mizuguchi T, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol* 2019;**20**(1):58.
- Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002;**18**(3):452–64.
- Lee C. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics* 2003;**19**(8):999–1008.
- Bolognini D. Browsable and searchable documentation for TRiCoLoR. <https://davebollo1993.github.io/tricolordoc>. 2019.
- Tenreiro MacHado JA. Shannon entropy analysis of the genome code. *Math Probl Eng* 2012;**2012**, doi:10.1155/2012/132625.
- Vaser R, Sović I, Nagarajan N, et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;**27**(5):737–46.
- Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**(18):3094–100.
- Bolognini D, Sanders A, Korbel JO, et al. VISOR: A versatile haplotype-aware structural variant simulator for short- and long-read sequencing. *Bioinformatics* 2020;**36**(4):1267–9.
- Bolognini D, Bartalucci N, Mingrino A, et al. NANOR: A user-friendly R package to analyze and compare nanopore sequencing data. *PLoS One* 2019;**14**(5):e0216471.
- Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
- Rausch T, Hsi-Yang Fritz M, Korbel JO, et al. Alfred: Interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics* 2019;**35**(14):2489–91.
- Pedersen BS, Quinlan AR. Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics* 2018;**34**(5):867–8.
- Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat Commun* 2019;**10**(1):4660.
- Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. *Front Genet* 2019;**10**:426.
- Dolle DD, Liu Z, Cotten M, et al. Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes. *Genome Res* 2017;**27**(2):300–9.
- Ferragina P, Manzini G. Opportunistic data structures with applications. In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*, Redondo Beach, CA. 2000, doi:10.1109/SFCS.2000.892127.
- Giner-Delgado C, Villatoro S, Lerga-Jaso J, et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat Commun* 2019;**10**(1):4222.
- Mikheenko A, Prjibelski A, Saveliev V, et al. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 2018;**34**(13):i142–50.
- Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* 2013;**14**(2):178–92.
- Richard GF, Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev* 2008;**72**(4):686–727.
- Bolognini D, Magi A, Benes V, et al. Supporting data for “TRiCoLoR: tandem repeat profiling using whole-genome long-read sequencing data.” *GigaScience Database* 2020. <http://dx.doi.org/10.5524/100790>.