Accurate detection of tandem repeats from error-prone sequences with EquiRep

Zhezheng Song[1,†], Tasfia Zahin[1,†], Xiang Li[1], and Mingfu Shao[1,2,*]

[1] Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA

[2] Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

Abstract. A tandem repeat is a sequence of nucleotides that appear as multiple contiguous, near-identical copies arranged consecutively. Tandem repeats are widespread across natural genomes, play critical roles in genetic diversity, gene regulation, and are associated with various neurological and developmental disorders. They can also arise in sequencing reads generated by certain technologies, such as those used for sequencing circular molecules. A key challenge in analyzing tandem repeats is reconstructing the sequence of the underlying repeat unit. While several methods exist, they often exhibit low accuracy when the repeat unit length increases or the number of copies is low. Furthermore, methods capable of handling highly mutated sequences remain scarce, highlighting a significant opportunity for improvement. We introduce EquiRep, a tool for accurate detection of tandem repeats from erroneous sequences. EquiRep estimates the likelihood of positions originating from the same location in the unit through self-alignment, followed by a novel refinement approach. The resulting equivalence classes and consecutive position information are then used to build a weighted graph. A cycle in this graph with maximum bottleneck weight covering most nucleotide positions is identified to reconstruct the repeat unit. We test EquiRep on two applications, identifying repeat units from satellite DNAs and reconstructing circular RNAs from rolling-circular long-read sequencing data, using both simulated and raw sequencing datasets. Our results show that EquiRep consistently outperforms or matches state-of-the-art methods, demonstrating robustness to sequencing errors and superior performance on long repeat units and low-frequency repeats. These capabilities underscore EquiRep's broad utility in tandem repeat analysis.

Keywords: tandem repeats · error-prone long reads · equivalence classes · local alignment

[†]contribute equally to this work

[*]to whom correspondence should be addressed; email: mxs2589@psu.edu

Introduction

A tandem repeat is informally referred to as the appearance of multiple consecutive copies of the same sequence (termed as the repeat unit). Tandem repeats are commonly found in natural genomes, but they can also be introduced intentionally in certain sequencing protocols that produce reads composed of tandem repeats. Due to either mutations or sequencing errors, the observed sequences or reads are often not exact copies of the repeat unit but containing errors. Analyzing tandem repeats thus often requires to reconstruct the (unknown) unit from the erroneous, noisy sequences. Below we first describe two biological applications involving tandem repeats. We then formally formulate the problem and present our algorithm.

The human genome consists of a vast array of repetitive elements, and many of them arise from a process called tandem duplication. In this process, a segment of the DNA is replicated multiple times, creating consecutive approximate repeat units. The length of these repeat units vary from a few base pairs (called short tandem repeats or STRs) to a hundred base pairs (called variable number tandem repeats or VNTRs) and sometimes upto thousand base pairs in satellite DNAs. Tandem repeats make up about 8-10% of the human genome and have been closely linked to several neurological and developmental disorders like Huntington's disease, Friedreich's Ataxia, fragile X syndrome, etc (Hannan, 2018; Siwach and Ganesh, 2008; Usdin, 2008). The repeat tracks associated with many of these diseases appear longer in certain affected individuals than typically observed in the general population (Hannan, 2018; Siwach and Ganesh, 2008; Usdin, 2008). For example, the GAA unit associated with Friedreich's Ataxia appears 5-30 times normally, but 66 to over 1000 times in affected individuals (Campuzano et al., 1996). More recently, longer repeats copies (25-30bp) have been discovered to influence schizophrenia (Song et al., 2018) and Alzheimer's disease (De Roeck et al., 2018). Alpha satellite repeats of about 171 bp (i.e., the so-called monomers) are found to be abundant in centromeric regions of many organisms and are essential for studying genome stability and evolutionary dynamics (Logsdon et al., 2024; Melters et al., 2013). To analyze tandem repeats, a critical step often involves the accurate reconstruction of the unit from either assembled genome or unassembled (long) reads.

The rolling circle amplification (RCA) is a recently refined sequencing technique that amplifies circularized template molecules, producing numerous tandem repeat copies of the original template. RCA can yield long tandem repeat units, with sequences often exceeding 150 bp and even reaching several kilobases in certain contexts. RCA followed by PacBio or Oxford Nanopore Technologies (ONT) sequencing is a popular protocol adopted in many recent studies, specially for detection of full-length circular RNAs (Xin et al., 2021; Zhang et al., 2021; Liu et al., 2021). A crucial step in this process is the prediction of a consensus sequence derived from long reads, providing a highly accurate reconstruction of the original template (e.g., circular RNA). This step requires *in silico* intervention, and typically employs widely used tandem repeat detection

tools for consensus sequence prediction. It is important to emphasize that the reliability of circular RNA detection is therefore significantly influenced by the accuracy of the predicted consensus sequence during this intermediate step. Consequently, there is a pressing need for reliable tools capable of accurately predicting tandem repeat patterns of different kinds, accounting for the variability in unit length and copy number that may exist in different biological contexts. Addressing this gap is particularly essential for improving the accuracy and reliability of full-length circular RNA identification, especially considering that circular RNAs have emerged as promising biomarkers for numerous diseases (Rybak-Wolf et al., 2015; Kristensen et al., 2022; Wang et al., 2016).

Both above critical applications can be abstracted as this computational problem: given a sequence R, decide if R contains tandem repeats (with mutations and errors) of a unit, and if yes, construct the sequence of the unit. Many methods have been developed, mainly driven by the development of sequencing technologies. Tools include mreps (Kolpakov et al., 2003), RepeatMasker (`https://www.repeatmasker.org/`), and INVERTER (Wirawan et al., 2010) are primarily designed to detect small repeat units from relatively low error rate data such as short-read sequencing data. They often do not perform well with higher repeat lengths and/or lower frequencies. Other tools like DeepRepeat (Fang et al., 2022), tandem-genotypes (Mitsuhashi et al., 2019), and ExpansionHunter (Dolzhenko et al., 2019) emphasize more the quantification of tandem repeats than unit reconstruction. Tandem Repeat Finder (TRF) (Benson, 1999) is one of the most widely used tandem repeat detection tools. It is based on the idea of k-tuple matching and utilizes a probabilistic model followed by statistical analysis to make repeat predictions. It is also suitable for use in erroneous long reads given its ability to handle substitutions and indels. With the advent of third-generation sequencing and the resulting access to long-reads data, new tools such as TideHunter (Gao et al., 2019) and mTR (Morishita et al., 2021) began to emerge. TideHunter is an efficient tandem repeat detection and consensus calling tool tailored for RCA-based long reads sequences. However, it faces challenges in accuracy when dealing with repeat of small length. Similarly, mTR struggles with repeats of low copy numbers, mostly due to difficulty in finding a long cycle of short, infrequent k-mers. Despite the promising potential of long-reads in revealing novel disease-associated tandem repeats and in reconstructing full-length circRNAs, tools capable of managing high error rates are rare. Those currently available also struggle to achieve satisfactory accuracy in challenging settings (such as too short/long units and low copy numbers), as suggested by our experiments. Therefore, the task of accurately detecting tandem repeats from noisy sequences, particularly for longer units and low copy numbers, remains largely unresolved.

Here we present EquiRep, a new tool for reconstructing the tandem repeat unit from error-prone sequences. EquiRep stands out for its robustness against sequencing errors, as well as its effectiveness in detecting

repeats of low copy numbers. EquiRep employs a novel idea that identifies *equivalent* positions in the given sequence. This is achieved by self-local alignment followed by a critical refinement step that reduces the noises. The refined, equivalent positions are organized into equivalence classes. A graph is constructed where nodes are equivalence classes and the identification of unit can be formulated as searching for a cycle in the graph with maximized bottleneck weight. We then evaluate the accuracy of EquiRep compared to leading methods across a variety of datasets over the two aforementioned applications, reconstructing repeat unit from satellite DNA and circular RNAs from RCA data.

Results

We implemented the algorithm described in Methods section as a new tandem repeat reconstruction tool named EquiRep. We compare EquiRep to four other repeat detectors: TRF, mTR, mreps, and TideHunter. For a given input sequence, each of these methods can generate multiple repeat patterns as the output while EquiRep generates a single repeat pattern. If there are multiple predictions, we choose the unit corresponding to a criterion (for example, maximum copy number) best for the method as the final predicted sequence. We evaluate these methods both on simulated and real datasets as follows.

Evaluation with Simulated Random Sequences

The simulated random sequences are generated as follows: (1), generate a random string U constituting nucleotides (A,T,G,C) of length 5, 10, 50, 100, 200, 500, 1000, which serves as the ground truth repeat unit; (2), concatenate multiple copies of the unit U to generate a longer sequence, with frequency (number of copies) of the unit being 3, 5, 10, and 20; (3), introduce random errors—insertions, deletions, and substitutions at equal probabilities—at rates of 10%, 15% and 20% into the concatenated string to simulate real-world sequencing errors and mutations; (4), insert random strings, matching the length of the concatenated string (i.e., the repeat region), at both sides of the concatenated string.

For each of the settings (the combination of unit length, frequency of units, and error rate), we randomly and independently generate 50 sequences. We evaluate the methods' predictions as follows. Let T be a ground-truth repeat unit and let P be a prediction. We compute a rotation-aware edit distance between P and T. Specifically, since P may be a rotation of the T, we calculate the edit distance between T and all possible rotations of P, and take the minimum value, defined as the rotation-aware edit distance. For each setting, we analyze the 50 instances and report the following 3 metrics. First, we measure *accuracy* as the number of instances (out of 50) where the method predicts the exact ground-truth unit (i.e., rotation-aware edit

distance is 0). Second, we evaluate the proportion of *close predictions*, defined as cases where the rotation-aware edit distance is less than 10% of the true unit length. Third, we report the average of the normalized rotation-aware edit distance (distance divided by the unit length) across all 50 instances.

Fig. 1(A-G) compares the accuracy on simulated data at 10% error rate for various lengths and copy numbers. EquiRep consistently predicts a comparable or greater number of correct instances than other methods. The methods with performance closest to EquiRep appear to be mTR and TRF; however, both struggle to maintain accuracy with large unit lengths. The accuracy of EquiRep is significantly higher than any of the other methods for unit length 500 and 1000 bp which demonstrates the ability of our tool to predict longer tandem repeats. Fig. 2(A-G) compares the ratio of close predictions on simulated data at 10% error rate. The ratio for EquiRep is high regardless of the copy number and the trend tends to be consistent over the different unit lengths, unlike other methods. Fig. 3(A-G) compares the averaged normalized rotation-aware edit distance. Observe that EquiRep consistently achieves the lowest distance, indicating that even when its predictions are incorrect, they remain the closest to the true sequence.

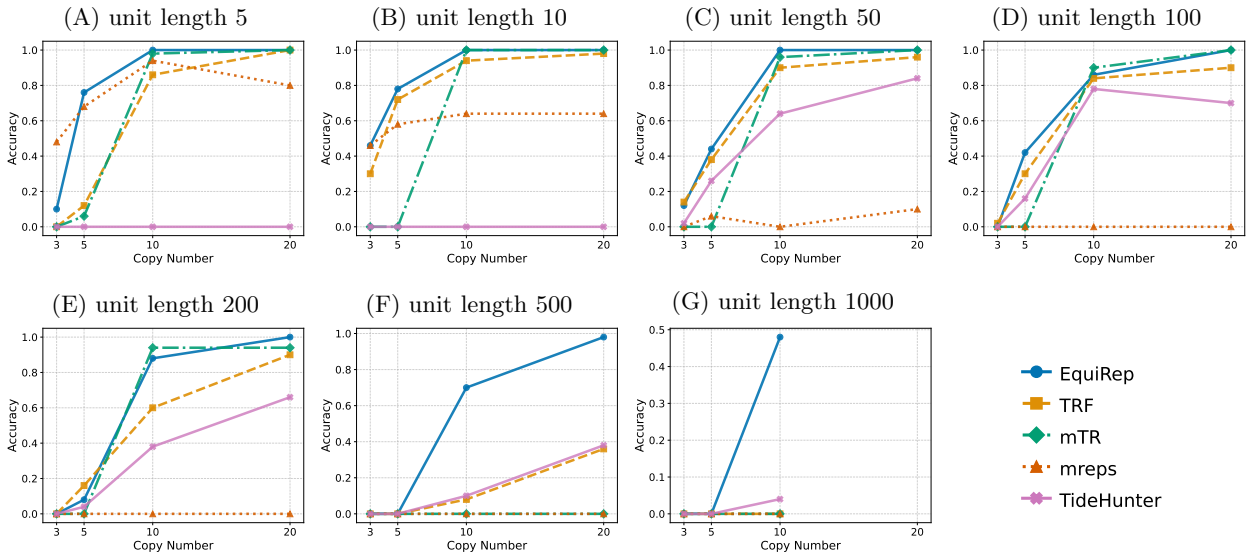

Fig. 1: Comparison of accuracy on simulated data at 10% error rate.

To better ilustrate the distributions of the normalized rotation-aware edit distances between the predicted unit and the ground-truth, we show the fine-grained plots for all simulated settings on data with 10% error rate, available in Supplementary Figures S5-S31.

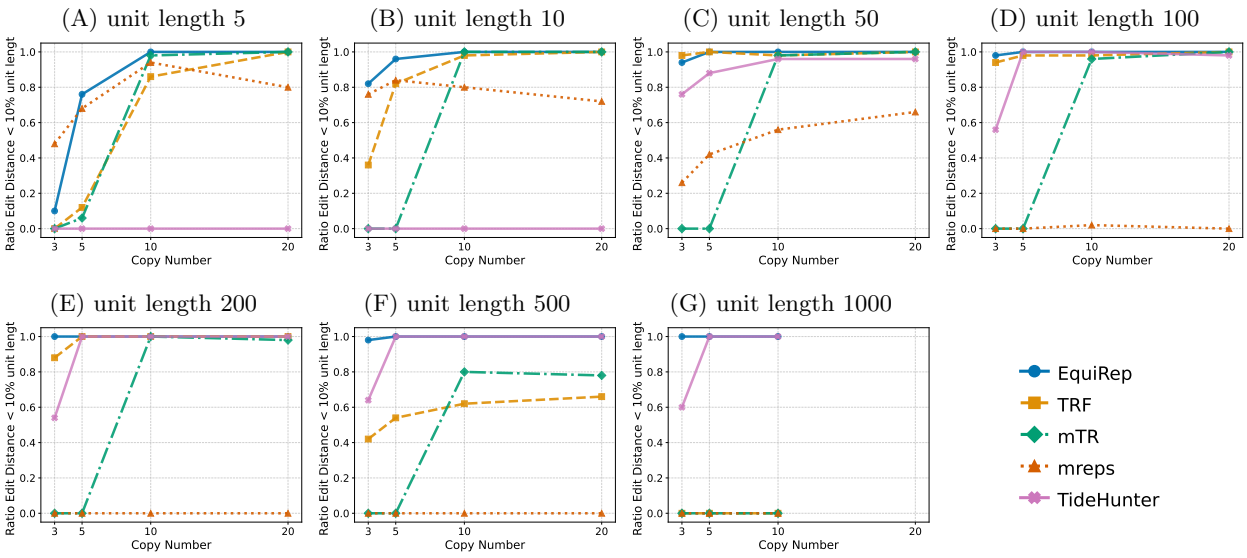

Fig. 2: Comparison of proportion of close predictions (rotation-aware edits less than 10% of the unit length) on simulated data at 10% error rate.

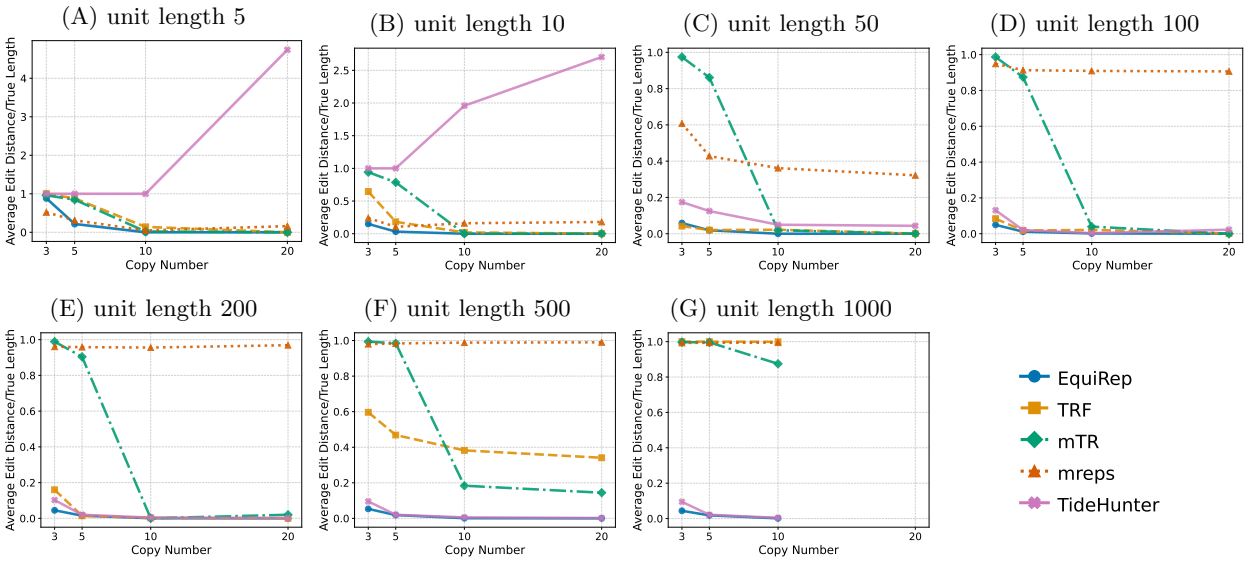

Fig. 3: Comparison of average normalized rotation-aware edit distance on simulated data at 10% error rate.

A comparison with another approach, dot2dot (Genovese et al., 2019), on simulated data with 10% error rate, is available at Supplementary Figures S32(A-G), S33(A-G), S34(A-G). EquiRep outperforms dot2dot drastically on all settings.

Results for 15% and 20% error rates are available in Supplementary Figures S35(A-G), S36(A-G), S37(A-G), and Supplementary Figures S38(A-G), S39(A-G), S40(A-G), respectively. For higher error rate, TRF, mreps, and TideHunter see a sharp decline in accuracy as the unit length exceeds 10 bp. Conversely, mTR's ability

to handle long, noisy reads allows it to achieve accuracy close to EquiRep; however, its performance drops when the unit length reaches 500 bp or longer. At such a long unit and with high sequencing errors, all methods struggle to accurately predict tandem repeats, particularly when the copy number is low. Overall, EquiRep outperforms other tool on the three metrics across different simulations.

Evaluation with Data Simulated with PBSIM2

To better mimic the real long reads, we evaluated our method using data simulated by PBSIM2 (Ono et al., 2020). To simulate, we first generate sequences containing repeats positioned in the middle with random sequences flanking both ends. The repeat configurations were consistent with those described in Subsection: Evaluation with Simulated Random Sequences, including repeat units of lengths 5, 10, 50, 100, 200, 500, and 1000, with each unit repeated 3, 5, 10, or 20 times. The following command `(pbsim --depth 1 --hmm_model PC64.model --accuracy-mean 0.90)` is subsequently used to simulate long reads using PBSIM2. Results were compared against the same set of alternative methods, detailed in Supplementary Figures S41(A-G), S42(A-G), S43(A-G). EquiRep consistently outperformed competing methods nearly all scenarios, highlighting its effectiveness on more realistic simulated reads.

Evaluation using Simulated Sequences with Recurring *k*-mers in a Unit

Genomic sequences are not pure random, often containing recurring substrings. We compare different methods on this scenario with simulations where the repeat unit itself contains recurring structures. In this setting, predicting the correct repeat sequence is challenging as methods may encounters difficulties in distinguishing between such recurring k-mers in a single unit and identical k-mers across multiple units.

We use this approach to simulate the above sequences. (1), for a given unit length $l \in \{50, 200, 500\}$, we generated a random k-mer of length $k \in \{5, 10, 20\}$, respectively; (2), we construct the repeat unit by concatenating the random k-mer 2 or 3 times. After these concatenations, any remaining positions within the unit (i.e., $l - 2k$ for 2 concatenations and $l - 3k$ for 3 concatenations) will be filled with random nucleotides; (3), we concatenate multiple copies of the repeat unit to generate a longer sequence, with frequency of units being 3, 5, 10, 20; (4), we introduce random errors at rates of 10% and 20%; (5), at the end we insert random strings, matching the length of the concatenated string at both ends.

The same evaluation metrics for the previous simulations are also used here. Supplementary Figure S44(A-C) indicates accuracy (the ratio of fully correct instances) of EquiRep exceeds or is equal to other methods when the simulations have 2 copies of a k-mer within the unit at 10% error rate. Supplementary Figure S45(A-C) shows that almost for all instances the edits predicted by our method are less than 10% of the unit length.

Again, EquiRep achieves the lowest averaged distance as illustrated in Supplementary Figure S46(A-C). Supplementary Figures S47(A-C), S48(A-C), S49(A-C) demonstrate the results for data with 20% error rate. There is a drastic decline in accuracy for all methods except mTR and EquiRep.

Nearly all repeat units generated by EquiRep have edits below 10% of the unit length for copy numbers above 10, which highlights the reliability of our predictions specially in challenging erroneous settings.

We also tested all methods on another set of data with 3 copies of repeating k-mers within the repeat unit, shown in Supplementary Figures S50(A-C), S51(A-C), S52(A-C) (for error rate of 10%) and in Supplementary Figures S53(A-C), S54(A-C), S55(A-C) (for error rate of 20%). EquiRep is able to make better or similar predictions in all cases indicating that its algorithm is least affected by the presence of embedding k-mers within repeat units.

Evaluation using Human Satellite DNA Data

We then test all methods on reconstructing repeat unit for satellite DNA in human Chromosome 5 (Paar et al., 2007). This known satellite DNA consists of 13 units (i.e., 13 monomers) each of which is of size around 171bp. To construct the input sequence for methods to predict, we concatenate the 13 monomers into a string denoted as (x). To create more testing instances, we introduce flanking regions on both sides of the concatenation denoted as (axa), and introduce errors of 1%, 5%, and 10% to (x) and (axa). To evaluate the predicted unit by different methods, we calculate the normalized rotation-aware edit distance between the predicted unit with each of the 13 known monomers and report the averaged distance.

Table 1 shows the results. EquiRep consistently maintains a lower normalized distance, outperforming or matching all other tools. The values for EquiRep are similar to mTR when the input sequences have flanking regions at either end (axa) but our method is about 87% better than mTR when just the repeat region is provided (x). Although TideHunter and TRF exhibit accuracy levels similar to ours, they fall short at higher error rates, where EquiRep excels with an 87% improvement.

Evaluation using *C. elegans* Centromere ONT Data

We adopted a dataset reported in (Yoshimura et al., 2019) that studied the assembly of *C. elegans* genome using Nanopre long-reads data. We collected the raw long reads that are aligned to centromere (listed in its Supplementary Figure S4). Each of the long reads may contain more than 1 repeating regions. Since our current method does not support detecting multiple repeating regions in a single input sequence, we manually extract the rough region with repeats. Specifically, we first generate a dot plot for each long read, observe

Table 1: Averaged normalized rotation-aware edit distance on human satellite DNA data.

Error Rate (%)	Pattern	EquiRep	mTR	TRF	mreps	TideHunter
0	**x**	0.1260	0.9960	0.1274	0.9492	0.1305
0	**axa**	0.1255	0.1260	0.1274	0.9737	0.1305
1	**x**	0.1251	0.9960	0.1408	0.9492	0.1282
1	**axa**	0.1251	0.1260	0.1408	0.9492	0.1282
5	**x**	0.1282	0.9843	0.2267	0.9204	0.1489
5	**axa**	0.1269	0.1264	0.2267	0.9263	0.1489
10	**x**	0.1363	0.9960	0.9960	0.9370	1.0550
10	**axa**	0.1251	0.1498	0.9960	0.9664	1.0550

the repeating regions, and then manually cut out these regions and pipe them to each of the methods. The ground-truth sequence of the unit is available, which are obtained by curating from PacBio HIFI datasets.

Table 2 presents the normalized rotation-aware edit distance between the predicted units and the ground truth. We report the average value across all cases. EquiRep achieves the second-best performance. For each method, we also report the number of cases where the normalized rotation-aware edit distance is below 0.2, indicating high-quality predictions. EquiRep performs well in 7 out of 13 cases, while the top-performing methods, mTR and TRF, achieve good predictions in 8 cases.

Evaluation with Rolling Circle Amplification (RCA) Data

The set of real data is a RCA based ONT sequencing protocol from isocirc (Xin et al., 2021) that has been used to detect a catalogue of full-length circular RNAs from 12 human tissues. We consider a subset of 101 sequences from prostate tissue long-read ONT data (obtained from the NCBI Gene Expression Omnibus [GEO; https://www.ncbi.nlm.nih.gov/geo/] under accession number GSE141693) for analysis. It is difficult to evaluate the repeats from the RCA based long reads data due to lack of reliable ground truth, so we evaluate this data in two different ways. Firstly, we use a dot plot analysis. Dot plots have served as a common approach for visualizing and identifying the structural patterns of sequences such as repeats. We first align the input sequence to itself with LASTZ (Harris, 2007) using specific parameters designed for generating dot plots. The alignment program generates a dot file which can be converted to an image file for visualization using a simple R (R Core Team, 2021) script. The dot file can be used to estimate the repeat unit length (but not sequence of the unit). We treat this estimate as a benchmark for comparing the predictions of EquiRep and other tools. We report the number of predictions that fall within 5%, 20%, 50%, and 80% error range of the true length. For the second approach, we first concatenate copies of the unit predicted to get a string A which is longer than the input sequence. Then we get the "semi-edit distance"

Table 2: Performance on raw ONT long reads from *C. elegans* centromere. Numbers are the normalized rotation-aware edit distance between the predicted units and the ground truth unit. The averaged normalized rotation-aware edit distance and the number of instances where a method achieves a rotation-aware edit distance less than 0.2 is summarized at the bottom.

Read Name/Region	**Unit Length**	**EquiRep**	**mTR**	**TRF**	**mreps**	**TideHunter**
SRR7594463.177832.regionA	26	0.9615	0.0385	0.0385	0.9615	0.7692
SRR7594463.177832.regionB	27	0.1111	0.1481	0.0000	0.9259	0.9259
SRR7594463.179860.regionA	27	0.9630	0.4074	4.9630	0.9630	4.8148
SRR7594463.179860.regionB	166	0.0904	0.0663	0.0783	0.9940	0.0904
SRR7594463.83311.regionA	166	0.0542	0.0241	0.0482	0.9940	0.0361
SRR7594463.83311.regionB	27	0.1481	0.6296	0.0741	0.9630	0.9259
SRR7594463.64356.regionA	226	0.0133	0.0044	0.0265	0.9956	0.0133
SRR7594463.64356.regionB	27	0.0741	0.1111	0.1111	0.9630	0.8148
SRR7594463.141714.regionB	27	0.5926	0.5185	0.5556	0.9630	3.1481
SRR7594463.82476.regionA	27	1.5556	0.5556	0.5556	0.9630	1.0741
SRR7594463.176233.regionA	27	0.8889	0.0741	0.2593	0.9630	0.8519
SRR7594463.176233.regionB	94	0.1596	0.1277	0.0745	0.9681	0.1383
SRR7594463.189890.regionB	94	0.4362	0.4149	0.8830	0.9894	0.4149
Average		0.4653	0.2400	0.5898	0.9690	1.0783
Count (< 0.2)		7	8	8	0	4

which is the smallest edit distance between any substring of A and the input sequence. The idea behind this metric is that, if the prediction is accurate, then the multiple concatenation of it should match the input sequence very well. We record the smallest edit distance and report the number of instances on which a method has a ratio (semi-edit-distance)/(input-sequence-length) less than or equal to 0.1, 0.2, 0.3, 0.5, 0.8.

Table 3 compares different methods in terms of the predicted repeat unit length, and Table 4 compares the normalized semi-edit-distance. In both metrics, EquiRep demonstrates high accuracy, consistently outperforming mTR, TRF, and mreps. The results are also comparable to TideHunter, which is specifically optimized for RCA-based analysis. Given that the exact repeat sequences for this dataset are not available, similar metric values in the table can be interpreted as comparable accuracy. It should be noted that while TideHunter excels on RCA data, its accuracy diminishes on shorter unit repeats as indicated by the simulation results. This highlights that EquiRep is adaptable to a broad range of complex sequences and versatile for various applications.

In above analysis of the RCA datasets, we observed that many repeat units exceed 1000 bp in length. This is consistent with the fact that many expressed circular RNAs are themselves longer than 1000 bp.

Table 3: Performance on RCA data: number of predicted repeat lengths within error ranges of the true length and number of no repeats found (out of 101 instances).

Error Range	EquiRep	mTR	TRF	mreps	TideHunter
0.95 to 1.05 (5%)	98	5	68	1	101
0.8 to 1.2 (20%)	100	5	68	1	101
0.5 to 1.5 (50%)	100	5	68	1	101
0.2 to 1.8 (80%)	101	9	69	1	101
#norepeat	0	18	30	7	0

Table 4: Performance on RCA data: number of predicted repeat units with ratio of edit distance to input length less than various percentages (out of 101 instances). SED = semi-edit-distance.

SED/Length	EquiRep	mTR	TRF	mreps	TideHunter
≤ 0.05 (5%)	0	0	0	0	0
≤ 0.1 (10%)	67	5	52	0	73
≤ 0.2 (20%)	99	5	68	1	101
≤ 0.3 (30%)	101	5	68	1	101
≤ 0.5 (50%)	101	28	69	40	101
≤ 0.8 (80%)	101	83	71	94	101

These observations also support the use of longer unit lengths (e.g., 500 bp and 1000 bp) in our simulated experiments (Section: Evaluation with Simulated Random Sequences).

Analysis of sensitivity of EquiRep to parameters

We conducted experiments to analyze the sensitivity of EquiRep to its three key parameters: (1), the score threshold (default: 25) used to identify significant paths from the initial matrix D; we tested alternative values, 0, 10, and 50; (2), the window size (default: 7) used for identifying local maxima in initial matrix D; we tested two other choices, 5 and 9; (3), the number of iterations (default: 5) of iterative matrix refinement; we tested two other values, 1 and 10. To assess the effect of a choice of a parameter, we make it the only change to the default setting of EquiRep, and then compare the variant with the default EquiRep. The same simulated data, used in Section: Evaluation with Simulated Random Sequences, with 10% error rate was used here to obtain the results. We also used the same three metrics in the evaluation.

The results corresponding to the 3 parameters were given in Supplementary Figures S56(A-G), S57(A-G), S58(A-G), Supplementary Figures S59(A-G), S60(A-G), S61(A-G), and Supplementary Figures S62(A-G), S63(A-G). S64(A-G), respectively. We can conclude that EquiRep is not sensitive to any of them, justifying its default choices.

Comparison of Running Time

Supplementary Table S1 presents the runtime of all methods on the simulated data from the Section: Evaluation with Simulated Random Sequences. with a 10% error rate. On average, mTR had the longest runtime, followed by EquiRep. TRF, mreps, and TideHunter were significantly faster. As noted in the Discussion, several modules in EquiRep are parallelizable, and we are optimistic about further improving its computational efficiency.

EquiRep is well-suited for processing a large number of error-prone long reads on multi-core servers, as it operates on individual reads, allowing efficient batch processing that fully utilizes available cores. It is also likely that, in large-scale long-reads dataset, the majority of the long reads do not contain repeating regions. Fast filtering strategies, such as the seed-chaining procedure used in Step 1 of EquiRep, can quickly discard such reads, leaving only a small subset that requires full processing by the complete EquiRep algorithm.

Discussion

In this paper, we present EquiRep, a robust and accurate tool for repeat detection. By leveraging a unique approach of grouping nucleotide positions into equivalence classes, EquiRep effectively builds a weighted graph to reconstruct repeat units with high accuracy. Our method addresses key challenges in detecting both short and long tandem repeats from highly erroneous sequences, areas where existing tools often fall short. EquiRep was applied to two applications: reconstructing the repeat unit from satellite DNAs and reconstructing the circular RNAs from rolling circular long reads. Through extensive testing using both simulated and real datasets, EquiRep outperforms or matches current state-of-the-art methods, demonstrating its robustness to sequencing errors and complex repeat patterns.

The task that EquiRep solves—reconstructing the repeat unit from erroneous sequence—is a general abstraction that can potentially be applied to other scenarios. One such application is to call circular consensus sequencing (CCS) read from PacBio SMRT (Single Molecule Real-Time) sequencing raw data, which produces multiple copies (with errors) of the circularized fragment. Several methods have been developed for calling CCS reads including PacBio official consensus caller, DeepConsensus (Baid et al., 2023). We leave the comparison with these methods and the adaptation of EquiRep for CCS read generation as future work.

We demonstrated that EquiRep can be used to reconstruct the basic repeating unit of satellite DNA, known as the monomer. It is well known that satellite DNA is often organized into higher-order repeat (HOR) units, where each HOR unit comprises multiple monomers, and these HOR units are themselves repeated in tandem. Currently, EquiRep does not capture this two-level structure of satellite DNA; it only reconstructs

the repeat unit at the lower-level, i.e., the monomer. As part of future development, we intend to extend EquiRep to identify and reconstruct HOR structures as well. This enhancement would enable the analysis of more complex, nested repeat architectures and make EquiRep particularly well-suited for characterizing satellite repeats in complete, Telomere-to-Telomere (T2T) assemblies.

We are optimistic that the computational efficiency of EquiRep can be largely improved. Currently, the self-local alignment step presents a bottleneck in runtime. By improving this step, possibly through adapting more efficient alignment algorithms or parallel processing, we can substantially reduce its runtime. The second time-consuming step in EquiRep is matrix refinement. Matrix operations are inherently parallelizable, and the sparse property of the matrix can be leveraged to achieve acceleration. While parallelization can improve performance, this approach benefits all tools when provided with additional resources. Therefore, to improve EquiRep's runtime from a design perspective—not just through scaling—we aim to streamline the pipeline itself. For instance, we are exploring faster local alignment strategies and considering eliminating redundant steps, such as performing path-finding only once rather than twice as in the current design. We plan to explore these directions to make EquiRep more efficient and scalable for practical use.

We also aim for improving EquiRep's accuracy. The framework of EquiRep allows it to be improved in several ways. One approach is to enhance matrix refinement, which is crucial for producing accurate equivalence classes. The current method considers three mutually supportive pairs, but it can be extended to account for insertions and deletions. More precise modeling of insertions and deletions using equivalence classes, rather than single positions, is expected to improve node splitting, a key step in rescuing over-combines. Initial predictions of unit length might also help with guiding the search for repeat units within a specified range. Finally, improved heuristics for identifying cycles that combine both weights and optimal positional coverage would enable the weighted graph to represent complex repeat patterns more accurately. We plan to explore these strategies to enhance EquiRep's accuracy, which we expect will lead to improved performance on real datasets such as satellite repeats.

We realize that for short repeats ($\leq$ 6 bp), there is often no clear notion of a true sequence due to their imperfect nature. In such cases, where detecting expansions and contractions rather than identifying a single consensus sequence might be more meaningful, EquiRep may have limited utility. For moderately long repeats (10 - 200 bp) found in telomeric or centromeric regions, as well as coding repeats like those in *CEL* or *MUC1*, a more defined repeat structure exists, and mutations within the repeat units can have important biological implications. While EquiRep is applicable in such contexts, its current inability to automatically detect and resolve multiple repeat regions within a sequence introduces challenges for practical use. We will carefully take these factors into account as we continue to develop and refine the tool, with the goal of broadening

its applicability and improving its usability. For very large repeats ($\geq$ 500 bp) in RCA data, TideHunter demonstrates performance in both speed and accuracy. However, TideHunter is specifically optimized for RCA applications and does not perform as well in more general scenarios, particularly when dealing with shorter repeat lengths. In contrast, EquiRep is designed as a more versatile tool, aiming to provide robust performance across a broader range of repeat detection tasks and repeat size ranges.

There is a methodological similarity between EquiRep and some multiple sequence alignment approach, such as Cactus (Paten et al., 2011b,a), as both use the concept of equivalent positions. This similarity arises naturally: in multiple sequence alignment, an ancestral sequence is assumed, and the observed nucleotides or residues that correspond to the same ancestral position are considered "equivalent". EquiRep uses a similar intuition as the (unknown) number of copies are assumed to be mutated from the same repeat unit. The key difference lies in how these equivalent positions are constructed. Cactus derives equivalences from pairwise alignments, whereas EquiRep recognizes that the aligned positions obtained from the initial self-alignment are often inaccurate to serve as reliable equivalences. To address this, EquiRep introduces a novel, matrix-based iterative algorithm for more accurate reconstruction. Furthermore, EquiRep includes a heuristic that can split incorrect equivalence classes caused by over-combination. In contrast, Cactus produces smaller equivalence classes as the multiple alignment, without employing a similar correction mechanism. On top of these, we note that the two approaches are solving different tasks (multiple sequence alignment vs. reconstructing the repeat unit) with different input data (multiple sequences vs. one sequence).

Methods

Given an error-prone (long) sequence/read R, EquiRep employs a 4-step approach to determine the sequence of the true repeat unit U in it (if any).

Identifying substring S with repeating structure: From the input long read, this step determines the repeating region that potentially consists of multiple (mutated) repeats of a unit (See Supplementary Figure S1).

Constructing classes of equivalent positions $\mathcal{C}$: This step is the core part of the EquiRep framework. Equivalence classes are formed from equivalent positions using diagonal-free self local alignment and a critical refinement step (See Supplementary Figure S2). Details of the diagonal-free self alignment is available in Supplementary Note 1.

Constructing candidate units from $\mathcal{C}$: A weighted graph is created using equivalence classes as nodes and edges representing the connections between positions. A cycle with maximized bottleneck weight is identified to generate a candidate unit (See Supplementary Figure S3). More candidates are generated using

heuristics to handle false combinations and small unit sizes (See Supplementary Figure S4).

Selecting the optimal unit: Among the multiple candidate units, the one that best satisfies a defined criterion is selected as the predicted repeat unit.

An extended version of the Methods with full descriptions of all the steps is provided as a Supplemental Methods section.

Software availability

The EquiRep source code is freely available at GitHub (`https://github.com/Shao-Group/EquiRep`) and as Supplemental Material. The scripts, evaluation pipelines, and instructions that can be followed to reproduce the experimental results of this work are also available at GitHub (`https://github.com/Shao-Group/EquiRep-test`) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgment

This work is supported by the US National Science Foundation (2145171 to M.S.) and by the US National Institutes of Health (R01HG011065 to M.S.).

Author contributions: All authors designed and implemented the methods. Z.S. and T.Z. conducted the experiments. All authors discussed the results and approved the final manuscript.

References

Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, Belyaeva A, Töpfer A, Wenger AM, Rowell WJ, et al.. 2023. Deepconsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nature Biotechnology* **41**: 232–238.

Benson G. 1999. Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research* **27**: 573–580.

Campuzano V, Montermini L, Molto MD, Pianese L, Cossée M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, et al.. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic gaa triplet repeat expansion. *Science* **271**: 1423–1427.

De Roeck A, Duchateau L, Van Dongen J, Cacace R, Bjerke M, Van den Bossche T, Cras P, Vandenberghe R, De Deyn PP, Engelborghs S, et al.. 2018. An intronic vntr affects splicing of abca7 and increases risk of alzheimer's disease. *Acta neuropathologica* **135**: 827–837.

Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, Emig-Agius D, Gross A, Narzisi G, Bowman B, et al.. 2019. Expansionhunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**: 4754–4756.

Fang L, Liu Q, Monteys AM, Gonzalez-Alegre P, Davidson BL, and Wang K. 2022. Deeprepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome biology* **23**: 108.

Gao Y, Liu B, Wang Y, and Xing Y. 2019. Tidehunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. *Bioinformatics* **35**: i200–i207.

Genovese LM, Mosca MM, Pellegrini M, and Geraci F. 2019. Dot2dot: accurate whole-genome tandem repeats discovery. *Bioinformatics* **35**: 914–922.

Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics* **19**: 286–298.

Harris RS. 2007. *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University.

Kolpakov R, Bana G, and Kucherov G. 2003. mreps: efficient and flexible detection of tandem repeats in dna. *Nucleic acids research* **31**: 3672–3678.

Kristensen LS, Jakobsen T, Hager H, and Kjems J. 2022. The emerging roles of circRNAs in cancer and oncology. *Nature Reviews Clinical Oncology* **19**: 188–206.

Liu Z, Tao C, Li S, Du M, Bai Y, Hu X, Li Y, Chen J, and Yang E. 2021. circfl-seq reveals full-length circular rnas with rolling circular reverse transcription and nanopore sequencing. *elife* **10**: e69457.

Logsdon GA, Rozanski AN, Ryabov F, Potapova T, Shepelev VA, Catacchio CR, Porubsky D, Mao Y, Yoo D, Rautiainen M, et al.. 2024. The variation and evolution of complete human centromeres. *Nature* **629**: 136–145.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al.. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology* **14**: 1–20.

Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, Oma Y, Kino Y, Mitsuhashi H, and Matsumoto N. 2019. Tandem-genotypes: robust detection of tandem repeat expansions from long dna reads. *Genome biology* **20**: 1–17.

Morishita S, Ichikawa K, and Myers EW. 2021. Finding long tandem repeats in long noisy reads. *Bioinformatics* **37**: 612–621.

Ono Y, Asai K, and Hamada M. 2020. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics* **37**: 589–595.

Paar V, Basar I, Rosandic M, and Gluncic M. 2007. Consensus higher order repeats and frequency of string distributions in human genome. *Current genomics* **8**: 93–111.

Paten B, Diekhans M, Earl D, John JS, Ma J, Suh B, and Haussler D. 2011a. Cactus graphs for genome comparisons. *Journal of Computational Biology* **18**: 469–481.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, and Haussler D. 2011b. Cactus: Algorithms for genome multiple sequence alignment. *Genome research* **21**: 1512–1528.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Giusti S, Hanan M, Behm M, Bartok O, Ashwal-Fluss R, et al.. 2015. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Molecular Cell* **58**: 870–885.

Siwach P and Ganesh S. 2008. Tandem repeats in human disorders: mechanisms and evolution. *Front. Biosci* **13**: 4467–4484.

Song JH, Lowe CB, and Kingsley DM. 2018. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. *The American Journal of Human Genetics* **103**: 421–430.

Usdin K. 2008. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome research* **18**: 1011–1019.

Wang F, Nazarali AJ, and Ji S. 2016. Circular RNAs as potential biomarkers for cancer diagnosis and therapy. *American Journal of Cancer Research* **6**: 1167–1176.

Wirawan A, Kwoh CK, Hsu LY, and Koh TH. 2010. Inverter: integrated variable number tandem repeat finder. In *Computational Systems-Biology and Bioinformatics: First International Conference, CSBio 2010, Bangkok, Thailand, November 3-5, 2010. Proceedings*, pp. 151–164. Springer.

Xin R, Gao Y, Gao Y, Wang R, Kadash-Edmondson KE, Liu B, Wang Y, Lin L, and Xing Y. 2021. isoCirc catalogs full-length circular RNA isoforms in human transcriptomes. *Nature communications* **12**: 266.

Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, Smith CL, Edgley ML, Rougvie AE, Fire AZ, et al.. 2019. Recompleting the caenorhabditis elegans genome. *Genome research* **29**: 1009–1022.

Zhang J, Hou L, Zuo Z, Ji P, Zhang X, Xue Y, and Zhao F. 2021. Comprehensive profiling of circular rnas with nanopore sequencing and ciri-long. *Nature biotechnology* **39**: 836–845.

Accurate detection of tandem repeats from error-prone sequences with EquiRep

Zhezheng Song, Tasfia Zahin, Xiang Li, et al.

Genome Res. published online August 21, 2025
Access the most recent version at doi:10.1101/gr.280750.125