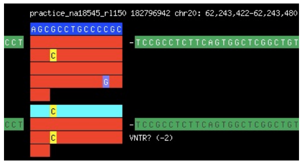
Testing the accuracy and speed of VNTRseek, a genetic variation detector, using a restricted read dataset

Sara Filler[1,2], Gary Benson, Ph.D.[3,4]

[1]Providence College, [2]Boston University BRITE REU Program, Summer 2021,
[3]Department of Biology, Boston University, [4]Department of Computer Science, Boston University

Abstract



```
practice_na18545_rl150 182796942 chr20: 62,243,422-62,243,480
AGCGCCTGCCCCGC
CCT  -TCCGCCTCTTCAGTGGCTCGGCTGT
C
G
C
CCT  -TCCGCCTCTTCAGTGGCTCGGCTGT
C    VNTR? (-2)
```

Figure 1. TR from the reference set with a mapping read pictured below that is a possible VNTR.

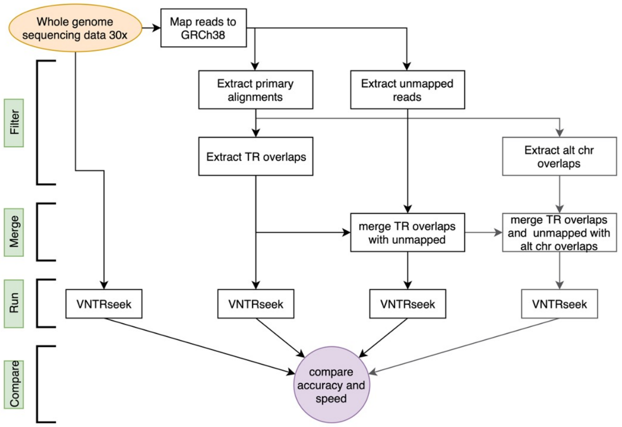
DNA Tandem Repeats (TRs) are common genomic features where a segment of the genomic code is repeated as two or more adjacent copies. Variable number of tandem repeats (VNTRs) are TR loci where the copy number varies in the population. VNTRs have been linked to neurodegenerative disorders and cancers, exhibit population-specific alleles, and can drive phenotypic change by affecting gene expression.[1]

Thus, VNTR genotyping may be helpful with disease prognosis and predicting an individual's drug response. To accelerate further studies of VNTRs and their functional importance, existing tools need to be optimized. VNTRseek is a program developed by the Benson Lab to detect genetic variation at TR loci[2]. The first step of the program processes sequencing reads to identify TRs. This step is the slowest, as the program usually uses whole genome sequencing (WGS) data consisting of hundreds of millions of reads. The goal of this project was to filter the reads in the input dataset in order to increase the speed of the first step while retaining VNTR detection sensitivity. Reads were filtered to retain only those which overlapped with known TR loci. Variations of this dataset included unmapped reads and reads mapping to alternate chromosomes. While further validation is needed, our methods successfully increased VNTRseek efficiency while maintaining acceptable VNTR detection sensitivity.

Major Questions

- If input dataset to VNTRseek is filtered to only reads mapping to TR loci…
  - Will all of the original VNTRs still be detected?
  - Will computation time be reduced?
  - Will adding unmapped reads or reads aligned to alternate chromosomes to the filtered dataset make a difference?

Methods


Whole genome sequencing data 30x
Map reads to GRCh38
Extract primary alignments
Extract unmapped reads
Extract TR overlaps
Extract alt chr overlaps
merge TR overlaps with unmapped
merge TR overlaps and unmapped with alt chr overlaps
VNTRseek
VNTRseek
VNTRseek
VNTRseek
compare accuracy and speed
Filter
Merge
Run
Compare

- Mapped WGS data at >30x coverage for two sample genomes to 2 different assemblies of reference genome GRCh38 using BWA-MEM[3,4]
- Tools used: BWA-MEM, SAMtools, BEDTools, VNTRseek
  - Ran each version of restricted read dataset 3 times through VNTRseek for runtime data
- Analyzed VNTRseek output files containing variant information (VCF files) via program we developed to find all similarities/differences between TRs and VNTRs detected

Results

Figure 2. Percentage of reads used for each filtered dataset as compared to total reads in the unfiltered dataset.

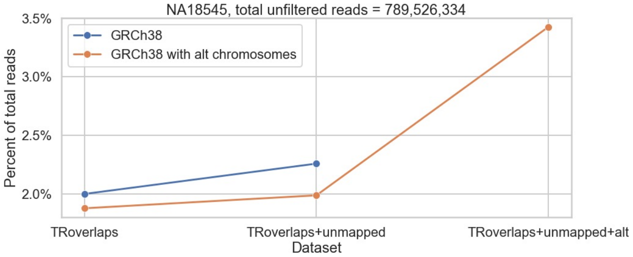
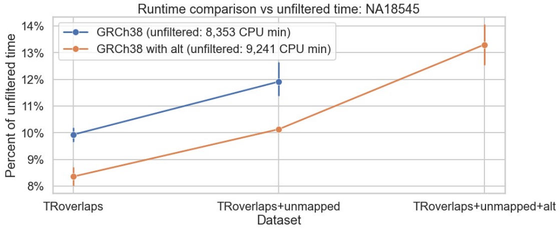
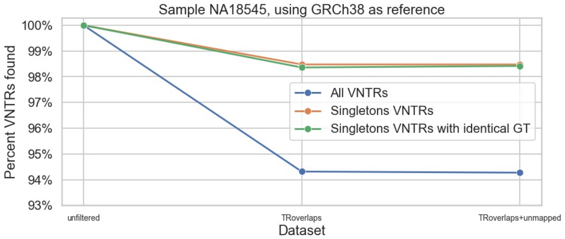
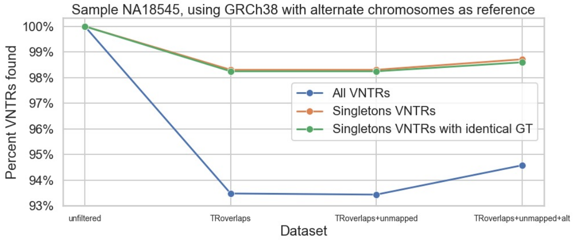


Figure 3. Average percentage of runtime for each variation of read filtering method as compared to the unfiltered dataset.



Figures 4, 5. Percentage of VNTRs found for all datasets used in each mapping reference genome. Out of all VNTRs, singletons are those which are high-confidence because their sequences are not similar to any other TR loci. Most singletons found had identical genotypes (GT).





Conclusion

- Results suggest read filtering is promising, increased program efficiency by at least 85% while retaining sensitivity of greater than 93% for all VNTRs and 98% for singleton VNTRs
- Regardless of mapping reference genome used, detection of singleton VNTRs remains nearly the same
- Using the GRCh38 assembly with alternate chromosomes would be preferable in future work
  - including reads mapping to alternate chromosomes in the filtered dataset yielded a slight increase in VNTR detection
  - Reads from more than 2600 sample genomes have already been mapped to this assembly by the 1000 Genomes Project Consortium and are available as CRAM files
- Future work should test read filtering on more datasets and continue to analyze which reference genome features and mapping methods should be used to further filter input datasets to detect VNTRs most efficiently
- Read filtering methods should be streamlined into a script to minimize the time taken to prepare the restricted read dataset and ensure reproducibility

References

1. Marzieh Eslami Rasekh, Yözen Hernández, Samantha D Drinan, Juan I Fuxman Bass, Gary Benson (2021). Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences. *Nucleic Acids Research*, Volume 49, 4308–4324
2. Yevgeniy Gelfand, Yozen Hernandez, Joshua Loving, Gary Benson (2014). VNTRseek—a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Research*, Volume 42, 8884–8894
3. Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P., & 1000 Genomes Project Consortium (2017). Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience*, *6*(7), 1–8.
4. The following datasets were obtained through the European Nucleotide Archive, as part of the 1000 Genomes Project Phase 3, run accessions:
   https://www.ebi.ac.uk/ena/browser/view/ERR3239357 (NA18545)
   https://www.ebi.ac.uk/ena/browser/view/ERR3240114 (HG00096)

This work was funded, in part, by NSF grant DBI-1949968, awarded to the Boston University Bioinformatics BRITE REU program.