Identifying Possible Variable Number of Tandem Repeats (VNTRs) Using Read Mapping Coverage

Noah Fields[1,2], Gary Benson[3,4]

[1]Boston University STaRS Program, Summer 2021, Boston MA, [2]Department of Bioengineering, Stanford University, Palo Alto, CA
[3]Department of Biology, Boston University, Boston, MA, [4]Department of Computer Science, Boston University, Boston, MA

Abstract

Tandem repeats (TR) are sequences of two or more DNA base pairs repeating in such a way that the copies lie adjacent to each other on the chromosome. In humans, tandem repeats are associated with a number of diseases and illnesses including diabetes and cancer. The Benson Lab is interested in detecting genetic variation at these TR loci, because this variation affects gene expression. TRs that vary in the number of repeated copies are called Variable Number of Tandem Repeats (VNTRs), and the Benson Lab has devised VNTRseek, a software program to detect VNTRs in sequencing reads, to find VNTRs. The goal of this project was to look at whether mapped read coverage (average frequency at which sequencing reads map to a location) at TR loci may provide insight into whether a TR locus is a VNTR. TR loci were also separated into size and variant status to see if a pattern arose in read mapping coverage. No statistically significant differences were found relating mapped read coverage to genetic variation at TR loci.

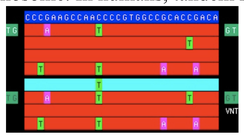

Fig 1. An example of a VNTR from the reference set of TR loci. The reference (dark blue) has four copies (red bars) while the first read (light blue) has 3 copies (red bars). The green on either side is the flanking sequence outside of the TR array.

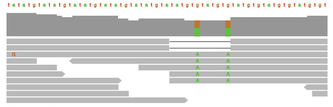

Fig 2. The top row represents the reference genome GRCh38. The dark gray below represents the frequency at which reads map to those nucleotides and the lighter gray segments represent the actual sequencing reads.

Methods

Sequencing reads from sample HG00096 sequenced by the New York Genome Center were aligned to the GRCh38 reference genome using the Burrows-Wheeler Aligner software program (BWA). VNTRseek was given the sequencing reads and a Tandem Repeat reference set to identify variant and non-variant TR loci. We examined alignments of sequencing reads to the reference genome to see how many and where sequencing reads were mapping to TR loci regions. Specifically, we looked at TR Loci that are Singletons and Non-variants. Singletons are TR Loci that do not resemble other regions in DNA. A randomly generated distribution of regions in DNA was created for comparison purposes. Analysis of mapped read coverage was done using Python.In part of our study, sections of sequencing reads that were deemed not similar to GRCh38 were 'clipped.' Analysis on clipped regions in Singleton Non-variant TR Loci was performed in addition to mapped read coverage.

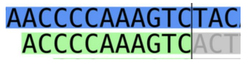

Fig 3. The blue represents the GRCh38 reference genome and the green represents a sequencing read. The read aligns perfectly to the reference genome until the last three nucleotides (gray), which are a clipped region in this read.

Results

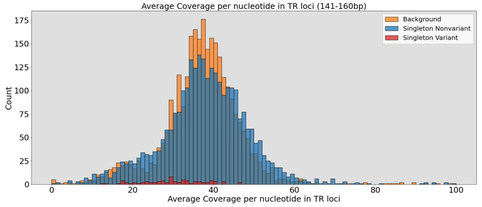

Fig 4. This histogram shows the average mapped read coverage per nucleotide in TR Loci in three distributions. The average mapped read coverage per nucleotide is 38.59 reads/bp for Singleton Non-variant TR Loci, 30.93 reads/bp for Singleton Variant TR Loci, and 38.84 reads/bp for the background distribution.

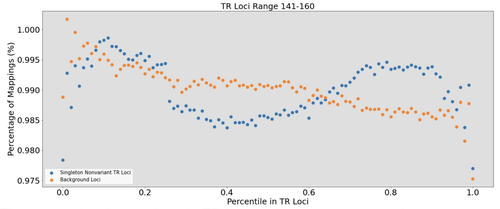

Fig 5. The scatter plot shows where in TR loci reads are most commonly mapping.

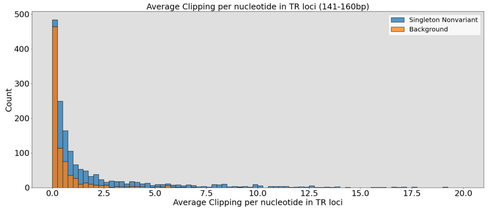

Fig 6. The histogram shows the average clipping per nucleotide in TR Loci in Singleton Non-variant Loci and our background distribution. On average, each nucleotide is clipped from reads 2.73 times in Singleton Non-variant TR Loci, and 0.93 times in the background regions.

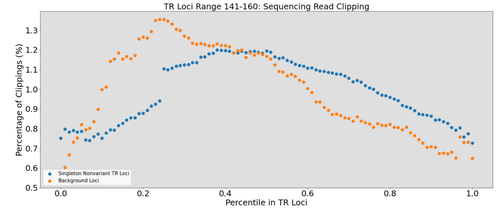

Fig 7. The scatter plot shows where in TR loci reads are most commonly clipped.

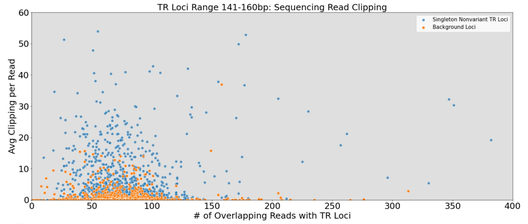

Fig 8. This figure shows the average clipping per read as a function of how many reads mapped to a specific location. Each data point represents a Singleton Non-variant TR Loci or background TR Loci. The average clipping per read in Singleton Non-variant TR Loci is 2.43 nucleotides, while the average clipping per read in background loci is 0.36 nucleotides.

Discussion and Future Directions

- Our results did not show a statistically significant difference in average sequencing read coverage amongst the three TR loci groups: Singleton Non-variant, Singleton Variant, and the background distribution. Additionally there was no significant difference in average clipping per nucleotide in the TR loci amongst Single Non-variant TR loci and our background distribution.
- There was no statistically significant difference in where reads are mapping in the TR Loci versus the background distribution. Additionally, there was no statistically significant difference in where reads were clipping in the TR Loci versus the background distribution.
- The mean of average clipping per read was much higher in Non-variant TR Loci than in our background distribution.

Future work should analyze data similarly and stratify data based on tandem repeat pattern length. This may allow for more statistically significant findings which may be diluted with our current analysis. We also only use one sample for our sequencing reads data set; sequencing read data from more samples could allow for stronger data and analysis in read mapping coverage. Lastly, work should be done to examine the Singleton Non-variant TR Loci which showed a high amount of clipping per read.

References

1. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.
2. Yevgeniy Gelfand, Yozen Hernandez, Joshua Loving, Gary Benson, VNTRseek—a computational too to detect tandem repeat variants in high-throughput sequencing data, *Nucleic Acids Research*, Volume 42, Issue 14, 18 August 2014, Pages 8884–8894.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.
4. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2. doi: 10.1093/bioinformatics/btq033.

This work was funded, in part, by **NHLBI 5R25HL118693**, awarded to the Boston University STaRS Program.