

Noise-Cancelling Repeat Finder: Uncovering tandem repeats in error-prone long-read sequencing data

Robert S. Harris^{1,*}, Monika Cechova¹, Kateryna D. Makova^{1,2}

¹Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA

²Center for Medical Genomics, The Pennsylvania State University, University Park, PA 16802, USA

*To whom correspondence should be addressed

Running head: Deciphering repeats from long sequencing reads

ABSTRACT

Summary

Tandem DNA repeats can be sequenced with long-read technologies, but cannot be accurately deciphered due to the lack of computational tools taking high error rates of these technologies into account. Here we introduce Noise-Cancelling Repeat Finder (NCRF) to uncover putative tandem repeats of specified motifs in noisy long reads produced by Pacific Biosciences and Oxford Nanopore sequencers. Using simulations, we validated the use of NCRF to locate tandem repeats with motifs of various lengths and demonstrated its superior performance as compared to two alternative tools. Using real human whole-genome sequencing data, NCRF identified long arrays of the (AATGG)_n repeat involved in heat shock stress response.

Availability and implementation

NCRF is implemented in C, supported by several python scripts. Source code, under the MIT open source license, and simulation data are available at <https://github.com/makovalab-psu/NoiseCancellingRepeatFinder>, and also in bioconda.

Contact

rsharris@bx.psu.edu

Supplementary Information

Supplementary data includes 7 notes, 2 figures, and 4 tables.

1 Introduction

Long tandem repeat (TR) arrays are associated with heterochromatin and play critical roles in the human genome. For instance, $(TTAGGG)_n$ repeats protect telomeres (Blackburn and Gall, 1978), $(AATGG)_n$ repeats are implicated in heat shock response (Goenka *et al.*, 2016), and alpha satellites with 171-mer repeated units bind histones participating in chromosome segregation during cell division (Masumoto *et al.*, 2004). The lengths of heterochromatin-associated TRs differ across populations (Altemose *et al.*, 2014; Wevrick and Willard, 1989) and change with aging and environmental exposure (Zhang *et al.*, 2015; Goenka *et al.*, 2016). Despite these important features of TRs, their length variation has been understudied due to the lack of experimental and computational techniques that can capture their full length.

Long TRs cannot be studied with short sequencing reads, but can be profiled with long-read technologies (Pacific Biosciences, or PacBio, and Oxford Nanopore, or Nanopore). However they are difficult to decipher because such technologies have high error rates and error profiles different from those of short-read Illumina sequencing (Schirmer *et al.*, 2016). To our knowledge, no tool currently exists to identify TR arrays in long, error-prone sequencing reads. Tools solving similar problems, primarily developed to work with short reads or assembled genomes, have limitations when applied to this use case (Lower *et al.*, 2018). Some of them fail to model asymmetric rates of insertions vs. deletions (e.g. Tandem Repeats Finder, or TRF (Benson, 1999), and Tantan (Frith, 2011)), while others do not permit high sequencing error rates (e.g. short read mappers bwa (Li and Durbin, 2009) and bowtie (Langmead, 2010)). HMMER (Eddy, 1998) reports single instances, but not repeats, of a motif. PacmonSTR (Ummat and Bashir, 2014) requires annotations of TRs in an assembled genome, while many long TRs are present in unmapped reads (Altemose *et al.*, 2014; Novák *et al.*, 2017, 2010). General purpose aligners, e.g., Minimap2 (Li, 2018), permit parameterizations for different long-read sequencing technologies, but are not designed to find TRs.

To address the decoding of TR arrays in error-prone long sequencing reads, we developed Noise-Cancelling Repeat Finder (NCRF), which identifies arrays of user-specified TR motifs in sequencing reads or assembled genomes. NCRF supports high and asymmetric rates of short insertions and deletions observed in long-read sequencing data. As a result, its performance is superior to alternative tools.

2 Developing NCRF

NCRF utilizes a modified Smith-Waterman algorithm (Smith and Waterman, 1981) aligning a read to any number of tandem copies of a specified motif sequence (Fig. S1, Supplementary Note S1A). A modified version of Dynamic Programming (DP) uses only one copy of the motif but allows loops from the end back to the beginning, so that one column of the DP matrix represents any number of tandem copies of the motif. A modified scoring recurrence supports asymmetry in insertion vs. deletion rates and high prevalence of short indels (). Resulting alignments are local in the read but global in the looped motif, and can begin and end at any position in the motif. Internal post-alignment processing removes segments with high error density (Supplementary Note S1B).

The reported alignments identify sequence segments that putatively align to perfectly repeated copies of a motif (Supplementary Note S1A). However, segments containing a mix of variants of the motif, or of a similar but different motif, may also be reported. Such mixes are consistent with the known evolutionary signatures of heterochromatic repeats (Plohl *et al.*, 2008). One can run an optional consensus filtering step to identify TR arrays with a single dominant motif, i.e. a motif whose each position matches over the array at least half the time (Supplementary Note S1C). Intervals reported for more than one motif can be identified with an optional overlap-detection step (Supplementary Note S1E).

Implementation details are presented in Supplementary Note S1.

3 Error profiles of long-read sequencing technologies

To guide our simulations and tune NCRF scoring parameters, we derived error rates for two common long-read sequencing technologies — PacBio and Nanopore (Supplementary Note S2). We observed whole-genome error rates of 15% with insertions being more common than deletions for PacBio (mismatches=1.7%, insertions=8.9%, deletions=4.3%), and error rates of 16% but favoring deletions for Nanopore (mismatches=4.6%, insertions=3.8%, deletions=7.7%; Table S1).

4 Analysis of simulated reads and tool comparison

We next simulated PacBio and Nanopore sequencing reads for a mock genome containing specific TRs with periods 5, 10, 20, 40, 80, and 171, length at least 500 bp, and mean length 1,000 bp (Supplementary Note S3). Prior to filtering (Supplementary Note S1C), NCRF discovered 99% and 95% of the specified TRs in PacBio and Nanopore reads, respectively (Fig. 1A and Table S2). However, false discovery rates were frequently high, particularly for the longer motifs, justifying the need for consensus filtering. Filtering reduced false discovery rates to approximately 1% for the 5-, 10-, and 20-mers, and below 4% for longer motifs. The reduction in false discovery rates due to filtering was accompanied by a reduction in true positive rates; by about 1% for the 5-, 10-, 20-, and 40-mers, and by a higher percentage for the 80- and 171-mer.

We next compared the performance of NCRF (applying consensus filtering) to that of two alternative tools — TRF (Benson, 1999) and Minimap2 (Li, 2018) (Fig. 1A, Supplementary Notes S4-S5). True positive rates were higher for NCRF than TRF for longer motifs, but were similar between the two tools for shorter motifs. TRF had low false discovery rates (less than 1%), however it failed to find any 171-mer repeats in Nanopore reads. Minimap2 had lower true positive rates and higher false discovery rates than NCRF across the board, except for true positive rates for the 171-mer. Surprisingly, both TRF and Minimap2 occasionally reported overlapping intervals for the same motif (Table S2). Several other tools were considered for this evaluation, but rejected after preliminary investigation (Supplementary Note S7).

5 Applying NCRF to Real Sequencing Data

Lastly, we applied NCRF to investigate perfect repeats of the (AATGG)_n repeat in publicly available PacBio and Nanopore sequenced data (Zook *et al.*, 2016; Jain *et al.*, 2018) subsampled to a common read length distribution (Supplementary Note S6) and generated for the same individual. The resulting densities of (AATGG)_n for the two technologies are reported in Fig. 1B. Searching for (AATGG)_n repeats with length of at least 500 bp in 16.5 Gb of PacBio sequenced reads, NCRF identified 8,883 repeats. 9,766,859 bp of the reads were part of these repeats, leading to the density of 0.6 bases per kilobase sequenced. In Nanopore reads, we found 9,947 repeats, totalling 35,561 kb or 2.2 bases per kilobase. Additional applications of NCRF to real sequencing data are presented in (Cechova *et al.*, 2018), where potential reasons behind differences in density between technologies are discussed.

6 Conclusions

To our knowledge, NCRF is the first tool designed specifically to identify TR arrays in noisy sequencing data, accounting for the unique characteristics of the long-read technologies. We anticipate it will accelerate research of heterochromatin-associated TR arrays and will aid in unraveling their functions in the genome.

ACKNOWLEDGEMENTS

We thank Paul Medvedev and Marzia Cremona for critical reading of the manuscript.

FUNDING

Funding was provided by the Eberly College of Sciences, The Huck Institute of Life Sciences, and the Institute for CyberScience, at Penn State, as well as, in part, under grants from the Pennsylvania Department of Health using Tobacco Settlement and CURE Funds. The department specifically disclaims any responsibility for any analyses, responsibility, or conclusions. M.C. was supported by the National Institutes of Health (NIH)-PSU funded Computation, Bioinformatics and Statistics (CBIOS) Predoctoral Training Program (1T32GM102057-0A1)

Conflict of Interest: none declared.

REFERENCES

Altemose,N. *et al.* (2014) Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput. Biol.*, **10**, e1003628.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Blackburn,E.H. and Gall,J.G. (1978) A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *J. Mol. Biol.*, **120**, 33–53.

Cechova,M. *et al.* (2018) High inter- and intraspecific turnover of satellite repeats in great apes. *bioRxiv*.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Frith,M.C. (2011) A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.*, **39**, e23.

Goenka,A. *et al.* (2016) Human satellite-III non-coding RNAs modulate heat-shock-induced transcriptional repression. *J. Cell Sci.*, **129**, 3541–3552.

Jain,M. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.

Langmead,B. (2010) Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics*, **Chapter 11**, Unit 11.7.

Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Lower,S.S. *et al.* (2018) Satellite DNA evolution: old ideas, new approaches. *Curr. Opin. Genet. Dev.*, **49**, 70–78.

Masumoto,H. *et al.* (2004) The role of CENP-B and alpha-satellite DNA: de novo assembly and epigenetic maintenance of human centromeres. *Chromosome Res.*, **12**, 543–556.

Novák,P. *et al.* (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Novák,P. *et al.* (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from

unassembled short reads. *Nucleic Acids Res.*, **45**, e111.

Schirmer,M. et al. (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, **17**, 125.

Ummat,A. and Bashir,A. (2014) Resolving complex tandem repeats with long reads. *Bioinformatics*, **30**, 3491–3498.

Wevrick,R. and Willard,H.F. (1989) Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. *Proc. Natl. Acad. Sci. U. S. A.*, **86**, 9394–9398.

Zhang,W. et al. (2015) A Werner syndrome stem cell model unveils heterochromatin alterations as a driver of human aging. *Science*, **348**, 1160–1163.

Zook,J.M. et al. (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*, **3**, 160025.

FIGURES

Figure 1. (A) Comparison of discovery performance of NCDF, TRF, and Minimap2 on simulated PacBio (upper panel) and Nanopore (lower panel) reads. Solid bars are true positive rates (TPR), crosshatched bars are false discovery rates (FDR). **(B)** Distribution of observed lengths of $(AATGG)_n$ arrays (with and without consensus filtering) observed in PacBio and Nanopore reads. Reads were subsampled to a similar length distribution of 16.5 Gb for both technologies (see Supplementary Note S6). Filtered and unfiltered results for PacBio are very similar.



