Tackling Indistinguishable TR’s and Verifying the Accuracy of VNTRseek

Rahul Ramesh[1], Marzie Rasekh[2], Yozen Hernandez[2], Gary Benson[2,3]

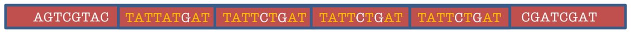
Boston University BRITE REU Program[1], Boston University Graduate Program in Bioinformatics[2],Laboratory for Biocomputing and Informatics[3], Boston University, Boston, MA 02215, USA

Abstract

DNA Tandem Repeat (TR) loci consist of two or more adjacent copies of a pattern sequence of nucleotides in a genome. Variable Number of Tandem Repeats (VNTRs) are polymorphic TRs which have a variable number of copies in the population. VNTRs have been located in genes associated with diseases ranging from depression to Fragile X syndrome so identifying them is beneficial. VNTRseek is a mapping software that aligns, or maps, high-throughput DNA sequencing reads that contains TRs to a genome reference TR set in order to identify mini-satellite VNTRs. “Indistinguishable” TRs is a classification placed on highly similar mini-satellite TRs that occur at multiple loci in a genome. Although high-throughput sequencing techniques have progressed, accurately mapping reads containing these indistinguishable TRs is still difficult because their flanking sequences and patterns are almost identical to multiple loci in the reference. Using a newer form of sequencing data, this project had two aims: 1) to verify the accuracy of VNTRseek mappings of indistinguishable TRs and 2) to identify the correct mappings when errors occurred. Using 10x Genomics linked-read molecular barcode data from the well-studied NA12878 genome, a female Caucasian, we obtained a more precise mapping and clarified the exact loci for indistinguishable TRs. After intersecting VNTRseek’s mapping with the barcode regions, our current results show that VNTRseek maps to the appropriate barcode region 97.7 percent of all reads that contain any TR, not only VNTR’s, and 77 percent of all the Indistinguishable TR reads. Errors in mapping VNTR’s were also observed. In order to understand these misaligned mappings done by VNTRseek, we have also collected allele information (number of pattern copies) to provide another avenue of analysis. We expect that this analysis will lead to refinement of VNTR detection and increased ability to detect genes affected by VNTR copy number variation.

Background and Motivation

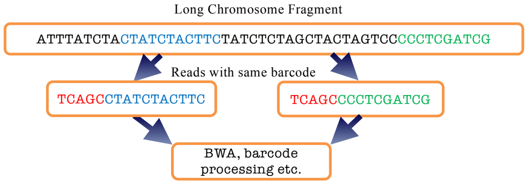
Example of TR:



AGTCGTAC TATTATGAT TATTCTGAT TATTCTGAT TATTCTGAT CGATCGAT

Repeated copies may not be identical

10x pipeline:



Long Chromosome Fragment

ATTTATCTACTATCTACTTCTATCTCTAGCTACTAGTCCCCCTCGATCG

Reads with same barcode

TCAGCCTATCTACTTC

TCAGCCCCTCGATCG

BWA, barcode processing etc.

Reads from the same molecule should map to the same region

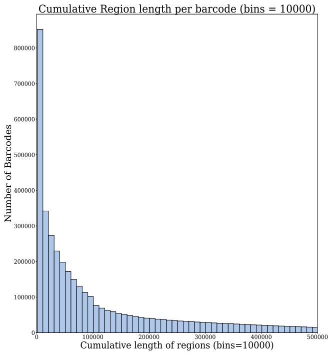
Indistinguishable TR’s

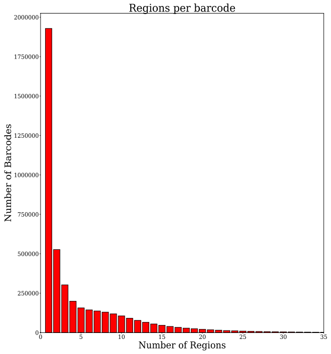
- TR’s that map to more than one place (Indistinguishables) are very abundant in the genome and mapping their locus accurately is important to understanding VNTR’s
- Using 10x data can provide us a better mapping in order to find VNTR’s

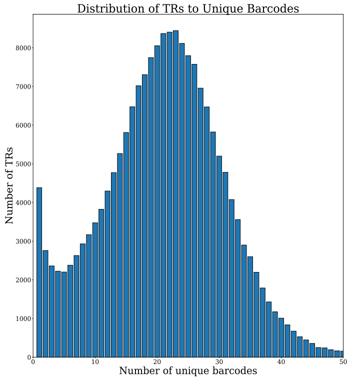
Methods

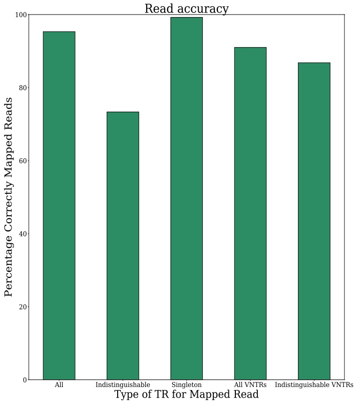
- Ran data from NA12878 through VNTRseek tool to map VNTR’s to reference genome
- Identified regions in the reference genome based on the barcode
- Intersected read coordinates from VNTRseek with barcode regions
- Correct and Incorrect reads counted and grouped by TR and allele
- For Indistinguishables analysis:
- Compared VNTRseek indistinguishable identified TR’s with their location and size and various other info for analysis.

Results





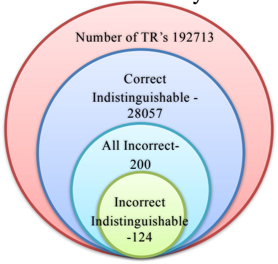




TR Data Table

Class	TR’S With	Total	Count w/valid allele	Count w/invalid allele	0/0	0/1	1/1	1/2
All	At least 1 correctly mapped read	192513	188796	3717	185784	1565	1259	87
All	At least one incorrectly mapped read	22023	21870	153	21155	435	180	22
All	Only incorrectly mapped reads	200	62	138	45	1	16	0
Indistinguishable	Only incorrectly mapped reads	124	46	78	34	1	11	0
Indistinguishable	At least one incorrectly mapped read	4819	4683	136	4420	201	54	8
Indistinguishable	At least 1 correctly mapped read	28057	27348	709	26532	520	269	27

TR Accuracy


Number of TR’s 192713
Correct Indistinguishable - 28057
All Incorrect- 200
Incorrect Indistinguishable -124

- Almost all of the TR’s had at least one correctly mapped read
- 21870 TR’s had both correctly and incorrectly mapped reads
- VNTRseek had an 98 percent accuracy in all its mapping when compared to the barcodes provided by 10x genomics. Reads mapped to VNTRs had 96 percent accuracy. Reads mapped to Indistinguishable TR’s had 86 percent accuracy.

Conclusion

- VNTRseek performed very well on its mapping of reads to VNTR’s and indistinguishable TR’s.
- Mapping of reads to Indistinguishable TR’s is better than anticipated previously

References and Acknowledgments

Yevgeniy Gelfand, Yozen Hernandez, Joshua Loving, Gary Benson; VNTRseek—a computational tool to detect tandem repeat variants in high-throughput sequencing data, *Nucleic Acids Research*, Volume 42, Issue 14, 18 August 2014, Pages 8884–8894

Zheng GX, Lau BT, Schnall-Levin M, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol.* 2016;34(3):303-11.

This work was funded, in part, by NSF grant DBI-1559829, awarded to the Boston University Bioinformatics BRITE REU program, and NSF grant IIS-1423022.