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\(^1\). 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, VNTRReduce, to reduce the number of pattern copies in a reference chromosome below VNTRseek’s detection threshold\(^1\). 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 alignment to include only those with flanking sequences and without clipped sequences or large edit distances. We generated a TR-reduced version of the human reference genome and a table of updated TR indices in the reduced genome.

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.

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 VNTRReduce command line program and in-progress filtering script may be used with any reference genome to find single pattern copy candidate reads in a sample genome dataset.

References and Acknowledgements