Single Copy VNTR Identification

Katherine Duchinski[1,2], Marzie Rasekh[3], Gary Benson[3]

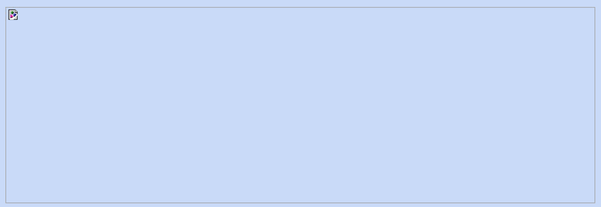
[1]College of Charleston, [2]Boston University Bioinformatics BRITE REU Program, Summer 2018, [3]Boston University Graduate Program in Bioinformatics

Abstract

Tandem repeat DNA sequences (TRs) consist of two or more adjacent copies of a nucleotide pattern. Variable number of tandem repeats (VNTRs) are TRs in which the copy number may vary among individuals. VNTRs are useful as genomic markers and have been associated with several genetic disorders and neurological diseases[1-4]. Repeat detection depends on the presence of multiple copies of the nucleotide pattern. We hypothesized that some individuals may have only one copy of the pattern, but due to limitations of current software, this has not, to our knowledge, been observed. Through novel command line software, we identified candidate single copy VNTRs in chromosome 21.

Background and Objective

VNTRs have been discovered in every chromosome of the human genome and have been observed to increase and decrease copy number across generations[5]. However, little is known about how VNTRs propagate or whether they can decrease to fewer than two copies of the pattern sequence, as current software cannot detect < 2 pattern copies. Copy loss may explain apparently heterozygous loci where one allele seems to be "missing." The goal of this project was to test the hypothesis that individuals may have less than two copies of a VNTR pattern sequence.



Methods

We created a Python command line program, VNTReduce, to reduce the number of pattern copies in a reference chromosome below VNTRseek's detection threshold[5]. The program outputs a reduced-TR copy of the reference chromosome and an updated table of the TR region indices. Using this program we produced reduced-TR versions of each chromosome in HG38. We used BWA to align a Chinese individual's genome (HG005-NA24631) to HG38 with reduced chromosome 21[6]. We filtered the alignment map for reads that spanned a TR region, including 20 nucleotides of left and right flank but excluding those with clipped sequences or edit distances larger than pattern size[7-8]. We considered these reads candidates for evidence of VNTR copy loss.

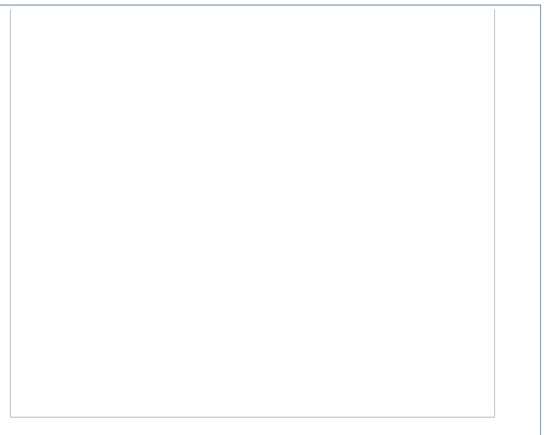


Figure 2
Diagram of the computational workflow associated with this study.

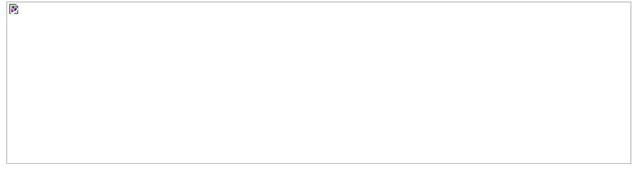


Figure 3
Alignment of a read with one pattern copy to a reference with multiple copies vs. a reference with one copy.

Results and Discussion

We generated a TR-reduced version of the human reference genome and a table of updated TR indices in the reduced genome. We represented repeat regions in lowercase letters in order to qualitatively evaluate correctness of the algorithm. VNTReduce calculates the number of nucleotides that the algorithm should delete and compares this to the size difference in the input and output files. Because these figures agreed for each chromosome, we can quantitatively conclude that the program ran as expected.

We discovered several overlapping TR regions. In these, some nucleotides were considered, simultaneously, the end of one pattern and the beginning of another, distinct pattern. In all such cases (86 in chromosome 21), both TRs were entirely excised from the new reference genome.

We identified at least one single-copy candidate read for the majority of VNTRs in chromosome 21 (Figure 4).We observed that extra pattern copies may appear, for instance, as deletions in the flank (Figure 5), so further read filtering is necessary.

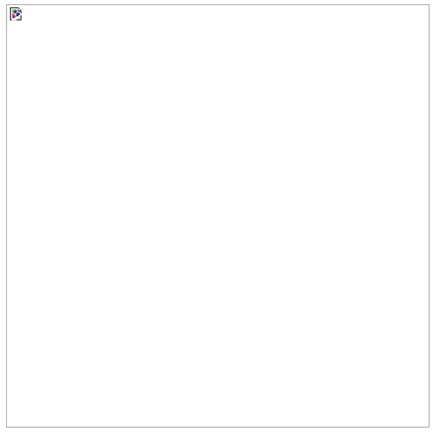


Figure 4
Nested Venn diagram of TRs in chromosome 21 with potential single copies. The reads were filtered to include only those with flanking sequences and without clipped sequences or large edit distances.

* = imperfection in alignment
\- = gap in alignment (deletion)

```
reference flank       *******     *  ****  *  reference single copy        *
GAGCTGGTGACACATCAGATGAAA-T-AGATGAAATAGACAGAGG
GAGCTGGTG-------AGA-G---GTCAGATGAAATAGAC-GAGG
read copy 1                                   read copy 2
```

Figure 5
Alignment result for a reduced TR reference (top string) and a candidate read (bottom string). This is not a true single copy--the aligner is reading an extra copy as deletions in the left flanking sequence.

Conclusions and Future Work

- Sample reads were aligned to a large proportion of copy-reduced TRs. Further investigation will potentially identify single pattern copies VNTRs in the sample genome, which will expand our knowledge of VNTR behavior.
- Loci of candidate single copy VNTRs will be compared with loci of anomalous "missing" alleles.
- The VNTReduce command line program and in-progress filtering script may be used with any reference genome to find single pattern copy candidates in a sample genome dataset.

References and Acknowledgements

[1]Huntington's disease collaborative research group. (1993). A novel gene containing a trinucleotide repeat…. Cell, 72, 971–983.
[2]Lasky-Su, J.A., Faraone,S.V., Glatt,S.J. and Tsuang,M.T. (2005). Meta-analysis of the association…. Am. J. Med. Genet. B Neuropsychiatr. Genet., 133B, 110–115.
[3]Lesch,K.P., Bengel,D., Heils,A., … Murphy,D.L. (1996) Association of anxiety-related traits with a polymorphism…. Science, 274, 1527–1531.
[4]Verkerk, A., Pieretti, M., Sutcliffe, J., et al. (1991) Identification of a gene (FMR-1)… exhibiting length variation in fragile X syndrome. Cell, 65, 905–914.
[5]Gelfand, Y., Hernandez, Y., Loving, J., & Benson, G. (2014). VNTRseek — a computational tool…. Nucleic Acids Research, *42*(14), 8884–8894.
[6]Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *00*(00), 1–3.
[7]Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics.
[8]Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics.
—,

This research was conducted during the Boston University BRITE Research Experience for Undergraduates 2018 and supported by NSF grants DBI-1559829 and IIS-1423022.