# Compressing DNA Read Sequences to Reduce Database Size


Julia Schulte[1,2] and Gary Benson[2,3,4]

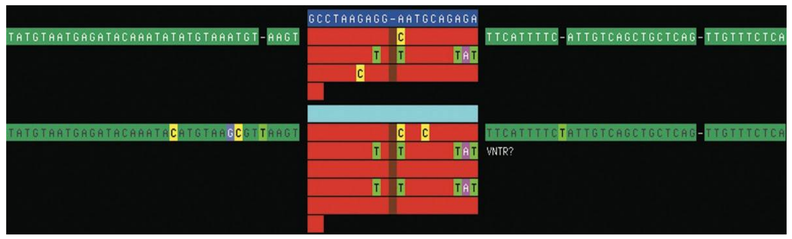
[1]Luther College, [2]Bioinformatics BRITE REU Program, Summer 2022, Boston, MA

[3]Department of Biology, Boston University, Boston, MA, [4]Department of Computer Science, Boston University, Boston, MA


## Abstract


A tandem repeat (TR) is two or more adjacent copies of a nucleotide pattern sequence. Variable number tandem repeats (VNTRs) are TRs that vary in the number of pattern copies among individuals in a population. They are known to affect gene expression, disease states, and chromatin structure. The Benson lab has developed the computational tool VNTRseek to detect tandem repeat variants in high-throughput sequencing data. The output of this program is a large database with information on where the sequencing reads align with a reference sequence. With aspirations of continuing to use this software on thousands of samples there is a storage dilemma. The goal of this project was to reduce the size of the read data stored in our database through two methods. One aims to convert triplet DNA letters into single letters from the American Standard Code for Information Interchange (ASCII) alphabet, which reduces sequence size by two thirds. The second method encodes only the differences between the read and reference by realigning the two sequences and only noting discrepancies. Both methods can effectively reduce database size, the second method having an even greater compression ability than encoding by triplets. Next steps would be implementing this pipeline into the larger VNTRseek tool so that initial data output is already in a reduced form.


## Background and Motivation



- TR's are two or more adjacent copies of a nucleotide pattern sequence. When the number of copies varies among individuals, they are called Variable Number Tandem Repeats
- Benson lab developed the software tool VNTRseek to detect VNTRs which outputs a database containing read sequences that have TRs inside and where they align to the TR reference set
- The size of a single database output by this program often exceeds 3Gb which is far too large for practical purposes
- The reads in this database have already been aligned to their reference sequences and can easily be realigned which would allow storage of only the reference sequence and where the read differs

ACGTTGGCGTGTCAGGTAACGCTATCAGGTAACGCTATCAGGTAACGCTATCAG-GTTATCG

6SG30IA46SCTATCA43D57SAT

- Considering the widespread reach and application of VNTR detection, this tool is going to continue to be frequently used making it important to reduce its storage impact

## Methods

Two primary methods of database reduction:

ACGTTGGCGTGTCAGGTAACGCTATGTTTG

```
Code['ACG'] = '!'
Code['CAG'] = '%'
Code['CTA'] = '&'
Code['GCG'] = '#'
Code['GTA'] = '^'
Code['TGT'] = '$'
Code['TTG'] = '@'
```

DNA Triplets

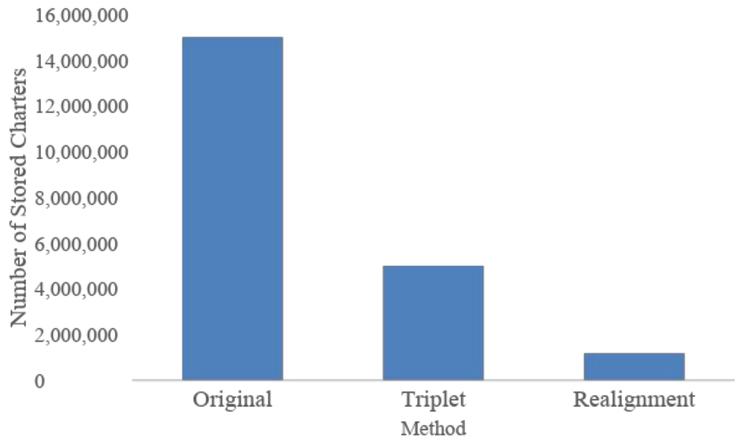
= !@#$%^&$@

Encoding Alignments:

```
     6                    30                    46      43   57
     *                    *                     ** * *  *    **
ACGTTGGCGTGTCAGGTAACGCTATCAGGTAACGCTATCAGGTAACGCTATCAG--GTTATCG READ
ACGTTGACGTGTCAGGTAACGCTATCAGGT--ACGCTATCAGGTAACGTAACCCGTGTTTACG REFERENCE
```

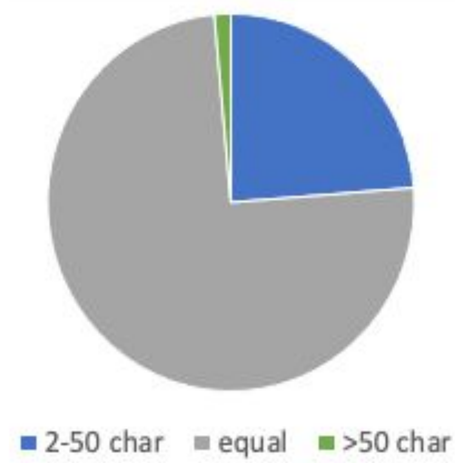
6SG30IA46SCTATCA43D57SAT

## Results




- Results are based on a sample of 100,000 reads
- Average length of read strings: Original – 150char, Triplet – 50char, Realignment – 11.8 char
- Triplet method used 3 times less space then the original and realignment method used 12 times less space

## Results




- ~75% of all reads were the exact same as their reference sequences
- ~1% of samples by realignment are bigger than they would be if stored using the triplet method

## Conclusion

- Storing only where the read and references differ effectively reduces database size making this data more accessible

## Future Work

- Implementation of the data into a database structure using SQL needs to take place in order for this data to be accessible and useable
- Further compression may be able to take place, one example is for the 1% larger than 50 characters using realignment, triplet method could be performed instead
- Reducing runtime is also a priority. The code takes around 2 hours to process 100,000 samples which is too slow when it will be used on millions of samples

## References and Acknowledgements


Gelfand, Y., Hernandez, Y., Loving, J., Benson, G. (2014) VNTRseek—a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Research, Vol. 42, No.14,* pg. 8884-8894.

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

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



Acknowledgement:

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



PhDPosters.com
Your Research Poster, Printed For Less