





Investigating Variable Number Tandem Repeats (VNTRs) Overlapping Coding Regions in the Genome

Final Report

Maha Alsuhaymi
alsuhaimi.maha@gmail.com

Mentor: Lisa Strug, Delnaz Roshandel
lisa.strug@utoronto.ca, delnaz.roshandel@sickkids.ca

August 27, 2025

Abstract

Variable Number Tandem Repeats (VNTRs) are a short sequence of nucleotides, with a length of 6 base pairs or more, repeated in tandem across the genome. They are a significant source of human genetic variation. Their repetitive structure and long length make them difficult to genotype using short-read sequencing. As a result, their contributions to various diseases, such as cystic fibrosis (CF), remain poorly understood. CF is an autosomal recessive disease caused by loss-of-function variation in transmembrane conductance regulator gene (*CFTR*) affecting multiple organs including the lungs. CF severity varies in individuals with the same CF-causing *CFTR* genotype, due in part to modifier genes. Genome-wide association studies (GWAS) of CF lung disease have identified a locus at the beginning of chromosome 5 containing approximately 100 VNTRs. The majority of these VNTRs are in linkage disequilibrium with GWAS single nucleotide polymorphisms (SNPs), suggesting they may be the variants driving GWAS signals. In particular, VNTRs overlapping coding regions i.e. exons or exon-intron boundaries may affect gene protein products. Here we aim to identify VNTRs intersecting protein coding regions of the genome with a specific focus on CF lung disease GWAS locus on chromosome 5 to identify the VNTRs that can change protein products of the genes and particularly could be responsible for CF lung disease GWAS signal on chromosome 5. VNTR boundaries will be defined using the adotto repeat catalog v.0.0.2 (https://zenodo.org/records/7987365). Coding gene locations and structures i.e. exons and introns will be mapped using the MANE release 1.0 (Matched Annotation from NCBI and EMBL-EBI) database. Computational analyses will identify VNTRs overlapping coding genes and their exons, as well as exon-intron boundaries. For VNTRs that overlap coding genes, we will investigate their potential effects on splice sites and assess whether they could disrupt protein sequences, for example by damaging canonical splice sites or causing frameshifts. The results will identify deleterious VNTRs that can damage protein products of the genes. They will also help evaluate whether VNTRs within the CF lung disease GWAS locus on chromosome 5 could be responsible for the observed GWAS signals.

Contents

Abstract **1**

1 Introduction **3**
1.1 Key Assumptions . . . 3
1.2 Objectives and Goals . . . 3

2 Methodology **4**
2.1 Tools and Libraries . . . 4
2.2 Data Sources . . . 4
2.3 Analysis Workflow . . . 5

3 Results Discussion **5**
3.1 Overview of VNTR Distribution Across Genomic Features . . . 5
3.2 Chromosome-Specific Distribution Patterns . . . 6
3.3 Exon-Intron Boundary VNTRs and Splicing Implications . . . 6

4 Reflections on Challenges **7**
4.1 Technical Challenges . . . 7
4.2 Process Challenges . . . 7

5 Conclusion and Future Work **7**

1 Introduction

Cystic fibrosis (CF) represents one of the most common life-threatening autosomal recessive disorders, affecting approximately 70,000 individuals worldwide. Although CF is caused by mutations in the *CFTR* gene, patients with identical *CFTR* genotypes often exhibit dramatically different disease severities. This phenotypic variability suggests the involvement of modifier genes that influence disease progression and patient outcomes. Genome-wide association studies (GWAS) have successfully identified genetic loci associated with the severity of CF lung disease, including a significant signal at chromosome 5. However, traditional GWAS approaches focus primarily on single nucleotide polymorphisms (SNPs) and often overlook more complex genetic variants such as Variable Number Tandem Repeats (VNTRs). VNTRs, consisting of short DNA sequences repeated in tandem, represent a substantial source of human genetic variation, but remain largely understudied due to technical limitations in their detection and analysis using conventional short-read sequencing technologies. The identification of disease-modifying VNTRs has profound clinical implications. Understanding which VNTRs contribute to CF lung disease severity could lead to improved prognostic tools, enabling clinicians to better predict disease progression and tailor treatment strategies to individual patients. Furthermore, these findings could reveal new therapeutic targets and inform the development of personalized medicine approaches for CF patients.

1.1 Key Assumptions

This analysis operates under several key assumptions and acknowledges important limitations.

- The adotto repeat catalog v.0.0.2 provides accurate and comprehensive VNTR boundary definitions for the human genome
- VNTRs overlapping protein-coding regions are more likely to have functional consequences than those in non-coding regions
- MANE database release 1.0 accurately represent the canonical gene structures
- VNTRs in linkage disequilibrium with GWAS SNPs may represent the true causal variants underlying observed association signals

1.2 Objectives and Goals

The primary objectives of this research are:

1. **VNTRs Identification**: Comprehensively catalog all VNTRs intersecting protein-coding genes across the human genome using computational analysis in R
2. **Functional Impact Assessment**: Evaluate the potential functional consequences of identified VNTRs on gene products, including effects on splicing, protein sequence integrity, and overall gene function
3. **R-based Analysis**: Develop and implement a robust computational framework using R programming for VNTR analysis and visualization

2 Methodology

2.1 Tools and Libraries

This analysis was performed using R programming environment (R Studio) with several specialized bioinformatics packages and data manipulation libraries:

Core R Packages:

- `GenomicFeatures` - For extracting genomic features from transcript databases
- `txdbmaker` - For creating transcript records and extracting introns coordinates
- `rtracklayer` - For importing and exporting genomic annotation files
- `GenomicRanges` and `plyranges` - For handling genomic interval operations and overlap calculations

Data Manipulation and Visualization:

- `dplyr` and `tidyr` - For data manipulation and transformation
- `ggplot2` - For creating publication-quality visualizations

2.2 Data Sources

VNTRs Annotations: The primary VNTRs catalog was obtained from the adotto repeat catalog v.0.0.2 (available at https://zenodo.org/records/7987365). This catalog provides comprehensive VNTR annotations across the human genome (GRCh38/hg38) including:

- Chromosomal coordinates (start and end positions)
- Repeat unit sequences
- Motif length

Gene Annotation Dataset: Gene structure annotations were obtained from the MANE (Matched Annotation from NCBI and EMBL-EBI) release 1.0 database, which provides:

- Canonical transcript annotations for protein-coding genes
- Exon coordinates
- Coding sequence (CDS) regions
- Gene symbols and identifiers

2.3 Analysis Workflow

The analysis followed a multi-step approach implemented in R:

Step 1: Data Importing

1. Import VNTRs catalog and MANE annotations into R using `rtracklayer`
2. Convert datasets to `GRanges` objects for efficient genomic operations

Step 2: Genomic Overlap Analysis

1. Identify VNTRs overlapping with coding regions using `plyranges::join_overlap_inner()`
2. Classify overlaps by genomic feature type:
    - Genic VNTRs (overlapping with genes)
    - Exonic VNTRs (overlapping with exons)
    - Intronic VNTRs (overlapping with introns)
    - Exon-intron boundary VNTRs (overlapping with both exons and introns)

Step 4: Analysis and characterization of VNTRs

1. Analyze VNTRs motif lengths to identify those divisible by 3 vs. non-divisible by 3
2. Calculate descriptive statistics for VNTR distributions

Step 5: Data Visualization and Reporting Generate autosome-wide distribution bar charts using `ggplot2`

3 Results Discussion

3.1 Overview of VNTR Distribution Across Genomic Features

The comprehensive analysis of VNTRs intersecting protein-coding genomic regions revealed a total of 218,881 VNTRs across the human genome (Table 1).

Table 1: Distribution of VNTRs overlapping genomic features

Genes	Exons	Introns	Exon-Intron Boundaries	Total Number
100,312	4,804	84,095	887	218,881

The distribution analysis demonstrated that VNTRs are predominantly located within introns (84,095 VNTRs, 38.4%), followed by genes more broadly (100,312 VNTRs, 45.8%), exonic regions (4,804 VNTRs, 2.2%), and exon-intron boundaries (887 VNTRs, 0.4%). This distribution pattern is consistent with the genomic organization of human genes, where intronic sequences comprise the majority of gene territory.

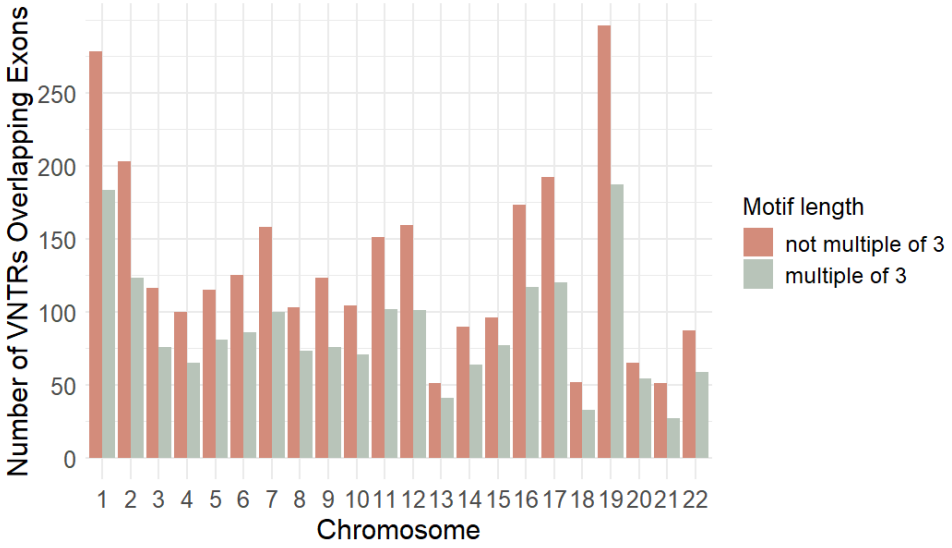


Figure 1: Motif Length "multiple of 3" vs "not multiple of 3" VNTRs Overlapping Exons per Chromosome.

3.2 Chromosome-Specific Distribution Patterns

The chromosome-wise distribution of exonic VNTRs revealed notable heterogeneity across the genome (Figure 1). Chromosome 19 exhibited the highest density of exonic VNTRs (approximately 275 VNTRs), followed by chromosomes 1 (approximately 200 VNTRs).

The analysis of motif length revealed an important functional distinction: VNTRs with repeat units that are multiples of 3 nucleotides (shown in green) versus those that are not multiples of 3 (shown in salmon). This classification is crucial because VNTRs with repeat units that are multiples of 3 are less likely to cause frameshift mutations when they expand or contract, as they maintain the reading frame of the protein-coding sequence.

3.3 Exon-Intron Boundary VNTRs and Splicing Implications

The identification of 887 VNTRs overlapping exon-intron boundaries represents a particularly important finding (Figure 2). These VNTRs show an uneven distribution across chromosomes, with chromosome 19 again showing the highest count (approximately 80 VNTRs).

VNTRs at exon-intron boundaries pose unique functional risks:

1. **Splice Site Disruption**: VNTRs overlapping canonical splice sites may interfere with proper intron removal, leading to aberrant splicing patterns
2. **Splicing Enhancer/Silencer Effects**: VNTRs may create or destroy splicing regulatory elements, affecting splice site recognition
3. **Alternative Splicing Modulation**: Repeat expansions may alter the local sequence context and influence alternative splicing decisions

The chromosome-specific distribution of boundary VNTRs generally correlates with the overall gene density, suggesting that these potentially deleterious variants are distributed proportionally to coding regions across the genome.

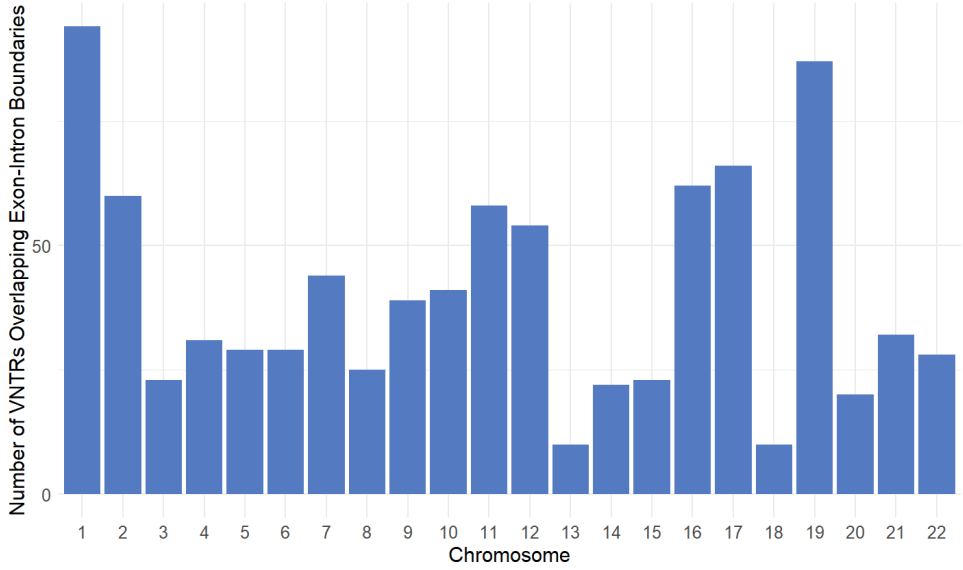


Figure 2: VNTRs Overlapping Exon-Intron Boundaries per Chromosome.

4 Reflections on Challenges

4.1 Technical Challenges

Computational Efficiency: Initial genome-wide overlap operations with 200,000+ VNTRs resulted in excessive memory usage and long processing time. The solution involved transitioning to the `plyranges` package, implementing chunk-based processing, reducing the processing time to minutes.
Intron Coordinate Extraction: Deriving accurate intron boundaries from transcript annotations proved complex because of multiple forms and splicing patterns. This was addressed using `txdbmaker` to create a transcript database and `GenomicFeatures::tidyIntrons()` for systematic coordinate extraction.

4.2 Process Challenges

Time Constraints: Completing comprehensive autosome-wide analysis within one month required significant scope management.
Learning Curve: Mastering specialized bioinformatics packages (`GenomicRanges`, `plyranges`, `txdbmaker`) required significant time investment. This was addressed through systematic package documentation review, incremental complexity increase, and extensive testing with simple cases before full dataset application.

5 Conclusion and Future Work

This comprehensive R-based analysis successfully identified and characterized 218,881 VNTRs across protein-coding genomic regions, revealing 4,804 exonic VNTRs and 887 exon-intron boundary VNTRs with potential functional significance. The most important finding is the dramatic enrichment of VNTRs on chromosome 19, which contains ∼5.7× higher density of exonic VNTRs. The motif length analysis revealed that 60%

of exonic VNTRs have repeat units not divisible by 3, indicating substantial frameshift potential that could impact protein function. This characterization provides the first catalog of functionally relevant VNTRs in human coding regions, establishing a foundation for disease association studies.

However, this analysis has significant limitations that must be addressed in future work. The computational predictions of VNTR functional impact lack experimental validation. Additionally, the focus on coding regions may miss important regulatory VNTRs in promoters or enhancers. To overcome these limitations, future work will include:

1. experimental validation of overlapping VNTRs.
2. integration of population genomics data to assess VNTRs and disease associations, particularly focusing on the CF lung disease GWAS locus on chromosome 5.

These next steps will transform the current descriptive catalog into a functionally annotated resource.