Finding *de novo* tandem repeats using VNTRseek

Gisselle Hidalgo[1,2], Gary Benson, Ph.D.[3,4]

[1]Kean University, Union, NJ, [2]Boston University Bioinformatics BRITE REU Program, Summer 2022, Boston, MA
[3]Department of Biology, Boston University, Boston, MA, [4]Department of Computer Science, Boston University, Boston, MA

Abstract

VNTRseek is a computational tool developed by the Benson Lab to identify sequencing reads containing tandem repeats (TRs) and map them to a set of reference TRs. This project aimed to establish a new computational pipeline to identify potential *de novo* TRs. It comprises Python programs and software tools for processing sequencing data, including SAMtools, BEDtools, and SQLite. With the use of VNTRseek output files as input for my pipeline, these candidate *de novo* TR reads are extracted and aligned to the reference genome and reference set to locate any TR sequences. BEDtools was utilized to determine which reads did not overlap with the reference TRs. The start and end of these reads with our target size TRs were used to locate those sequences in the reference genome. These location ranges were then processed through Tandem Repeat Finder, a TR finder program created by the Benson Lab, to determine whether a TR was found at those locations. Identifying *de novo* TRs can help determine which genomic regions have the propensity for variability and provide a more complete compendium of variable TRs in the human genome.

Introduction and Background

- Tandem repeats are sequences of two or more DNA bases repeated identically or approximately on a chromosome.
  - Some TR unstable expansions are known to cause genetic diseases such as Huntington's Disease and fragile x syndrome.
- *De novo* TRs are, in this case, where the reference sequence does not indicate a TR.
  - My pipeline will extract these *de novo* TRs and I will perform an alignment to location to determine if there are known TRs at candidate locations.
- VNTRseek is a computational tool designed by the Benson Lab to find tandem repeats (TRs) and map them to a reference set of TRs.
- Variable number tandem repeats (VNTRs) are TR loci that vary in copy number among individuals (shown in Figure 1).

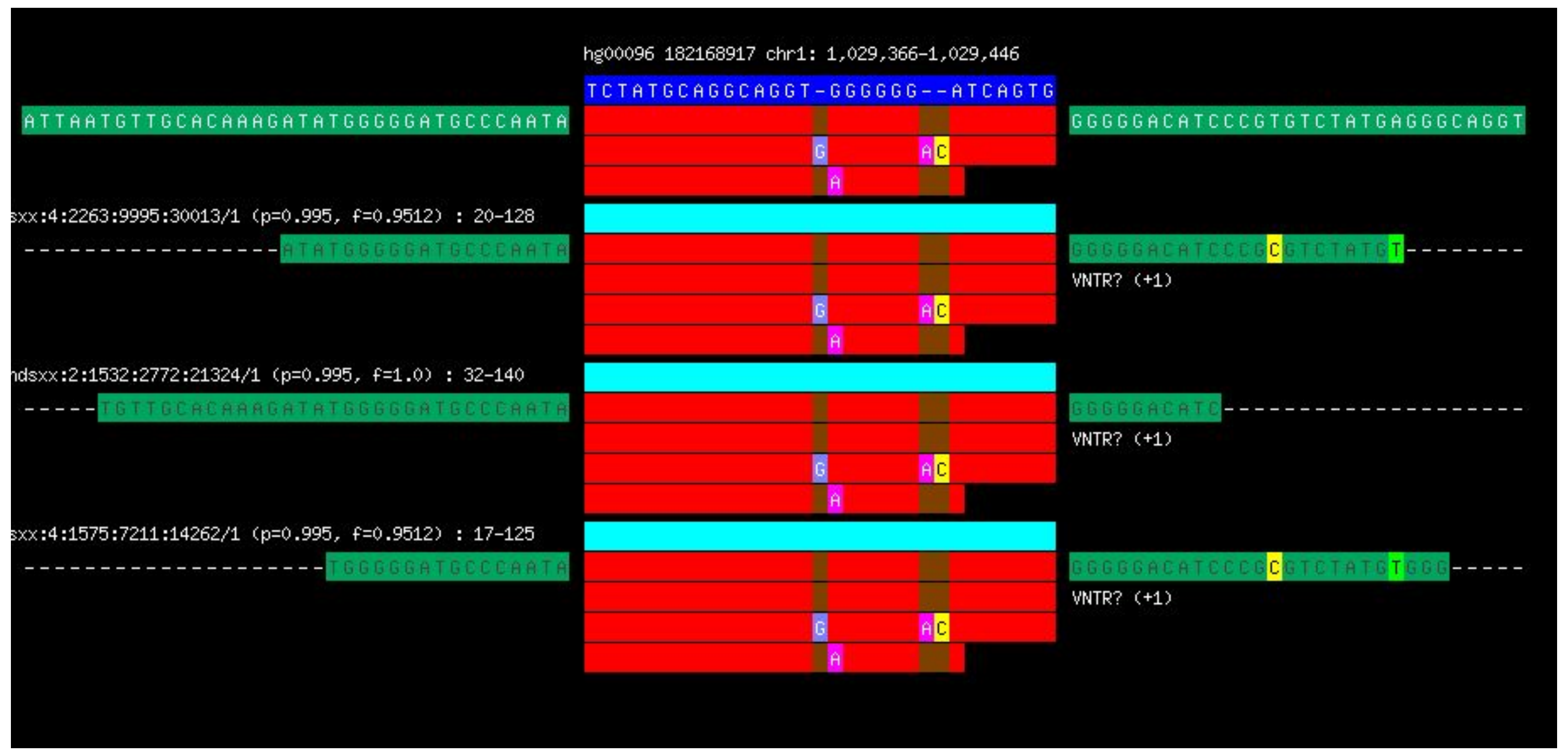
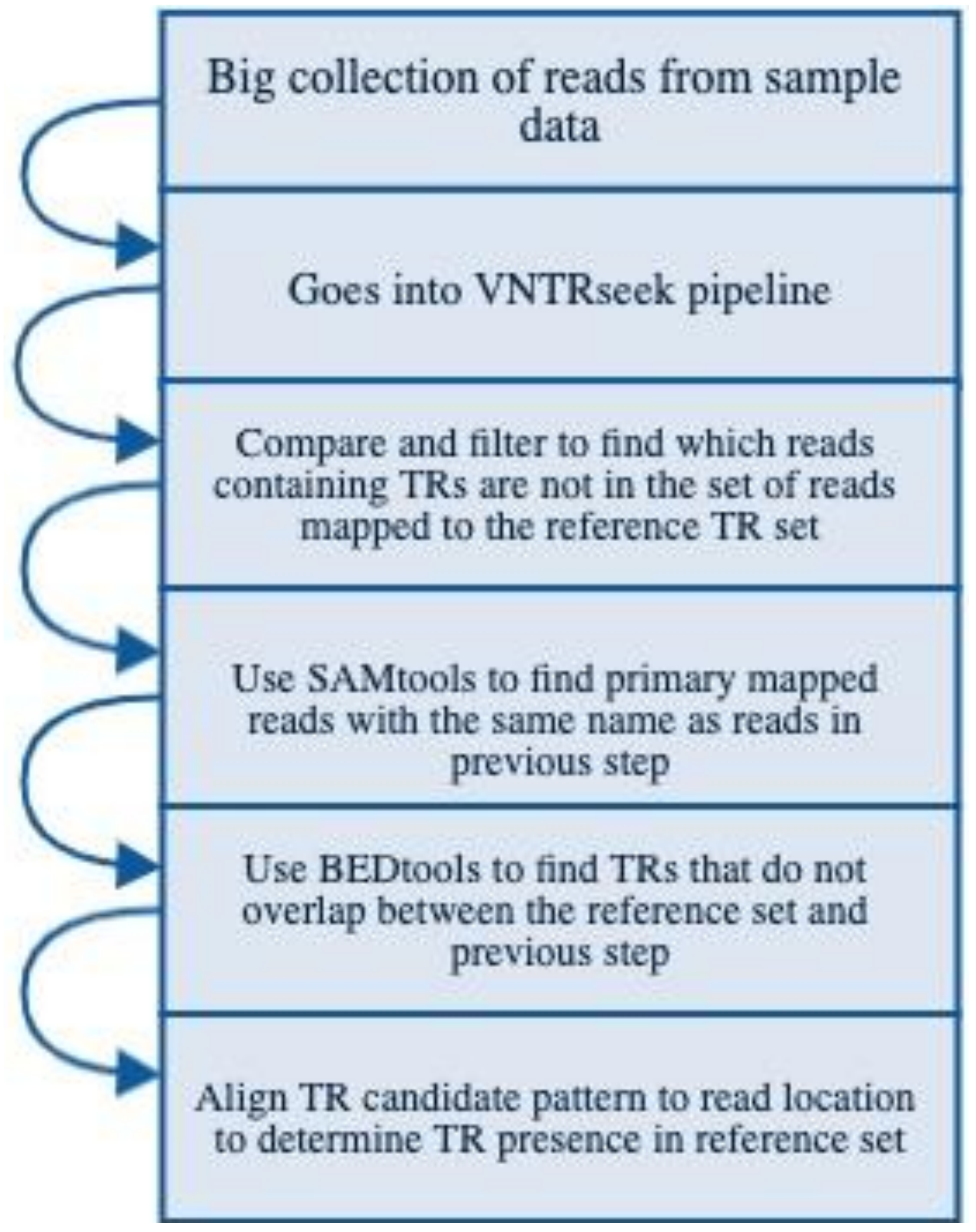

hg00096 182168917 chr1: 1,029,366-1,029,446
TCTATGCAGGCAGGT-GGGGGG--ATCAGTG
ATTAATGTTGCACAAAGATATGGGGGATGCCCAATA
GGGGGACATCCCGTGTCTATGAGGGCAGGT
G
AC
A
sxx:4:2263:9995:30013/1 (p=0.995, f=0.9512) : 20-128
ATATGGGGGATGCCCAATA
GGGGGACATCCCGCGTCTATGT
VNTR? (+1)
ndsxx:2:1532:2772:21324/1 (p=0.995, f=1.0) : 32-140
TGTTGCACAAAGATATGGGGGATGCCCAATA
GGGGGACATC
VNTR? (+1)
sxx:4:1575:7211:14262/1 (p=0.995, f=0.9512) : 17-125
TGGGGGATGCCCAATA
GGGGGACATCCCGCGTCTATGTGGG
VNTR? (+1)

Figure 1 (above): Output from VNTRseek showing VNTRs in an individual.

Methods

On the right is a figure showing a simplified version of the pipeline created for this project.

- Tools used:
  - BEDtools, SAMtools, VNTRseek, SQLite
- Files types used:
  - SAM
    - Text file format containing alignment information of sequences mapped against reference sequences
  - BAM
    - Binary versions of SAM files
  - BED
    - Tab-delimited files describing features by chromosome, start, end, name, score and strand.
    - DB
      - database files containing fasta reads.
  - Files produced from previous step used as input for the next step


Big collection of reads from sample data

Goes into VNTRseek pipeline

Compare and filter to find which reads containing TRs are not in the set of reads mapped to the reference TR set

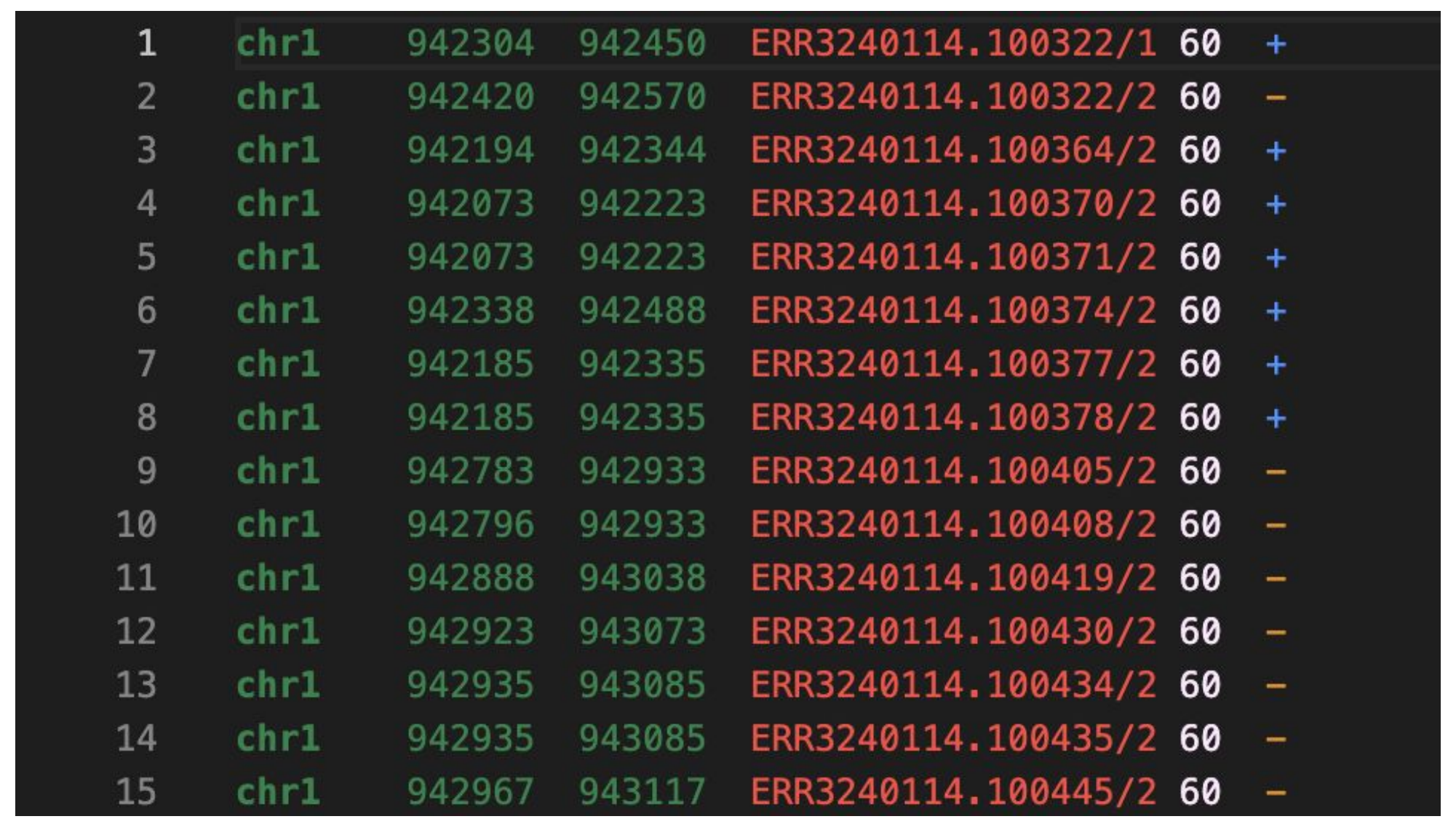
Use SAMtools to find primary mapped reads with the same name as reads in previous step

Use BEDtools to find TRs that do not overlap between the reference set and previous step

Align TR candidate pattern to read location to determine TR presence in reference set

Figure 2 (above): Simplified pipeline steps

Results


```
chr1    942304  942450  ERR3240114.100322/1 60  +
chr1    942420  942570  ERR3240114.100322/2 60  –
chr1    942194  942344  ERR3240114.100364/2 60  +
chr1    942073  942223  ERR3240114.100370/2 60  +
chr1    942073  942223  ERR3240114.100371/2 60  +
chr1    942338  942488  ERR3240114.100374/2 60  +
chr1    942185  942335  ERR3240114.100377/2 60  +
chr1    942185  942335  ERR3240114.100378/2 60  +
chr1    942783  942933  ERR3240114.100405/2 60  –
chr1    942796  942933  ERR3240114.100408/2 60  –
chr1    942888  943038  ERR3240114.100419/2 60  –
chr1    942923  943073  ERR3240114.100430/2 60  –
chr1    942935  943085  ERR3240114.100434/2 60  –
chr1    942935  943085  ERR3240114.100435/2 60  –
chr1    942967  943117  ERR3240114.100445/2 60  –
```

Preliminary results show over 3,100 candidate *de novo* tandem repeats in bed file format. We focused on chromosome 1 to narrow the results down.

Figure 3 shows the chromosome 1 candidate reads with start and end positions for the repeats.

Figure 3 (above): Chromosome 1 candidate *de novo* reads.

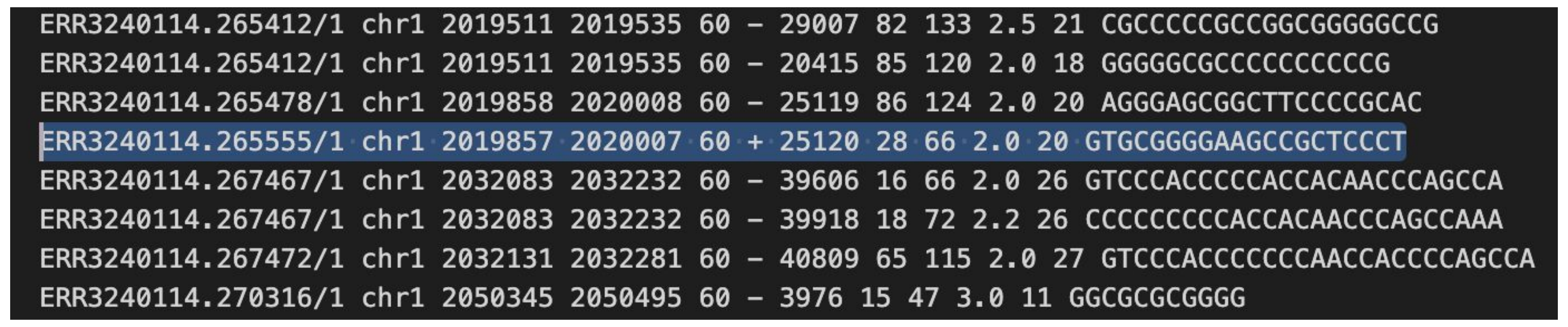
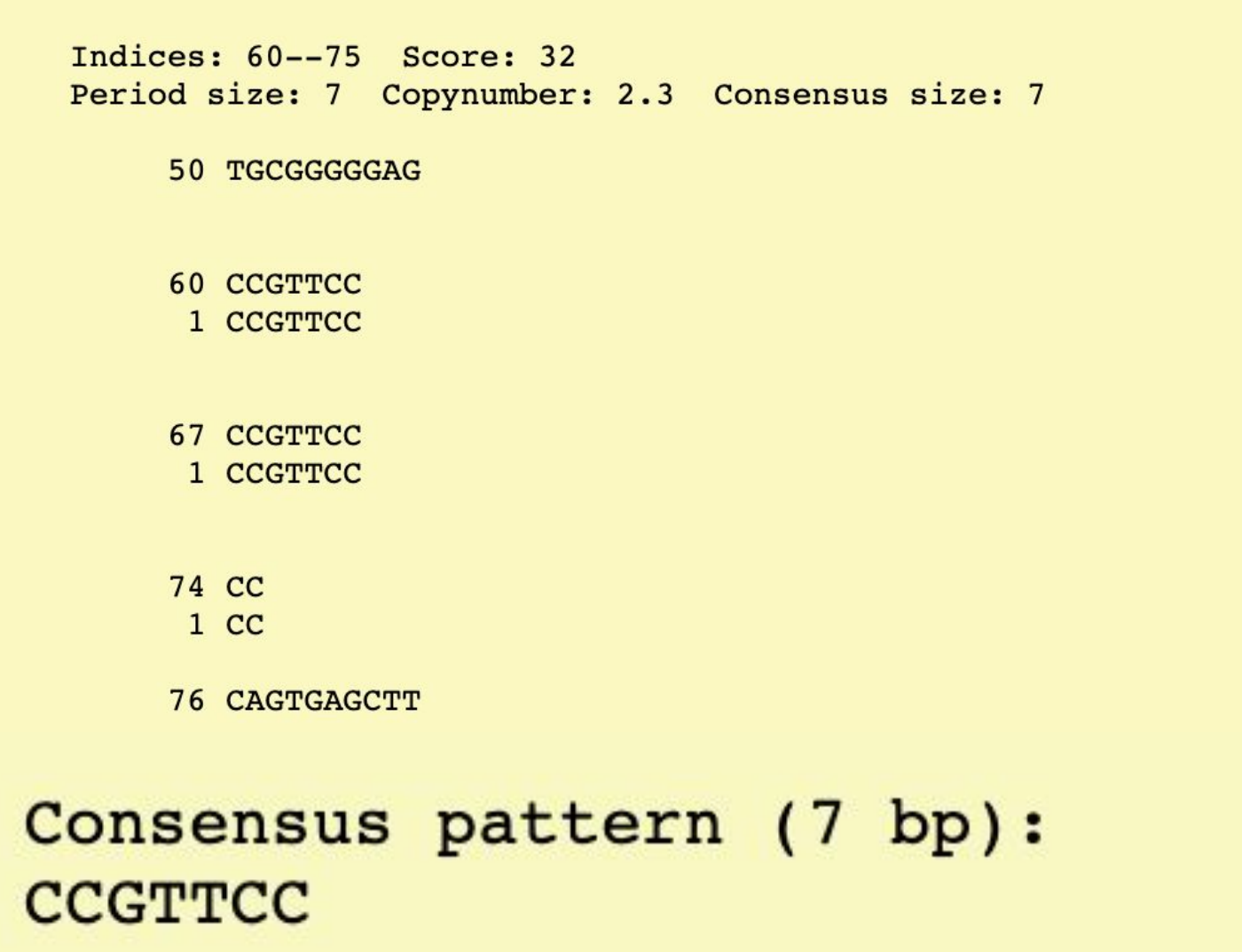

```
ERR3240114.265412/1 chr1 2019511 2019535 60 – 29007 82 133 2.5 21 CGCCCCCGCCGGCGGGGGCCG
ERR3240114.265412/1 chr1 2019511 2019535 60 – 20415 85 120 2.0 18 GGGGGCGCCCCCCCCCCG
ERR3240114.265478/1 chr1 2019858 2020008 60 – 25119 86 124 2.0 20 AGGGAGCGGCTTCCCCGCAC
ERR3240114.265555/1 chr1 2019857 2020007 60 + 25120 28 66 2.0 20 GTGCGGGGAAGCCGCTCCCT
ERR3240114.267467/1 chr1 2032083 2032232 60 – 39606 16 66 2.0 26 GTCCCACCCCCACCACAACCCAGCCA
ERR3240114.267467/1 chr1 2032083 2032232 60 – 39918 18 72 2.2 26 CCCCCCCCCACCACAACCCAGCCAAA
ERR3240114.267472/1 chr1 2032131 2032281 60 – 40809 65 115 2.0 27 GTCCCACCCCCCCAACCACCCCAGCCA
ERR3240114.270316/1 chr1 2050345 2050495 60 – 3976 15 47 3.0 11 GGCGCGCGGGG
```

Figure 4 (above): Highlighted candidate read of interest with corresponding information.

We focused on specific reads with pattern sizes equal to or greater than 20 bp, shown above in Figure 4. Those start and end positions were extracted from the reference genome and run through Tandem Repeats Finder. From the read of interest, it was seen that there was a consensus pattern of CCGTTCC which partially matches the pattern highlighted in Figure 4. This read shows a TR in the reference genome that was not found in the Tandem Repeats Database.


```
Indices: 60--75  Score: 32
Period size: 7  Copynumber: 2.3  Consensus size: 7

      50 TGCGGGGAG


      60 CCGTTCC
       1 CCGTTCC


      67 CCGTTCC
       1 CCGTTCC


      74 CC
       1 CC

      76 CAGTGAGCTT

Consensus pattern (7 bp):
CCGTTCC
```

Figure 5 (above): Read of interest with consensus pattern and number of repeats found.

Conclusions

- The pipeline was successful in filtering which reads were not present in the reference set of TRs.
  - Allowed me to determine which reads could possibly be *de novo*.
  - Had preliminary results of 3,100 reads.
- Found at least one read from the candidate reads bed file and performed an alignment to location to determine if a TR was present at that location.
- It was seen there was a partial match to the original pattern size.
- Many reads chosen for analysis contained TRs not found within the read start and end positions in the reference.
- It is plausible that there are more reads with unmapped TRs in this file in the other chromosomes.
- Further alignment to location will be done to analyze the rest of the candidate reads.

Future Work

- I would like to look at chromosomes where genetic diseases are commonly found (e.g. X chromosome).
- I would also like to see if this program works with other species, more specifically, *Drosophila*.

References

M. Rasekh, Y. Hernandez, S. Drinan, J. Fuxman-Bass, and G. Benson (2021). Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences. *Nucleic Acids Research*, 2021, Vol. 49, No. 8, 4308-4324.

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*, Vol. 42, No. 14, 8884–8894.

G. Benson,"Tandem repeats finder: a program to analyze DNA sequences" *Nucleic Acids Research* (1999) Vol. 27, No. 2, pp. 573-580.

Yevgeniy Gelfand, Alfredo Rodriguez and Gary Benson TRDB—The Tandem Repeats Database. (2007) *Nucleic Acids Research*, 35:D80-D87.

Acknowledgement

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