2019

# A study of the variability of minisatellite tandem repeat loci in the human genome based on high-throughput sequencing data

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES AND

COLLEGE OF ENGINEERING

Dissertation

# A STUDY OF THE VARIABILITY OF MINISATELLITE TANDEM REPEAT LOCI IN THE HUMAN GENOME BASED ON HIGH-THROUGHPUT SEQUENCING DATA

by

**YOZEN HERNANDEZ**

B.A., Hunter College - City University of New York, 2009
M.S., Boston University, 2013

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader

Gary Benson, PhD
Associate Professor, Computer Science/Biology

Second Reader

Stefano Monti, PhD
Associate Professor, Department of Medicine

*If your confusion leads you in the right direction,*
*the results can be uncommonly rewarding.*
—Haruki Murakami, *Hardboiled Wonderland and the End of the World*

# Acknowledgments

There are many people to thank here for their support, their patience, and their love as I undertook what I consider to be, thus far, the greatest challenge of my life.

To my wife, Sharifah, who has had to bear the largest share of the burden of putting up with me during this time, I owe my deepest thanks. She has been my main support, through even the toughest times. She has shared all my pain, and guided me even while carrying her own academic and personal burdens. It is safe to say I would not have been able to complete this degree without her.

To my family, immediate and extended, I thank them for their endless encouragement and ceaseless positivity, all of which they selflessly provided even when my ego was bruised and my spirit worn by the constant toil and feelings of homesickness. When I would fall into depression, when I would give in to cynicism and hopelessness, they would be there to pick me back up and show me that I had love and a home, no matter what.

I thank my friends, old and new – and I have made many of the new during the course of this program – for their support, attention, and counsel. In particular for those who experienced the ritual of graduate life, as there was something cathartic in sharing our pain, or joy. Indeed, seeing many of them graduate with their degrees did nothing to diminish my self-esteem. Rather, I saw hope there, and I shared in the celebration of their accomplishment. Special thanks in particular to Owens O'Brien, Anar Enhsaihan, Harold Gomez, and Drs. Ania Tassinari, Joshua Loving, Nacho Caballero, and Dan Gusenleitner for all the laughs, inspiration, and advice. Finally, to Marzie Rasekh, for her invaluable knowledge and experience with both tandem repeats and machine learning. Her intellectual contribution to this work – the mlZ classifier – and her willingness to conduct run after run of in-development VNTRseek trials, finally gave this project the last push it needed, and it is better for it.

To the Bioinformatics program at BU's administrative team, thank you for all your support, and for making this a top-tier program. Thank you Dave, for always being on top of my program-related deadlines, and for you assistance on issues I'm sure you were not actually responsible for, but you took the time to help me anyway. Thank you Johanna, for so much, but also for helping me book a room for my committee meetings, even when I was cutting it close. Thank you, Caroline, for your help when I needed it most after my daughter was born, and I was having a difficult time with my work/life balance. Thank you to Mary Ellen, who as the program's system administrator, was basically an IT superhero, and also a kindred spirit for me as that was one of the many hats I've worn too.

My thesis committee I thank for their time and invaluable insight. Particularly their time, as I can only speculate on the difficulty managing classes, students, and numerous PhD committees atop other commitments. Special thanks in particular to Dr. Gary Benson, my mentor, for all of his patience, support, and advice. I can see why people often compare the PhD advisor/student relationship to family, as I feel as though he has become a part of mine. Also special thanks to Dr. Weigang Qiu for being my earliest advisor and supporter in academic life, and for being an amazing human being. He too made me feel as though I was a part of his family, and taking his class and working in his lab inspired me to pursue this course in my life.

Finally, to my daughter – though she will probably never see this. Thank you, my dear Amira. You have perhaps had to put up with the most: a father who has had to choose between spending time with his dissertation work, or his child; the stress and tension at home as the future looked uncertain; the mercurial mood of a person who needed to balance work and life while figuring out how this whole parenthood thing worked. I only hope that from this I am better experienced to imbue in you the qualities of tenacity, self-confidence, and diligence. To work hard to achieve what

you want, no matter the challenges.

Yozen Hernandez

PhD Candidate

Bioinformatics Department

# A STUDY OF THE VARIABILITY OF MINISATELLITE TANDEM REPEAT LOCI IN THE HUMAN GENOME BASED ON HIGH-THROUGHPUT SEQUENCING DATA

**YOZEN HERNANDEZ**

Boston University, Graduate School of Arts and Sciences and College of Engineering, 2019

Major Professor: Gary Benson, PhD
Associate Professor of Computer Science & Biology

ABSTRACT

Variable Number Tandem Repeats (VNTRs) are repetitive sequences of DNA which exhibit polymorphism in the number of copies of the repeating pattern. As with the better known SNPs, CNVs, and other mutations, VNTRs are a form of variation in the genome. Diseases such as Fragile X syndrome, and even behavioral disorders, such as ADHD, have been attributed to VNTR polymorphisms, where changes in copy number affect chromosome and protein structure, and gene expression. Microsatellite (TRs with a pattern length $< 7nt$) VNTRs are well-characterized and have been used for DNA fingerprinting. Minisatellite VNTRs (pattern length $\geq 7nt$), however, are a relatively understudied source of genetic variation; computational complexity and the lack of specialized tools available make detecting and studying them difficult. The traditional method for examining these features involves targeted amplification and gel electrophoresis to distinguish array lengths. In this work, I discuss our effort to discover a comprehensive set of VNTRs using VNTRseek, a tool developed in Dr. Gary Benson's lab for detecting VNTRs in silico using whole genome sequencing

reads. I further discuss the curation and analyses we have performed in order to build a researcher-oriented tool, the VNTRdb, which allows other researchers access to this work and enables them to perform similar analyses. Having a tool with which VNTRs can be detected with relative ease, alongside a well-curated resource for VNTR alleles, will help promote further research into how they may be related to complex diseases, natural variation, or other areas of study.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | | |
|---|---|---|
| AJ | ..... | Population code, Ashkenazi Jewish |
| API | ..... | Application Programming Interface, a set of methods for interacting with, e.g., a library |
| bp | ..... | Base Pairs, a unit of measurement |
| CEU | ..... | Population code, Central Europeans in Utah |
| CEPH | ..... | Centre d'Etude du Polymorphisme Humain, a research institute |
| CGL | ..... | Copy Gain/Loss. Refers to change in copy number in a VNTR with respect to the reference. |
| CHB | ..... | Population code, Han Chinese from Beijing |
| DNA | ..... | DeoxyRibonucleic Acid, macromolecule used as the means for transferring heritable genetic information in biological systems. |
| ED | ..... | Edit distance, measure of dissimilarity between two strings (or sequences). |
| FL | ..... | Flank length |
| GIAB | ..... | Genome In A Bottle Project |
| GRC | ..... | Genome Research Consortium |
| GRCh19 | ..... | Human Genome reference, build 19. Produced by the GRC. |
| GRCh38 | ..... | Human Genome reference, build 38. Produced by the GRC. |
| LCS | ..... | Longest common subsequence, measure of the longest subsequence common to two strings (or sequences). |
| LINEs | ..... | Long Interspersed Nuclear Elements |
| LTRs | ..... | Long Terminal Repeats |
| NCBI | ..... | National Center for Biotechnology Information |
| nt | ..... | Nucleotides, a unit of measurement equivalent to bp but indicating single-stranded sequence |
| Ref TR | ..... | Reference TR (also possibly as ref-TR) |
| Refset | ..... | Reference set |
| REST | ..... | Representational State Transfer, a software |

| | | |
|---|---|---|
| | | architecture used to define how a web service can be interacted with. |
| SINEs | ..... | Short interspersed nuclear elements |
| SRA | ..... | NCBI Sequence Read Archive |
| SP | ..... | Support, as in "Read support" |
| TP, FP | ..... | True positive, false positive |
| TR | ..... | Tandem Repeat |
| TRDB | ..... | The TR DataBase |
| TRF | ..... | Tandem Repeats Finder, a software tool written by Dr. Gary Benson |
| UTR | ..... | UnTranslated Region, a section of genetic sequence flanking protein coding sequences. |
| VCF | ..... | Variant Call Format, a standardized file format for reporting variants. |
| VNTR | ..... | Variable Number TR |
| VNTRDB | ..... | The VNTR DataBase |
| WD, WS | ..... | Weighted distance, weighted score |
| WGS | ..... | Whole-Genome Sequencing |
| WGS500 | ..... | A collaborative project involving the Wellcome Trust Centre for Human Genetics |

# Chapter 1

# Introduction

## 1.1 Motivation and Tandem Repeat Polymorphism

The completion of the Human Genome Project initiated a new age of information – that of genetic information. Since then, the cost of genome sequencing has dropped dramatically and a growing number of individuals are having their own genomes sequenced. The increasing availability of genetic data has necessitated the development of new methods and tools to process them, which has in turn led to an increased understanding of our own genetic code in ways which were not previously possible – specifically, by elucidating the exact genetic sequence at specific genomic loci on an individual basis.

Genetic information, stored chemically in the macromolecule DNA, can be represented easily in textual form using an alphabet consisting of four symbols derived from the names of the constituent nucleotides: A, C, G, and T. Living systems use DNA to transmit information from one generation to another to increase their fitness in meaningful units we call genes. Changes in the genetic code can impact the fitness of the next generation in a positive way (beneficial), a negative way (deleterious), or can have no immediately obvious effect at all (neutral). Knowing the genetic sequence of humans overall, and on an individual level, can inform us of the ways in which changes to the genetic code impact everything from our appearance to our health, or to our response to our environment or pharmaceuticals.

**Tandem Repeats** (TRs) are regions of genetic sequence which repeat sequen-

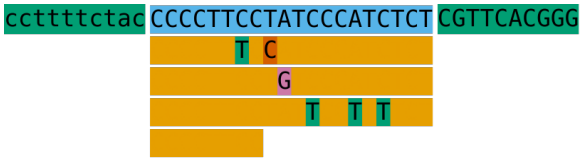



**Figure 1·1:** An example of a TR with inexact copies. This TR, located at chr10:11254877-11254944 on GRCh38/hg38, shows sequence differences from the consensus sequence (in blue, top) in each copy (shown in gold and stacked for clarity). The differences are highlighted in different colors with the single-letter code for the the nucleotide filled in. The non-repeating flanking sequence is shown at either side of the TR in green.

tially – i.e., in tandem. For example, the sequence `AGGTCTAAGGTCTAAGGTCTAAGGTCTA` consists of 4 identical and sequential repetitions of the underlying **pattern** `AGGTCTA`. The sequence of repeats is itself referred to as the **array**, and the number of copies simply as the **copy number**. TRs which consist of short ($< 7$ nucleotides ($nt$)) patterns are termed **microsatellites**[1], while longer ($\geq 7nt$) patterned TRs are called **minisatellites**.

TRs can have patterns which are very long, ranging into the thousands of bases. However, for the purposes of this work, we limit our study of TRs to minisatellites which can which can fit inside a sequencing read. Copies of a pattern are not always exact, as TR arrays may have many internal sequence changes. Copies are not always complete, either, and so TRs can have a fractional number of copies (figure 1·1). To account for variation in copies of the pattern, we generate a **consensus pattern** for each TR, and may refer to the consensus pattern as simply the pattern.

About 50% to as much as 69% of the human genome is repetitive (Treangen and Salzberg, 2011; Koning et al., 2011), with about 3% of the genome consisting of TRs (Treangen and Salzberg, 2011). Of these, some proportion exhibit polymorphism in a population, and are known as **Variable Number Tandem Repeats** (VNTRs).

[1]Microsatellites are also sometimes called Short Tandem Repeats (STRs) or Simple Sequence Repeats (SSRs), with different disciplines preferring different terms.

In the domain of polymorphic TRs, we generally speak of alleles in terms of the number of tandem copies (in this work we do not consider substitutions or short indel variations when distinguishing alleles).

Polymorphism in microsatellite TRs has been extensively studied and several diseases such as Fragile X syndrome (Adinolfi et al., 1999), Huntington's disease (MacDonald et al., 1993), myotonic dystrophy (Fu et al., 1992), and Friedreich's ataxia (Campuzano et al., 1996) have been associated with changes in the typical copy number. In these diseases, large increases in the copy number – or **expansions** – result in a disease phenotype.

The inherent variability of microsatellites has also been exploited in order to determine the provenance of DNA samples in a practice known as DNA fingerprinting – e.g., forensics teams are able to search for matches between trace DNA left at a crime scene and samples provided by one or more suspects by comparing the lengths of a select set of microsatellite loci. The set is chosen with the assumption that the likelihood of two individuals sharing the exact same profile is extremely low (Kimpton et al., 1993) (though appropriate set selection and issues with sample collection are among the points of controversy with the method (Murphy, 2018)). An analagous methodology for bacteria is called Multiple Loci VNTR Analysis (MLVA) and is one of the tools employed to trace the origin of pathogenic microbes, with particular success in *Mycobacterium tuberculosis* (Blouin et al., 2012) and others (Belkum, 2007; van den Berg et al., 2007; Kendall et al., 2010; Haguenoer et al., 2011; Pourcel et al., 2011; Zaluga et al., 2013; Chalker et al., 2015; Parvej et al., 2019).

Genotypes for both micro- and minisatellites have been challenging to detect. Software specialized in polymorphic microsatellite detection and typing include tools such as lobSTR (Gymrek et al., 2012), popSTR (Kristmundsdóttir et al., 2017), hipSTR (Willems et al., 2017), RepeatSeq (Highnam et al., 2013), and others (McIver

et al., 2011; Fondon et al., 2012; McIver et al., 2013). All must deal with the ambiguities repeat rich regions present when the repeat region exceeds the length of the sequencing read, as these reads have a reduced information content and can align in multiple locations within a repeat. Each of these tools employ strategies to deal with these challenges. LobSTR uses a preprocessing step where the repeat sequence for each target microsatellite is determined and then the Fast Fourier Transform of the sequence is calculated to characterize the sequence. It uses non-repeating flanking sequence information to anchor the repeat, and builds a noise profile of the stutter noise of the sequence using statistical learning. HipSTR begins with building a model of the stutter noise profile, and uses that model and an HMM to realign candidate reads to the candidate haplotypes. PopSTR begins with a step which determines which reads are considered to be informative. Reads containing a repeat are selected, and those which cross the barrier between flanking sequence and repeat sequence, and which also have a mate pair that has been mapped a fixed distance away from the microsatellite, are considered informative. PopSTR also requires knowledge of population data for the microsatellite alleles.

All of these methods rely on non-repeating alignment sequence to determine adequate mapping, but they all also make assumptions about the profile of the repeat sequence. Their models have been developed and designed with short repeating patterns in mind. They may be less well suited towards detection of longer ($\geq 7bp$) minisatellite loci as some studies suggest that variability is dependent on copy number and pattern length, though the latter to a lesser degree (Legendre et al., 2007; Ames et al., 2008).

Minisatellites have, therefore, been much more challenging to study. Well-known software tools used in the aligning and mapping of DNA sequences from reads to genomes, such as BWA (Li and Durbin, 2009b), have a great deal of difficulty aligning

sequences with insertions or deletions larger than 5bp even when using tools such as GATK (McKenna et al., 2010), which can help remap misaligned reads (Gymrek et al., 2012). Building *de novo* genome assemblies is also challenging, as *de novo* assembly is usually performed when no other reference of the target genome exists. Sample data will carry the allelic diversity of the population, and highly homologous duplicated sequences are difficult to reconstruct from short reads which provide less information than longer reads which may be able to span the repeating sequence. Repeats longer than the reads may be collapsed together, leading to gaps in the assembly (Treangen and Salzberg, 2011; Steinberg et al., 2014). Even in bacterial genomes, where repeats comprise somewhere between 5% and 10% of the total genome size, repeats confound assembly in particular when short reads are used (Acuña-Amador et al., 2018). Developers of mapping and genome assembly software have all developed strategies to reduce the effects repeat sequences have on the quality of their results, such as using information from mate pairs and de Bruijn graphs, calculating statistics from the known read depth and comparing that to repeat regions (which will have a much higher apparent read depth), simply using longer reads, or some combination of these and other strategies (Treangen and Salzberg, 2011; Chin et al., 2013; Steinberg et al., 2014; Acuña-Amador et al., 2018).

In a recent (Jan 2019) paper, Audano et al. (2019) examine structural variants (SVs) in the human genome including VNTRs. Their approach discovered over 50 thousand VNTRs, but the analysis was limited to a relatively small sample set of 15 human genomes, and their focus was broader than just VNTRs which our study specializes in.

Without a specialized tool for the detection of polymorphic minisatellite loci, researchers may be challenged with drawing conclusions about possibly complex diseases from incomplete or inadequate data. For example, several known VNTRs with

documented changes in phenotype have been found in exon regions, including VNTRs in DRD4 (Grady et al., 2003; Wang et al., 2004; Leung et al., 2017), where some alleles are thought to be involved in Attention Deficit/Hyperactivity Disorder; PER3 (Ebisawa et al., 2001; Benedetti et al., 2008; Golalipour et al., 2017), where copy changes have been associated with age of onset of Bipolar Disorder, and even Multiple Sclerosis; and GP1BA (Simsek et al., 1994; Mikkelsson Jussi et al., 2001; Cervera et al., 2007), where some evidence suggests that one allele is involved in Aspirin Treatment Failure and increased risk of Ischemic stroke. VNTRs have also been found in promoter regions, such as in the human insulin gene (Bell et al., 1982, 1984) where alleles have been found to vary in frequency in different populations, and have been implicated in type 2 diabetes, atherosclerosis (Owerbach et al., 1982), and hypertriglyceridemia (Jowett et al., 1984).

In this study, we detect VNTRs across a variety of genomic regions, such as protein coding exons, introns, untranslated regions (UTRs), promoter regions, and others (see chapter 3).

## 1.2 Summary of dissertation

In chapter 2, I describe VNTRseek (Gelfand et al., 2014) and how the VNTRseek reference set is produced. In chapter 3, I describe our analysis of 370 WGS samples from 368[2] individuals by VNTRseek. In chapter 4, I describe an online database constructed to disseminate our VNTR results. Finally in chapter 5 I describe the conclusions of this work, as well as future directions which may build upon it.

The aims of this dissertation are as follows:

1. Improving the methodology of VNTRseek reference set refinement. VNTRseek is the main tool used in this work. My first aim focuses on improving how the

[2]Paired tumor and normal samples were sequenced for two individuals, so they are technically represented twice.

reference set and parameter set selection should be performed. I evaluate how various conditions and parameters affect the outcome of VNTRseek results on simulated data (chapter 2).

2. Comprehensive study of VNTRs in the human genome using high-throughput whole genome sequencing data. In this aim I worked on the wide-ranging analysis of 370 samples. I describe VNTR genomic location and context, the variability in terms of the number of alleles, the impact the quality of the input data has on the results, and the challenges in processing increasingly large amounts of data with a young technology (chapter 3).

3. VNTRdb – A database of VNTRs meant to facilitate the distribution and analysis of VNTR data in the human genome. In this aim, I worked on the creation of a user-friendly, researcher-oriented tool to disseminate our curated catalog of VNTRs via VNTRdb. VNTRdb was created to be independent of species, built on modern web technology, and with the analyses that researchers would want to conduct in mind (chapter 4).

# Chapter 2

# Validation of VNTRseek, improving the methodology of VNTRseek reference set refinement, and optimization of VNTRseek

## 2.1 How VNTRseek works

As discussed earlier in chapter 1, VNTRseek is a software pipeline designed to detect TRs and VNTRs in a set of sequencing reads. It was designed for whole-genome sequencing (WGS) reads, as discussed below, but could be used for exome or RNA sequencing reads as well. VNTRseek takes as input: 1) a set of reference TRs, and 2) a set of sequencing reads in either gzip-compressed FASTQ (or FASTA) format or BAM format.

The pipeline runs in 20 steps (table 2.1). Steps are numbered 0-19, with 0 functioning as a preparation step where a MySQL (VNTRseek versions older than 1.10) or SQLite (starting with version 1.10) database is initialized. Step 1 scans the reads with TRF (Benson, 1999), which can detect TRs in the read data. The default TRF parameters are: `2 5 7 80 10 50 2000` (match weight, mismatch penalty, indel penalty, match probability, indel probability, minimum score, maximum period size). All of these parameters are hard-coded in the main pipeline script, and are not configurable via command line interface or configuration file. Instead, a user is required to edit the pipeline script to change them. Detected **read TRs** are filtered for pattern

| Step | Description |
|---|---|
| 0 | Database initialization |
| 1 | Run TRF |
| 2 | Renumber read TRs |
| 3 | Remove redundant read TRs |
| 4 | Calculate read TR profiles and cluster with PSEARCH |
| 5 | Join clusters produced in parallel in step 4 |
| 6 | (Unused) |
| 7 | Prior to v1.10: Insert reference TR data into database. v1.10+: (Unused). |
| 8 | Insert informative reads into database |
| 9 | Write out flanking sequences for all clusters. Required for next step. |
| 10 | Align flanks |
| 11 | Prior to v1.10: Update indistinguishable reference TRs. v1.10+: (Unused). |
| 12 | Record mapping information and rank read TRs using flank alignment results. |
| 13 | Calculate edges between clustered TRs |
| 14 | (Optional after v1.09) Generate index files to remove PCR duplicates |
| 15 | (Optional after v1.09) Remove PCR duplicates using data from step 14 |
| 16 | Remove multiply mapped read TRs |
| 17 | Compute VNTRs and allele support |
| 18 | (Unused) |
| 19 | Produce output VCF, distribution tables, and tex format report |

**Table 2.1:** Table of steps in VNTRseek.

length (minimum $7nt$) and amount of flanking sequence, as we require a minimum of $10nt$ non-repeat sequence on both sides of the repeat to aid in mapping the reads to a unique location on the genome. We cannot detect TRs where the pattern is repeated less than 1.8 times, as below that threshold TRF does not report TRs.

Step 2 is a simple renumbering step, where all detected read TRs are assigned a unique identifier. This is necessary because step 1 is parallelized by having a user-determined number of instances of TRF processing reads, and each instance assigns read IDs starting from "1". Step 3 eliminates cyclically redundant TRs – TRs which become identical upon alphabetic rotation of their profiles (see below) – in the reference and read TRs.

In step 4, Read TRs are assigned to candidate reference TRs using PSEARCH, a tool written for VNTRseek. These pairings are based on partial matching of the read TR and reference TR consensus patterns using **spaced seed indexing** (Ma et al.,

2002; Mak and Benson, 2009), a fast method for determining matching "words" in sequences. Profiles are built from each TR array's sequence alignment. These profiles are a sequence of $l$ **standard vectors**, where $l$ represents a column in the multiple alignment of the copies in the array. Each vector represents the counts $n_\sigma$, of the five possible characters $\sigma \in \{A, C, G, T, -\}$ in a column where "-" represents a gap. N's are ignored. The counts are converted proportionately (normalized) so that they sum to 10:

$$\left( \sum_{\sigma \in \{A,C,G,T,-\}} n_\sigma \right) = 10$$

A vector of normalized counts is then replaced by the closest "standard composition vector" via Euclidean distance, where a standard composition vector contains five positive integers or zero that sum to 10. The standard composition vectors are all concatenated in the order of the columns of the TR multiple alignment to produce the normalized profile. A profile of the reverse complement is produced using the same process, starting with a reverse complement of the TR sequence. The Euclidean distance score between normalized vectors is itself converted to a weighted distance score, $WD$ which is then converted to a weighted "pseudo-similarity" score ($WS$) as described in (Gelfand et al., 2014) with scores ranging from 0 to 100.

Reference TR to read TR pairings are then confirmed by two more alignment steps: 1) a **longest common subsequence** (LCS) comparison of consensus patterns, and 2) a **profile alignment** of the TR array (Gelfand et al., 2014).

Steps 5 through 13 consist primarily of data transformation or filtering steps. VNTRseek uses a clustering algorithm to group TRs together based on a similarity score (below) and reduces the number of alignments that need to be performed. Clusters from step 4 are joined in step 5 since step 4 is a parallel step and the results of each parallel execution must be merged. Reads which contain TRs of

interest are stored in the database and a final alignment is computed: an **edit-distance alignment** of the flanking sequence, which measures the dissimilarity of two sequences. Read TR to reference TR mappings are confirmed when a threshold score is exceeded on all of the alignments, which are also hard-coded defaults: the LCS must be at least 85% of the shortest sequence and the profiles must share at least 88% similarity ($WS \geq 88$). In versions earlier than 1.09, flanking sequences were required to have an edit distance score $ED \leq 10\%$ of the combined flank lengths of the read TR to be considered passing. In versions 1.09 or greater, flanking sequence alignments can have as many errors as determined by the formula $ED \leq min(8, 0.4 * len)$. In other words, the smaller of 8, or 40% of the alignment length. Both flanks are required to meet this criteria.

PCR duplicates are removed in steps 14 and 15, with PCR duplicate removal becoming an optional step in v1.10+ due to the increasing popularity of PCR-free sequencing techniques.

Step 16 finds and removes reads which: contain one TR that maps to multiple reference TRs with identical scores, contain two TRs that map to reference TRs that are too far apart, or contain more than two TRs.

Step 17 computes allele support and VNTR calls. The copy number of the read TRs and reference TRs are compared and alleles are called when the number of reads supporting ($RS$) an allele exceed a user-defined threshold (default $RS \geq 2$).

VNTRseek reports alleles in terms of the integral copy change *with respect to the reference* – i.e., the copies gained/lost ($CGL$). Therefore, if read TRs are detected with the same number of copies as their paired reference TR, the allele would be given as "0". An increase with respect to the reference of one copy would be given as "1", and similarly a loss of one copy would be given as "-1". Partial copy changes are rounded to the nearest whole number if the difference is at least 0.8 copies.

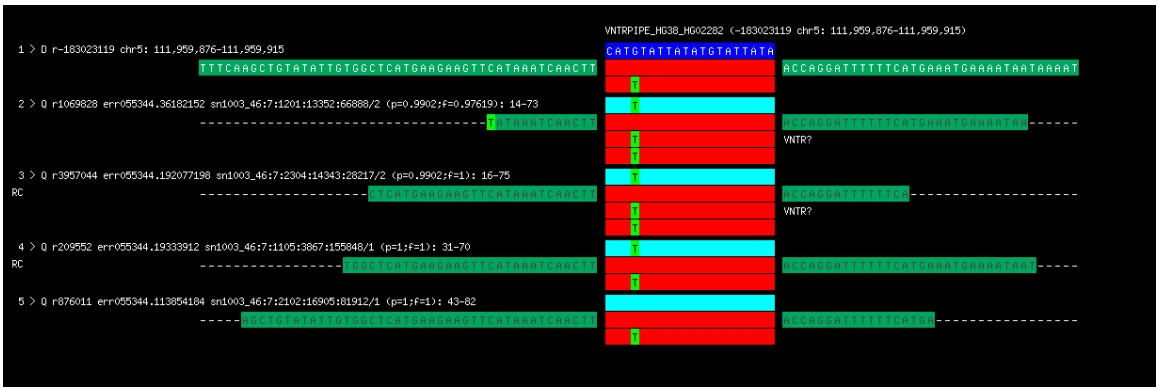



**Figure 2·1:** An example of a heterozygous VNTR call in sample HG02282 as visualized by VNTRview. The genotype at this locus would be reported by VNTRseek as "0/1", meaning the reference allele and an allele with one copy gain, with respect to the reference, were detected.

In final step 19, genotypes are reported in a VCF file. Results can also be extracted from a database used by the pipeline for intermediate actions, and a web-based visualization tool – VNTRview – is provided to allow an interactive exploration of the results[1] (figure 2·1).

## 2.2 Reference set selection and refinement

Reference set selection is a critical part of VNTRseek genotyping. Reference TR loci may share a high degree of both sequence and flanking sequence similarity, making them difficult to distinguish computationally. In some instances, the flanking sequence shares a high degree of similarity with the TR consensus sequence, which may result in a false positive alternate allele call when the read start is internal to the TR. We refer to such TRs as **indistinguishables**: families of related TRs which are difficult to distinguish. The methods we use to determine indistinguishable TRs are described further in subsection 2.2.1.

[1] At the time of this writing, the current release of VNTRview is only compatible with versions 1.08 and 1.09.x of VNTRseek.

Depending on the application, a specific set of target TRs should be selected so that the results are limited only to read TRs which map to the targets. Since we are interested in a whole-genome look at polymorphic TRs, we began with a list of all TRs in the Genome Research Consortium's reference Human genome, assembly builds 37 and 38 (GRCh37 and GRCh38, also known as hg19 and hg38, respectively)(Church et al., 2011; Schneider et al., 2017).

We developed this method on GRCh37 using TRF version 4.07, which detected 1,188,939 repeats. We filtered this initial set on the average TRF score, removing those with average score per alignment column $\leq 1.3$. We removed TRs with an overlap greater than 20% of the TR's length with common interspersed repeat elements such as SINEs, LINEs, LTRs and DNA transposons, as found by RepeatMasker (Smit et al., 2013). We also removed redundant TRs using the redundancy elimination tool of TRDB (Gelfand et al., 2007). If two or more TRs overlap by more than 50% of their length, the TR with the longer array length is kept. In case of a tie, the TR with the longest pattern size is kept. We further filtered this set to remove microsatellites, keeping those with a pattern size $\geq 7bp$. The remaining TRs in this **filtered set** numbered 230,671 (refset230671).

### 2.2.1 Parameters for calling indistinguishables

As mentioned previously, indistinguishable TRs are TRs which are either evolutionarily related and are difficult to distinguish by their sequences, or can be confused for one another depending on the occurrence of the repeat sequence within a window of a given length (eg, as in a fixed-length read). We do not eliminate indistinguishable TRs from our reference set. Instead, we mark them to indicate low-confidence allele calls.

We determine indistinguishable TRs by mapping the TRs in our filtered set back to our initial set of all TRs in the reference genome (the **unfiltered set**). The mapping

is performed by executing the same procedure that the VNTRseek pipeline uses to map read TRs to reference TRs. Initially, we used the same thresholds VNTRseek uses in a standard analysis ($WS \geq 88\%$ and $ED \leq 10\%$; see section 2.1) in order to determine if a TR is indistinguishable, with one minor difference: only one flank needed to pass the $ED$ score test. TRs were called indistinguishable if they clustered with at least one other reference TR.

However, no well-defined methodology existed to determine optimal parameters for indistinguishable TR classification. This section describes the development of such a method: a well-defined, reproducible procedure to identify an indistinguishable TR set. Our goal was to maximize the detection of problematic TRs while minimizing the loss of distinguishable TRs, using a reasonable parameter set. We evaluate the ocurrence of problematic TRs based on the number of false postive calls in the results of a VNTRseek trial where the input read set consists of unmodified sequences drawn from the reference TR set, and the unfiltered reference set is used as the input reference set.

Profile alignments (mentioned above) are useful for comparing repetitive sequences which have some short sequence variation. The source of the variation can either be due to instrument error or naturally occurring mutations. An ideal $WS$ cutoff should allow us to call sequences which are evolutionarily related but genomically dispersed as indistinguishable, while simultaneously allowing sequences which have either diverged enough or converged by chance to a similar sequence to still be labeled as distinguishable. Apart from the 88% cutoff used for a typical VNTRseek run, we evaluated three other thresholds for $WS$: 91%, 93% and 95%.

Sequencing read length limits our ability to distinguish TR sequences in some cases, particularly for long TRs. Non-repetitive flanking sequences have a better chance of mapping uniquely to a genomic location and are an important factor in

distinguishing TR loci. VNTRseek allows the user to choose a minimum flank length ($FL$) needed to distinguish TRs. Shorter values of $FL$ can lead to more TRs appearing indistinguishable as shorter sequences have lower complexity, but flank length restrictions can lead to a TR being thrown out if it is simply too long to be spanned along with the required flanking sequence. We evaluated several values of $FL$ which might be used in a typical analysis of VNTRseek given the current most frequently available read lengths: 10, 20, and 50nt. Flanking sequence edit distance, $ED$, was also varied: 10%, 20%, and $min(8, 0.4 * len)$, the latter value being used in a standard analysis. The lower $ED$ thresholds allow us to examine the effect of being stricter with reference vs reference comparisons compared to read vs reference. We also evaluated requiring both flanks to pass the $ED$ test, in addition to just one flank.

Input comprised sequence data from TRs in refset230671, with arrays converted into their profile representation, and with flanking sequences of length sufficient to satisfy the FL requirement. We ran the indistinguishable search procedure for each profile alignment cutoff, at each flank length, requiring either one or both flanks to pass the $ED$ test for a total of 48 trials (we did not run trials for $ED \leq min(8, 0.4 * len)$ at this stage). We produced Venn diagrams of the results for each cutoff, grouping indistinguishable TRs by flank length (figure 2·2). This allowed us to visualize the change in indistinguishable counts across parameter sets, as well as the size of the set called in common across all flank lengths (the intersect). The total number of indistinguishables across all flank lengths (the union) cannot be displayed easily on a Venn Diagram, so figure 2·2 has that figure annotated below the percent cutoff used.

Requiring both flanks to be below the $ED$ threshold has the largest effect on the number of TRs called indistinguishable, as both flanks must share a high degree of similarity with their corresponding flank to make a call. Increasing the $WS$ cutoff decreased the number of indistinguishable calls, with an average of nearly half the

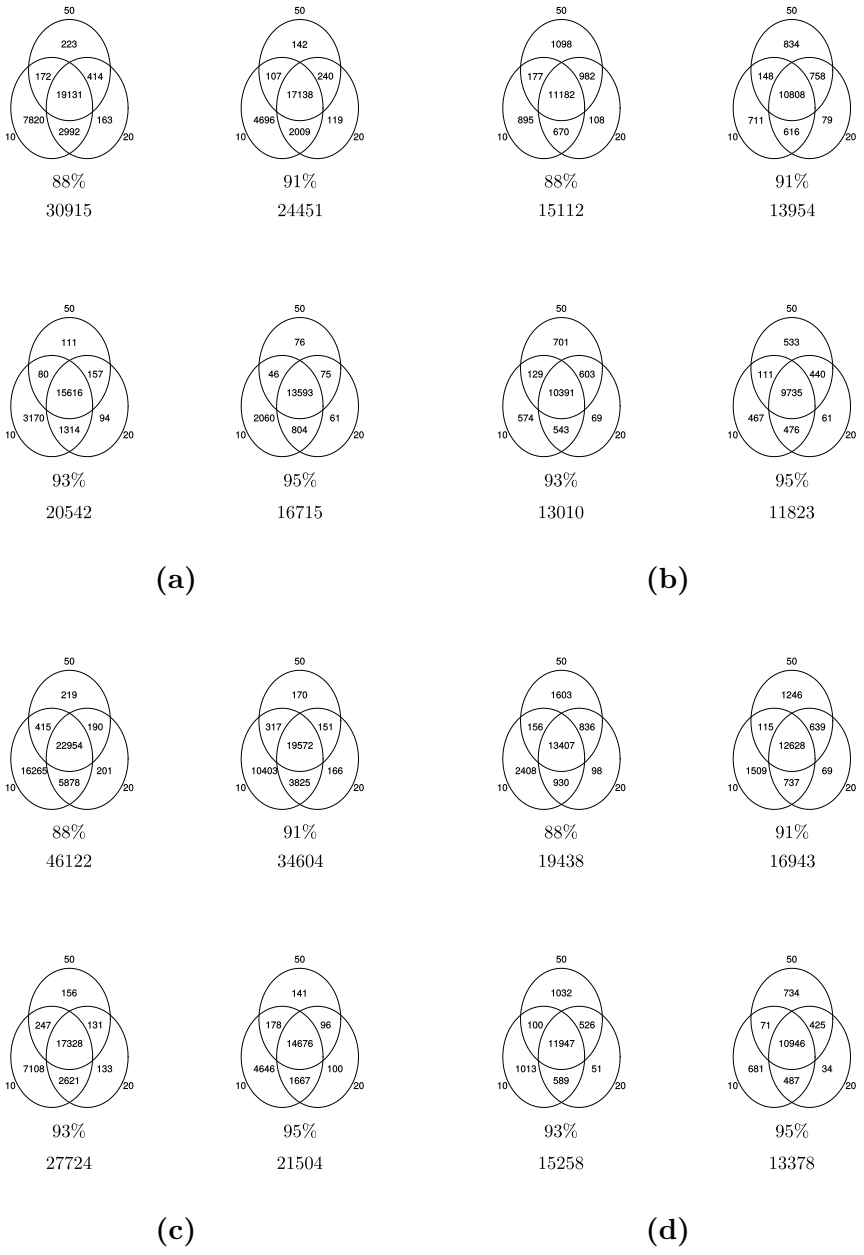



**Figure 2·2:** Overlap of indistinguishable calls at $WS$ thresholds of 88%, 91%, 93%, and 95% and $ED$ thresholds either below 10% the shared flank length (**a** and **b**) or 20% (**c** and **d**). We varied how many flanks were taken into consideration in scoring, either one (**a** and **c**) or both (**b** and **d**). The $WS$ thresholds are annotated below each diagram, and the total number of indistinguishable calls is annotated below that.

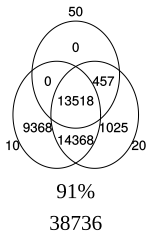



**Figure 2·3:** Overlap of indistinguishable calls at $WS = 91\%$ and $ED \leq min(8, 0.4 * len)$. Both flanks were required to pass the $ED$ cutoff.

number of indistinguishables called in total when comparing the largest and smallest values of $WS$ with all other parameters fixed. $ED$ level had the next highest impact, with the number of additional indistinguishable calls dropping with increasing $WS$. Flank length however, did not have a linear effect on the number of TRs removed, as some TRs appear indistinguishable at short lengths but are distinguishable at longer lengths. As a compromise, we took the union of indistinguishable calls across all flank lengths.

We chose a $WS$ threshold of 91%: an increase in the $WS$ threshold makes our classification of indistinguishables more stringent, and taking the union means that we will consider all TRs called indistinguishable, even if called in just one of the flank lengths (table 2.2). We decided that the cutoff for $ED$ score should match the VNTRseek cutoff used in typical analyses, which was $ED \leq min(8, 0.4 * len)$ (figure 2·3). We now require both flanks meet this criteria.

TRs which have been clustered together (see 2.1) with other indistinguishable TRs are considered indistinguishables themselves, even if they were not mapped to another TR in the unfiltered reference set. We chose to change this criteria and instead base indistinguishable calling solely on a TR having at least one link to another TR, given the parameter set. Repeating the procedure with flank lengths 10, 20, and 50nt and taking the union, results in a final, indistinguishable set.

| Reference set | Parameter set | | Size of Union |
|---|---|---|---|
| | $ED$ | Flanks Req. | |
| refset230671, GRCh37 | 10% | Single | 24451 |
| | 10% | Double | 13954 |
| | 20% | Single | 34604 |
| | 20% | Double | 16943 |
| | $min(8, 0.4 * len)$ | Double | 38736 |

**Table 2.2:** Indistinguishable TRs by parameter set. $ED$ column gives the maximum $ED$ score allowed for flank alignments. The 'Flanks Req.' column indicates whether one flank ("single") or both flanks ("double") were required to meet the $ED$ threshold. All are run with a 91% $WS$ threshold. The final row is the set of indistinguishables chosen.

### 2.2.2 Elimination-based indistinguishable calling

As discussed in subsection 2.2.1, the indistinguishable TRs are difficult to make accurate calls for due to the high degree of similarity these loci share with each other. This difficulty is in part related to the length of sequence being compared. For most modern sequencing platforms, read length is fixed at 100, 101, 125, 148, 150, or 250 nt. Given the heterogeneity of read length, we cannot expect to capture all cases in which a TR may appear indistinguishable with the above method. That TR array length can be polymorphic in the case of VNTRs adds another layer of complexity.

The method discussed in this section is meant to address the issue of TRs which result in a VNTR call simply due to their position within a read. It simulates cases where TRs are spanned by a read with asymmetric flanking sequence lengths (as opposed to some fixed length as above), or only partially spanned with reads starting internal to the TR.

Sliding windows of length $L$ (one of 100, 125, 150, or 250 nt) were drawn over all TRs in refset230671. The first window, $R_0$, starts one read length before the start of the TR ($TRstart - L$) and the last window, $R_{L+TRlength}$, starts at the base immediately following the last base of the TR ($TRstop + 1$).

Sets of sliding windows were drawn for all the above possible read lengths. Each set was run through a full run of VNTRseek with reference set refset230671 and at least 10bp flanks required (up to a maximum of 50bp), but only one read required to support an allele call. At pipeline completion, all reference TR loci generating reads which supported a non-reference allele call were retrieved, as well as any reference TR loci which were called as VNTRs. These retrieved loci were removed from the reference set because of their ability to cause false positive VNTR calls.

We consider this method to be more effective at eliminating problematic TRs which were not caught by our initial indistinguishable calling method, as it can capture TRs which are problematic only due to the combination of read length and position within a read.

## 2.3 Validation using simulated reads and VNTRs

We validated VNTRseek performance in order to demonstrate its ability to correctly identify variants in simulated data. We simulated reads as well as VNTR alleles, and conducted several VNTRseek trials. We then compare the detected VNTRs and simulated VNTRs, and collected various statistical measures such as the sensitivity, specificity, and positive predictive value.

Reads were simulated for the Roche/454 Sequencing GS FLX platform. The 454 GS FLX sequencer was used to produce the Watson (Wheeler et al., 2008) and KB1 genomes (Schuster et al., 2010), which were used in our pilot studies with VNTRseek (Gelfand et al., 2014). Read locations were determined by a 64-bit Mersenne Twister pseudo random number generator[2] (Matsumoto and Nishimura, 1998) assuming a diploid genome with chromosome lengths and sequences matching the GRCh37 human genome reference sequence. Read lengths were drawn randomly from a normal

[2]Available at www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt64.html

distribution as generated using an implementation of the Box Muller Polar transform method[3], assuming a mean length of 261 nt and a standard deviation of 27 nt. The program which generated the read locations and read lengths was written by Dr. Gary Benson.

To introduce sequencing errors, we produced a simulated read set which was then modified by applying error rates derived from empirical data on the 454 GS FLX sequencer, as described in (Huang et al., 2012). The error rates were homopolymer length-dependent, with indel error types (overcalls and undercalls) at much higher frequency than substitution errors, and with the probability of error increasing with homopolymer length. Error rates were applied on a per-homopolymer basis, with each homopolymer of length $n \geq 1$ having a random chance of being over or undercalled, and homopolymers of length one having a random chance of undergoing substitution.

To test VNTR detection we simulated VNTRs by creating a modified reference set. 1118 randomly selected reference TRs (5% random selection frequency) had their copy number changed by one or two copies added or removed. In each case, existing copies within a selected reference TR were randomly selected for duplication or removal. These modifications only affected the copy count and repeat profile of the sequences.

Using the above parameters, we produced six read sets: three read sets drawn from the reference genome with no mutations, and a second group of three read sets produced by modifying the reads of the first three by mutating the sequences as described previously. All six sets were run on VNTRseek (min flank required = 10, max flank considered = 50, read support $\geq 1$). We compared mapped locations to read origins, and called VNTRs to simulated VNTRs. Average results for all 6 sets are shown in table 2.3 while tables 2.4 to 2.7 (taken from (Gelfand et al., 2014)) show the results for a typical pair of sets.

[3]Described in www.taygeta.com/random/gaussian.html

| Read | Read Mapping | | Reference TR Mapping | | VNTR Calling | | |
|---|---|---|---|---|---|---|---|
| Set | Sen | Spec | Sen | Spec | Sens | Spec | PPV |
| Exact | 97.5% | 99.6% | 96.9% | 99.2% | 95.8% | 100%* | 96.2% |
| Error | 90.2% | 99.5% | 94.7% | 99.0% | 90.9% | 100%* | 91.6% |

**Table 2.3:** Average accuracy measures for three simulated read sets generated from the reference genome (Exact) and three sets obtained by introducing errors into the exact reads (Errors). Read Mapping is the accuracy of assigning reads to the correct reference TRs, Reference TR Mapping is the accuracy with which reference TRs are assigned reads, VNTR Calling is the accuracy of calling VNTRs in a modified reference set where 1118 randomly selected reference TRs (approximately 0.5% of the total) were modified by adding or subtracting one or two pattern copies. PPV is positive predictive value, the fraction of called VNTRs that were correct. *Specificity for VNTR calling is slightly less than 100%, see tables 2.4 to 2.7.

| Randomly Generated Reads | | | | |
|---|---|---|---|---|
| | Reference TR Spanning | | Other TR Spanning | |
| Read Set | Generated | Correctly Mapped | Generated | Incorrectly Mapped |
| Exact | 855,782 100% | 834,633 97.5% | 1,607,291 100% | 7,048 0.4% |
| Errors | | 771,335 90.1% | 1,654,643 100% | 7,575 0.5% |

**Table 2.4:** Results for two typical simulated read sets generated from the reference genome and mapped back to the reference TRs. Reads in one set exactly match the reference while reads in the other contain simulated sequencing errors. Reference TR Spanning reads (positive set) are those that spanned the locus of a reference TR including at least twenty nucleotides of flanking sequence on each side. Other TR Spanning reads (negative set) are those that contained a spanned TR, but not a reference TR. Incorrectly mapped in this group means the read was mapped to a reference TR. Incorrectly mapped TR spanning reads are examined further in table 2.5.

| Reference TR Spanning Reads Not Correctly Mapped | | | | | | | |
|---|---|---|---|---|---|---|---|
| Read Set | All | Mapped Incorrectly | Failed TRF | Ties | Multiple Loci | PCR Dupes | Failed Scoring |
| Exact | 21,149 | 454 | 545 | 13,030 | 2,556 | 2,755 | 1,809 |
| Errors | 84,447 | 783 | 33,281 | 12,187 | 2,275 | 2,258 | 33,663 |

**Table 2.5:** Fate of reference TR spanning reads (see: table 2.4) not mapped correctly. A very small number of reads mapped to the wrong reference. Otherwise, a read was discarded if 1) TRF failed to detect a TR, 2) (Ties), the same TR mapped to more than one reference with equal score, 3) (Multi), different TRs in the read mapped to two references that were not close enough together to be spanned by the read or mapped to three or more references no matter their spacing, 4) it was eliminated as a PCR duplicate, 5) TR profile or flank scores failed to meet the thresholds.

| Reference TRs | | | | | | | |
|---|---|---|---|---|---|---|---|
| Read Set | With Spanning Reads | With Mapped Reads | | | Without Spanning Reads | With Mapped Reads | |
| | | Only Correct | Other | None | | None | Any |
| Exact | 209,519 100% | 203,071 96.9% | 1,923 0.9% | 4,525 2.2% | 20,787 100% | 20,639 99.3% | 148 0.7% |
| Errors | | 198,365 94.7% | 2,783 1.3% | 8,371 4.0% | | 20,597 99.1% | 190 0.9% |

**Table 2.6:** Reference TR results for two typical simulated read sets, one exact and the other with the same reads with introduced errors. Out of 230,306 reference TRs, 209,519 had at least one spanning read in the simulated data sets (first column). Sensitivity (percent in third column) is measured as the ratio of reference TRs with only correctly mapped reads (third column) to reference TRs with spanning reads (second column). Specificity (percent in seventh column) is measured as the ratio of unspanned reference TRs which had no reads mapped to them (seventh column) to all unspanned reference TRs (sixth column).

| Read Set | Generated VNTRs | | | | Unmodified TRs | | |
|---|---|---|---|---|---|---|---|
| | 2-span | Called | | Not | | Called | |
| | | Correct | Incorrect | Detected | All | VNTR | PPV |
| Exact | 913<br>100% | 875<br>95.8% | 1<br>0.1% | 38<br>4.2% | 229,188<br>100% | 35<br>0% | –<br>96.0% |
| Errors | 913<br>100% | 830<br>90.9% | 1<br>0.1% | 83<br>9.1% | 229,188<br>100% | 76<br>0% | –<br>91.5% |

**Table 2.7:** VNTR results for a modified reference set and two typical simulated read sets, one exact and the other with the same reads with introduced errors. 1118 randomly selected reference TRs (approximately 0.5% of the total) were modified by adding or subtracting one or two pattern copies. 913 of these had at least two spanning reads in the simulated read sets (the minimum required to call an allele). Sensitivity is the ratio of correctly called VNTRs to the total VNTRs with two spanning reads (column 3). Specificity is the ratio of unmodified TRs not called as VNTRs to all unmodified TRs (column 6). Given the large negative set size, an important measure is positive predictive value (PPV), the ratio of true VNTR calls to all VNTR calls (column 8). In the read set with errors, 8.5% of the VNTR calls were incorrect (approximately 1 out of 12). When subdivided, this corresponds to approximately 1 in 20 incorrect calls for singletons and 1 in 2.2 incorrect calls for indistinguishables. (For both Exact and Errors rows, Called Correct, Incorrect, and Not Detected add to 914 because in each case, one indistinguishable VNTR was called with both the correct number of copies and an incorrect number of copies. PPV reflects a reduction of one in the correct calls.)

Overall, VNTRseek performs very well, with high ($\geq 90\%$) sensitivity and specifity for both read mapping and VNTR calling, even for read sets with errors (tables 2.3 and 2.6). Given the overwhelmingly large negative set (unmodified reference TRs) compared to the positive set (simulated VNTRs), the positive predictive value (PPV) is a more appropriate measure to convey VNTRseek performance on VNTR detection, as it is the fraction of VNTR calls which were correct. For exact reads, the average PPV was 96.2% and for reads with simulated errors it was 91.6%. Approximately 1 in 20 singleton VNTR calls were wrong, while approximately 1 in 2.2 indistinguishable calls were wrong, demonstrating the importance of highlighting these classes in the output.

## 2.4 Improvements to VNTR Calling Software

In chapter 3 I discuss a comprehensive analysis of over 370 human WGS samples. The initial portion of the study was conducted on 350 low-coverage ($\leq$ 24x) samples from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2012, 2015), 17 genomes from the Illumina Platinum Genomes data set (Eberle et al., 2017a), and 3 genomes comprising a trio of the Yoruban ethnic group (see chapter 3 for more details). Despite the majority of the input comprising lower coverage data, the total time for download and analysis of these data sets was over 4 months. The data was analyzed on a modern compute cluster, with a typical node having 16 CPU cores available for use. At peak performance, 4 analyses were run simultaneously with 8 processing cores used by each.

Execution time for a single example analysis of approximately 26x coverage (797,113,367 reads at 101nt) was 24.29 hours. With the optimizations described later on in subsection 2.4.1, this average analysis was reduced to a 9.45 hour execution time, with still more room for improvement.

Inefficient computation accounted for part of the lengthy total processing time. Frequent runtime errors attributed to differences in the nature of the input data, lengthy processing times due to remotely executing analyses while using a locally hosted database server, and various technical issues on the compute cluster all resulted in many lost hours with little advancement in data processing. Fortunately, each of these setbacks also presented an opportunity to address issues in VNTRseek, which were only exposed when executing on a shared and distributed system with a networked file system, or when using data produced by a new source for which we had not anticipated any significant difference in file format.

### 2.4.1 Improvements to VNTRseek performance

VNTRseek takes advantage of multiprocessor units by parallelizing input processing with TRF, profile-based clustering, and within cluster alignments. Typical bottlenecks in throughput are TRF and PSEARCH (CPU intensive), and reading and writing files to and from the disk (I/O intensive). A third bottleneck was the relative instability of the software as we moved from a single-machine environment with low ($\approx 5x$) coverage genomes to much larger datasets, which often resulted in many lost hours of analysis as we tried to determine the issue.

We could address CPU bottlenecks by improving the algorithms used in the two software tools which expend the most CPU cycles. Improvements to TRF are planned, notably an implementation of the algorithm developed in Loving et al. (2014) which leverages the inherent bit-parallelism of vector operations in modern CPU architectures. But these are not in the scope of this dissertation thesis. PSEARCH may benefit from a code refactor to determine computational bottlenecks in that tool. However, after a simple code analysis we determined that the benefits from making improvements to these algorithms and their code bases would be far outweighed by the undertaking itself, particularly due to the code complexity, and the potential to

negatively affect the wide userbase of TRF.

We instead decided to resolve I/O bottlenecks, as typical workflows for our analysis involved using compute clusters with large numbers of compute cores and plenty of available memory. Storage, however, is much more limiting, both because disk space can be limited and hard to request but also because the file systems tend to be networked, specifically using NFS. NFS disk mounts may have a number of issues with file locking and synchronization, making optimization of IO operations a priority.

The first area we targeted was the input WGS data reading code. This code was responsible for reading in sequence data in gzip compressed files in either FASTA or FASTQ format, and then the read information was passed in to TRF. TRF output is then passed into a conversion program called trf2proclu via UNIX pipes, which also calculated the TR profiles (see 2.1).

However, this pipeline had a number of issues: all FASTA and FASTQ code was written from scratch with no testing framework, the pipeline from TRF to trf2proclu appeared to break periodically with no way to gracefully handle problems, and the code was too tightly coupled to allow us to expand to accept other file formats. This code was refactored to decouple the various components.

Code which reads in sequencing files is now separated into functions by input file format, and is called by code which determines the format. The output of these functions is always a Perl hash variable which contains the header and sequence of each read record. This enables us to add more file formats with ease, or simply change the implementation without affecting other parts of the code. Further, custom written code was removed in favor of using `seqtk` (https://github.com/lh3/seqtk) a well-known, well-tested, and widely used tool for reading both FASTA and FASTQ files, and converting between them. This means we can use the same function for both FASTA and FASTQ formats. We also added a function which could be used to read

SAM/BAM files using samtools (Li and Durbin, 2009a), broadening the appeal and usability of VNTRseek.

The pipeline between TRF and trf2proclu was simplified, and input and output was checked for potential errors, such as signals from the operating system which result in program termination, error in the input format, or duplicate sequences. The latter was a particularly necessary improvement, as VNTRseek assumes that read headers are unique, in order to distinguish reads. Some paired end data sets do not add segment information into their headers (e.g., /1 or /2 tags at the end of the header) instead relying on the fact that the pairs are distributed in their own file to indicate the difference. However, input is streamed together in VNTRseek, so file information is not seen by the time the reads reach downstream processes in the pipeline. Sequence reading functions were coded so that if headers gave no indication of their segment, a unique identifier was added to the header before being passed to TRF.

Some clusters or operating systems place limits on the run time of a process. VNTRseek was modified so that after 1 million processed records, the TRF and trf2proclu processes are terminated and new ones started to avoid hitting this limit. 1 million was chosen as a reasonable number which appeard to avoid the issue on our system, but the user may choose to change this value. We still, however, recommend that users using FASTA/FASTQ archives first split the input into files with around 1 million reads each anyway, due to the fact that the TRF step of the pipeline will only parallelize based one the number of files it detects. An alternate method for parallizing this step is being considered, but has not been implemented while the performance gain is analyzed.

VNTRseek versions 1.09.x and earlier are notorious for generating large numbers of files, many of which turn out to be redundant in production scenarios. We refac-

tored other areas of the code which wrote out many intermediate files in favor of smaller or fewer writes, no output unless given a specific instruction, or a cleanup procedure following a step where the output is no longer needed. trf2proclu, for example, generated large numbers of files, which could reach gigabytes of storage usage in large data sets. About half of the files, by number, were redundant as they contained the same information as another output file type. The information in these files was necessary, however, for calculations further down the pipeline. We moved the place in which these calculations take place to within trf2proclu itself, and merged the data following the parallel step, eliminating the need for these files.

Step 8 in the pipeline required rereading all input files, sequentially, in order to retrieve sequence data for all reads which spanned supported VNTR alleles. This step was an $O(n)$ search with $n$ being the number of reads in the input set. Storing every single read would be impractical, so this solution seemed appropriate. We searched for a different solution which would either eliminate or reduce the need to do an $O(n)$ search, or reduce it to $O(m)$, where it is expected that $m \ll n$ as $m$ represents the number of VNTR allele supporting reads. Thanks to the earlier code refactoring for step 1, we are able to determine at completion of a TRF/trf2proclu pipeline which reads span a VNTR allele and record these reads in a compact form. At step 8, only these files are read, drastically improving performance.

VNTRseek depends on a database to record its results, read sequences, reference set, and other data. However, most compute clusters do not provide a full relatioal database management system for users, particularly not for long term use. In versions 1.09.x and earlier, we simply hosted our database locally using MySQL. However, pushing potentially large amounts of data over the network, sometimes with many simultaneous analyses, quickly proved to be impractical. While MySQL was an ideal solution for management of multiple analyses as a queuing system, performance would

degrade quickly with too much traffic. We put in effort to port VNTRseek over to SQLite, which is an ideal solution for applications which need to store data locally in the same way it might use a file to keep some data. While SQLite could not function as a substitute for the large number of intermediate files VNTRseek produces, due to SQLite not being designed for parallel writes, it was suitable to replace MySQL for the output of each analysis. SQLite also allowed us to reduce redundancy as some reference data, which was recalculated on each analysis, could now be calculated once and shared among all concurrent analyses which utilized the same reference set transparently.

Following these changes, the average performance improvement of a typical run of VNTRseek – measured as the difference in runtimes of an analysis both post and pre optimizations, and with identical input and environment, divded by the pre optimization runtime – is $\approx 60\%$.

### 2.4.2 Enabling repeat detection in centromere regions using TRF

While using TRF version 4.08 to scan the GRCh38 reference genome shortly after its release, we encountered a bug where TRF would slow down significantly during the analysis of centromere regions, and then hang. As a result, we simply excluded centromeres from further analysis, and our reference set for GRCh38 does not include TRs from these regions.

Following the majority of the work described in chapter 3, we investigated the cause. As a C program, TRF must allocate buffers in memory to store data such as the sequence being analyzed, alignment matrices, and so on. After following memory allocations and analyzing points in the input data which resulted in the program hang, we determined that some of our buffers were simply too small to read very long TRs in the centromere regions, including one exceeding 5 million bp. Since it would be impractical to constantly allocate large amounts of memory, some allocations being

impossible on 32-bit systems, portions of TRF were refactored so that any array which needed to be increased in size was instead coverted to a dynamically allocated array. The array size was also made to be configurable with a reasonable default, and documentation was written to include guidance on how to use the new option.

# Chapter 3

# Comprehensive study of VNTRs in the human genome using high-throughput whole genome sequencing data

| Dataset | Read Length (bp) | Coverage | Samples In Set | Citation |
|---|---|---|---|---|
| 1000 Genomes | 100–101 | 5–24x | 330 | (Consortium, 2012) |
| CEPH 1463 family | 100 | 48–109x | 16 | (Inc., 2014; Eberle et al., 2017b) |
| HapMap Y117 Yoruban trio | 250 | 76–77x | 3 | (See table 3.6) |
| WGS500 samples | 100 | 26–112x | 8 | (Taylor et al., 2015) |
| CHM1 | 150 | 42x | 1 | (Chaisson et al., 2014) |
| CHM13 | 250 | 137x | 1 | (Huddleston et al., 2017) |
| GIAB Ashkenazi trio | 250 | 64–74x | 3 | (Zook et al., 2016) |
| GIAB Chinese trio | 148–250 | 117–355x | 3 | (Zook et al., 2016) |
| GIAB NA12878 | 148 | 306x | 1 | (Zook et al., 2016) |
| Tumor/Normal Samples | 101 | 41–95x | 4 | (Drmanac et al., 2010) |

**Table 3.1: Datasets.** Data from 370 genomes was used. Coverage values refer to "read coverage" – the product of the number of reads and the average read length, divided by the haploid genome size, as in the Lander/Waterman equation (Lander and Waterman, 1988). All values are approximate. Some values for the CEPH 1463 and WGS500 samples, as reported here, are higher than those stated in the original sources because replicates for the same genome were combined when available. See Data Section for URIs of data.

## 3.1 Introduction

With the increasing availability of whole genome sequencing (WGS) data, researchers are faced with an ever growing source of data to mine for information. Most modern WGS pipelines for human data involve mapping to a reference genome directly after the sequencing experiment. While the most commonly used tools for genome mapping or assembly may be adequate for detecting single point mutations, they are far less capable of correctly mapping repetitive DNA making genotyping variability in repeat sites challenging, and as a result repetitive regions may end up being removed due to poor mapping quality (Gymrek, 2017). In a previous paper (Gelfand et al., 2014), our lab introduced a tool for genotyping polymorphic tandem repeats known as Variable Number Tandem Repeats (VNTRs). These loci vary in copy number, and present issues for most aligner software which is typically run after a sequencing experiment, such as BWA-MEM (Li, 2013).

In this paper, we present the results of a wide-ranging, multi-year study into the variability of VNTRs in the human genome using WGS publicly available data. While other tools similar to VNTRseek have been developed since our earlier publication, such as adVNTR (Bakhtiari et al., 2018a), this paper presents what is, to our knowledge, the most comprehensive catalog of polymorphic minisatellites in the human genome to date, pooling results from over 300 publicly available samples.

Other tools for polymorphic TR typing include tools such as lobSTR (Gymrek et al., 2012), popSTR (Kristmundsdóttir et al., 2017), and hipSTR (Willems et al., 2017) which were developed to detect mircrosatellite tandem repeats (pattern size $\leq 6bp$) rather than minisatellite repeats (pattern size $\geq 7bp$), which are the focus of this paper. Similar to lobSTR, VNTRseek uses flanking sequences to disambiguate mapping. popSTR requires prior knowledge of population data, and hipSTR has a preprocessing step in which a profile of the stutter noise of the repeats is built.

VNTRseek does not have either of these requirements.

## 3.2 Materials and Methods

### 3.2.1 WGS datasets

Ten datasets were used in this study comprising 370 whole genome sequencing samples (Table 3.1): 330 individuals from the 1000 Genomes Project Consortium (2012); 16 of the 17 member CEPH 1463 family as sequenced by Illumina for their Platinum Genomes set Inc. (2014); Eberle et al. (2017b); a Yoruban trio (HapMap Y117) sequenced using a PCR-free technique; eight individuals from the WGS500 project, a large-scale craniosynostosis disease association study Taylor et al. (2015) consisting of two trios (unaffected parents and affected child) plus an unaffected couple whose affected child has no publicly available data; a Chinese trio (CHB), an Ashkenazi Jewish trio (AJ), and sample NA12878, the remaining CEPH 1463 family member, as sequenced by the Genome in a Bottle (GIAB) Consortium Zook et al. (2016); two hydatidiform mole (CHM) cell line genomes which are essentially haploid: CHM1 Chaisson et al. (2014) and CHM13 Huddleston et al. (2017), sequenced by The Genome Center at Washington University School of Medicine; and tumor/normal pairs (breast invasive ductal carcinoma cell line/lymphoblastoid cell line) from two unrelated individuals, HCC1187 and HCC2218 (Drmanac et al., 2010). Coverage ranged from approximately 5x, in several 1000 Genomes samples, to 355x, in the GIAB Chinese trio child. 358 of our samples were sequenced with read length 100–101bp. The remaining 12 consisted of either 148 bp or 250 bp reads. Input data consisted of sequencing data produced on the Illumina platform in FASTQ format.

### 3.2.2 TR reference set

Human reference genome GRCh38 (hg38) (Lander et al., 2001) was used to produce a **reference set** of TRs in the Tandem Repeats Database (TRDB) (Gelfand et al., 2007) with the Tandem Repeats Finder (TRF) software (Benson, 1999) and four quality filtering steps as described in (Gelfand et al., 2014). Centromere regions were excluded from the reference set. The result was a set of 228,486 reference TRs (refset228486). The TRs were classified into two subcategories, singletons and indistinguishables. A **singleton** TR appears to be unique in the genome based on a combination of its repeat pattern and flanking sequence. An **indistinguishable** TR belongs to a family of genomically dispersed TRs which share highly similar patterns and flanking sequence and can therefore produce misleading genotype calls.

Indistinguishable TRs were identified using the procedure described in (Gelfand et al., 2014) i.e., each TR array from the refset228486 was converted into a single simulated read and all simulated reads were mapped to the original unfiltered TR set using VNTRseek (Gelfand et al., 2014). Any TR which mapped to a different locus was labeled indistinguishable. 37,200 TRs were identified as indistinguishable ($\sim$ 16.3%). Indistinguishable TRs were not removed from the reference set, but genotype calls in the output of VNTRseek were flagged if the locus was indistinguishable.

To reduce the number of singleton false positive VNTR calls in this study, two methods were used to eliminate problematic TR loci from the reference set. The first involved detecting false mappings of simulated reads and is described in subsection 2.2.2. This procedure was conducted for each of the read lengths 100bp, 150bp, and 250bp and produced three separate reference sets (table 3.2), available at https://dx.doi.org/10.5281/zenodo.1491907. The second method is described in subsection 3.2.5.

| Read Length (bp) | Singleton TRs Removed | Final Reference Set Size | Expected Genotyped |
|---|---|---|---|
| 100-101 | 1,704 | 226,782 | 153,293 |
| 148 | 1,976 | 226,510 | 168,742 |
| 250 | 4,812 | 223,674 | 177,864 |

**Table 3.2: Modification of the reference set to reduce false positive TRs.** The original reference set contained 228,486 TR loci, labeled as singleton or indistinguishable. Using simulated reads generated from the reference set, singleton TRs that were called as false positive VNTRs or those which generated reads leading to such a result were removed (see Materials and Methods). The "Expected Genotyped" column is the number of singleton TR loci for which the sum of array length and minimum flank lengths did not exceed the read length (for the 100/101bp set, 100 bp was used as read length).

### 3.2.3 TR Annotation

Reference TRs were annotated with genomic context features in hg38 using the R packages "GenomicFeatures" (Lawrence et al., 2013) and "VariantAnnotations" (Obenchain et al., 2014), from Bioconductor version 3.2.3 (Huber et al., 2015). The packages allow annotation of regions using information from the UCSC genome browser (Rosenbloom et al., 2015) based on interval overlaps.

A copy of the NCBI RefSeq (curated) interval set (Pruitt et al., 2014), downloaded from the UCSC browser downloads server (URL: http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/ncbiRefSeqCurated.txt.gz), was converted into GTF format using the UCSC `genePredToGtf` utility (URL: http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/genePredToGtf), and imported into R using the `makeTxDbFromGFF` function from the "GenomicFeatures" package. Non-protein coding genes were filtered out from the final interval list. For all RefSeq protein coding sequences, a TR was classed as: 'coding', 'intron', '3′ UTR', or '5′ UTR' if the TR was completely contained by the specified region; 'splice site' if it overlapped the first or last two nucleotides of an intron; 'promoter' if any portion of it overlapped a

region extending from 2000bp upstream to 200bp downstream of a transcription start site. Some intragenic TRs overlapped multiple classes, either because they spanned multiple regions, or because they could be multiply classified due to alternative gene splicing/transcription start sites. To eliminate overlapping classes, these were labeled "Promoter and other intragenic" or just "Other intragenic" depending on whether a promoter was one of the classifications; and 'intergenic' if it did not overlap any of the proceeding regions. Table 3.4 summarizes the annotations. VariantAnnotations was unable to map 1,212 TRs (54 of which were VNTRs) for unknown reasons and these were excluded from the table.

We retrieved the GO terms of genes overlapped by "coding" VNTRs, and counted the number of VNTRs for each term. VNTRs from the 3 most frequent GO terms were selected and were then searched for in the ClinVar (Landrum et al., 2016), dbSNP (Sherry, 2001), and PubMed (Noa, 2017) databases for any supporting evidence of a previously annotated VNTR. If an indel variant was found in any of the above databases, with a size change expected for the matching locus in our dataset, we recorded the annotation and any identifiers.

### 3.2.4 VNTR Detection

Read sets were processed with VNTRseek (Gelfand et al., 2014) (https://github.com/yzhernand/VNTRseek) using default parameters: a minimum flanking sequence length of 10 nt on each side of the array, a maximum flank length of 50 nt, and at least two reads mapped with the same array copy number required to make an allele call. Output from VNTRseek included two VCF files containing genotype calls, one reporting all detected TR and VNTR loci, and the other limited to VNTR loci only. VCF files contained two specialized FORMAT fields: SP, for number of reads supporting each allele, and CGL, for number of copies gained or lost with respect to the reference. For example, a CGL of $-1$ indicated an allele with one less copy

compared to the reference, and a CGL of 0 indicated the reference allele.

### 3.2.5 Refinement of allele and genotype calls

Marzie Rasekh, a PhD student in the Benson Lab, developed a method to eliminate likely false positive allele calls and refine genotype calls, which is called mlZ and summarized briefly here. mlZ is a machine learning approach based on comparison of the expected and observed number of reads supporting ($RS$) an allele, the zygosity as given by VNTRseek, whether or not a gain or loss of one copy is observable, and other features derived from the model of the expected read support.

Expected read support was determined from a combination of theoretical and observed read support distributions. Theoretical distributions were modeled using the read length, read coverage, fragment length distribution of a sequencing experiment, and simulated fragments placed randomly through the genome. The simulated fragments could span heterozygous or homozygous TR loci, and both distributions were modeled.

Observed read support from VNTRseek results were then sorted into 10bp bins by the observed array length. Outliers above 3.5 standard deviations (sd) in the homozygous distribution, or below 3.5 sd in the heterozygous distribution, were removed. The Z-scores for each allele in each distribution were then included as features, along with all previously mentioned features, in a decision tree which produced a final score and a revised genotype call.

| Dataset | Samples | Genotyped | Singleton | Raw VNTRs | | | Refined VNTRs | | | Reported |
|---|---|---|---|---|---|---|---|---|---|---|
| | | TRs | TRs | All | Singleton | Multi | All | Singleton | Multi | VNTRs |
| 1000 Genomes | 330 | 181,594 | 153,275 | 5,140 | 3,569 | 4 | | | | 3,565 |
| CEPH 1463 | 16 | 180,496 | 152,588 | 2,552 | 1,651 | 11 | | | | 1,640 |
| WGS500 | 8 | 177,282 | 149,989 | 2,077 | 1,407 | 10 | | | | 1,397 |
| Tumor/Normal | 4 | 177,592 | 150,545 | 2,044 | 1,292 | 7 | | | | 1,285 |
| CHM1 | 1 | 186,209 | 159,753 | 1,763 | 1,284 | 167 | | | | 1,117 |
| CHM13 | 1 | 197,502 | 170,140 | 2,912 | 2,155 | 356 | | | | 1,799 |
| Yoruban Trio | 3 | 204,189 | 175,005 | 5,686 | 4,388 | 57 | 5,385 | 4,242 | — | 4,242 |
| Ashkenazi Trio | 3 | 204,270 | 175,074 | 4,867 | 3,712 | 39 | 4,582 | 3,569 | — | 3,569 |
| Chinese Trio | 3 | 206,934 | 176,980 | 6,798 | 5,035 | 167 | 3,991 | 3,083 | — | 3,083 |
| NA12878 | 1 | 193,185 | 164,994 | 3,788 | 2,635 | 52 | 2,176 | 1,670 | — | 1,670 |
| Combined | 370 | 211,079 | 180,127 | 13,205 | 9,932 | 698 | 11,248 | 8,457 | 505 | 7,952 |

**Table 3.3: TRs and VNTRs detected, by dataset.** Genotyped TRs is the number of distinct TR loci genotyped in at least one individual within a dataset. Singleton TRs are those not annotated as indistinguishable in the reference set. Raw VNTRs are called by VNTRseek. Refined VNTRs are called in a post-processing step following mlZ analysis. Multis reported under refined VNTRs are only called in the remaining genomes, as mlZ processing ignores multis. Within those categories, "All" includes both indistinguishables and singletons. Multi are singleton loci for which at least three alleles were detected in a single individual (two in the haploid samples). Reported VNTRs is the number of singleton VNTRs minus the multis from the initial VNTR calls.

## 3.3 Results

### 3.3.1 TRs and VNTRs Detected

370 sequencing read datasets were analyzed with VNTRseek to discover minisatellite VNTRs. Table 3.3 summarizes our results. A total of 211,079 TR loci were genotyped across all samples (92.2% of the TRs in refset228486). 871 loci would not have been detected since they were not included in any of the read-length-specific reference sets. 89.7% (14,826) of the remaining 16,536 loci could not be detected because their arrays were too long to fit within the longest reads in our data sets, even with a loss of one copy. 13,205 of the genotyped loci were called as VNTRs. Of these, 3,273 were indistinguishables and were removed from further analyses (except in subsection 3.3.2) leaving 9,932 Singleton VNTRs. An additional 698 loci (hereafter referred to as "multi") were genotyped with more than $n$ alleles in at least one genome (where $n$ is the ploidy of the sample) and were also removed, for a final count of 9,234 singleton, non-multi VNTRs. Three VNTRs were called "multi" in over 5% of our sample set and were not flagged as indistinguishable by VNTRseek: 182621445, 182713833, and 183258087. One was classified by RepeatMasker as "simple repeats" and all were found within a segmental duplication. The alleles detected for each were consistent in the sense that they were also called in over 5% of our samples, with one exception: the +3 allele of TRID 183258087 which was detected only twice.

### 3.3.2 Genotype and allele refinement

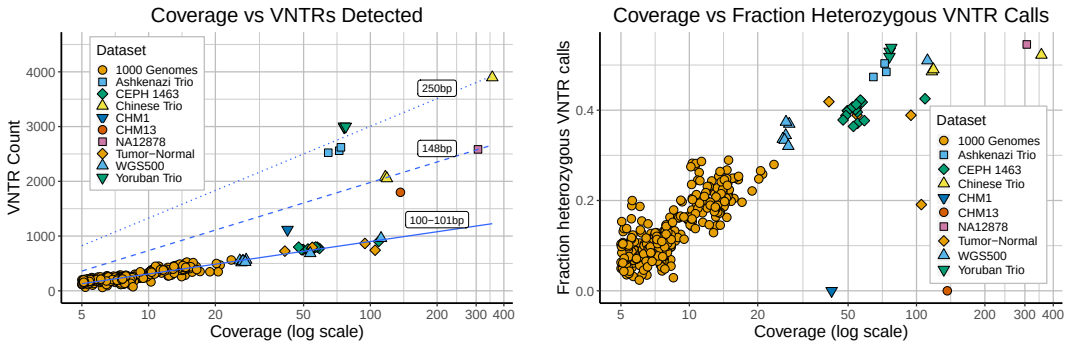
VCF output from ten high-coverage ($> 100x$), PCR-free, long-read-length (read length $\geq 148$) genomes were post-processed using MLZ. Indistinguishables are included, but multi TRs are ignored by mlZ and not processed. The results post-processing are given in columns 8-11 for the last 5 rows of table 3.3.

### 3.3.3 Relationship of detection to coverage and read length

The ability to genotype TR loci was strongly dependent on coverage (Supplementary Figure 3.S2). Assuming a locus could be genotyped if the read length was at least as long as the reference array length plus the minimum flanking sequence lengths, the percentage of singleton TRs genotyped ranged from a low of 23.85% for one of the 1000 Genomes samples (HG01437, 101 bp reads, ∼6x coverage) to a high of 98.02% for the Chinese trio child (NA24631, 250 bp, ∼355x). The lowest percentage for the 250 bp samples was 95.79% (CHM13, ∼137x).

VNTR discovery was directly related to both coverage and read length. Figure 3·1a shows a linear relationship between the log of the coverage and the number of VNTRs detected. Samples from the low coverage 100/101 bp 1000 Genomes dataset yielded an average of 262 VNTRs and the highest number of VNTRs detected in the the 100/101 bp samples was 961. In contrast, the longer read datasets produced more VNTRs. Longer reads can span longer arrays (see subsections 3.3.1 and 3.3.8) and they also increase the overall probability of detecting shorter arrays and alleles that have gained in length relative to the reference. In the 250 bp samples, VNTRseek detected between 1,799 and 3,897 VNTRs (up to 2,849 VNTRs after the refinement from subsection 3.3.2). Ploidy also has an effect. Both CHM1 and CHM13 have fewer VNTRs than expected given their coverage and read length. In these haploid samples, heterozygous loci in the underlying diploid genomes will often exhibit only the reference allele and will therefore not be counted as VNTRs.

Coverage also affected the ability to detect heterozygosity at a VNTR locus. Figure 3·1b shows a linear relationship between the log of the coverage and the proportion of VNTRs that were genotyped as heterozygous. For the highest coverage genomes, the proportion reached an apparent maximum at just over 50%. Notably, one of the cancer cell line samples (HC1187) had a significantly reduced proportion of het-




**(a)** **(b)**

**Figure 3·1: Influence of coverage and read length on genotyping.** **(a)** Relationship between log of the coverage and raw VNTRs detected. Regression lines were drawn for samples with the same read length. Points for the 100/101 bp read datasets are well clustered around the lower trend line. Much higher VNTR counts were obtained for samples with read length $> 101$ bp. Both haploid genomes, CHM1 (150 bp) and CHM13 (250 bp) have fewer VNTRs than expected because heterozygous loci with one reference allele will appear to be VNTRs only about half the time on average. **(b)** Relationship between coverage and heterozygous VNTR calls. Low coverage reduces the probability of finding both alleles when a locus is heterozygous, leading to erroneous homozygous genotype calls. The fraction of loci that were called heterozygous peaked at just over 50% for the high coverage genomes. The haploid CHM1 and CHM13 genomes should have no heterozygous loci (the few singleton loci with more than one allele were classified as "multi" and not used in this figure). One cancer cell line sample had a significantly reduced number of heterozygous VNTR calls indicating possible wide-spread loss of heterozygosity.

erozygous calls, despite having a comparable number of detected TRs, normalized by coverage, with respect to the corresponding normal sample, possibly reflecting widespread loss of heterozygosity mutations.

Other factors influencing heterozygous VNTR calls are discussed in subsections 3.3.6 and 3.3.2.

### 3.3.4 Sample support for VNTR calls.

Figure 3·2a shows the distribution of the number of samples that supported each VNTR genotype call. Close to one-third of the VNTR loci (3,117) were detected as variant in only one genome sample, suggesting that sampling was not extensive enough, that many were rare variants, or that many were artifactual.

Supporting the limited sampling hypothesis, 73% of single sample loci (2,277) were observed in long-read samples, of which there were only nine, and 68% of those (1,549) had array lengths too long to be detected with 100 bp reads.

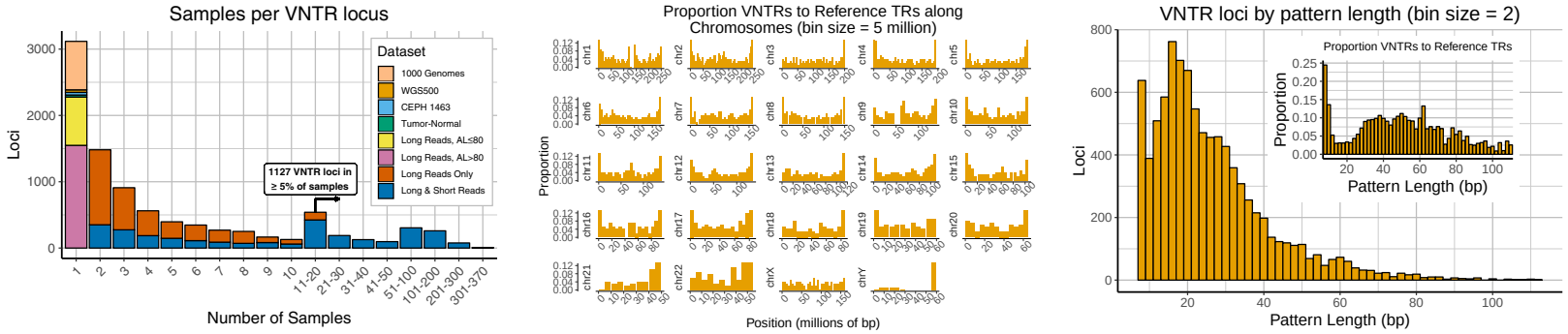
Supporting the rare variants hypothesis, for the 1000 Genomes samples, the average number of VNTR loci/sample not found in any other genome was only 2.52. Out of a random sample of 10 of these, all appear to be accurate calls (Supplementary Figures 3.S6-3.S15).

1,127 VNTR loci were detected in at least 5% of the samples (19 samples) and can be considered common variants. This is likely an underestimate, again because many of the loci could be detected only in the long read samples.

### 3.3.5 Distribution of VNTR loci

Across all chromosomes, an average of 60.9 reference set TRs and 3.1 VNTRs were present every 1 Mb (Supplementary Table 3.S1). VNTR calls exhibited a bias towards the chromosome ends (centromere TRs were excluded from this study). This was true even when accounting for the fact that the proportion of reference set loci was also

biased towards the chromosome ends (Figure 3·2b and Supplementary Figure 3.S1). In four chromosomes at least 40% of the reference TR loci were located within the first and last 10Mb (CHRs 19, 20, 21, 22, Supplementary Figure 3.S1a). For VNTR loci, the bias was even more pronounced, with 11 chromosomes having 40% of the VNTR loci located within the first and last 10Mb (CHRs 7, 8, 10, 12, 13, 16, 17, 18, 19, 20, 21, 22, Supplementary Figure 3.S1b and Table 3.S1).

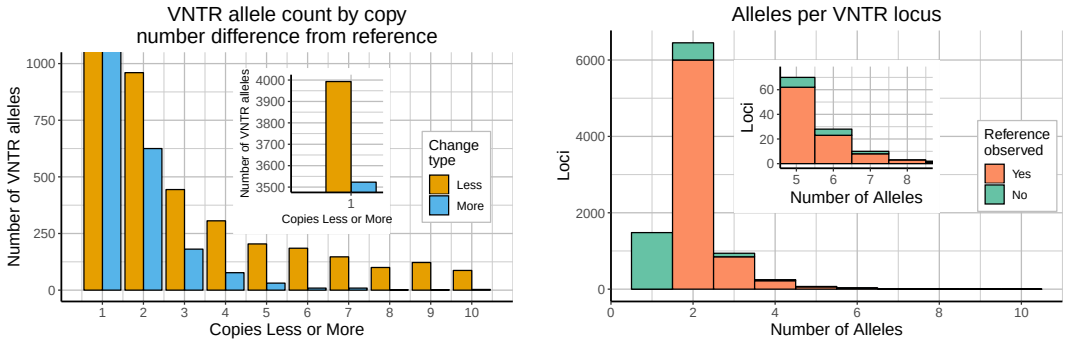



**(a)** **(b)** **(c)**

**Figure 3·2: VNTR locus characteristics.** **(a)** Number of samples in which a VNTR locus was observed. Colors in the leftmost bar indicate detection with 100 bp reads, distinguished by datasets, or long reads (250 bp, 148 bp), distinguished by the array length of the observed VNTR alleles (AL = array length). The "Long Reads $AL \leq 80bp$" category counts loci with only alleles $AL \leq 80bp$ that were not seen in 100/101 bp sets. An array length greater than 80 bp could not be observed in the short reads. Colors in the remaining bars indicate detection with long reads alone (at most 9 samples) or with both long and short reads. Note binning for number of samples $> 10$. Close to one-third of the VNTR loci were detected as variant in only one genome sample (left-most bar). However, nearly half of those were observed in only the 12 long-read samples and with array lengths too long to be detected with short reads. (Shortest array length too long in the case of multiple alleles at the same locus.) Using a cut-off of $\geq$5% of samples ($\geq$18 samples) for the definition of common variants, 1,127 loci were in this category. **(b)** Ratio of VNTR loci to reference TR loci along the chromosomes, binning every 5MB. VNTRs are more common towards chromosome ends, both in actual counts (Supplemental Figure 3.S1) and in proportion to the number of reference TRs (shown here). **(c)** VNTR pattern length distribution (bin size = 2). **Inset:** Note that the shortest pattern TRs (7-10 bp) are most likely to be variable, in comparison to their representation in the reference set, and that pattern lengths around 20 are least likely to be variable.

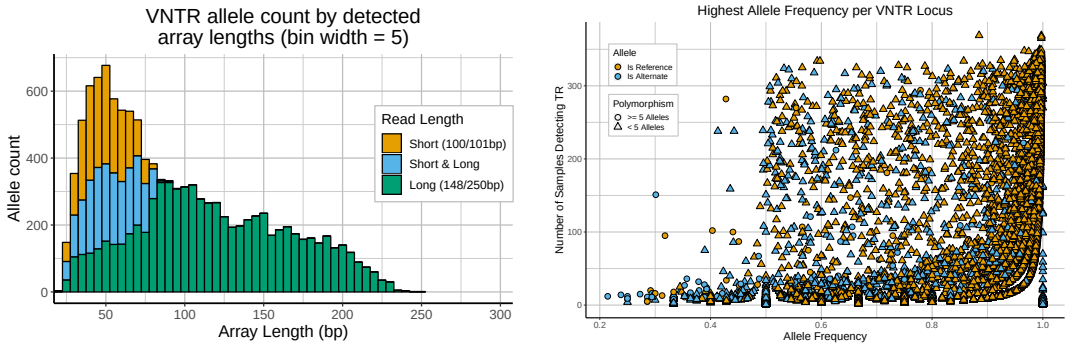
### 3.3.6 VNTR locus and allele characteristics

VNTRs with pattern lengths between 7 bp and 112 bp were detected (Figure 3·2c). The bulk of the pattern lengths (87.7%) were $\leq$ 40 bp, and only 5% of patterns were longer than 54 bp. Compared with the distribution of reference TR pattern sizes, TRs with very short patterns (7-10 bp) were overrepresented in the VNTRs and those with pattern sizes around 20 bp were underrepresented.




**(a)** **(b)**

**Figure 3·3: VNTR allele characteristics.** **(a)** Change in the number of copies in a VNTR allele relative to the reference. Approximately 64% of variant alleles exhibited a one-copy change relative to the reference (inset shows the top of the first two bars). Overall, a decrease in copy number was more frequently detected than an increase. Limited read length favored loss detection, but TR reference set bias towards arrays with fewer copies favored gain detection. **(b)** Number of alleles detected per locus across all datasets. Leftmost bar: 1,442 loci were detected with just one allele, a variant. Second bar: for 70% of the loci, two alleles were detected, and in the vast majority of those, one of the alleles was the reference allele. Overall, no reference allele was detected in 2,066 loci. Inset: in 110 loci, five or more alleles were observed.




**(a)** **(b)**

**Figure 3·4: VNTR allele characteristics (continued). (a)** Number of VNTR alleles (including reference allele) by length of detected array. **(b)** Frequency of the most commonly detected allele at a VNTR locus compared to sample representation. Allele frequency was determined by counting alleles in each sample in which the locus was genotyped. For a locus typed as homozygous (respectively heterozygous), the count for the allele was two (one). Blue symbols represent loci where the most frequent allele is a variant. Circles represent loci with at least five detected alleles.

A total of 11,667 variant alleles were detected. Of these, approximately 64% exhibited a single copy gain or loss with respect to the reference (Figure 3·3a) with loss being slightly more frequent (3,993 loss vs 3,523 gain), as it was overall. Two opposing conditions influenced detection of gain or loss. Fixed read length favored loss detection because longer arrays had a lower probability of being spanned by a read. For example, with the 100 bp reads, 20% of the reference TRs consisted of arrays that could be detected following a single copy loss, but not a single copy gain. The TR reference set, however, favored gain detection overall because a high proportion of the reference TRs contained very few copies. 77.3% contained $\leq 2.8$ copies. At this limit, loss of a single copy would cause the allele to have fewer pattern copies than the minimum required for detection by TRF, and would therefore not be detectable by VNTRseek (Supplementary Figure 3.S3). In the cases where read length had little effect on detection (*i.e.*, short array lengths), a clear bias towards gain was apparent (Supplementary Figures 3.S4, 3.S5).

Figure 3·3b shows the number of alleles detected per VNTR locus across all samples. In 22% of loci (2,066), no reference allele was found. Although absence of reference alleles may have been be due to low coverage or few long read samples, the presence of cases with high sample coverage suggests that the reference allele could be rare or incorrect. For example, 137 no-reference VNTR loci were found in 10 or more samples, and 23 were found in 100 or more samples.

Figure 3·4b displays the relationship between locus allele frequencies and sample coverage. For many loci (3,240), a variant allele had the highest frequencies. Alleles and frequencies for each VNTR locus are given in the supplementary material.

**Genome Context.** 4,849 singleton, non-multi VNTRs overlap genes from the UCSC RefSeq table (Pruitt et al., 2014; Rosenbloom et al., 2015) by at least one bp, including potential promoter regions and UTRs. Breaking down the overlap of

| Annotation | Singleton TRs | VNTRs | Percentage |
|---|---|---|---|
| Intergenic | 87,502 | 4,259 | 4.87 |
| Promoter | 3,948 | 314 | 7.95 |
| 5'UTR | 149 | 8 | 5.37 |
| Coding | 1,466 | 55 | 3.75 |
| Splice site | 535 | 70 | 13.08 |
| Intron | 80,295 | 4,170 | 5.19 |
| 3'UTR | 926 | 26 | 2.81 |
| Promoter and other intragenic | 2,803 | 240 | 8.56 |
| Other intragenic | 593 | 38 | 6.41 |

**Table 3.4: TR and VNTR annotations, by RefSeq gene features.** Shown are the number of TRs and VNTRs that overlap a given gene feature. Percentage is the ratio of VNTRs to TRs in each category. VNTRs are overrepresented in the splice site and promoter categories relative to other categories. Note, the "Intergenic" annotation applies to all TRs that do not overlap another category. "Promoter and other intragenic" applies to TRs which overlap the promoter region of a gene, or a neighboring gene, and one of the intragenic regions. Some intragenic TRs overlapped multiple classes, either because they spanned multiple regions, or due to alternative gene splicing/transcription start sites. These are labeled as "Other intragenic".

TRs by common classes of genomic regions, we see that the majority of these overlap 'intergenic' regions, followed by 'intron', and 'promoter' regions, in decreasing order (table 3.4). 75 VNTR loci overlapping a coding site (the 55 from the "Coding" row plus 20 more from the "Other intragenic" row in table 3.4), one has a pattern size which is not a multiple of 3 according to TRF. This TR primarily overlaps exon 1 of the gene LYSMD4 (Entrez ID 145748) and was detected in one sample, NA24631, with a loss of one copy. Such a copy number change would result in a reading frame shift. An alternative transcript of this gene has a transcription start site (TSS) further downstream than other transcripts for the same gene, placing it directly in the middle of the reference location of this TR. However, this allele is eliminated by mlZ post-processing due to poor read support.

Among the intragenic VNTRs, three VNTRs overlap variants annotated in the ClinVar database. One is a 45-bp VNTR in intron 5 of USH1C and is implicated

in Usher Syndrome 1 and 1C, though there is conflicting evidence of pathogenicity (ClinVar Variation ID: 20181, TRID: 182325055) (Savas et al., 2002). Another is a 30-bp VNTR in the promoter region of MAOA (ClinVar ID: 9968, TRID: 183311386), where a lower copy number is associated with Autism Spectrum Disorder (ASD) and antisocial behavior (Cohen et al., 2003). The last is a 12-bp VNTR in the 5' flanking region of CSTB (ClinVar ID: 55956, TRID: 182814480), where a large increase in pattern copies is associated with Unverricht-Lundborg syndrome (also known as EPM1), a neurodegenerative disease (Lafreniére et al., 1997). Wild-type alleles for the CSTB VNTR are two to three copies (reference has 3 copies), while pathogenic alleles have over 40 copies. The CSTB VNTR is the only one short enough to be observed in genomes outside the 250bp samples, and overall we detect the benign two and three copy alleles. Likewise, for the USH1C VNTR, only benign alleles are observed as the reportedly pathogenic allele would be too long to be spanned by our longest reads. A potentially pathogenic allele is observed in one individual of the Y117 family for the MAOA VNTR, where a 2 copy VNTR is associated with ASD or antisocial behavior. The individual is heterozygous at that locus and the pathogenic allele is supported by 8 reads. ClinVar classifies these VNTRs as "microsatellite" loci while dbSNP classifies the USH1C and MAOA variants as "indels" (the CSTB VNTR does not appear in dbSNP).

The majority of commonly detected loci (see 3.3.4) are found in intergenic or intronic regions. Genes of potential interest among the intronic VNTRs are ZNF544 (a zinc finger protein, involved in regulation of RNA polymerase II), TP53 (a tumor supressor protein), and PCDH15 (a calcium-binding protein in which mutations may result in hearing loss and Usher Syndrome Type 1F). A selected list of these can be seen in supplementary table 3.S2.

| Trio | Loci genotyped in all | All heterozygous | Inconsistent | All heterozygous, all different | Inconsistent |
|---|---|---|---|---|---|
| Y117 | 3,485 | 353 | 3 | 56 | 1 |
| WGS500 Trio 1 | 221 | 21 | 2 | 0 | NA |
| WGS500 Trio 2 | 187 | 26 | 0 | 0 | NA |
| CEPH Trios | (904, 1013) | (35, 61) | 0 | (0, 3) | 0 |
| GIAB AJ Trio | 2,901 | 268 | 0 | 30 | 0 |
| GIAB Chinese Trio | 2,302 | 194 | 2 | 9 | 0 |

**Table 3.5: Consistency with Mendelian inheritance of VNTR genotypes in trios.** Only loci detected in all members of a trio were considered (column 2). When all genotypes at a locus are called as heterozgyous, only 7 loci are inconsistent. Requiring that all genotypes be different as well further significantly reduces the number of loci under consideration and yields only 1 locus as inconsistent. CEPH family results are summarized, with the lowest and highest values seen throughout all 13 trios given in parenthesis.

### 3.3.7 Consistency of Genotype Inheritance.

**Consistency with Mendelian inheritance.** Consistency means that the genotype of a child can be explained as one allele from the mother and one from the father. It was evaluated for all trios in the datasets, i.e., the two trios in the WGS500 dataset, the Ashkenazi, Yoruban, and Chinese trios, and all 13 possible trios from the CEPH 1463 family. A locus was considered for evaluation if it was detected in all members of the trio and called heterozygous in all. A second stricter criterion additionally required that all three genotypes had to be different. Failure to take these criteria into consideration could lead to false interpretations of consistency. For example, in violation of the all called heterozygous criterion, let the mother be A|B, the father be B|C and the child be A|C. If both parents are detected as heterozygous, but only allele A is detected in the child, then the genotype appears to be A|A and the result appears to be inconsistent because the father has no A allele. Similar situations arise when only one allele is detected in one of the parents. In violation of the all heterozygous and all different criteria, let the parents be as above, the child be A|B, and all called

heterozygous. In this case the data are consistent with Mendelian inheritance, but do not exclude the possibility that VNTRseek is systematically categorizing single alleles $A$ and $B$ from two independent loci as two alleles from the same locus.

Under both criteria, only a handful of loci were inconsistent (Table 3.5). No inconsistencies were found in the Ashkenazi and CEPH trios and one WGS500 trio. The other WGS500 trio, showed two inconsistencies as did the Chinese trio. The Yoruban trio showed three inconsistencies under the lenient criterion and one under the strict criterion.

The single Yoruban inconsistency at the strict criterion (locus: TRID 182759931) appears to be an artifact. The maternal and paternal genotypes were -1/+2 and 0/+2, respectively, while the child's genotype was +1/+2. However, the -1 and +1 alleles had support of only 2 reads each while the other alleles had support consistent with the coverage and yielding a homozygous genotype in the mother and child (+2/+2 in the mother – 29 reads, +2/+2 in the child – 33 reads) and a heterozygous genotype in the father (0/+2 – 13/16 reads). Closer examination of the mapping alignments showed that the -1 and +1 reads did not fully span the arrays, and the left ends were mapped incorrectly into the left flank with a number of errors below our threshold.

### 3.3.8 Characteristics of the reference set that potentially preclude allele detection

TRs which contain fewer than 2 copies of their pattern would not have copy losses detected by VNTRseek because the minimum copy number TRF can detect is 1.8. Approximately 75% of our reference set has a reference copy number of 2.7 or below. Therefore, whole copy loss in these TRs would be invisible to us. Figure 3·3a shows how this may be significant, as the most common variant observed is a loss of one copy with respect to the reference. Additionally, TR alleles with array length plus minimum flanking sequence longer than a read would not be detected by VNTRseek.

16% (36,631) of the singleton, non-"multi" TRs in our reference set have a reference array length longer than 80bp (the maximum array length a 100/101bp read can span) and 0.04% (10,248) are longer than 230bp (the maximum array length for the 250bp read sets).

## 3.4 Discussion

We present a thorough investigation of VNTRs in 370 WGS data sets, and detail how these variants may be significant sources of variation by placing them in a genomic context and demonstrating their variability.

Given the limitations of both our methods, and the available data, it is clear that more variants are likely yet to be discovered. Analyzing these will require longer reads, higher quality data, and in some cases novel methods. Our lab is already investigating methods to analyze variation in loci where TRF and VNTRseek are unable to detect TRs. However, we argue that the data presented here are compelling enough and complete enough so as to act as a resource for further study. The variants we discovered span all regions of the genome, both coding and non-coding, and most do not appear in curated databases. We hope that our work enables others to pursue yet another avenue of research as we uncover more insight into the variability of the human genome.

The variants discovered here will be made available both in data repositories such as dbSNP, and our own managed resource, a database of VNTRs.

### 3.4.1 Data

Data for the 370 genome samples used in this study were obtained from the URLs in table 3.6.

| Dataset (individuals) | URL or Accession numbers |
|---|---|
| 1000 Genomes (330) | ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/ |
| CEPH 1463 (16) | http://www.illumina.com/platinumgenomes/ |
| HapMap Y117 trio (3) | https://www.ebi.ac.uk/ena/data/view/PRJEB4252 NCBI BioProject: PRJEB4252 |
| WGS500 (8) | https://www.ebi.ac.uk/ena/data/view/PRJEB9151 |
| CHM1 (1) | https://www.ncbi.nlm.nih.gov/sra/SRX652547 |
| CHM13 (1) | NCBI SRA: SRR1997411, SRR3189741, SRR3189742, and SRR3189743 |
| GIAB AJ Trio (3) | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio |
| GIAB Chinese Trio (3) | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio |
| GIAB NA12878 (1) | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NIST_NA12878_HG001_HiSeq_300x/ |
| Tumor/Normal Pairs (4) | https://basespace.illumina.com/projects/38600562 |

**Table 3.6: Links to dataset sources.** Datasets used in this study were collected from publicly available sources. URLs are for the repositories containing the data, or the specific project or experiment page with download links. In all cases, gzipped FASTQ files were used.

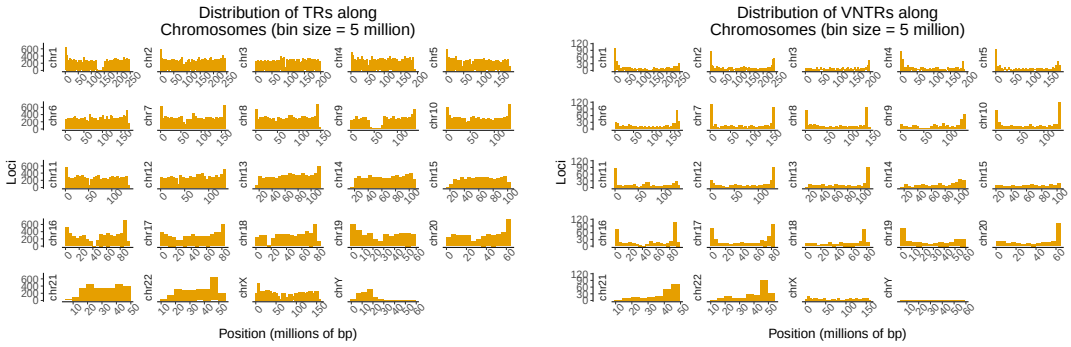
| Chr | Avg. per Mb | | Fraction in chr ends | | Total | |
|---|---|---|---|---|---|---|
| | TRs | VNTRs | TRs | VNTRs | TRs | VNTRs |
| chr1 | 57.18 | 2.65 | 0.14 | 0.29 | 14,238 | 661 |
| chr2 | 64.84 | 3.01 | 0.14 | 0.31 | 15,757 | 731 |
| chr3 | 61.31 | 2.22 | 0.10 | 0.22 | 12,200 | 441 |
| chr4 | 68.91 | 2.93 | 0.13 | 0.34 | 13,162 | 559 |
| chr5 | 63.31 | 2.73 | 0.15 | 0.39 | 11,522 | 497 |
| chr6 | 66.19 | 3.04 | 0.16 | 0.35 | 11,319 | 519 |
| chr7 | 68.34 | 3.40 | 0.22 | 0.41 | 10,934 | 544 |
| chr8 | 68.92 | 3.27 | 0.21 | 0.42 | 10,062 | 477 |
| chr9 | 53.53 | 2.68 | 0.24 | 0.39 | 7,441 | 373 |
| chr10 | 69.34 | 3.88 | 0.27 | 0.48 | 9,291 | 520 |
| chr11 | 62.67 | 2.92 | 0.20 | 0.38 | 8,523 | 397 |
| chr12 | 66.37 | 3.01 | 0.20 | 0.43 | 8,893 | 403 |
| chr13 | 61.70 | 2.46 | 0.23 | 0.44 | 7,095 | 283 |
| chr14 | 53.08 | 2.40 | 0.19 | 0.33 | 5,733 | 259 |
| chr15 | 43.79 | 1.76 | 0.16 | 0.28 | 4,467 | 180 |
| chr16 | 65.82 | 4.19 | 0.39 | 0.60 | 5,990 | 381 |
| chr17 | 66.85 | 4.95 | 0.38 | 0.60 | 5,615 | 416 |
| chr18 | 67.21 | 3.35 | 0.32 | 0.48 | 5,444 | 271 |
| chr19 | 76.54 | 4.59 | 0.45 | 0.59 | 4,516 | 271 |
| chr20 | 69.08 | 4.14 | 0.44 | 0.60 | 4,490 | 269 |
| chr21 | 66.57 | 3.81 | 0.40 | 0.67 | 3,129 | 179 |
| chr22 | 56.24 | 4.47 | 0.45 | 0.61 | 2,868 | 228 |
| chrX | 46.81 | 1.43 | 0.16 | 0.19 | 7,349 | 224 |
| chrY | 17.33 | 0.16 | 0.25 | 0.33 | 1,005 | 9 |
| Average | 60.91 | 3.06 | Total | | 191,043 | 9,092 |

**Table 3.S1: Distribution of singleton TRs and VNTRs per chromosome.** Reference TRs and VNTRs are not distributed uniformly in the chromosomes. Both are overrepresented in the first and last 10 Mb of the chromosomes (excluding telomeres) listed as "chr ends" above, with the VNTR proportion more pronounced than the TR proportion. Note that percentages in the chromosome ends naturally increases as the chromosome size decreases.

## 3.5 Supplementary Material

| TRID | Gene | Location | dbSNP rs | ClinVar ID | Sample Calling | PMIDs | Comment |
|---|---|---|---|---|---|---|---|
| 183169331 | IRF5 | Exon 6/9, Intron 5/7, Intron 6/9 | rs60344245 | | AW_CRS_1631 | 23049601; 15805103 | WGS500 trio unaffected mother and others. Is multi in CHM13 only. |
| 182388468 | KRT2 | Exon 1 | rs763805940 | | HG005 | 9804344 | GIAB Chinese child |
| 182318145 | DRD4 | Exon 3 | rs765323854 | | HG006 | 24229552 | GIAB Chinese father |
| 182318121 | CDHR5 | | | | HG005 | | GIAB Chinese child. Entrez gene says repeats are non polymorphic |
| 182574350 | GP1BA | Exon 2, Intron 2 | rs886038267 | 255466 | HG006 | 1577776; 26191334 | GIAB Chinese father. ClinVar says likely benign |
| 183311386 | MAOA | Exon 1 | | 9968 | NA19238 | 12919132 | Y117 Mother. ClinVar: Pathogenic; risk factor |
| 182325055 | USH1C | Intron 5, Intron 4 | rs55983148 | 20181 | NA19240 | 11810303 | Y117 child. ClinVar: Conflicting interpretations of pathogenicity |
| 182814480 | CSTB | Promoter | rs386833438 | 55956 | HG03616 | 9054946 | Sample: 1000 Genomes, BEB population. ClinVar: Pathogenic. 0 and -1 alleles detected. Literature has +1 as pathogenic. |
| 182468054 | NPAS3 | | rs1038697388 | | HG02941 | | Sample: 1000 Genomes, ESN population. -1 allele observed, and supported by dbSNP. |
| 182610081 | FOXK2 | Exon 1 | rs779355780 | | NA19239, NA19240 | | Sample: 2 members of Y117 pedigree (Yoruban trio). -1 allele observed and supported by dbSNP. |
| 182168889 | HES4 | Intergenic | rs36126598 | | | | We find alleles +2 to +6, dbSNP lists a deletion and duplication. |
| 183178235 | NOS3 | Intron 5, Intron 4 | rs869109213 | | NA19239 | | Downstream variant. Associated with smoking-dependent risk for coronary artery disease |
| | | | rs61722009 | | NA19240 | 23176758; 17018701; 9535806 | |
| 182328935 | BDNF | Intron 1 | rs67192910 | | CHM13, NA19238, NA19239, AJ Trio, NA24631, NA24695 | | rs931222868 is also a possible match in dbSNP for a -2 allele. We only see the -1 allele. |

**Table 3.S2: Intragenic or gene-proximal VNTRs observed as polymorphic in external databases.** The closest gene to the locus is given in the "Gene" column, and the location relative to the gene is given in the "Location" column. If the gene is known to have multiple possible transcripts, and the VNTR locus is internal to the gene, the "Location" column will have a comma separated list of locations. The closest relevant entry in dbSNP is given in the column of the same name. Should the variant have been submitted to ClinVar, then its ID in ClinVar will be given as well. Samples in which the locus has been observed are listed in the "samples" column. Some VNTRs are mentioned in literature as being associated with some phenotype, and relevant PMIDs are listed in the "PMIDs" column. Notes regarding these VNTRs can be found in the last column, "Comments".




**(a)** **(b)**

**Figure 3.S1: TR and VNTR distribution along chromosomes.** **(a)** Counts of reference TR loci per 5MB. **(b)** Counts of VNTR loci per 5MB. VNTRs are more common towards chromosome ends, both in actual counts and in proportion to the number of reference TRs.

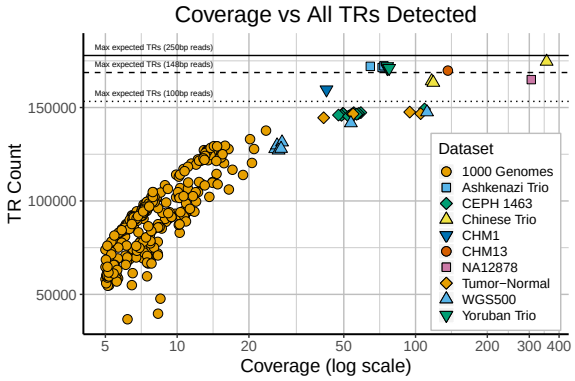
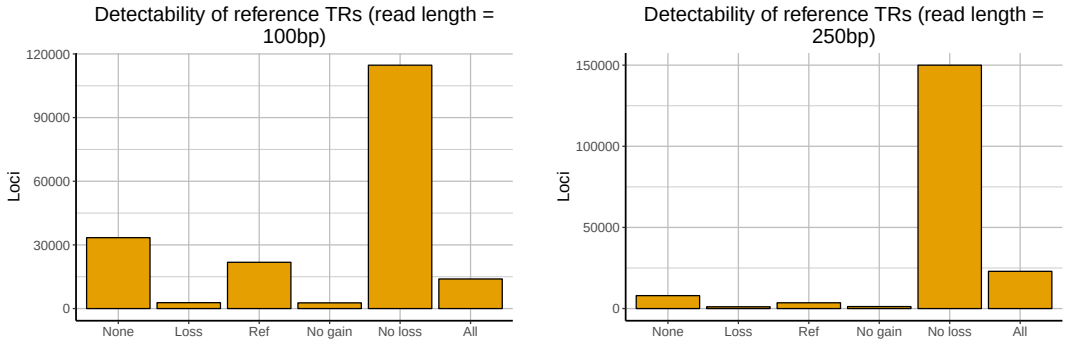



**Figure 3.S2:** Coverage vs Singleton TRs genotyped. Note the log scale on the x-axis. The fraction of loci that could be genotyped was strongly related to the read coverage, until approaching the maximum possible for a given read length (horizontal lines).




**(a)** **(b)**

**Figure 3.S3: Limitations of the reference set with respect to the ability to detect copy gain or loss.** We divide the reference set by the detectability of alleles up to a copy change of $\pm 1$. Loci in the "None" category cannot be detected. The "Loss" category comprises loci which can only be spanned by a read with a copy loss of 1. "Ref" means that only the reference allele can be spanned. "No gain" means reference and -1 alleles can be spanned, and "No Loss" means the reference and +1 alleles can be spanned, but the -1 allele cannot be detected by TRF. "All" indicates that at least the reference, -1, and +1 alleles are detectable. **(a)** 100/101 bp reads. **(b)** 250 bp reads

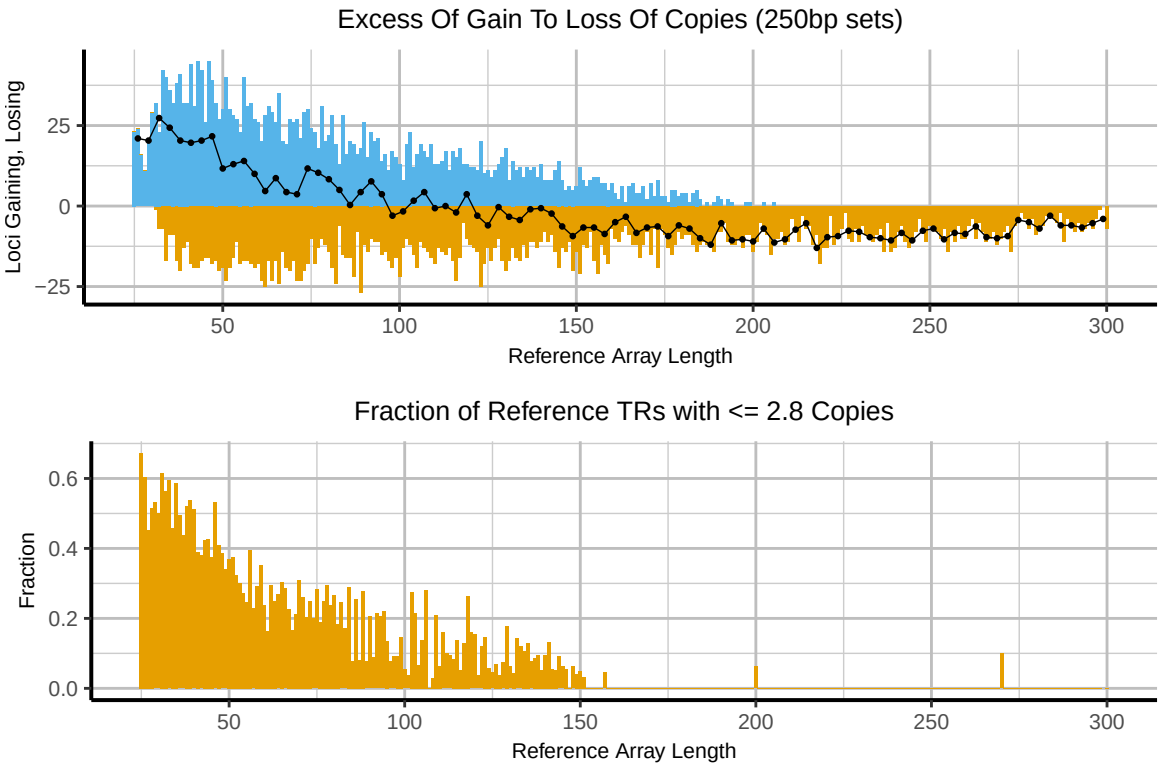



**Figure 3.S4: Gain and loss of copies for 250bp reads.** Top: At each reference array length (bin size = 1), the numbers of loci which gained (positive) and lost (negative) copies relative to the reference are shown. The black line and points are the averages of the two values. Bottom: Fraction of loci that have two few copies ($\leq 2.8$) for loss to be detected by TRF and VNTRseek. Gain is clearly dominant at shorter array lengths where the effect of read length on the ability to detect gain is minimal and gain can only be detected for a large fraction of the loci. At longer array lengths, loss dominates as gain increasingly would make the arrays longer than the read length.

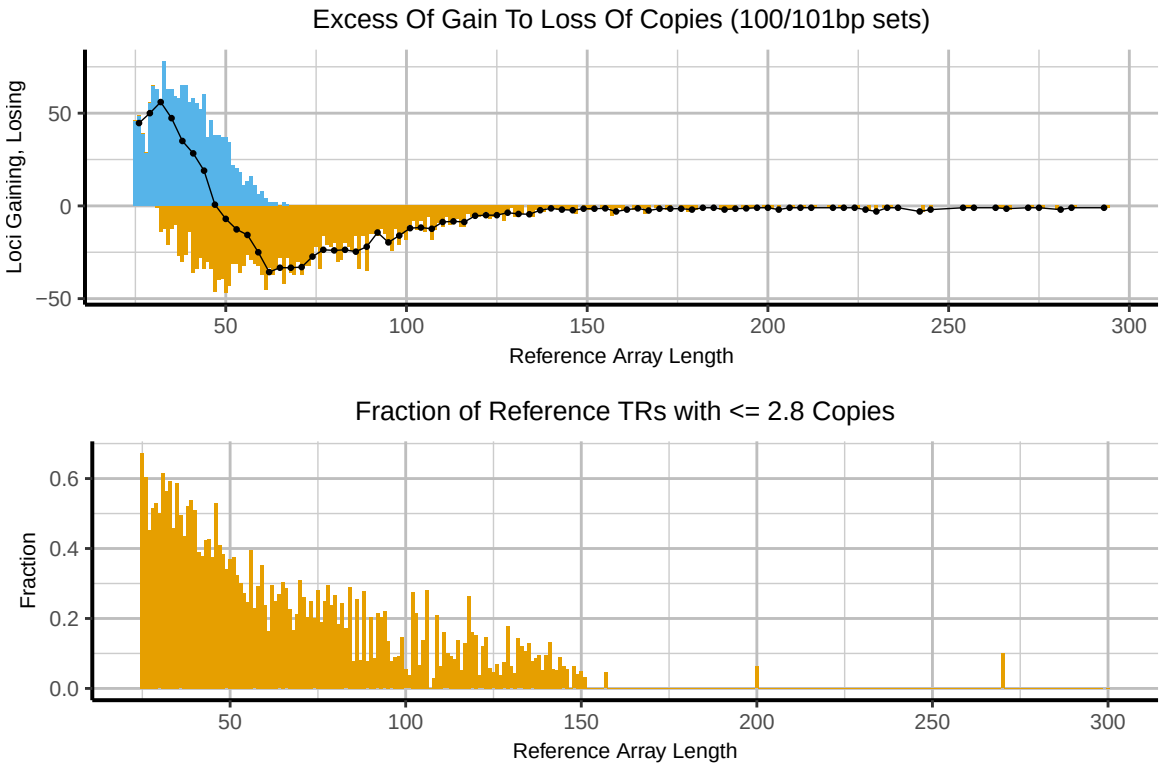



**Figure 3.S5: Gain and loss of copies for 100/101bp reads.** The effect is similar to that observed for the 250bp reads.

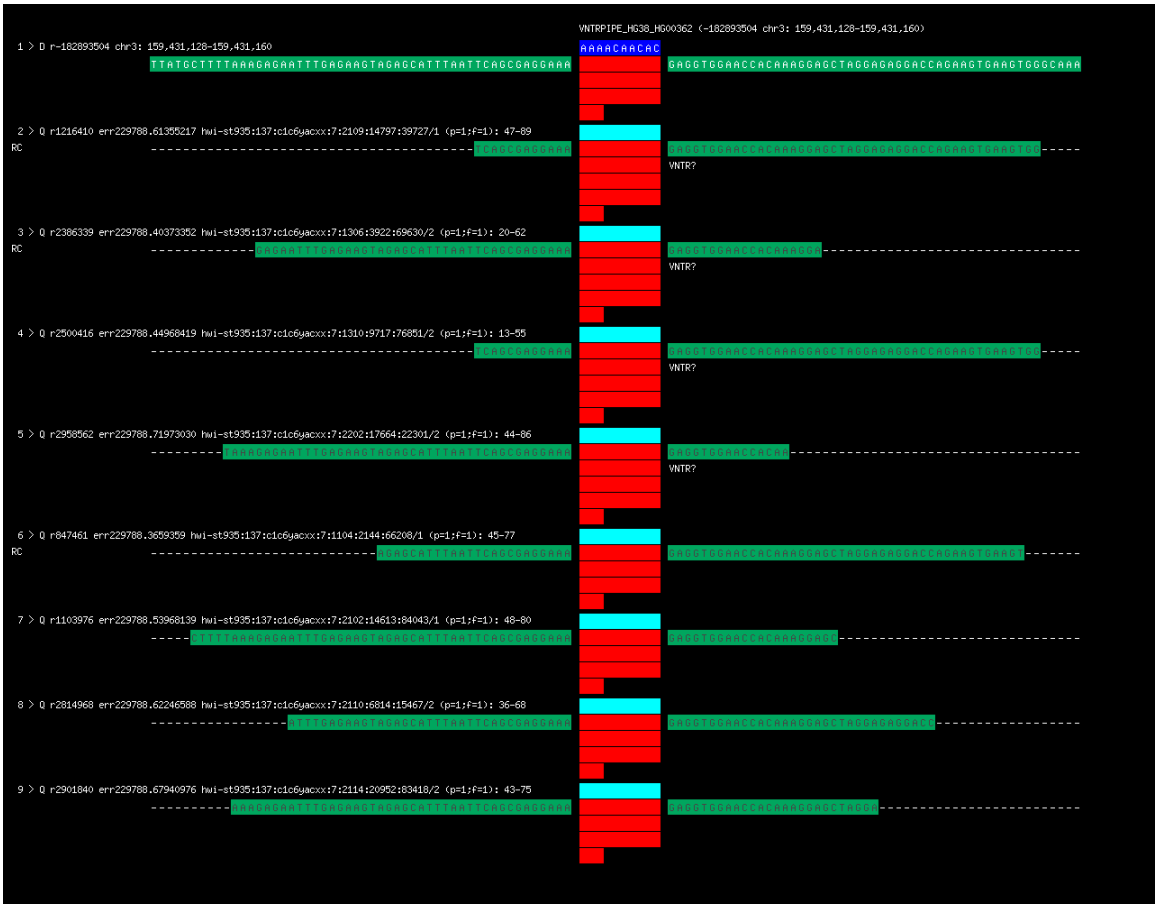



**Figure 3.S6:** VNTR unique to the 1000 Genomes HG00362 sample. Genotype is heterozygous with two observed alleles, 3.3 copies (reference) and 4.3 copies. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.


**Figure 3.S7:** VNTR unique to the 1000 Genomes HG00236 sample. Genotype is heterozygous with two observed alleles, 5.4 copies (reference) and 3.4 copies. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.

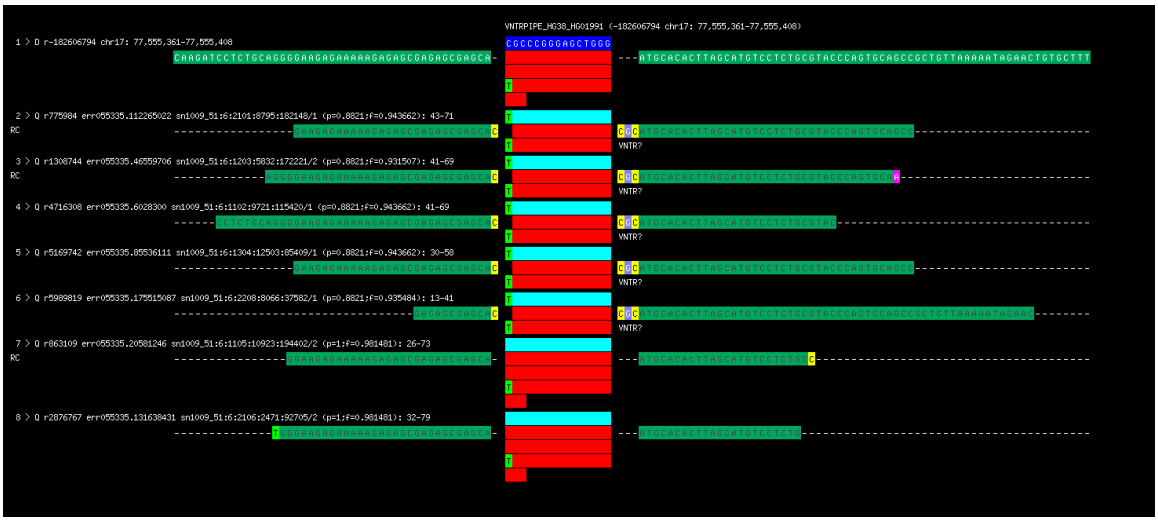



**Figure 3.S8:** VNTR unique to the 1000 Genomes HG01991 sample. Genotype is heterozygous with two observed alleles, 3.2 copies (reference) and 2.2 copies. Due to the way TRF detects the TRs, the ends of the TR are slightly inaccurate in the 2.2 copy reads because of the small number of copies. The 2.2 copy allele has lost the first copy present in 3.2 copy allele. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.

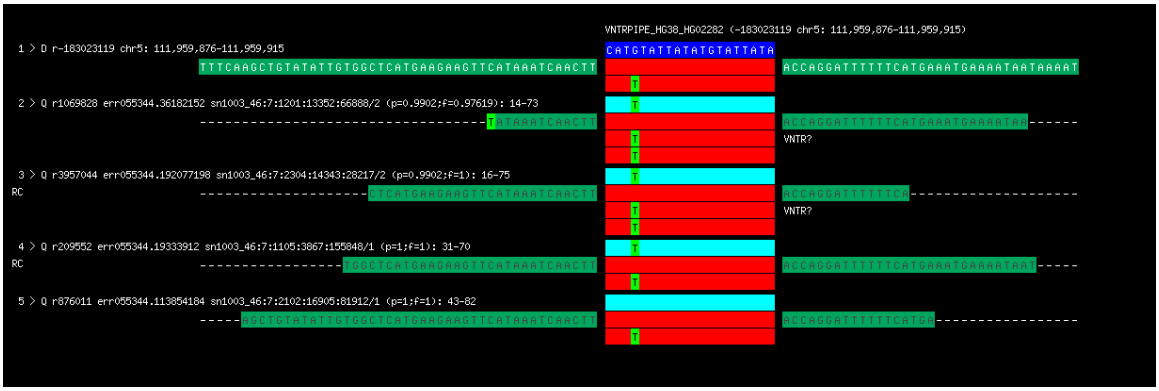



**Figure 3.S9:** VNTR unique to the 1000 Genomes HG02282 sample. Genotype is heterozygous with two observed alleles, 2 copies (reference) and 3 copies. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.

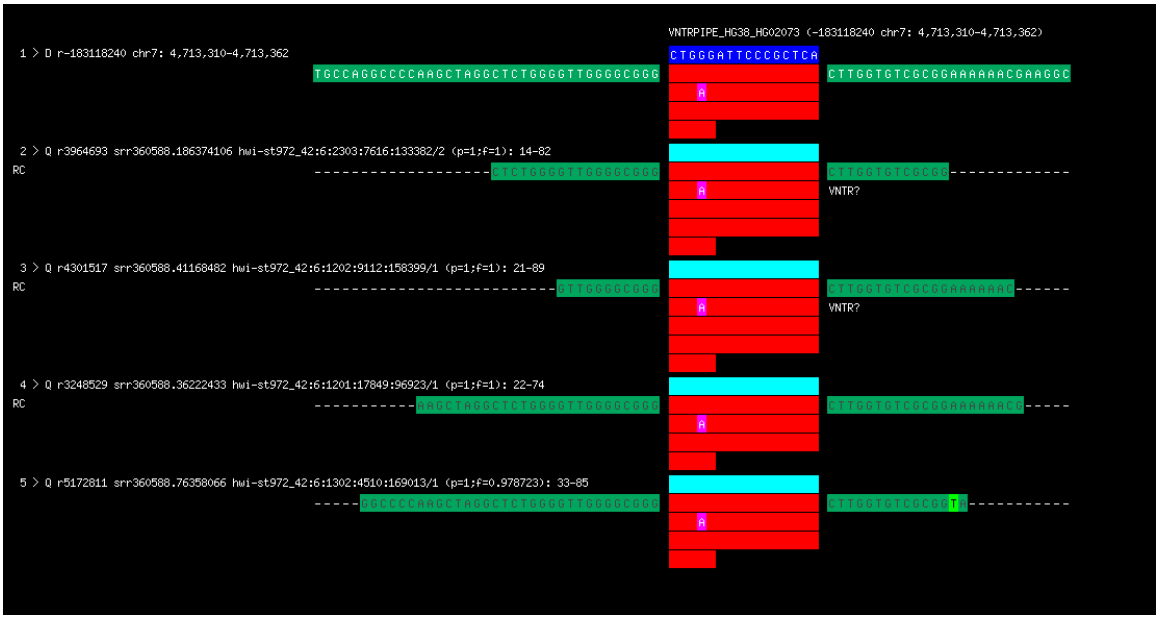



**Figure 3.S10:** VNTR unique to the 1000 Genomes HG02073 sample. Genotype is heterozygous with two observed alleles, 3.3 copies (reference) and 4.3 copies. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.


**Figure 3.S11:** VNTR unique to the 1000 Genomes HG02073 sample. Genotype is heterozygous with two observed alleles, 2.8 copies (reference) and 3.8 copies. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.

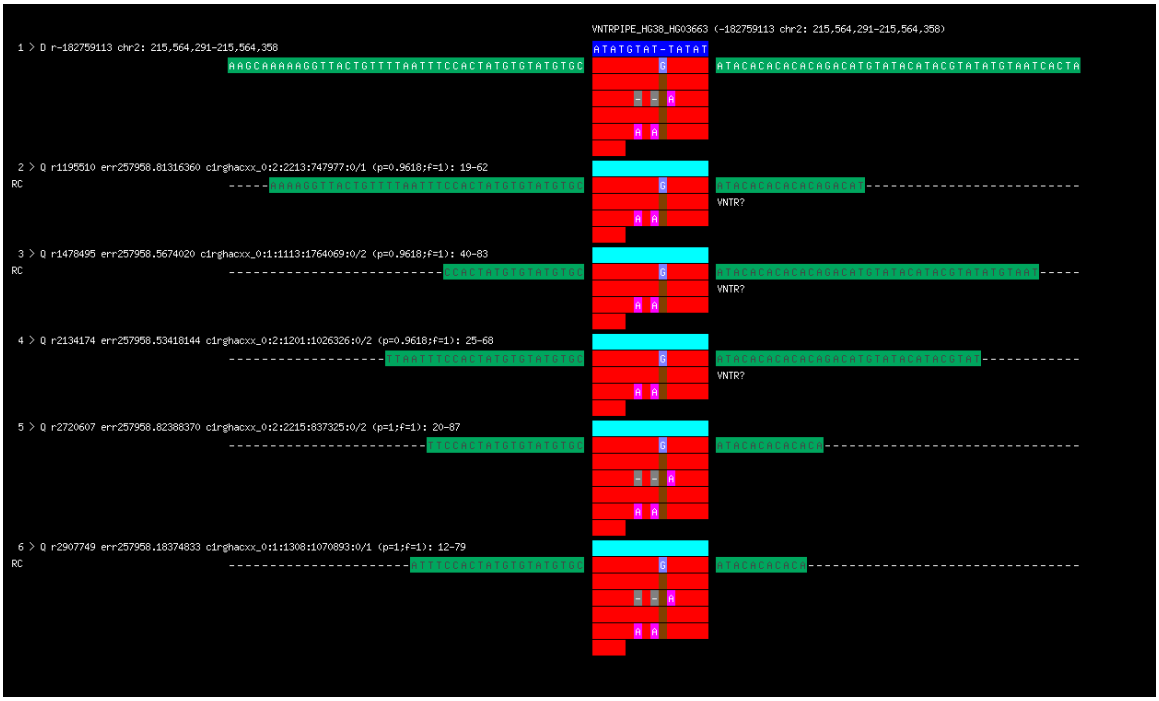



**Figure 3.S12:** VNTR unique to the 1000 Genomes HG03663 sample. Genotype is heterozygous with two observed alleles, 5.3 copies (reference) and 3.3 copies. This TR occurs in a stretch of other short TRs. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Brown indicates a gap induced by an insertion relative to the reference. Within flanks, color other than green indicates difference from reference flanks.

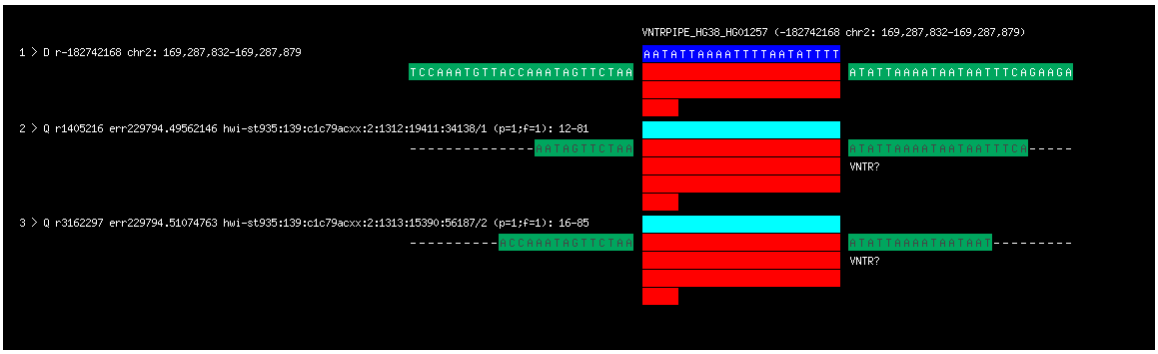



**Figure 3.S13:** VNTR unique to the 1000 Genomes HG01257 sample. Genotype is homozygous with one observed alleles, 3.2 copies. This TR occurs in an A/T rich region. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.

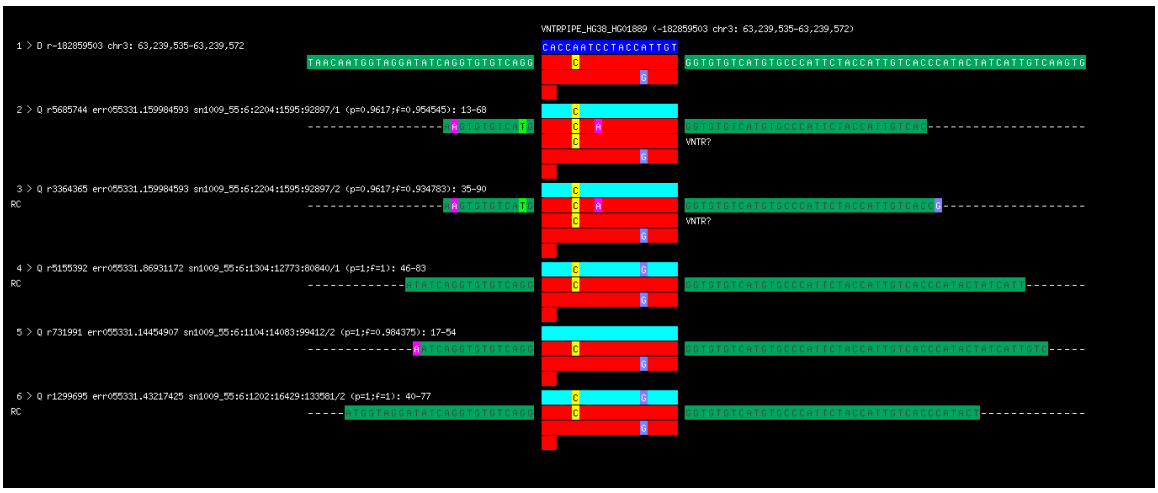



**Figure 3.S14:** VNTR unique to the 1000 Genomes HG01889 sample. Genotype is heterozygous with two observed alleles, 2.1 copies (reference) and 3.1 copies. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.

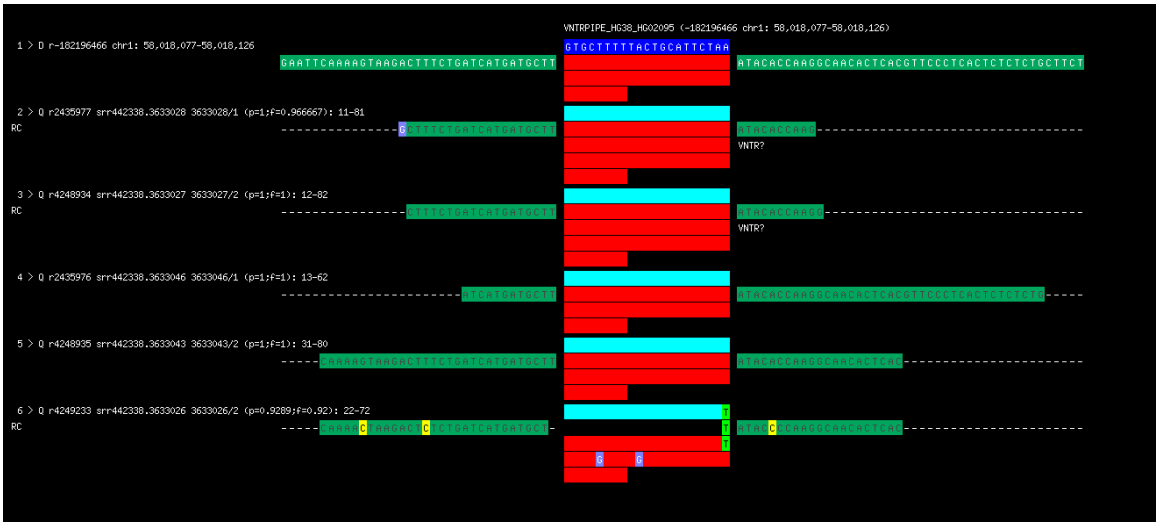



**Figure 3.S15:** VNTR unique to the 1000 Genomes HG02095 sample. Genotype is heterozygous with two observed alleles, 2.4 copies (reference) and 3.4 copies. Note the inaccuracy of the TR end in the last read due to the way TRF detects TRs. Reference is shown at the top, reads below. Green is flanking sequence. Blue is consensus sequence of reference. Aqua is consensus sequence of read. Red is tandem copies. Within arrays, color indicates match with reference consensus sequence. letters or dash indicates difference from consensus. Within flanks, color other than green indicates difference from reference flanks.

# Chapter 4

# VNTRdb – A database of VNTRs meant to facilitate the distribution and analysis of VNTR data in the human genome

VNTRdb is a database for Variable Number Tandem Repeats. VNTRdb is designed for the visualization, curation, and analysis of VNTRs. It is released under an open license and is written in Perl (with some parts in C for performance) using Mojolicious and SQLite. We developed the frontend of VNTRdb to be intuitive and straightforward to use, focusing on search and presentation of information. The initial data set available for analysis was produced in Hernandez et al. (2019). A preview of VNTRdb is currently available at http://orca.bu.edu/vntrdb.

## 4.1 Introduction

Variable Number Tandem Repeats (VNTRs) are polymorphic minisatellite loci, with alleles varying by the number of copies of the tandemly repeating pattern. Due to their instability (Jeffreys et al., 1985; Kimpton et al., 1993), VNTRs have proven to be effective for use as genetic markers and have been used to study genetic diversity (Hasan et al., 2012; Hernandez et al., 2019), and migration and breeding patterns (Wink, 2006), to identify and distinguish between bacterial strains (Blouin et al., 2012; Pourcel et al., 2011; Zaluga et al., 2013; Chalker et al., 2015; Parvej et al., 2019), and to establish paternity/familial relationships (Jeffreys et al., 1991). In the human genome, VNTRs have been detected both inter and intragenically (Brookes, 2013; Bakhtiari et al., 2018b; Audano et al., 2019; Hernandez et al., 2019) and some

have been tied to disease phenotypes (Brookes, 2013; Bell et al., 1982, 1984; Cervera et al., 2007; Leung et al., 2017).

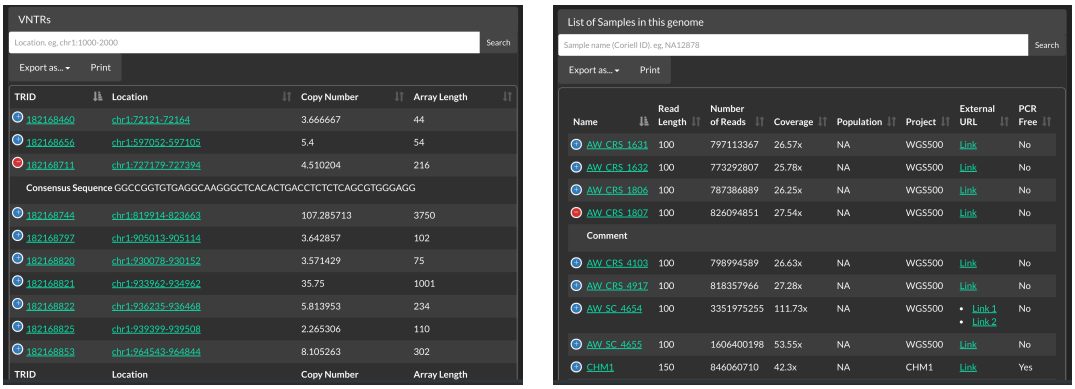
Current resources which track VNTRs are dbSNP (Sherry et al., 2001), dbVar (MacDonald et al., 2014), and the European Variation Archive (EVA) (Cook et al., 2016). Using dbSNP and dbVar for VNTRs requires additional filtering to exclude single nucleotide polymorphisms (SNPs) and overlapping variants of different lengths, or in the case of dbVar, using external tools such as the NCBI Variation Viewer or BEDTools (Quinlan, 2002). dbSNP and dbVar also place a greater focus on the exact sequence composition of the variation, while we consider VNTR alleles in the less restrictive sense of copy number change. Another database by the name of VNTRDB is described in literature (Chang et al., 2007), but it focused on bacterial VNTRs and is now apparently defunct.

At the time of this writing, VNTRdb is populated with data from an analysis of 370 whole-genome sequencing data sets from 368 individuals (chapter 3 and Hernandez et al. (2019)).

## 4.2 Database design and overview

VNTRdb is primarily written in Perl using the Mojolicious modern web framework (https://mojolicious.org). VNTRdb perfoms some on-the-fly sequence alignment when displaying TR diagrams, which is provided by code written in C and borrowed from VNTRview and VNTRseek (Gelfand et al., 2014). The backend database is constructed using SQLite and the subset of SQL which that system supports. Some SQL is used within the codebase, but we primarily rely on the SQL::Abstract package or the DBIx::Class object-relational mapping (ORM) library to generate SQL queries from possibly complex queries made to the server.

A user can browse the VNTRdb website by organism, and then browse the avail-




**(a)** **(b)**

**Figure 4·1:** Index of VNTRs (a) and samples (b) for human data in VNTRdb. The "+" icons indicate that the row can be expanded for further information, as it was unable to display the full contents of the row in the current screen size. VNTRs can be downloaded in BED and CSV formats (for any list of VNTRs), or in VCF format (samples only) from the menu on the upper left.

able data by genome sample, genomic location, or variant locus ID (figure 4·1). Variants are labeled by their ID in TRDB(Gelfand et al., 2007) and by their VNTRdb ID, and search can performed using chromosomal coordinates. VNTRdb supplies an Application Programming Interface (API), allowing programmatic access to the data from a script or third party resource.

We designed a simple REST API using the OpenAPI 2.0 specification (formerly known as Swagger, https://www.openapis.org), which allows us to describe the API as a formatted text file (either YAML or JSON) which is also machine readable. The VNTRdb website uses the API internally as well, meaning that significant parts of its functionality are exposed for external developers to use.

## 4.3 Typical use case examples

Suppose an analysis revealed the potential for a VNTR within the specific region, denoted using UCSC genome browser chromosomal coordinates, chr11:17527050-

17527210. A user would visit the page at http://orca.bu.edu/vntrdb/vntrs/Homo%20sapiens/hg38 and type in the region into the search bar above the list of VNTRs (figure 4·1a). The user will then be presented with a list of only those VNTRs with an overlap of at least one nt with the given region. Searching for VNTRs within a gene region requires knowing the coordinates of a gene on the reference assembly, although support for gene names (symbols) is planned.

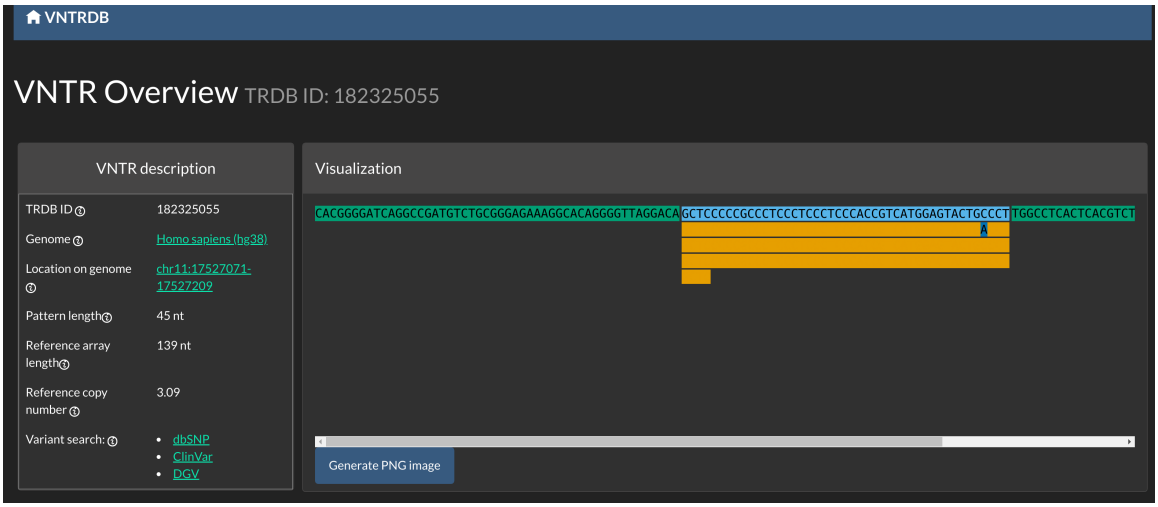
Alternatively, VNTRdb allows browsing of all samples, and their genotype calls at every TR reference locus. Researchers interested in particular samples can find them based on their Coriell ID, population, or other external ID, as in the cases of CHM1 (Chaisson et al., 2014), CHM13 (Huddleston et al., 2017), and samples from WGS500 (Taylor et al., 2015) which all have their own identifiers (figure 4·1b).

If a user is interested in the variability of a particular locus, the VNTR information page has a display which shows the different alleles that are in the database along with a multiple alignment between the reference sequence and the individual supporting sequences from each sample (figure 4·3). Both sample and VNTR record pages link out to relevant external sources, including the UCSC genome browser (Tyner et al., 2017), dbSNP, and ClinVar (figure 4·2). This enables users to check with more resources if the variant they have found or are interested in, has been described elsewhere, and to easily inspect the region further.

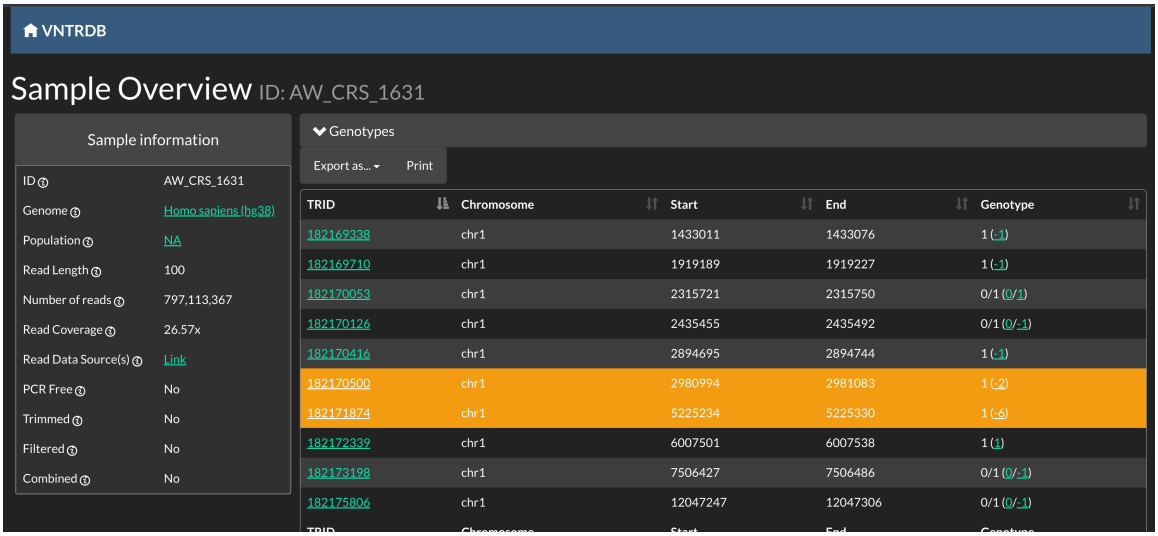
## 4.4 Conclusion and continued development

VNTRdb is a powerful tool for the discovery, analysis, and visualization of VNTR data. We currently do not have plans to allow submission of VNTRs via the web interface, but are planning a process for submission in some other form.

VNTRdb is under active development, and there are more features we would like to include. Searching for VNTRs by gene symbol or cytogenic location can be




(a)




(b)

**Figure 4·2:** VNTR and sample record pages. Links to external sources can be found here, as well as additional information on the sample or VNTR. Genotypes for samples can be downloaded in VCF format.

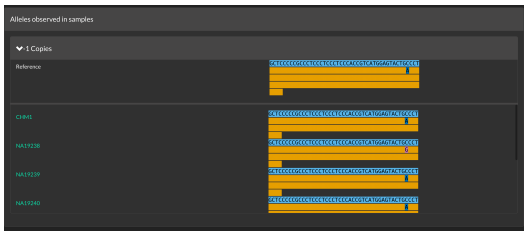



**Figure 4·3:** Example multiple alignment of allele and reference sequence. The pane is scrollable and multiple alleles are displayed in stacked panes, which are collapsible.

performed using data from the UCSC browser, but it may be included as a search option in VNTRdb (in hg38 only). Further curation of variants in the database is also planned, as one challenge we experienced in our research was the exact matching between dbSNP records and variants we detect.

VNTRdb was intentionally designed to be independent of the organisms represented by the data, though we currently only have human data available. It is developed under an open license and the server can also be deployed on a self-hosted server. Our deployment is meant to serve as a well-curated resource for use by others, but the availability of the code and data allows users to host mirrors. Contributions to the code are welcome.

## 4.5 Data availability

A preview of VNTRdb is currently available at `https://orca.bu.edu/vntrdb`. The source code will also be available on Bitbucket and GitHub. All data is available in VCF format (variants), BED format (VNTR loci), and as an SQLite database (all data).

## 4.6 Funding

National Science Foundation grants IIS-1423022 and IIS-1017621.

# Chapter 5

# Conclusions

## 5.1 Discussion

The work performed in this study contributes to our collective understanding of the diversity of a class of genetic mutations which have been historically poorly understood and understudied. We have developed a tool which automates the process of VNTR discovery, shown that it can perform with a high degree of accuracy, and have made significant improvements in its performance, increasing its usefulness and encouraging its adoption. Already we have received feedback from other researchers in the field who have either demonstrated interest in the technology, or have already deployed it in their research. This work has been presented both locally and abroad, and has been met with a positive reaction overall and interest.

The database discussed in chapter 4 complements the analysis from chapter 3 well by offering a customized solution to the discovery and further study of these variants by other researchers.

## 5.2 Future work

There are several exciting opportunities for further study in this area which can have a continuing impact on the ever-growing and ever-changing field of genetic sequencing and testing. Other methods we are developing in our lab can supplement the data generated by VNTRseek by focusing on array lengths outside the detection range of

VNTRseek, such as TR arrays which are too long to be spanned by a read, or novel TRs which would not be detected by VNTRseek since it relies on a target list of loci. Better alignment methods such as BitPal (Loving et al., 2014) (developed in our lab), highly parallel graphics processing-based programming methods, and more efficient storage and design patterns will further improve the performance of VNTRseek and TRF.

Outside of programming and the human genome, VNTRseek could provide an opportunity to further explore the mutational landscape of VNTRs in pathogenic microbes. In chapter 1 we discussed applications of microsatellite loci in disease tracking. These methods rely on slow "wet-lab" based technology which requires isolation of specific regions of the bacterial genome. Instead, VNTRseek offers a way of going from sequencing data directly to a profile of VNTRs which can be used to make a quick determination, or simply be used to inform on the population structure.

# References

(2017). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 45(D1):D12–D17.

Acuña-Amador, L., Primot, A., Cadieu, E., Roulet, A., and Barloy-Hubler, F. (2018). Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of Porphyromonas gingivalis reference strains. *BMC Genomics*, 19(1):54.

Adinolfi, S., Bagni, C., Musco, G., Gibson, T., Mazzarella, L., and Pastore, A. (1999). Dissecting FMR1, the protein responsible for fragile X syndrome, in its structural and functional domains. *RNA*, 5(9):1248–1258.

Ames, D., Murphy, N., Helentjaris, T., Sun, N., and Chandler, V. (2008). Comparative Analyses of Human Single- and Multilocus Tandem Repeats. *Genetics*, 179(3):1693–1704.

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., and Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, 176(3):663–675.e19.

Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V., and Bafna, V. (2018a). Targeted genotyping of variable number tandem repeats with advntr. *Genome Research*, 28(11):1709–1719.

Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V., and Bafna, V. (2018b). Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Research*, 28(11):1709–1719.

Belkum, A. V. (2007). Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). *FEMS Immunology & Medical Microbiology*, 49(1):22–27.

Bell, G. I., Horita, S., and Karam, J. H. (1984). A Polymorphic Locus Near the Human Insulin Gene Is Associated with Insulin-dependent Diabetes Melliitus. *Diabetes*, 33(2):176–183.

Bell, G. I., Karam, J. H., and Rutter, W. J. (1982). Properties of a polymorphic DNA segment in the 5' flanking region of the human insulin gene. *Progress in Clinical and Biological Research*, 103 Pt A:57–65.

Benedetti, F., Dallaspezia, S., Colombo, C., Pirovano, A., Marino, E., and Smeraldi, E. (2008). A length polymorphism in the circadian clock gene Per3 influences age at onset of bipolar disorder. *Neuroscience Letters*, 445(2):184–187.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580.

Blouin, Y., Hauck, Y., Soler, C., Fabre, M., Vong, R., Dehan, C., Cazajous, G., Massoure, P.-L., Kraemer, P., Jenkins, A., Garnotel, E., Pourcel, C., and Vergnaud, G. (2012). Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching Mycobacterium tuberculosis Clade. *PLOS ONE*, 7(12):e52841.

Brookes, K. J. (2013). The VNTR in complex disorders: The forgotten polymorphisms? A functional way forward? *Genomics*, 101(5):273–281.

Campuzano, V., Montermini, L., Moltò, M. D., Pianese, L., Cossée, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Cañizares, J., Koutnikova, H., Bidichandani, S. I., Gellera, C., Brice, A., Trouillas, P., Michele, G. D., Filla, A., Frutos, R. D., Palau, F., Patel, P. I., Donato, S. D., Mandel, J.-L., Cocozza, S., Koenig, M., and Pandolfo, M. (1996). Friedreich's Ataxia: Autosomal Recessive Disease Caused by an Intronic GAA Triplet Repeat Expansion. *Science*, 271(5254):1423–1427.

Cervera, A., Tàssies, D., Obach, V., Amaro, S., Reverter, J. C., and Chamorro, A. (2007). The BC Genotype of the VNTR Polymorphism of Platelet Glycoprotein Ib$\alpha$ Is Overrepresented in Patients with Recurrent Stroke Regardless of Aspirin Therapy. *Cerebrovascular Diseases*, 24(2-3):242–246.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., and Eichler, E. E. (2014). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–611.

Chalker, V., Pereyre, S., Dumke, R., Winchell, J., Khosla, P., Sun, H., Yan, C., Vink, C., and Bébéar, C. (2015). International Mycoplasma pneumoniae typing study: interpretation of M. pneumoniae multilocus variable-number tandem-repeat analysis. *New Microbes and New Infections*, 7:37–40.

Chang, C.-H., Chang, Y.-C., Underwood, A., Chiou, C.-S., and Kao, C.-Y. (2007). VNTRDB: a bacterial variable number tandem repeat locus database. *Nucleic Acids Research*, 35(suppl_1):D416–D421.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., and Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E. E., Weinstock, G., Mardis, E. R., Wilson, R. K., Howe, K., Flicek, P., and Hubbard, T. (2011). Modernizing Reference Genome Assemblies. *PLOS Biology*, 9(7):e1001091.

Cohen, I. L., Liu, X., Schutz, C., White, B. N., Jenkins, E. C., Brown, W. T., and Holden, J. J. A. (2003). Association of autism severity with a monoamine oxidase a functional polymorphism. *Clinical Genetics*, 64(3):190–197.

Consortium, T. . G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Cook, C. E., Bergman, M. T., Finn, R. D., Cochrane, G., Birney, E., and Apweiler, R. (2016). The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Research*, 44(D1):D20–D26.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M., and Reid, C. A. (2010). Human genome sequencing using unchained base reads on self-assembling dna nanoarrays. *Science*, 327(5961):78–81.

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H.-Y., Humphray, S. J., Halpern, A. L., Kruglyak, S., Margulies, E. H., McVean, G., and Bentley, D. R. (2017a). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, 27(1):157–164.

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H.-Y., Humphray, S. J., Halpern, A. L., Kruglyak, S., Margulies, E. H., McVean, G., and Bentley, D. R. (2017b). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, 27(1):157–164.

Ebisawa, T., Uchiyama, M., Kajimura, N., Mishima, K., Kamei, Y., Katoh, M., Watanabe, T., Sekimoto, M., Shibui, K., Kim, K., Kudo, Y., Ozeki, Y., Sugishita, M., Toyoshima, R., Inoue, Y., Yamada, N., Nagase, T., Ozaki, N., Ohara, O., Ishida, N., Okawa, M., Takahashi, K., and Yamauchi, T. (2001). Association of structural polymorphisms in the human period3 gene with delayed sleep phase syndrome. *EMBO reports*, 2(4):342–346.

Fondon, J. W., Martin, A., Richards, S., Gibbs, R. A., and Mittelman, D. (2012). Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing. *PloS One*, 7(3):e33036.

Fu, Y. H., Pizzuti, A., Fenwick, R. G., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., Jong, P. d., and Et, A. (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science*, 255(5049):1256–1258.

Gelfand, Y., Hernández, Y., Loving, J., and Benson, G. (2014). VNTRseek—a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Research*, 42(14):8884–8894.

Gelfand, Y., Rodriguez, A., and Benson, G. (2007). TRDB—The Tandem Repeats Database. *Nucleic Acids Research*, 35(suppl 1):D80–D87.

Golalipour, M., Maleki, Z., Farazmandfar, T., and Shahbazi, M. (2017). PER3 VNTR polymorphism in Multiple Sclerosis: A new insight to impact of sleep disturbances in MS. *Multiple Sclerosis and Related Disorders*, 17:84–86.

Grady, D. L., Chi, H.-C., Ding, Y.-C., Smith, M., Wang, E., Schuck, S., Flodman, P., Spence, M. A., Swanson, J. M., and Moyzis, R. K. (2003). High prevalence of rare dopamine receptor D4 alleles in children diagnosed with attention-deficit hyperactivity disorder. *Molecular Psychiatry*, 8(5):536–545.

Gymrek, M. (2017). A genomic view of short tandem repeats. *Current Opinion in Genetics & Development*, 44:9–16.

Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*.

Haguenoer, E., Baty, G., Pourcel, C., Lartigue, M.-F., Domelier, A.-S., Rosenau, A., Quentin, R., Mereghetti, L., and Lanotte, P. (2011). A multi locus variable number of tandem repeat analysis (MLVA) scheme for Streptococcus agalactiae genotyping. *BMC Microbiology*, 11:171.

Hasan, N. A., Choi, S. Y., Eppinger, M., Clark, P. W., Chen, A., Alam, M., Haley, B. J., Taviani, E., Hine, E., Su, Q., Tallon, L. J., Prosper, J. B., Furth, K., Hoq, M. M., Li, H., Fraser-Liggett, C. M., Cravioto, A., Huq, A., Ravel, J., Cebula, T. A., and Colwell, R. R. (2012). Genomic diversity of 2010 Haitian cholera outbreak strains. *Proceedings of the National Academy of Sciences*, 109(29):E2010–E2017.

Hernandez, Y., Rasekh, M., Loving, J., Gelfand, Y., and Benson, G. (2019). An Analysis of the Diversity of Minisatellite Tandem Repeats in the Human Genome.

Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., and Mittelman, D. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research*, 41(1):e32.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, 12(2):115–121.

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., Peluso, P., Boitano, M., Chin, C.-S., Korlach, J., Wilson, R. K., and Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, 27(5):677–685.

Inc., I. (2014). Platinum Genomes. http://www.illumina.com/platinumgenomes/. Accessed: 2015-10-17.

Jeffreys, A. J., Turner, M., and Debenham, P. (1991). The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework. *American Journal of Human Genetics*, 48(5):824–840.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314(6006):67–73.

Jowett, N. I., Williams, L. G., Hitman, G. A., and Galton, D. J. (1984). Diabetic hypertriglyceridaemia and related 5' flanking polymorphism of the human insulin gene. *British Medical Journal (Clinical research ed.)*, 288(6411):96–99.

Kendall, E. A., Chowdhury, F., Begum, Y., Khan, A. I., Li, S., Thierer, J. H., Bailey, J., Kreisel, K., Tacket, C. O., LaRocque, R. C., Harris, J. B., Ryan, E. T., Qadri, F., Calderwood, S. B., and Stine, O. C. (2010). Relatedness of Vibrio cholerae O1/O139 Isolates from Patients and Their Household Contacts, Determined by Multilocus Variable-Number Tandem-Repeat Analysis. *Journal of Bacteriology*, 192(17):4367–4376.

Kimpton, C. P., Gill, P., Walton, A., Urquhart, A., Millican, E. S., and Adams, M. (1993). Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *Genome Research*, 3(1):13–22.

Koning, A. P. J. d., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics*, 7(12):e1002384.

Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B., and Halldórsson, B. V. (2017). popSTR: population-scale detection of STR variants. *Bioinformatics*, 33(24):4041–4048.

Lafreniére, R. G., Rochefort, D. L., Chrétien, N., Rommens, J. M., Cochius, J. I., Kälviäinen, R., Nousiainen, U., Patry, G., Farrell, K., Söderfeldt, B., Federico, A., Hale, B. R., Cossio, O. H., Sørensen, T., Pouliot, M. A., Kmiec, T., Uldall, P., Janszky, J., Pranzatelli, M. R., Andermann, F., Andermann, E., and Rouleau, G. A. (1997). Unstable insertion in the 5' flanking region of the cystatin b gene is the most common mutation in progressive myoclonus epilepsy type 1, epm1. *Nature Genetics*, 15(3):298–302.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D.,

Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., Bastide, M. d. l., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–239.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., and Maglott, D. R. (2016). Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLOS Computational Biology*, 9(8):e1003118.

Legendre, M., Pochet, N., Pak, T., and Verstrepen, K. J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research*, 17(12):1787–1796.

Leung, P. W.-l., Chan, J. K. Y., Chen, L. H., Lee, C. C., Hung, S. F., Ho, T. P., Tang, C. P., Moyzis, R. K., and Swanson, J. M. (2017). Family-based association study of DRD4 gene in methylphenidate-responded Attention Deficit/Hyperactivity Disorder. *PLOS ONE*, 12(3):e0173748.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem.

Li, H. and Durbin, R. (2009a). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H. and Durbin, R. (2009b). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Loving, J., Hernandez, Y., and Benson, G. (2014). BitPAl: a bit-parallel, general integer-scoring sequence alignment algorithm. *Bioinformatics*, 30(22):3166–3173.

Ma, B., Tromp, J., and Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445.

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., and Scherer, S. W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(D1):D986–D992.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N., MacFarlane, H., Jenkins, B., Anderson, M. A., Wexler, N. S., Gusella, J. F., Bates, G. P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A.-M., Lehrach, H., Buckler, A. J., Church, D., Doucette-Stamm, L., O'Donovan, M. C., Riba-Ramirez, L., Shah, M., Stanton, V. P., Strobel, S. A., Draths, K. M., Wales, J. L., Dervan, P., Housman, D. E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J. J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F. S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D., and Harper, P. S. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983.

Mak, D. Y. F. and Benson, G. (2009). All hits all the time: parameter-free calculation of spaced seed sensitivity. *Bioinformatics (Oxford, England)*, 25(3):302–8.

Matsumoto, M. and Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator. *ACM Trans. Model. Comput. Simul.*, 8(1):3–30.

McIver, L. J., Fondon, J. W., Skinner, M. A., and Garner, H. R. (2011). Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics*, 97(4):193–199.

McIver, L. J., McCormick, J. F., Martin, A., Fondon, J. W., and Garner, H. R. (2013). Population-scale analysis of human microsatellites reveals novel sources of exonic variation. *Gene*, 516(2):328–334.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

Mikkelsson Jussi, Perola Markus, Penttilä Antti, and Karhunen Pekka J. (2001). Platelet Glycoprotein Ib$\alpha$ HPA-2 Met/VNTR B Haplotype as a Genetic Predictor of Myocardial Infarction and Sudden Cardiac Death. *Circulation*, 104(8):876–880.

Murphy, E. (2018). Forensic DNA Typing. *Annual Review of Criminology*, 1(1):497–515.

Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., and Morgan, M. (2014). Variantannotation: a bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14):2076–2078.

Owerbach, D., Billesbølle, P., Schroll, M., Johansen, K., Poulsen, S., and Nerup, J. (1982). Possible Association Between Dna Sequences Flanking the Insulin Gene and Atherosclerosis. *The Lancet*, 320(8311):1291–1293.

Parvej, M. S., Nakamura, H., Alam, M. A., Wang, L., Zhang, S., Emura, K., Kage-Nakadai, E., Wada, T., Hara-Kudo, Y., and Nishikawa, Y. (2019). Host range-associated clustering based on multi-locus variable-number tandem-repeat analysis, phylotypes, and virulence genes of atypical enteropathogenic Escherichia coli strains. *Appl. Environ. Microbiol.*, pages AEM.02796–18.

Pourcel, C., Minandri, F., Hauck, Y., D'Arezzo, S., Imperi, F., Vergnaud, G., and Visca, P. (2011). Identification of Variable-Number Tandem-Repeat (VNTR) Sequences in Acinetobacter baumannii and Interlaboratory Validation of an Optimized Multiple-Locus VNTR Analysis Typing Scheme. *Journal of Clinical Microbiology*, 49(2):539–548.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D., and Ostell, J. M. (2014). Refseq: an update on mammalian reference sequences. *Nucleic Acids Research*, 42(D1):D756–D763.

Quinlan, A. R. (2002). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., and Kent, W. J. (2015). The ucsc genome browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681.

Savas, S., Frischhertz, B., Pelias, M. Z., Batzer, M. A., Deininger, P. L., and Keats, B. J. (2002). The *ush1c* 216G→A mutation and the 9-repeat vntr(t,t) allele are in complete linkage disequilibrium in the acadian population. *Human Genetics*, 110(1):95–97.

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., Harden, G., Hubbard, T., Pelan, S., Simpson, J. T., Threadgold, G., Torrance, J., Wood, J. M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy, A. M., Durbin, R., Wilson, R. K., Flicek, P., Eichler, E. E., and Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864.

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V., Wang, Q., Burhans, R., Riemer, C., Wittekindt, N. E., Moorjani, P., Tindall, E. A., Danko, C. G., Teo, W. S., Buboltz, A. M., Zhang, Z., Ma, Q., Oosthuysen, A., Steenkamp, A. W., Oostuisen, H., Venter, P., Gajewski, J., Zhang, Y., Pugh, B. F., Makova, K. D., Nekrutenko, A., Mardis, E. R., Patterson, N., Pringle, T. H., Chiaromonte, F., Mullikin, J. C., Eichler, E. E., Hardison, R. C., Gibbs, R. A., Harkins, T. T., and Hayes, V. M. (2010). Complete Khoisan and Bantu genomes from southern Africa. *Nature*, 463(7283):943–947.

Sherry, S. T. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

Simsek, S., Bleeker, P. M. M., Schoot, C. E. v. d., and Borne, A. E. G. K. v. d. (1994). Association of a Variable Number of Tandem Repeats (VNTR) in Glycoprotein Ib$\alpha$ and HPA-2 Alloantigens. *Thrombosis and Haemostasis*, 72(5):757–761.

Smit, A. F., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0.

Steinberg, K. M., Schneider, V. A., Graves-Lindsay, T. A., Fulton, R. S., Agarwala, R., Huddleston, J., Shiryev, S. A., Morgulis, A., Surti, U., Warren, W. C., Church, D. M., Eichler, E. E., and Wilson, R. K. (2014). Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Research*, 24(12):2066–2076.

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A. R., Attar, M., Babbs, C., Becq, J., Beeson, D., Bento, C., Bignell, P., Blair, E., Buckle, V. J., Bull, K., Cais, O., Cario, H., Chapel, H., Copley, R. R., Cornall, R., Craft, J., Dahan, K., Davenport, E. E., Dendrou, C., Devuyst, O., Fenwick, A. L., Flint, J., Fugger, L., Gilbert, R. D., Goriely, A., Green, A., Greger, I. H., Grocock, R., Gruszczyk, A. V., Hastings, R., Hatton, E., Higgs, D., Hill, A., Holmes, C., Howard, M., Hughes, L., Humburg, P., Johnson, D., Karpe, F., Kingsbury, Z., Kini, U., Knight, J. C., Krohn, J., Lamble, S., Langman, C., Lonie, L., Luck, J., McCarthy, D., McGowan, S. J., McMullin, M. F., Miller, K. A., Murray, L., Németh, A. H., Nesbit, M. A., Nutt, D., Ormondroyd, E., Oturai, A. B., Pagnamenta, A., Patel, S. Y., Percy, M., Petousi, N., Piazza, P., Piret, S. E., Polanco-Echeverry, G., Popitsch, N., Powrie, F., Pugh, C., Quek, L., Robbins, P. A., Robson, K., Russo, A., Sahgal, N., van Schouwenburg, P. A., Schuh, A., Silverman, E., Simmons, A., Sørensen, P. S., Sweeney, E., Taylor, J., Thakker, R. V., Tomlinson, I., Trebes, A., Twigg, S. R. F., Uhlig, H. H., Vyas, P., Vyse, T., Wall, S. A., Watkins, H., Whyte, M. P., Witty, L., Wright, B., Yau, C., Buck, D., Humphray, S., Ratcliffe, P. J., Bell, J. I., Wilkie, A. O. M., Bentley, D., Donnelly, P., and McVean, G. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, 47(7):717–726.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Treangen, T. J. and Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, 13(1):36–46.

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C. M., Gibson, D., Gonzalez, J. N., Guruvadoo, L., Haeussler, M., Heitner, S., Hinrichs, A. S., Karolchik, D., Lee, B. T., Lee, C. M., Nejad, P., Raney, B. J., Rosenbloom, K. R., Speir, M. L., Villarreal, C., Vivian, J., Zweig, A. S., Haussler, D., Kuhn, R. M., and Kent, W. J. (2017). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research*, 45(D1):D626–D634.

van den Berg, R. J., Schaap, I., Templeton, K. E., Klaassen, C. H. W., and Kuijper, E. J. (2007). Typing and Subtyping of Clostridium difficile Isolates by Using Multiple-Locus Variable-Number Tandem-Repeat Analysis. *Journal of Clinical Microbiology*, 45(3):1024–1028.

Wang, E., Ding, Y.-C., Flodman, P., Kidd, J. R., Kidd, K. K., Grady, D. L., Ryder, O. A., Spence, M. A., Swanson, J. M., and Moyzis, R. K. (2004). The Genetic Architecture of Selection at the Human Dopamine Receptor D4 (DRD4) Gene Locus. *American Journal of Human Genetics*, 74(5):931–944.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. a., and Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–6.

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and *de novo* STR variations. *Nature Methods*, 14(6):590–592.

Wink, M. (2006). Use of DNA markers to study bird migration. *Journal of Ornithology*, 147(2):234–244.

Zaluga, J., Stragier, P., Van Vaerenbergh, J., Maes, M., and De Vos, P. (2013). Multilocus Variable-Number-Tandem-Repeats Analysis (MLVA) distinguishes a clonal complex of Clavibacter michiganensis subsp. michiganensis strains isolated from recent outbreaks of bacterial wilt and canker in Belgium. *BMC Microbiology*, 13:126.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., Saghbini, M., Dzakula, Z., Hastie, A., Cao, H., Deikus, G., Schadt, E., Sebra, R., Bashir,

A., Truty, R. M., Chang, C. C., Gulbahce, N., Zhao, K., Ghosh, S., Hyland, F., Fu, Y., Chaisson, M., Xiao, C., Trow, J., Sherry, S. T., Zaranek, A. W., Ball, M., Bobe, J., Estep, P., Church, G. M., Marks, P., Kyriazopoulou-Panagiotopoulou, S., Zheng, G. X. Y., Schnall-Levin, M., Ordonez, H. S., Mudivarti, P. A., Giorda, K., Sheng, Y., Rypdal, K. B., and Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3:sdata201625.

# CURRICULUM VITAE