2021

Characterizing VNTRs in human populations

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES AND

COLLEGE OF ENGINEERING

Dissertation

CHARACTERIZING VNTRS IN HUMAN POPULATIONS

by

MARZIEH ESLAMI RASEKH

B.S., University of Isfahan, 2009
M.S., Bilkent University, 2015

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2021

Approved by

First Reader

Gary Benson, PhD
Associate Professor of Biology and Computer Science

Second Reader

Juan I. Fuxman Bass, PhD
Assistant Professor of Biology

من چو از کان معانی یک جوم   همچنین جو جو بدان کان می روم
من چو از خورشید کیوان ذره‌ام   ذره ذره سوی کیوان می روم
این سخن پایان ندارد لیک من   آمدم زان سر به پایان می روم

I am a grain from the field of meaning, therefore, I go, *grain by grain*, to that field.
I am a particle from the Sun of Saturn, therefore, I go, *little by little*, toward Saturn.
This discourse has no end, yet I, have come from one end and and am going toward
The End.

– *Book of Shams*, Rumi

In the loving memory of my grandmother,

Ameneh Razmjoo

Acknowledgments

Pursuing a PhD degree was not something I planned. It was an adventure that started with a spontaneous trip to Turkey with my very dear friend, Sara Tatari. She encouraged me to apply to Bilkent University for my Masters.

Being inspired by the great scientists I worked with at Bilkent University, Can Alkan (my Master's advisor), Öznur Taştan, and Özlen Konu, I made a change of course and decided to pursue my doctorate degree in Bioinformatics at Boston University. I acknowledge their great role in introducing me to the many exciting aspects of this field and training me to do research. Can, Öznur, and Özlen, I truly appreciate your mentorship and am fortunate to have been able to work with you.

At Boston University, my advisor, Gary Benson, provided me with the opportunity to work in his lab. Throughout these years, he encouraged my scientific curiosity and supported me to work on my research ideas, all while guiding me towards improving my methods. He made my work possible by ensuring I had the means I required. Gary, I thank you for your mentorship, guidance, and patience.

I thank Yözen Hernandez, a previous student in our lab, for his help in setting me up in the lab and helping me with VNTRseek.

I thank Samantha Drinan, who performed the experimental validations throughout this thesis. It has been a great experience coauthoring three papers with you! You have added a lot to my work.

I thank my Thesis Committee for the guidance and academic advice they have provided me during these years. To Trevor, Stefano, Eric, and Juan, I thank you for ensuring the successful completion of my dissertation.

I thank the admins of the Bioinformatics program for their academic support. Dave King, Johanna Vasquez, and Caroline Lyman, and especially Mary Ellen, I thank you for all your help.

I thank Thomas Gilmore, Xiaoling Zhang, and Alberto Cruz-Martín for their scientific advice on this thesis. Tom, you have been a great mentor and friend to me, and for that, I am deeply grateful.

I thank Xiao Ni and Douglas Robinson for giving me the opportunity to do an internship at Novartis and gain experience in the biotech industry. This internship taught me a lot and helped me choose my future path.

I thank my family for their support and unconditional love throughout my doctorate degree and my whole life. Especially my sister, Maryam, my mother, Pouran, and my father, Abbas – for your emotional guidance and – of course – financial support, I thank you!

To Elen Katz, nurse practitioner at Boston University Student Health Center who helped me overcome my many health challenges, I am truly grateful for your kind and professional care that was above and beyond.

Finally, living in Boston for five years has been one of the best experiences of my life, mainly because of the awesome people I have come to call friends.
My first mentor, Nastia Guronivich, who guided me through graduate life, I thank you.
Gracia Bonilla, a dear friend who showed me around the city, I thank you. Tanya Karagiannis, my dearest friend, with whom I spent long nights working in the office and discovered places in Boston, I thank you. And other friends I have met through the Bioinformatics program and became close to and made happy memories with: Sebastian Carrasco Pro, Aaron Chevalier, Jasmine Kwasa, Ana Leshchyk, Josh Lippai, Rebecca Panitch, Bashir Rastegarpanah, Abigail Sison, Ahmed Youssef, I thank you all.

To Amir Memar and Setare Navai, for being great friends from Esfahan, Iran to the States, I thank you. Mike Griffin, for your emotional support and for all the fun trips up North, I thank you my dear. Sheltering at home during the pandemic was not so hard with you!

I consider myself very fortunate and grateful to have had the opportunity to complete my PhD with your support and be surrounded by such amazing people. I will miss graduate life – but am looking forward to new adventures. I am excited for the next chapter(s) of my life.

CHARACTERIZING VNTRS IN HUMAN POPULATIONS

MARZIEH ESLAMI RASEKH

Boston University, Graduate School of Arts and Sciences and College of Engineering, 2021

Major Professor: Gary Benson, PhD
Professor of Biology and Computer Science

ABSTRACT

Over half the human genome consists of repetitive sequences. One major class is the tandem repeats (TRs), which are defined by their location in the genome, repeat unit, and copy number. TRs loci which exhibit variant copy numbers are called Variable Number Tandem Repeats (VNTRs). High VNTR mutation rates of approximately 10^{-4} per generation make them suitable for forensic studies, and of interest for potential roles in gene regulation and disease. TRs are generally divided into three classes: 1) microsatellites or short tandem repeats (STRs) with patterns <7 bp; 2) minisatellites with patterns of seven to hundreds of base pairs; and 3) macrosatellites with patterns of >100 bp. To date, mini- and macrosatellites have been poorly characterized, mainly due to a lack of computational tools. In this thesis, I utilize a tool, VNTRseek, to identify human minisatellite VNTRs using short read sequencing data from nearly 2,800 individuals and developed a new computational tool, MaSUD, to identify human macrosatellite VNTRs using data from 2,504 individuals. MaSUD is the first high-throughput tool to genotype macrosatellites using short reads.

I identified over 35,000 minisatellite VNTRs and over 4,000 macrosatellite VNTRs, most previously unknown. A small subset in each VNTR class was validated experimentally and in silico. The detected VNTRs were further studied for their effects on gene expression, ability to distinguish human populations, and functional enrichment. Unlike STRs, mini- and macrosatellite VNTRs are enriched in regions with functional importance, e.g., introns, promoters, and transcription factor binding sites. A study of VNTRs across 26 populations shows that minisatellite VNTR genotypes can be used to predict super-populations with >90% accuracy. In addition, genotypes for 195 minisatellite VNTRs and 22 macrosatellite VNTRs were shown to be associated with differential expression in nearby genes (eQTLs).

Finally, I developed a computational tool, mlZ, to infer undetected VNTR alleles and to detect false positive predictions. mlZ is applicable to other tools that use read support for predicting short variants.

Overall, these studies provide the most comprehensive analysis of mini- and macrosatellites in human populations and will facilitate the application of VNTRs for clinical purposes.

Contents

List of Tables

List of Figures

List of Abbreviations

CCS	. . . . . .	Circular Consensus Sequencing
CNV	. . . . . .	Copy Number Variation
eQTL	. . . . . .	Expression quantitative trait loci
GIAB	. . . . . .	Genome in a Bottle Consortium
IGSR	. . . . . .	International Genome Sample Resource
NGS	. . . . . .	Next Generation Sequencing
NYGC	. . . . . .	New York Genome Center
SD	. . . . . .	Segmental Duplication
SGDP	. . . . . .	Simon's Genome Diversity Project
SNP	. . . . . .	Single Nucleotide Polymorphic
SNV	. . . . . .	Single Nucleotide Variant
STR	. . . . . .	Short Tandem Repeat
TFBS	. . . . . .	Transcription Factor Binding Site
TPM	. . . . . .	Transcripts Per Kilobase Million
TR	. . . . . .	Tandem Repeat
TRDB	. . . . . .	Tandem Repeat Database
TRF	. . . . . .	Tandem Repeat Finder
VNTR	. . . . . .	Variable Number Tandem Repeat
WGS	. . . . . .	Whole Genome Sequencing

Chapter 1

Introduction

The human genome is about 3.2 billion base pairs long. About 3% of the genome comprises gene sequences. The non-coding genome, once called junk DNA, is now known to play an important role in biological processes (Carey, 2015; Biémont and Vieira, 2006; Ludwig, 2016). Over 50% of this non-coding genome consists of repetitive DNA. The analysis of certain types of repetitive DNA is the focus of this thesis.

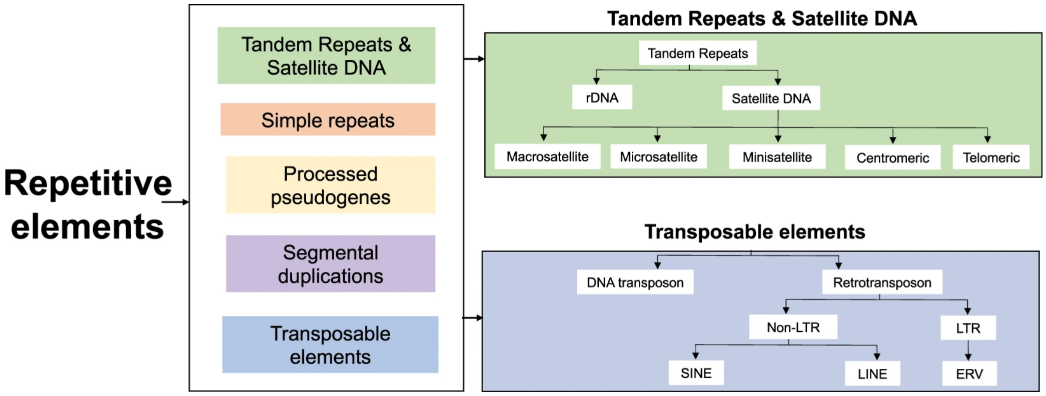


Figure 1·1: Classification of repetitive DNA in the human genome. Figure obtained from (Billingsley et al., 2019).

1.1 Repeats in the human genome

There are different classes of repeats in the human genome (International Human Genome Sequencing Consortium, 2001). This classification is depicted in Figure 1·1 and can be summarized as follows:

- Satellite DNA are small fragments of DNA that can move around in the genome.

Transposons are a class of satellite DNA that can change their position in the genome. Tandem Repeats (TRs) are another class of satellite DNA which consist of patterns that are repeated consecutively. Telomeres and centromeres of the DNA consist of AT-rich blocks of tandem repeats. TRs are also called simple repeats.

- Segmental Duplications (SDs) are duplications of >10 Kbp or longer with >90% similarity spread out in the genome (Vallente and Eichler, 2005).

- Processed pseudogenes are inactive copies of genes or small RNA.

The focus of this dissertation is tandem repeats. TRs can be defined by their location on the genome, the pattern (also called the repeat unit), and the number of copies. An example of a tandem repeat is given in Figure 1·2.

1.2 Sequencing technologies

Over the years many different techniques have been developed to sequence genomes. Today, the most common and cost-effective sequencing technology is paired-end sequencing with Illumina machines (Kulski, 2016). In this method, whole genome shotgun sequencing is used (Weber and Myers, 1997) to shear the DNA into small fragments of a few hundred bp, usually 250–550 bp. The fragments follow a normal distribution. Each fragment is read with a laser on each end from different strands for a fixed length. The two “reads” obtained from a fragment have fixed length and come from different strands of the DNA. Paired-end sequencing is commonly used for Whole Genome Sequencing (WGS). Read lengths of 100–250 bp are common for WGS. Paired-end sequencing produces short reads with very high per base precision (2%) (Schirmer et al., 2016; Ma et al., 2019).

Other technologies such as PacBio (Rhoads and Au, 2015) and nanopore(Jain

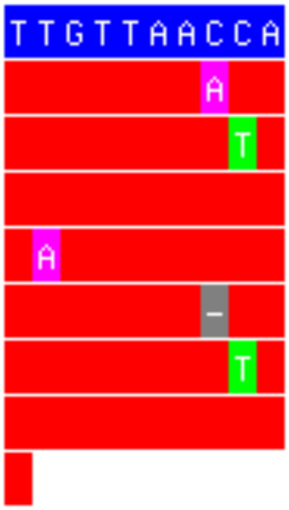


Figure 1·2: An example tandem repeat. A tandem repeat (TR) is characterized by its location on the genome, the pattern, and the number of copies. The blue box is the consensus pattern. The repeat units are shown as stacked red boxes. The mutations in each repeat unit compared to the pattern consensus are shown in different colors (pink for mutation to A, green for mutation to T, and gray for indel). This TR has pattern length 10 bp (TTGTTAACCA), as presented in the consensus (blue box above), and 7.1 copies (number of stacked red blocks). The repeats do not need to be exact, mutations may occur, neither do they have to be complete, partial repeats are allowed. The image was obtained from the Tandem Repeat DataBase (TRDB) (Gelfand et al., 2007).

et al., 2016; Jain et al., 2018) can sequence longer reads of thousands to hundreds of thousands of base pairs. However these reads have higher error rate (>10%). While Illumina errors consist mainly of SNPs, PacBio reads and nanopore reads can also include indels, causing downstream errors.

Recently, PacBio introduced a new method, Circular Consensus Sequencing (CCS), to overcome the low read quality of their long reads. In CCS, the molecules are made circular, and the circular molecules are sequenced many times and create a consensus read sequence (Wenger et al., 2019). Since the error rate of PacBio reads are about 10%, if the molecule is read 10 times, the average error of the consensus will become 1%, which is comparable to the error rate of Illumina reads. Clone sequencing (Duitama et al., 2012) and 10X Genomics WGS (Marks et al., 2019) combine the two methods by first fragmenting the DNA into large molecules, and then, producing short precise reads from the large molecules. In the future, producing longer reads with high accuracy will become possible.

1.3 Alignment of short reads

The reads produced by the different technologies should be mapped back to the reference genome. Alignment of reads to the genome is a matured methodology, and the most commonly used aligners are BWA MEM (Li, 2013) and Bowtie2 (Langmead and Salzberg, 2012) being the most common used for WGS. While alignment methods perform with over 95% accuracy genome-wide, their precision suffers on repeat-rich regions.

1.4 The human reference genome

The latest version of the human reference genome was published in 2013 and is named Genome Research Consortium human build 38 (GRCh38) from NCBI or Human Genome Build 38 (hg38) for short from UCSC and is $\sim 3.1 \times 10^9$ bp long `https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/`. The two versions are identical and only differ in the way the chromosome names are saved. In GRCh38, the chromosomes are written as chr1, chr2, ... while in hg38 the "chr" is dropped and the chromosomes are written as 1, 2, ... to save disk space. The previous version of the reference genome was published in 2009, and was labeled GRCh37 from NCBI and hg19 from UCSC.

Aligning reads are aligned to the reference genome, and because the reference genome was built from few individuals, biases are introduced (Ballouz et al., 2019).

Individuals with more similarity to the reference genome, i.e., Europeans, will have better alignment. More distance genomes, i.e., Africans, will have many unmapped reads (Sherman et al., 2019). To correct for this bias, pan-genomes have been introduced (Li et al., 2010) that present the genomes as a graph with all the possible variants as edges in it. GRAF implements workflows to align reads to pan-genomes and to detect variants which is available at `https://www.sevenbridges.com/graf/`.

1.5 Genomic variation

Differences between two DNA molecules are called genomic variants. Germline variation are fixed mutations in the gametes, such as sperm or egg cells of an individual compared to the reference genome. Differences between DNA of cells in one individual compared to the germline cells of the same individual are called somatic variation. While germline mutations are either inherited or occur at random during meiosis, somatic variation occurs during mitosis (cell division). Somatic variants are commonly studied in cancer tumors to understand the mutations that have caused the cell to become malignant. Genomic variants can be classified based on their length. Single nucleotide variants (SNVs) are mutations, insertions, or deletions of one base pair. Small insertion or deletions of $\leq$50 bp are called indels. Copy Number Variants (CNVs) are differences in copies of genes or sequences in one person's genome compared to the reference genome. Structural variation are deletions, insertions, inversions, or translocations of thousands to millions of base pairs. One form of genetic variation are TRs with different copies between individuals that are called Variable Number Tandem Repeats (VNTRs).

1.5.1 Homozygous vs heterozygous variants

The human genome is diploid, meaning that, except for the sex chromosomes in male individuals, there are two copies of each chromosome. However, the reference genome, which is the point of comparison, is haploid (one one copy of each chromosome). A homozygous variant is a locus on the genome where both chromosomes are the same but are different from the reference genome. Any heterozygous locus on a genome, by definition, is a genomic variant, because at least one copy of the chromosome must differ from the reference.

1.5.2 The genotype of a variant

A loci in the human genome has two "alleles", which is the observed sequence on each chromosome. The genotype of the diploid locus in an individual is presented as "allele$_1$/allele$_2$". At homozygous loci, only one allele is observed and at heterozygous loci two alleles are observed.

1.6 Population-wide studies of genomic variants

Genomic variation is common in the human genomes. Population-wide studies aim to find variants occurring in a large number of individuals in order to study the natural diversity among populations or to associate phenotypes to genotypes. With the goal of capturing the global picture of variation in human genomes, different consortiums have collected variants on large cohorts of individuals.

In 2008, The 1000 Genomes Project took up as an international effort with the goal of sequencing 1000 genomes (The 1000 Genomes Project Consortium, 2015). The goal was to sequence at least 1000 genomes from different populations and identify variants with minor allele frequencies as low as 1% across the genome. In 2015, the third and final phase published variants on 2,504 genomes from five super-populations and 26 sub-populations. Figure 1·3 shows a map of the populations presented in the 1000 Genomes Project. Although the 1000 Genomes Project was a breakthrough in genetic studies at its time, the majority of the genomes were sequenced at low coverage (<10X). So, in Phase 3 of the 1000 Genomes Project, 30 genomes were re-sequenced at higher coverage. These 30 genomes included two trios (child-father-mother) from the Yoruban and CEPH population, and another 24 genomes each from the remaining populations.

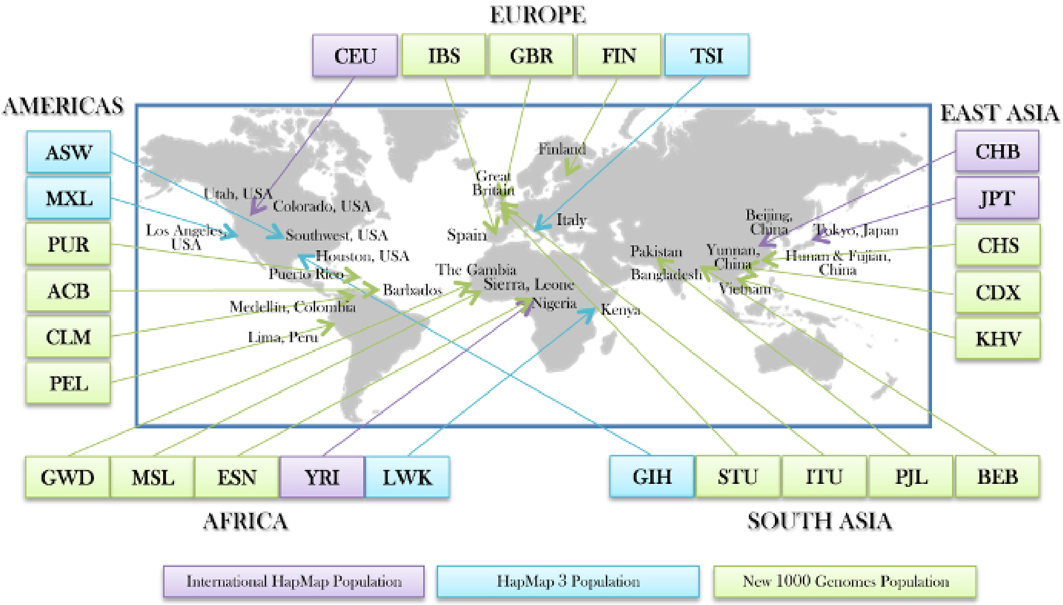


Figure 1·3: The populations from the 1000 Genomes Project. Total 2,504 unrelated individuals from 26 populations and five super populations were sequenced. Figure obtained from the Coriell website (`https://www.internationalgenome.org/cell-lines-and-dna-coriell`).

In 2019, the New York Genome Center (NYGC) resequenced the 2,504 genomes from the 1000 Genomes Project at higher coverage and longer read length. In addition 698 other genomes, which are related to the original samples, were sequenced.

In 2016, the Simon's Genome Diversity Project (SGDP) sequenced 300 individual genomes from 142 populations with high coverage with the aim to"maximize medical relevance by studying populations with large numbers of present-day people" (Mallick et al., 2016). Figure 1·4 shows a map of the populations selected for the SGDP. Many of these populations were not included in the 1000 Genomes Project, and have been under-represented in the population-wide studies.

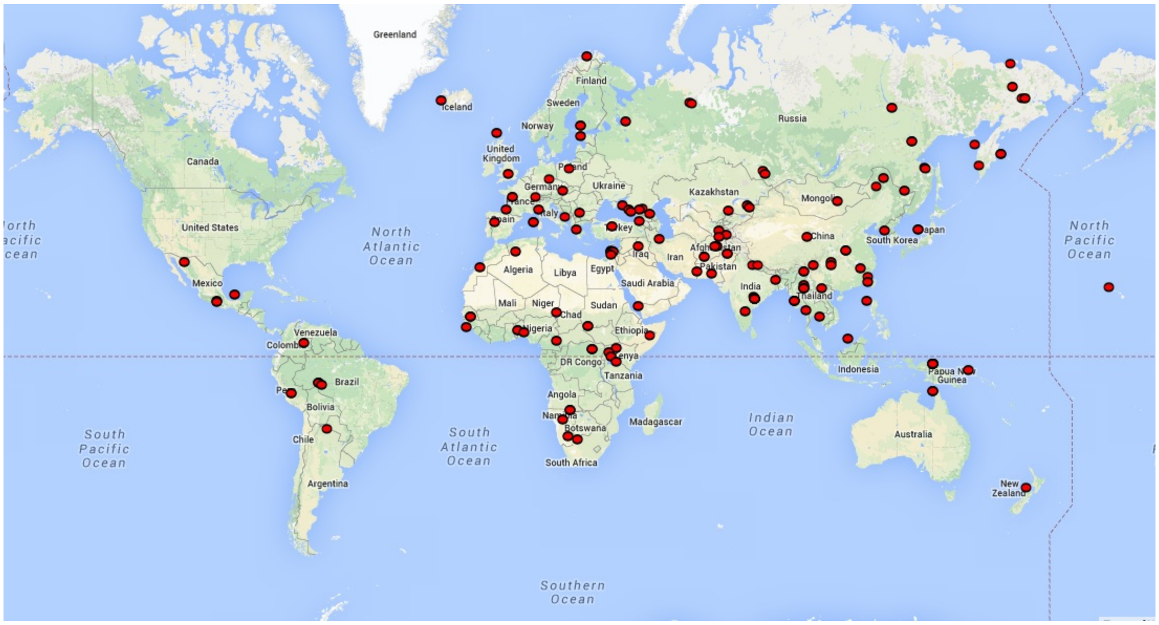


Figure 1·4: The genomes from Simon's Genome Diversity Project (SGDP). A total of 260 genomes were sequenced at read length 100 bp and coverage >40× from populations unrepresented in the 1000 genomes project (Mallick et al., 2016). The goal of this project was to include individuals proportional to the current world population.

The Personal Genome Project (PGP) (Church, 2005) sequences genomes and medical information of 100,000 volunteers publicly with the goal to expedite personalized medicine studies. Seven of the genomes from PGP were sequenced at very high coverage by the Genome In A Bottle consortium (GIAB) (Zook et al., 2016). International Genome Sample Resource (IGSR) (Fairley et al., 2020) provides a collection of publicly available genomes.

1.6.1 Common vs private variants

A common variant is a variation that occurs in >5% of the population under study. A private variant, on the other hand, occurs in <1% of the individuals. This definition was used in the 1000 Genomes Project to detect common SNPs vs. private SNPs. They found about 4–6 million SNPs per individual, which 3–4 million were common. Using the number of private SNPs, the SNP mutation rate is estimated to be about 10^{-9} per generation (The 1000 Genomes Project Consortium, 2015). Common variants can be used to study the normal spectrum of variation in human populations.

1.6.2 Major vs. minor allele

Alleles observed at one loci across the population can be studied to find major or minor alleles. Major allele is the most common allele at a given loci in the population. Minor allele is the second most common allele. Minor alleles are commonly associated with phenotypes (Asif et al., 2020; Panagiotou et al., 2010).

1.6.3 Heterozygosity and diversity

Variants in population-wide studies are also studied for heterozygosity. For a two allele system the heterozygosity is defined as $2pq$, where p is the probability of the major allele and q is the probability of the minor allele. The p and q can be used to calculate the Hardy Weinberg equilibrium, $(p+q)^2 = 1$, where p^2 is the frequency of homozygous occurrences of major alleles and q^2 is the frequency of homozygous frequency of the minor allele. Deviations from this equation could be used to detect genetic shift or selection in a population.

Higher heterozygosity in a population is linked to higher diversity and is linked to many human traits (Campbell et al., 2007). Heterozygosity and allele diversity have been associated with higher adaptive potential for quantitative traits (Vilas et al., 2015). Loss of heterozygosity in tumor tissue is linked to tumor progression (Cavenee, 1991; Schwarzenbach et al., 2012; Zuo et al., 2010).

1.7 De novo detection vs genotyping genomic variants

To detect variants in the genome, aligned reads are used. Detecting variants without any prior information of their position is called de novo variant detection. De novo methods are useful when the variant can occur at any location, such as SNPs or structural variations. When we know where the variants are located, such as minisatellites or testing for haplotypes, we can simply recruit reads originating from the region of interest. This is called genotyping. For genotyping, a reference set containing all the loci of interest should be provided.

Approaches to statistically infer variants from short WGS reads use four different "signatures": read depth, split read signatures, paired-end signatures, or local assembly. These techniques are summarized in Figure 1·5.

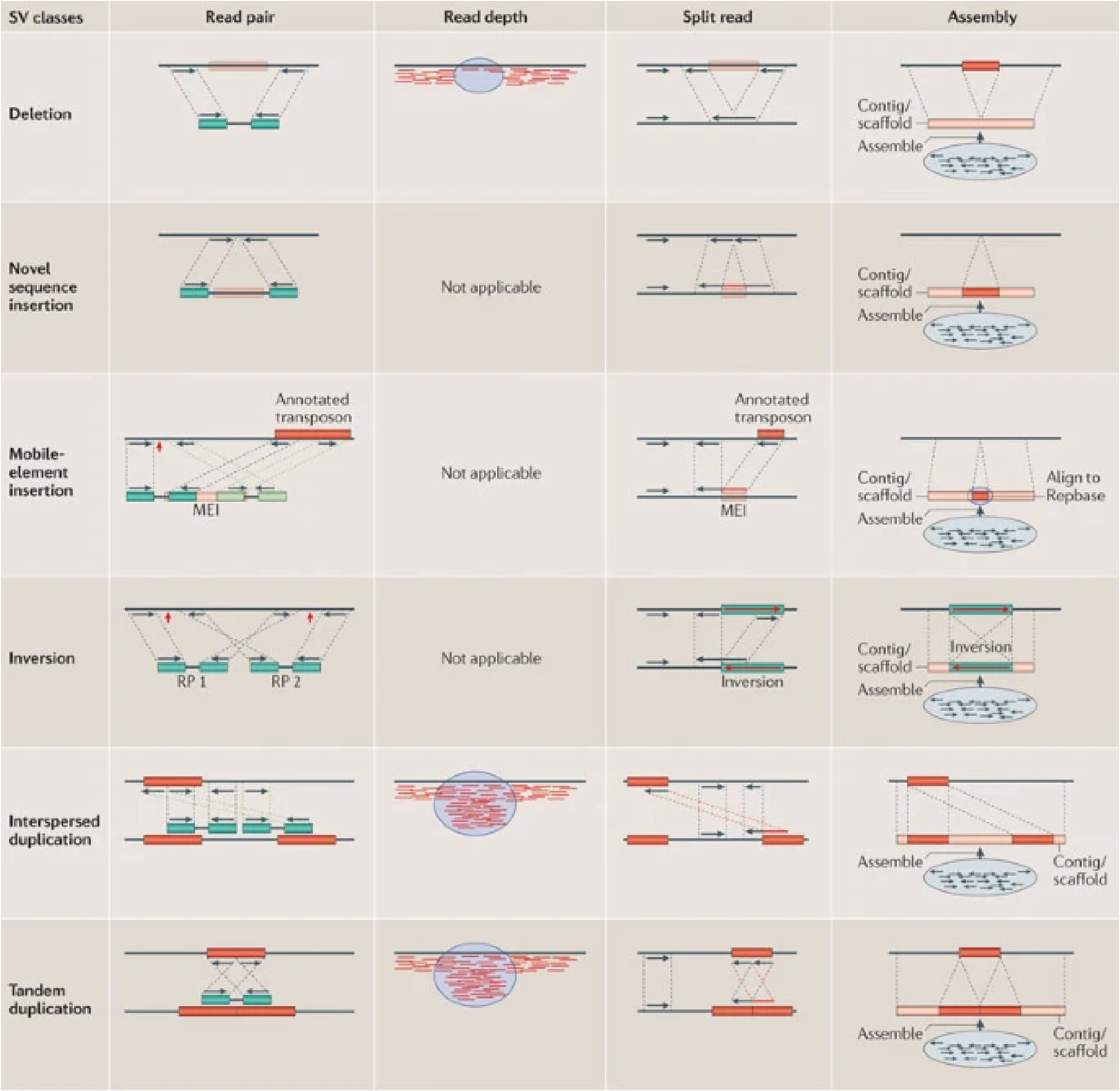
- *Read depth signature*: Comparing read depth of a region to the expected read depth is used to detect variants that cause a difference in read coverage, i.e. insertions, deletions, and CNV. This method can detect the change in read depth, but cannot find the position of the variant.

- *Split read signature*: When part of a read aligns to one position and another part aligns to another position, we detect a “split read”. Split reads can provide hints of a structural variation. BWA MEM provides split reads when a read partially aligns.

- *Paired end signatures*: Paired-end reads revolutionized structural variation detection. The distance that the pairs align to compared to the average fragment size, and the strands they align to can give a structural variation-specific signature. This method is commonly used in the state of the art structural variation detection tools such as LUMPY (Layer et al., 2014) and DELLY2 (Rausch et al., 2012).

- *Local assembly:* Another technique is to assemble the reads originating from the region of interest and locally assembling them using de bruijn graphs. The assembled sequence could be analyzed for the existence of certain structures, such as tips or bubbles, or compared to the reference genome.

Computational tools such as GATK (McKenna et al., 2010) and samtools mpileup (Li et al., 2009) have been developed for detecting short variants, i.e., SNPs and indels. Detecting small variants has matured and these tools perform with >90% precision and recall. In contrast, tools to call variants in repeat-rich regions such as balanced rearrangements, i.e. inversions and translocations, suffer from high false positives due to mapping errors. Structural variations breakpoints occur inside SDs and TRs, in which alignment tools perform poorly on these regions. Read support, defined as the number of reads showing evidence of the variant, is commonly used to measure the confidence of variant detection tools. Variants are required to be supported by a certain number of reads to be statistically reliable. Increasing the read support reduces the false positive calls and improves the precision, but also reduces sensitivity.



Nature Reviews | Genetics

Figure 1·5: Methods to detect large variants from short WGS reads. Paired end reads can be used to detect variants that are larger than the read length. In general four different signatures are used, paired-end read signature: pairs aligning too far or too close compared to the fragment size give a signature for the variant: the difference in read coverage at one region compared to the expected coverage; split read signature: when a read maps partially to two different positions, and local assembly: reconstructing the region using de bruijn graphs. Assembly can be used to find the sequence of an insertion. Image was obtained from (Alkan et al., 2011).

1.8 A review of minisatellites in the literature

Minisatellite TRs comprise more than a million base pairs in the human genome. These are defined as TRs with pattern length of 7-–100 bp. While many minisatellite loci appear to be monoallelic with regard to copy number, a significant fraction exhibit copy number variability and are called Variable Number Tandem Repeats (VNTRs). Changes in VNTR copy number have been proposed to arise by slipped strand mispairing (Taylor and Breden, 2000; Madsen et al., 1993; Levinson and Gutman, 1987), unequal crossover (Jeffreys et al., 1998; Debrauwère et al., 1999), and gene conversion (Jeffreys et al., 1998; Pâques et al., 2001).

1.8.1 De novo detection of minisatellites on the reference genome

Tandem Repeat Finder (TRF) (Benson, 1999) detects repeats of at least 1.8 copies in a given sequence. TRF uses parameters: match score, mismatch penalty, and gap penalty to find consensus patterns repeated at least 1.8 times in a sequence of nucleotides. The repeats are reported as the start and end position, the consensus pattern and its length, and the score. Minisatellites can be extracted by limiting the pattern size to 7–100 bp. The score can be used to filter out repeats that are not reliable.

Tandem Repeat DataBase (TRDB) (Gelfand et al., 2007) provides an online database of TRs on different versions of the reference genome. It also provides tools to filter the repeats and visualize them. TRs can also be visualized on the“Simple Tandem Repeats by TR” UCSC genome browser track (Kent, 2002).

1.8.2 Characteristics of minisatellites

VNTRs are highly mutable, with germline mutation rates estimated between 10^{-3} and 10^{-7} per cell division (Bustamante et al., 2013; Fu et al., 2016; Vogler et al., 2006; Verstrepen et al., 2005; Legendre et al., 2007). This mutation rate, which far exceeds that of SNPs, makes VNTRs useful for DNA fingerprinting (Panigrahi, 2018; Sinha et al., 2018; Imam et al., 2018).

VNTRs have also been predicted to have high heterozygosity, ranging from 43% to 59% (Denoeud et al., 2003), and the copy numbers of several VNTR loci have been shown to be population-biased in humans (Deka et al., 1991; Deka et al., 1992), suggesting that these VNTRs would not be useful for population wide studies.

The Eichler group (Sulovari et al., 2019) has examined TR loci on human and ape genome assemblies from PacBio sequencing data and has identified 1,584 human-specific VNTR loci proximal to 52 genes as candidate regions associated with disease. Additionally, by comparing VNTR loci situated in known gene enhancers with RNA sequencing data, the authors found that expansion of VNTRs correlated with up-regulation of the corresponding genes, suggesting that TR copy number can modulate enhancer activity.

More than half of previously identified human VNTR loci (Sulovari et al., 2019; Hancock and Santibáñez-Koref, 1998) are located near or within genes (Duitama et al., 2014) and some occur within coding exons (Duitama et al., 2014; Lancaster et al., 1990; Van Tol et al., 1992). Therefore, the potential effects of VNTRs on gene expression or the protein products are substantial. Indeed, VNTRs have been shown to contain binding sites for transcription factors such as NF-κB and Myc/HLH (Trepicchio and Krontiris, 1992; Krontiris et al., 1993), have been associated with changes in levels of gene expression (Sonay et al., 2015), including tissue-specific expression (Bakhtiari et al., 2020), and have been associated with gene splicing differences (Pacheco et al., 2019; De Roeck et al., 2018).

1.8.3 Minisatellites and disease

VNTRs have been proposed as drivers of phenotypic variation in evolution (Fondon and Garner, 2004; Laidlaw et al., 2007; Sulovari et al., 2019). For example, VNTR polymorphisms may play a role in neurogenesis and account for "human-specific cognitive traits" (Sonay et al., 2015).

Furthermore, minisatellite VNTRs have been associated with a variety of diseases (Antwi-Boasiako et al., 2018; Ksiazek et al., 2019; Cong et al., 2018; Ramírez-Patiño et al., 2013; Vairaktaris et al., 2007; Safarinejad et al., 2013; Ibrahimi et al., 2019), including neurodegenerative disorders (Marinho et al., 2019) such as Alzheimer's disease (Katsumata et al., 2019; Chang et al., 2019; De Roeck et al., 2018) and Huntington's disease (Scott et al., 1991; Krontiris et al., 1993), and other psychiatric conditions, such as PTSD (Hoxha et al., 2019), ADHD (Šerý et al., 2015; Grünblatt et al., 2019), depression (Van Assche et al., 2016), and addiction (Stolf et al., 2019).

1.8.4 Theoretical prediction of VNTRs in the literature

Denoeud et. al., (2003) examined 127 minisatellites with copy numbers ≥10 on chromosomes 21 and 22 of the human genome on 76 unrelated European genomes. They investigated the effect of the characteristics of the probability of TR polymorphism. The characteristics they studied were: pattern length, copy number, percent match of the repeat units, entropy of the consensus pattern, GC content of the array, Purine/Pyrimidine bias of the array (strand asymmetry for purines and pyrimidines), and a theoretical estimate called *HistoryR*. HistoryR (Benson and Dong, 1999) measures how ancient the TR is using the number of mutations accumulated in the repeat units. They found a strong correlation between HistoryR and GC content and probability of a TR being polymorphic, and a weak correlation between pattern size and entropy. They predicted that each individual would have ∼4,800 VNTRs.

In another work (Näslund et al., 2005), the authors developed a predictive model to predict the probability of a TR to be a VNTR from the following characteristics: the reference copy number, the match percentage of the repeat units, the entropy of the consensus pattern, and the GC content of the array. They found that minisatellites with higher copy numbers are more likely to be polymorphic, and that when the repeat units were more similar they were more likely to be an polymorphic VNTR, similar to the HistoryR measure discussed above. Overall, they predicted 29,224 (18.55%) out of 157,549 minisatellite TRs to be VNTRs corresponding to 9.1 VNTRs per million base pairs.

1.8.5 Methods to genotype VNTRs

Despite their biological significance, until recently, relatively few human VNTRs have been identified and studied in detail. The ever-increasing availability of accurate whole genome sequencing (WGS) data, however, provides an opportunity for high throughput, genome-wide VNTR genotyping. Furthermore, the emergence of PCR-free WGS datasets is reducing locus selection bias and enabling better filtering of false positive VNTR variants.

Nonetheless, genotyping variability in repeat sites remains challenging (Stolf et al., 2019; Gymrek, 2017; Tørresen et al., 2019). Few high-throughput tools are available for minisatellite genotyping. The adVNTR tool (Bakhtiari et al., 2018) trains a Hidden Markov Model for each VNTR locus of interest and has been used to predict variability in 2,944 VNTRs intersecting coding regions.

VNTRseek (Gelfand et al., 2014), developed in our lab, uses the Tandem Repeats Finder (TRF) (Benson, 1999) to detect and characterize TRs inside reads and then maps read TRs to TRs in a reference set. Because it builds pattern profiles before mapping, VNTRseek is robust in the presence of SNPs and small indels. The latest version of VNTRseek (`V1.10`) supports BAM/CRAM, FASTQ, and FASTA files, and processes a ∼100× coverage WGS dataset in a few hours on a machine with 16 processors and 4G RAM . The output is given in both VCF and SQLite format, which helps downstream data analysis.

1.9 A review on macrosatellites in the literature

Most macrosatellites are polymorphic and exhibit different copy numbers in different individuals (Schaap et al., 2013; Brahmachary et al., 2014). The high rate of polymorphism among macrosatellites is believed to allow the organism to evolve rapidly, for example, in order to adapt to environmental changes (Vinces et al., 2009; Verstrepen et al., 2005). Polymorphic macrosatellites have also been linked to diseases, including facioscapulohumeral muscular dystrophy (FSHD) (Geng et al., 2012; de Greef et al., 2009; Jones et al., 2012; Lemmers et al., 2012), immunodeficiency, centromeric region instability, facial anomalies syndrome (ICF) (Kondo et al., 2000), and cancer (Dumbovic et al., 2018; Balog et al., 2012). Many macrosatellites have high GC content and are often hot spots for methylation, which can result in reduced expression of the downstream gene (Huichalaf et al., 2014; Giacalone et al., 1992; Chadwick, 2009). Tandem repeats of CG-rich islands can trigger heterochromatization in a copy number-dependent manner, resulting in repeat-induced gene silencing (RIGS) (Garrick et al., 1998; Ye and Signer, 1996).

Despite the importance of macrosatellites in human biology, only a handful of macrosatellites have been studied, in part because of a lack of computational tools to characterize them in a high throughput fashion. In 2010, four macrosatellites were experimentally genotyped and characterized in the CEPH family (Tremblay et al., 2010), where it was found that macrosatellites are polymorphic across the genome and exhibit meiotic and mitotic instability. In a second study in 2013, six autosomal and two X chromosomal macrosatellites were genotyped in 270 HapMap individuals from Central Europe, Asia and Africa using Southern blotting (Schaap et al., 2013). Great variability in size for these loci was found, with a mitotic mutation rate of 0.4–2.2%. In 2014, Nanostring probes were used to genotype 15 macrosatellite loci in 165 HapMap samples and five primates (Brahmachary et al., 2014). The authors found that macrosatellite loci were highly polymorphic in the human genomes and that the true copy number of many macrosatellites is under-represented in the human reference genome. The authors also found association of macrosatellite copy numbers with nearby DNA methylation levels, which resulted in RIGS. Finally, they showed that macrosatellite loci are under selection, suggesting that they have biological roles.

Genotyping macrosatellites is difficult. Macrosatellites are too large to be genotyped by probes. They cannot be detected by typical assembly methods or SNP tagging (Burgner et al., 2003) since SNPs mutate at rate 10^{-8} rate and tandem repeats at 10^{-4}.

Long read technologies that use DNA polymerization such as PacBio and nanopore cannot be used to detect long tandem repeats due to high error rates (up to 15%) and GC content biases. NanoSatellite (De Roeck et al., 2019) is a tool that genotypes tandem repeats using long-read sequencing on PromethION directly using electric current data. It can improve tandem repeat genotyping bias on GC content in long reads and is more robust against mutations in repeat units. However, such long read datasets are not currently available for many applications.

Short NGS reads, on the other hand, are the most cost- and time-effective sequencing technology for variation detection. While macrosatellites fall in the category of copy number variants (CNV), high-throughput CNV methods such as CNVseq (Xie and Tammi, 2009), LUMPY (Layer et al., 2014), and CONY (Wei and Huang, 2020), rely on read depth and fail to genotype macrosatellites because repeat-rich regions cause ambiguity in aligning short reads and assembly.

1.10 Thesis rationale

A limited number of studies have characterized minisatellite and macrosatellite VNTRs in human populations. VNTRs play an important role in gene regulation and neurodegenerative diseases and disorders. However, to date the diversity of VNTRs has not been investigated in a large cohort of humans. The primary goal of this thesis is to provide a comprehensive collection of human VNTRs in a large cohort of humans from various populations. By identifying the VNTRs and their characteristics, we can better understand their functional role in human biology and provide a platform to incorporate them in studies of human disease.

In Chapter 2, VNTRs on a large cohort of human genomes are characterized on 2,800 WGS datasets (2,770 individual genomes) using VNTRseek. The VNTRs were characterized by their heterozygosity, allele diversity, and allele frequency. In Chapter 3, a population-wide study of VNTRs in 2,504 unrelated genomes is carried out to investigate the functional role of VNTRs on gene regulation and expression intensities. A model to predict ancestry from VNTRs will also be presented. Chapter 4 focuses on validating and confirming the VNTR predictions. The predictions are validated in silico and in vitro, and the consistency is evaluated across sequencing platforms and by Mendelian inheritance on related genomes. Also, the error rate of the predictions is quantified and the source of these errors are discussed. Chapter 5 delves deeper into the problem of type 1 and type 2 errors, and presents a computational tool named mlZ (machine learning on Z-scores) to detect false positive allele and infer false negative ones. The tool is tested thoroughly on simulated data and on real data. Chapter 6 presents a novel computational tool named MaSUD to detect macrosatellites using short NGS reads and it's application on the 1000 Genomes Project data. The dissertation will conclude in Chapter 7, with a discussion of limitations of our work, comparison to the literature, the impact of this thesis, and future directions.

Chapter 2

Characterization of VNTR minisatellites in humans

2.1 Introduction

Tandem repeats (TRs) are patterns of 7–100 bp repeated consecutively in the human genome. A TR with difference in copy number across genomes is called a Variable Number Tandem Repeat (VNTR). As described in Chapter 1, VNTRs have been associated with diseases and disorders and are predicted to play an important role in gene regulation. To characterize VNTRs in the human population, I downloaded 2,800 WGS datasets from 2,770 individuals and processed them with VNTRseek on the GRCh38 reference genome. We investigate the abundance of VNTRs, their allele diversity and frequency, heterozygosity, and their major alleles.

In Section 2.2 a brief description of the datasets used and the methodology will be discussed. The results section will follow in Section 2.3, first by summarizing the VNTRs detected by each data source, and then various characteristics, i.e., heterozygosity, frequency of alleles, common genotypes, of the VNTR loci will be investigated. Section 2.3.6 will summarize VNTRs in the tumor-normal paired tissues. Finally, in section 2.4 a short summary of the results will be given.

2.2 Materials and methods

This section provides details on the datasets used in this thesis. Also the filtering steps used to obtain the reference TR and how the VNTR genotyping was performed are described.

2.2.1 WGS datasets

Datasets comprising 2,801 PCR-free, WGS samples from 2,770 individuals were used in this study (Table 2.1). This datasets consist of the following:

- 30 individuals from the 1000 Genomes Project Phase 3 (The 1000 Genomes Project Consortium, 2015), including the Utah (CEU) and Yoruban (YRI) trios (mother-father-child);

- 2,504 unrelated individuals mostly overlapping with the 1000 Genomes Project, recently sequenced at >30× coverage by the New York Genome Center (NYGC) [1];

- 253 individuals from the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016);

- seven individuals sequenced by the Genome in a Bottle (GIAB) Consortium (Zook et al., 2016), including the Chinese (HAN) and Ashkenazi Jewish (AJ) trios and NA12878 (with ID HG001);

- two haploid hydatidiform mole cell line genomes, CHM1 (Chaisson et al., 2015) and CHM13 (Huddleston et al., 2017);

- tumor/normal tissues from two unrelated individuals with breast cancer (breast invasive ductal carcinoma cell line/lymphoblastoid cell line) from the Illumina Basespace public WGS datasets (Drmanac et al., 2010); and

[1] These data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S.

- the AJ child sequenced with PacBio Circular Consensus Sequencing (CCS) reads (Wenger et al., 2019).

Duplicates of 27 genomes were present in two datasets: 1000 Genomes and NYGC. One of these, NA12878, was also included in the GIAB dataset (with ID HG001).

Data source		Read	Read	Samples
Data source		Length (bp)	Coverage	in Set
1000 Genomes Phase 3 HC	YRI trio	250	71–73×	3
	CEU trio	250	55–63×	3
	Others	250	33–66×	24
New York Genome Center (NYGC)		150	29–101×	2,504
Simon's Genome Diversity Project (SGPD)		100	33–133×	253
Genome In A Bottle (GIAB)	AJ trio	250	61–69×	3
	HAN trio	148 / 250	111–333×	3
	NA12878 (HG001)	148	291×	1
Haploid genomes	CHM1	148	40×	1
	CHM13	250	128×	1
Illumina basespace	Tumor/Normal	101	38–88×	4

Table 2.1: WGS datasets. 2,801 publicly available WGS samples, for 2,770 individuals, were used in this study. Read coverage was calculated as 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 coverage values are approximate. The 1000 Genomes Phase 3 samples were released in 2015. The NYGC samples were released in 2020 by the New York Genome Center (NYGC). For the Simons Genome Diversity Project (SGDP), released in 2016, only datasets which were not present in the the 1000 Genomes datasets were used.

Overall, read coverage ranged from approximately 27×, in the PacBio sample to 333×, in the GIAB Chinese child. Besides the PacBio data, reads consisted of three lengths, 100/101 bp (257 samples), 148/150 bp (2,508 samples), and 250 bp (35 samples). All data were downloaded as raw fastq files, except for the PacBio data which were obtained as a BAM file with reads aligned to GRCh37. SRA links to the data are given in Table 2.1.

The majority of the population-wide analyses presented in Chapter 3 were performed on the 2,504 genomes from NYGC. The 253 genomes from SGDP provided insight into under-represented populations. The 27 genomes duplicated in the 1000 Genomes and NYGC datasets were used in Chapter 4 to measure consistency across sequencing platforms. The trios from the 1000 Genomes (CEU and YRI) and GIAB (AJ and Chinese HAN) datasets were used for analyzing Mendelian inheritance in Chapter 4. The cancer datasets were used to find possible changes in VNTRs in tumor tissues (Section 2.3.6). The haploid genomes were used for measuring false positive heterozygous calls (Chapter 4 and Chapter 5). The PacBio data was used for validation purposes only (Chapter 4).

2.2.2 The reference TR set

The 22 autosomes and sex chromosomes from the human reference genome GRCh38 were used to produce a reference set of TRs in TRDB (Gelfand et al., 2007)with the TRF software and four quality filtering steps as described in (Gelfand et al., 2014). In addition, centromere regions were excluded from the reference set. These filtering tools are available online in TRDB. Starting with 1,199,362 TRs found by TRF, we curated a filtered reference set with 228,486 TRs.

Using VNTRseek, the TRs were classified into two subcategories, singleton and indistinguishable. 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 that share highly similar patterns and flanking sequences, and can therefore produce misleading genotype calls. Indistinguishable TRs were identified using the procedure described in (Gelfand et al., 2014), performed by Yozen Hernandez, a former PhD student. Each TR array from the initially filtered reference set of 228,486 TR loci was treated as a single read and all such reads were mapped to the original unfiltered TR set using VNTRseek. Any TR that mapped to a locus other than its own was labeled indistinguishable. Indistinguishable TRs were not removed from the reference set, but genotype calls in the output of VNTRseek were flagged (and filtered) if the locus was indistinguishable. In the output VCF files the filtering field marks these TRs as "SC", meaning they did not pass the "Singleton Criterion" filter. Alleles from indistinguishable TR loci detected in a sample were filtered from that sample before further processing. Indistinguishable TRs (total 37,200 or about 16% of the reference set) were filtered.

Simulation testing revealed that some singletons produced false positive VNTRs. To minimize this issue, an additional filtering step was added to eliminate problematic singleton loci from the reference set. The following procedure was followed. For each reference TR, a sliding window equal in size to a specified read length was used to generate reads. The leftmost window ended one base upstream (-1 bp) of the TR array start position and the rightmost window began one base downstream (+1 bp) of the TR array end position, with the window moving in increments of one base. The combined simulated reads for all TRs were mapped back to the reference set with VNTRseek. A Singleton TR locus was then removed if:

- it was the source of at least one read resulting in a VNTR call, either at its own locus or another locus; or
- at least one read drawn from a different locus resulted in a VNTR call for the TR.

The procedure was repeated for three separate read lengths, 100 bp, 150 bp, and 250 bp, to produce three separate reference sets.

We assumed that genotyping was possible for a reference TR locus, given a particular read length, if the TR array length plus a minimum 10 bp flank on each side, would fit within the read. We will discuss the detection range of VNTRseek in Chapter 5.

2.2.3 Genotyping TRs and VNTRs

Each short read WGS dataset described in Section 2.2.1 was processed separately with VNTRseek using default parameters: minimum and maximum flanking sequence lengths of 10 bp and 50 bp, respectively, on each side of the array, and requiring at least two reads mapped with the same array copy number 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. The VCF files contained two specialized FORMAT fields: SP, for number of reads *supporting* each allele, and CGL, to indicate genotype by the number of copies *gained* or *lost* with respect to the reference. For example, a genotype of 0 indicated detection of only the TR reference allele (zero copies gained or lost), while 0,+2 indicated a heterozygous locus with a reference allele and an allele with a gain of two copies.

To remove clear inconsistencies, we filtered the VCF files to remove per sample VNTR loci with more alleles than the expected number of chromosomes. VNTR loci which reported more alleles in a diploid sample than the expected number of chromosomes were termed *Multis* in that sample. They correspond to loci with the following:

- three or more alleles on an autosomal chromosome,
- three or more alleles on chromosome X of a female individual,
- any allele on chromosome Y of a female individual, or
- two or more alleles on a sex chromosome of a male individual

For the two haploid samples, any locus that reported more than one allele, or any Y chromosome locus that reported any allele was termed *Multi*.

After Multi filtering, a TR locus was labeled as a VNTR if any remaining allele, different from the reference, was observed in any sample. The number of reference

TR alleles that could be genotyped using each of the read lengths in our data is summarized in Table 2.2.

Read Length(bp)	Reference Set Size	Reference Singletons	Singletons Removed	Final Ref. Set	Expected Genotyped (%)
100–101	228,486	191,286	1,704	226,782	153,293 (80.14%)
148–150	228,486	191,286	1,976	226,510	168,742 (88.21%)
250	228,486	191,286	4,812	223,674	177,864 (92.98%)

Table 2.2: Filtering out reference singletons to reduce false positive VNTRs. 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. 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 100/101 bp set, 100 bp was used as read length, for the 148/150 bp set, 150 bp was used). Percent is the Expected Genotyped out of all the original Reference Singletons.

2.2.4 Data availability

The reference TR set files, output VCF files, and the pre-processed data files along with the code to create figures and tables are published at: DOI 10.5281/zenodo.4065850

2.3 Results

In this section, a summary and characterization of the VNTR predictions is presented.

2.3.1 About one in five minisatellite TRs are variable in the human population.

WGS datasets from 2,770 human genomes were analyzed with VNTRseek to detect VNTRs. Overall, 184,315 out of 191,286 singleton reference TR loci (~96%) were genotyped across all samples (Table 2.3) while 5% of the loci had TR arrays too long to fit within the longest reads and could only be genotyped if they lost a sufficient number of copies.

Dataset	Samples	TRs Genotyped	Multis	VNTRs Detected
1000 Genome Project	30	178,395	366	8,761
NYGC	2,504	177,612	1,181	33,403
SGDP	253	156,803	221	9,944
GIAB	7	178,804	239	6,736
CHM1	1	159,563	175	1,118
CHM13	1	170,805	632	1,977
Tumor-Normal	4	150,531	21	1,291
Totals	2,800	184,315	-	35,638

Table 2.3: Summary of genotyping VNTRs by data source. TRs Genotyped is the number of distinct TR loci genotyped across all individuals within a dataset. (All other numbers are also per dataset.) Multis are TR loci genotyped in a single individual with more than the expected number of alleles. They could be artifacts or indicate copy number variation in a genomic segment. Multis were excluded from further analysis on a per sample basis. VNTRs Detected is the number of TR loci, excluding Multis, with a detected allele different from the reference.

A total of 5,198,392 VNTRs were detected, corresponding to 35,638 (19%) distinct VNTR loci, indicating an abundant occurrence of these variable repeats. Their occurrence within genes was common, totaling 7,698 protein coding genes, and 3,512 exons. The resulting genotypes were output in VCF format files and summarized for each genome (see Section 2.2.4).

2.3.2 The number of TRs and VNTRs genotyped depends on coverage and read length.

To determine the effect of coverage and read length on VNTR genotyping, two quantities were measured: 1) the percentage of reference singleton TRs that were genotyped and 2) the total number of singleton VNTRs that were detected in each genome. Only singleton loci were considered in all further analysis. Figure 2·1 shows that there was a strong positive correlation between coverage and the ability to genotype TRs. A strong correlation with read length was also apparent, however, the effect was larger, primarily due to the ability of longer reads to span, and thus detect, longer TR arrays. These results suggest that our VNTR numbers are under-counts.

VNTR detection was similarly dependent on coverage and read length, as shown in Figure 2·2 However, detection was also positively correlated with population, which seemed likely due to the evolutionary distance of populations from the reference genome, which is primarily European (Lappalainen et al., 2013; Günther and Nettelblad, 2019). For example, in the 250 bp trios with comparable coverage, the African Yoruban genomes (YRI) had the highest number of VNTRs, followed by the Ashkenazi Jewish genomes (AJ), and finally, the Utah genomes (CEU). Notably, within each trio, the VNTR counts were similar.

The haploid genomes CHM1 (150 bp) and CHM13 (250 bp) had greatly reduced VNTR counts relative to genomes with similar coverage and read length. This was because in these genomes, which are derived from haploid genomes, the parental heterozygous loci with one reference allele would appear to be VNTRs, on average, only about half the time.

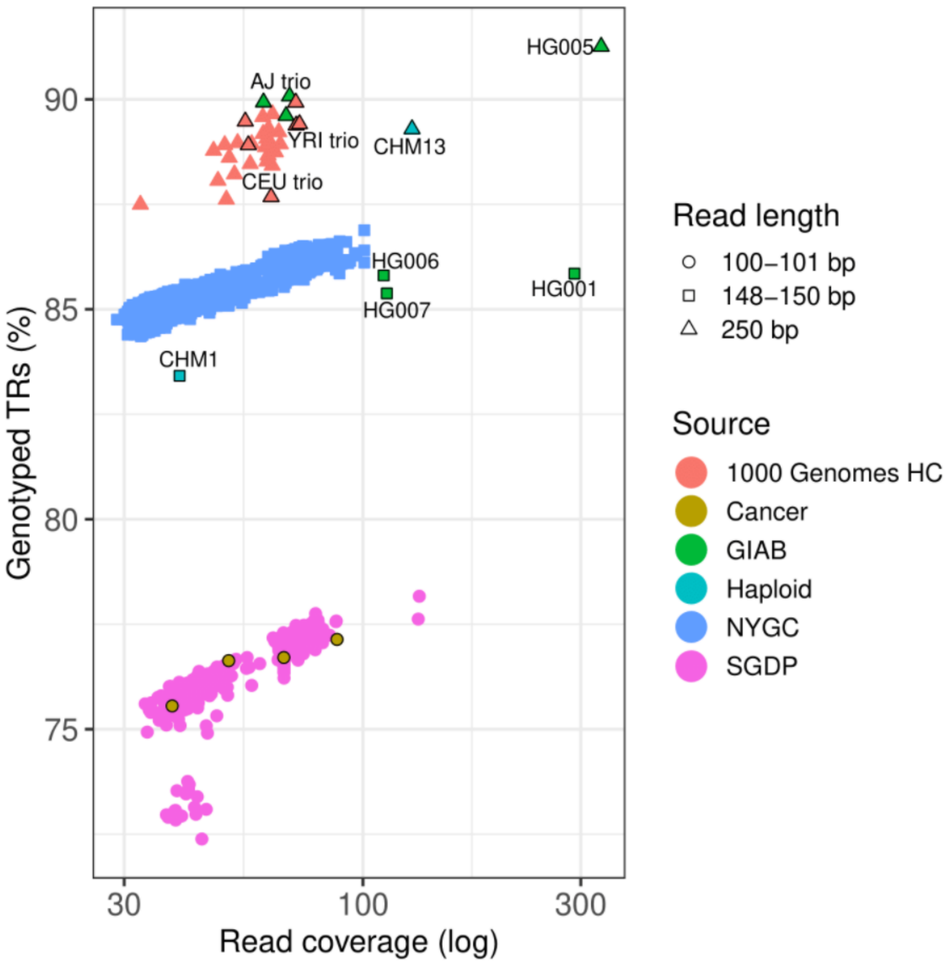


Figure 2·1: TR Genotyping sensitivity. This graph shows the relationship between coverage, read length, and the percentage of TRs in the reference set that were genotyped. Each symbol represents a single sample and specific samples are labeled. Increasing read length had the largest effect on sensitivity because many reference TR alleles could not be detected at the shorter read lengths (see Table 2.2).

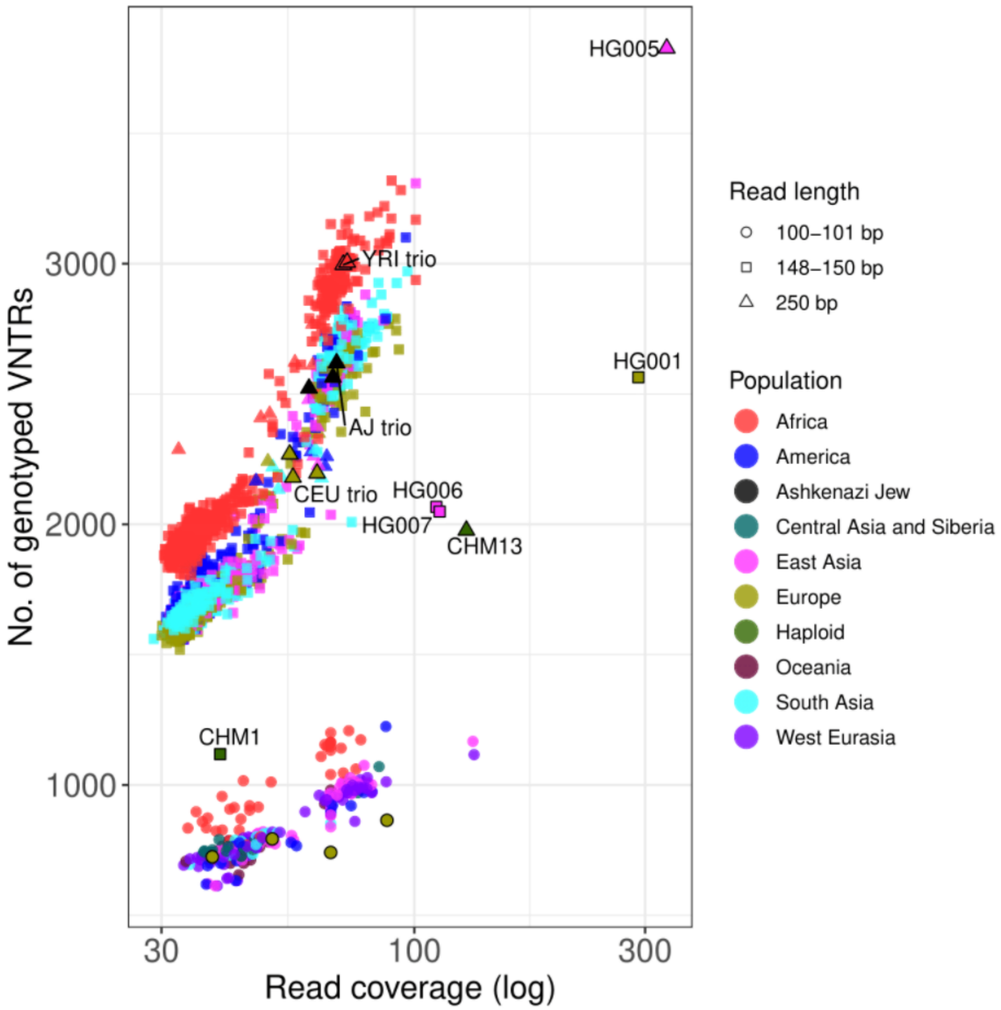


Figure 2·2: Number of VNTRs detected per genome. Graph shows the relationship between coverage, read length, and the number of VNTR loci detected. Read length and coverage both had large effects. Coloring of symbols shows that the population also had a strong effect, reflecting distance from the reference, which is primarily European. Note the reduced numbers for CHM1 (150 bp) and CHM13 (250 bp). Because they are haploid genomes, parental heterozygous loci with one reference allele would appear to be VNTRs, on average, only about half the time.

2.3.3 More than two alleles are common in VNTRs.

Two alleles were detected in the majority of VNTR loci across all datasets (Figure 2·3). However, at 10,698 loci (29%), three or more alleles were detected. In a substantial number of loci (5,395), the reference allele was never seen, but only 105 of these were in the VNTRseek detectable range for the 100 bp and 105 bp reads, which made up the bulk of our data.

In total, 4,084 VNTRs from NYGC the reference genotype (0/0) could be observed (array length ≤130 bp). And, in a total of 1,166 of these, the reference genome was not the major allele. These VNTRs with non-reference major alleles overlap with 731 genes, 540 of which encode proteins. A total of 91 of these VNTRs occurred inside 78 exons. In 124 VNTRs (overlapping 90 genes), the reference genotype was seen in less than 1% of the population. A list of VNTRs where the reference was never observed, while it was detectable is given in Table 2.4.

In the case of TR ids 182290167 and 182289941, which overlap with introns 9 and 13 of gene CTNNA3, all individuals had genotype -1/-1 (2,501) and +1/+1 (2,429), respectively. For TR id 182387900 overlapping with gene POU6F1, 2,503 people had genotype -1/-1 and the other person probably had an error. Other examples of VNTRs where the reference allele was never observed are: TR id 182501000 in the fourth intron of the gene HERC2, where every individual had -6/-6 and reference was never seen. Every genotyped individual had genotype -1/-1 for TR id 182967914overlapping the fifth intron of ASIC5. For TR id 182996967 overlapping intron ten of gene SPEF2, all genotyped individuals were -2/-2. These results suggest that the reference genome does not represent the major alleles. It has been previously shown that using pan-graphs with known VNTRs improves mappability of short reads (Lu et al., 2020).

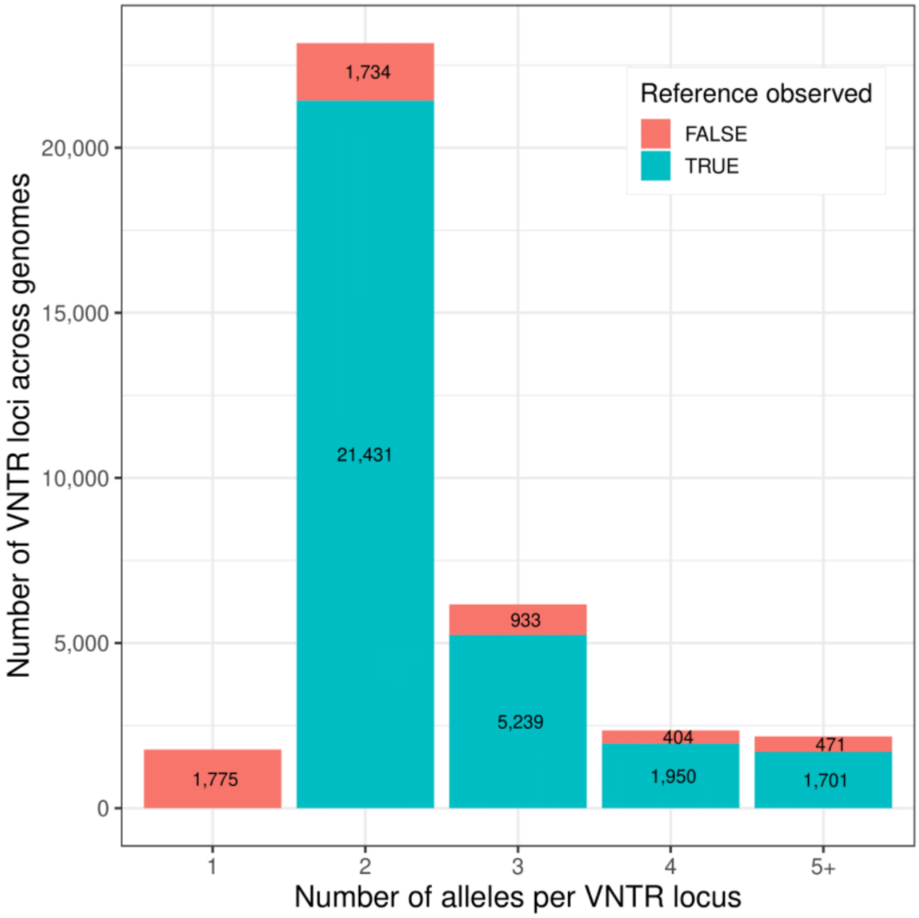


Figure 2·3: Alleles detected per VNTR locus. Each bar represents the specific number of alleles detected across all datasets. Coloring shows that proportion of loci where the reference allele was (TRUE) or was not (FALSE) observed

TRid	Major genotype	Count	Gene	Pattern size (bp)	Reference copy number	Array size (bp)
182170668	1/1	2,499	PRDM16	25	2.7	67
182227226	-1/-1	2,503		37	3.1	114
182249995	-1/-1	2,504		16	2.9	47
182289941	1/1	2,504	CTNNA3	13	4.4	59
182290167	-1/-1	2,504	CTNNA3	8	5.5	44
182331426	-2/-2	2,500	EHF	8	7.3	58
182387900	-1/-1	2,504	POU6F1	34	3.2	110
182400618	-1/-1	2,503		7	4.6	32
182413455	-2/-1	2,488		8	7.6	61
182416530	-1/-1	2,499	CABP1	18	4.2	76
182552008	1/1	2,503	AC007493.2	12	3.2	37
182713567	-1/-1	2,489		26	3.5	91
182759585	1/1	2,504	IGFBP2	9	4.3	39
182797532	1/1	2,502		7	5.4	41
182829368	-1/-1	2,504		17	3.4	57
182916461	-1/-1	2,504	AC006296.3	12	4.3	51
182996967	-2/-2	2,503	SPEF2	27	4.3	121
183017144	-1/-1	2,503	MCTP1	15	3.7	55
183080876	-1/-1	2,504		16	3.9	62

Table 2.4: Example of VNTRs where the reference genotype was not observed. VNTRs where the reference genotype was never detected, while over 2,400 individuals found another genotype are listed. In the second column, the major genotype detected in most individuals is given. The third column in the number of individuals with the major genotype. The genes/transcripts these VNTRs overlapped with are listed in the fourth column. In all these loci, the reference genotype was detectable, i.e. the reference array size was <130 bp (last column).

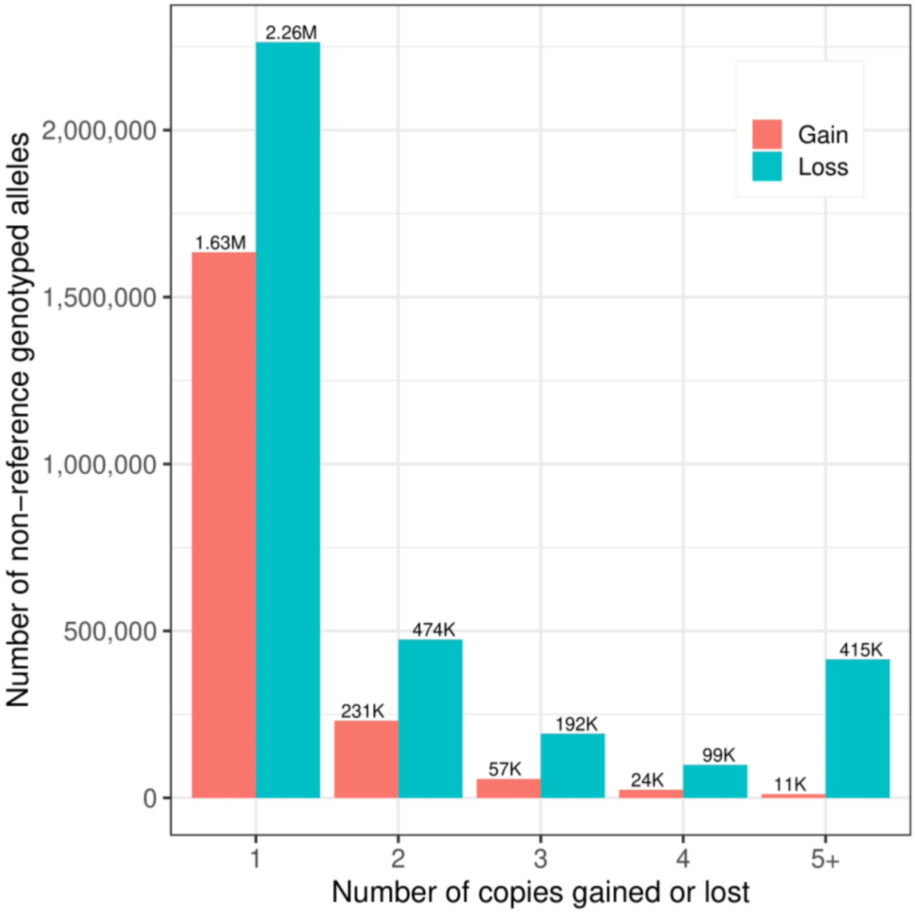


Figure 2·4: Number of VNTR alleles with gains or losses. Each bar represents a specific number of copies gained or lost in non-reference VNTR alleles relative to the reference allele. Loss was always more frequently observed.

2.3.4 Loss of VNTR copies relative to the reference is more common than gain of VNTR copies.

Overall, VNTRseek found approximately 1.8-fold more alleles with copy losses (3,444,128), with respect to the reference copy number, than gains (1,958,250). Loss of one copy (2,263,608) was the most common type of VNTR polymorphism (Figure 2·4).

The overabundance of VNTR copy loss may be an underestimate. Because VNTRseek requires a read to span a TR array in order for it to be detected, only limited gain in copies could be observed. Observing gain of one copy would have been possible in approximately 68%, 82%, and 92% of loci for samples with read lengths of 100 bp, 150 bp, and 250 bp, respectively. By contrast, the reference locus needed to have a minimum of 2.8 copies for a loss of one copy to be observed by TRF, and only 16% of the reference loci met this criterion. Higher observed copy loss could be explained by a bias in the reference genome towards including higher copy number repeats, or by an overall mutational preference for copy loss.

2.3.5 VNTRs have high heterozygosity

High heterozygosity in human populations suggests higher genetic variability and may have beneficial effects on a range of traits associated with human health and disease (Campbell et al., 2007). Since calculating heterozygosity for VNTRs is not straightforward (because of limitations on discovering alleles, especially within shorter reads), we used the percentage of detected, per-sample heterozygous VNTRs as an estimate for heterozygosity. At read length 250 bp, per-sample heterozygous VNTR loci comprised approximately 46–55% of the total, which is comparable to previous theoretical estimates of 43–59% (Denoeud et al., 2003). At shorter read lengths, the bottom of the range extended lower (∼38–57% for 150 bp reads, ∼29–51% for 100 bp reads, Figure 2·5), as expected, because longer alleles were undetectable if they did not fit within a single read.

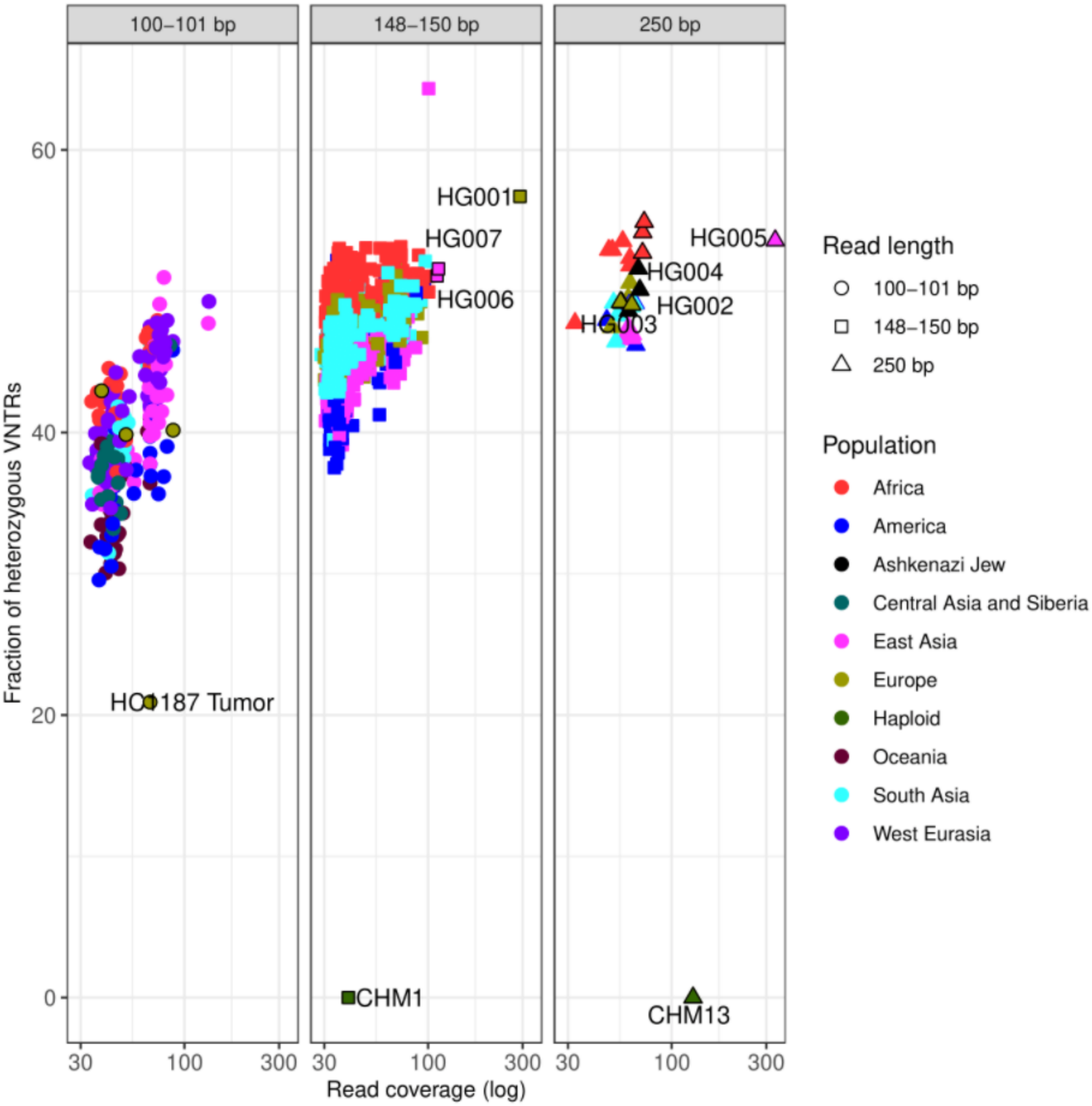


Figure 2·5: Fraction per sample of VNTR loci called as heterozygous. Samples are separated by read length. Higher read length and coverage provide more statistical power to detect heterozygous calls. Stratification by population is evident and is further displayed in Figures 2·6, 2·7, and 2·8

Interestingly, despite the previous comment, within genomes that were comparable in read length and coverage, the fraction of heterozygous loci clustered within populations (Figure 2·6, Figure 2·7, and Figure 2·8), with African genomes generally having the most heterozygous calls and East Asians the fewest. This result is consistent with previous findings of population differences in SNP heterozygosity among the Yoruban and the Ashkenazi Jewish individuals as compared to European individuals (López Herráez et al., 2009; Bray et al., 2010), and suggests that there is higher genomic diversity among African genomes, as has been previously noted (Edea et al., 2015).

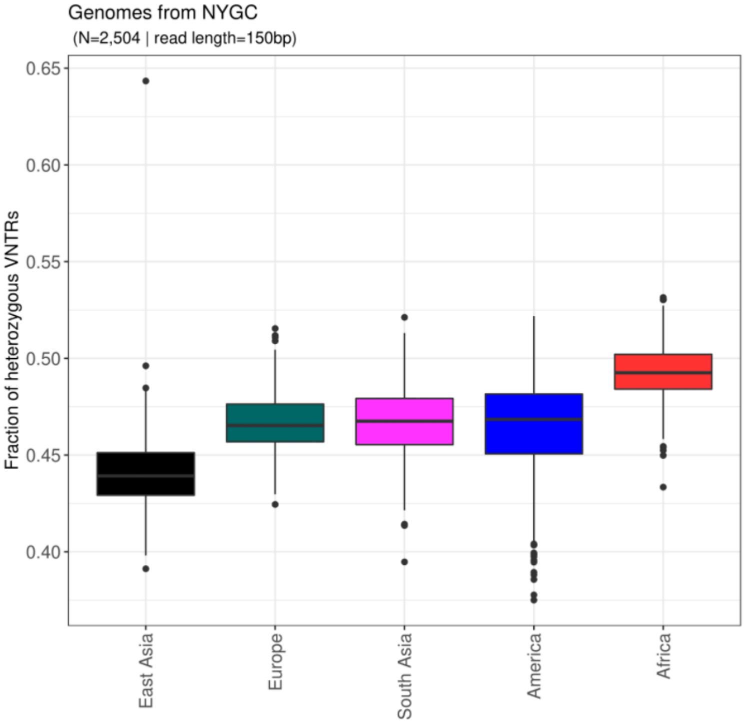


Figure 2·6: Heterozygous VNTR calls on the 2,504 NYGC 150 bp samples per superpopulation. Data are presented as box plots for each superpopulation showing the interquartile range and median (middle line). Africans had the highest percentages of heterozygous calls and East Asians had the lowest compared to the other superpopulations.

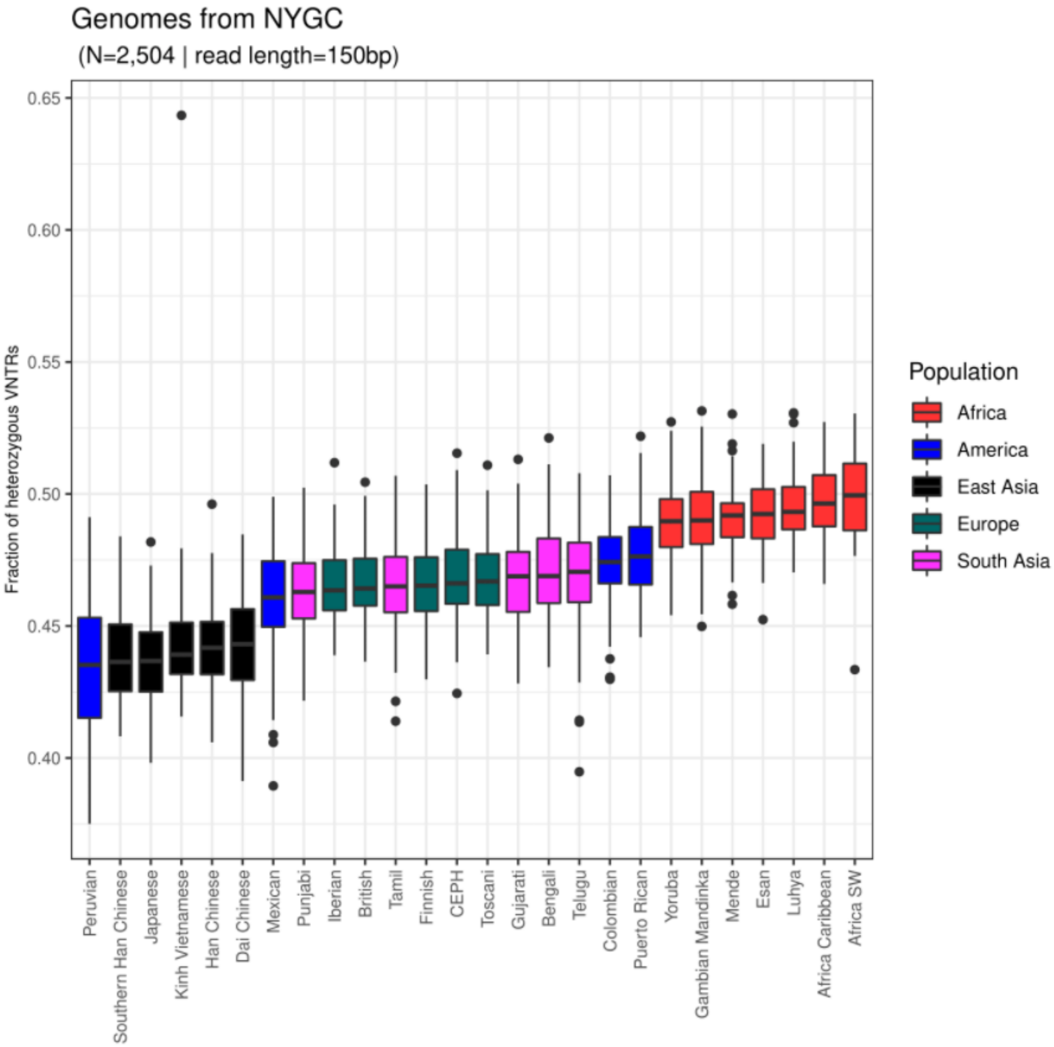


Figure 2·7: Heterozygous VNTR calls on the 2,504 NYGC 150 bp samples per subpopulation. Data are presented as box plots for each subpopulation showing the interquartile range and median (middle line). Among the admixed Americas populations, Peruvians had the lowest median percentage of heterozygous calls, similar to that of East Asians. The Peruvian population appears to have had the least mixing with African and European genomes in the NYGC dataset, based on this measure. South Asian and European sub-populations had similar frequencies of heterozygous calls.

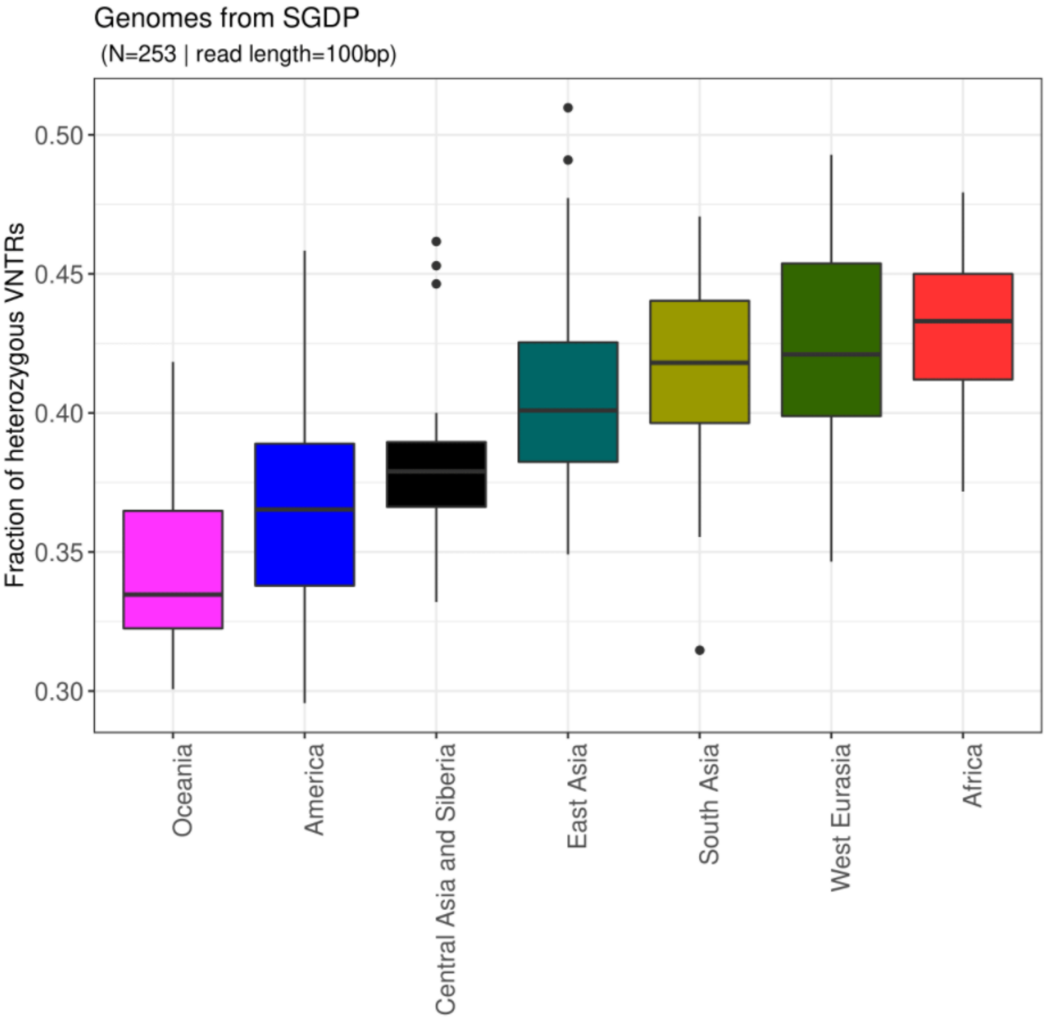


Figure 2·8: Heterozygous VNTR calls on the 253 SGDP 100 bp samples per superpopulation. Data are presented as box plots for each superpopulation showing the interquartile range and median (middle line). The SGDP aimed to sequence underrepresented populations. Although the overall percentage of heterozygous calls was lower than for the NYGC samples due to lower sensitivity with shorter reads, Africans still had the highest percentage of heterozygous calls. Interestingly genomes with Oceania ancestry had a very low percentage of heterozygous calls.

2.3.6 VNTRs in tumor vs normal tissue.

On the cancer genomes, we investigated the change in VNTR alleles and the change in heterozygosity. The data consisted of paired tumor-normal samples from individuals HC1187 and HC2218. A significant loss of heterozygosity (LOH) was observed in predicted VNTRs of one of the tumor tissues compared to its matching normal tissue (sample ID HC1187). The percentage of heterozygous VNTRs was approximately double that seen in the normal tissue (∼38% vs ∼19%) (Table 2.5). Extreme loss of heterozygosity in small variants has previously been reported in these samples by Illumina Basespace (Edea et al., 2015; Raczy et al., 2013) with the number of heterozygous small variants in HC1187 being four times lower in the tumor tissue compared to the normal. Taken together, these results suggest that VNTR LOH could be linked to tumor progression.

Knowledge of gene associations with somatic tumor mutations (VNTR alleles present in a tumor, but not normal tissue) could be useful as indicators of cancer prognosis and for therapy. In the HC2218 individual, somatic tumor mutations overlapped with lncRNAs (ACO73336.1, AC107959.2, AL355388.2), introns (C3orf67, COX17, DHRS3, DPP6, GAN, PCGF3, RGS12, SLC25A13, SLC6A19, TACR2, TEPP), and promoter regions (TRIM24, DUSP4). RGS12 is a known oncogene, TRIM24 has been associated with prognosis in breast cancer (Chambon et al., 2011; LU et al., 2017; Pathiraja et al., 2015) and over-expression of DUSP4 has been shown to improve the outcome of chemotherapy and overall survival (Balko et al., 2012; Menyhart et al., 2017).

In the HC1187 individual, somatic tumor mutations overlapped with lncRNAs (LINC01708, AC1058290.1, AC104596.1), exons (THNSL2), introns (AJAP1, SMAD1, FLT4, PTPN3, ADAMTSL2, ANO2, SOX5, SGCG, WDR72, NQO1, CCDC200, ARHGAP45, AC005258.1, PEAK3) and promoters (HFM1, TBK1, GNS, LEMD3, FGFR3, VIPR2). SMAD1, PTPN3, NQO1, and FGFR3 are known oncogenes, and TBK1 and FGFR3 have been used as treatment targets for HER2+ breast cancer (Deng et al., 2014; Long et al., 2020).

Sample	Coverage	Multis	Total	VNTRs	Heterozygous	Ratio
	(×)		TRs		VNTRs	(%)
HC1187 Normal	51	43	146,794	793	302	0.38
HC1187 Tumor	67	60	146,969	741	144	0.19
HC2218 Normal	38	40	144,710	724	293	0.4
HC2218 Tumor	88	57	147,812	864	328	0.37

Table 2.5: Ratio of heterozygous VNTRs in paired normal and tumor samples. The ratio of heterozygous calls in the tumor samples was lower than the paired normal tissue, suggesting loss of heterozygosity. In HC1187, the percentage was half that of the normal tissue. With same read length and higher coverage in the tumor samples, this finding cannot likely be attributed to artifacts.

Additionally, in both tumors a large number of loci exhibited loss of both alleles in comparison to the normal tissue (Table 2.6). Given that the coverage for the tumor samples was significantly higher than for the normal tissue, it is unlikely that these observations were due to artifacts. Also, the tumor samples did not show a higher percentage of filtered multi VNTRs (too many alleles) than the normal samples (1.37% and 1.23% in normal tissue vs 1.72% and 1.71% in tumor tissue).

2.4 Summary

Variable Number Tandem Repeats (VNTRs) are tandem repeat (TR) loci that vary in copy number across a population. Using VNTRseek, a program to detect VNTRs using short WGS reads, we analyzed human whole genome sequencing datasets from 2,770 individuals in order to detect minisatellite VNTRs, i.e., those with pattern sizes ranging from 7 bp to 126 bp, and with array lengths up to 230 bp. A total of 5,209,412 human VNTRs corresponding to 35,638 loci were identified, corresponding to 19% of the TR loci. At higher read length and coverage VNTRseek had sensitivity

Genotype		HC1187	HC2218	Description
Normal	**Tumor**			
AA	AA	391	349	No change
	—	75	32	Loss of both alleles in tumor
	AB	9	31	Allele mutation in tumor, or failure to detect in normal
	BB	2	1	Allele mutation in tumor
AB	AB	109	232	No change
	—	116	39	Loss of both alleles in tumor
	AA	90	40	Loss of one allele (LOH) in tumor
	AC	1	0	Allele mutation in tumor
—	AA	103	127	Failure to detect in normal
—	AB	36	84	Failure to detect in normal

Table 2.6: Comparison of VNTR alleles in paired normal and tumor samples. Notation: AA means homozygous; AB and AC mean heterozygous; — means not detected. Due to significantly higher coverage in the tumor samples, we assumed the tumor genotyping was likely correct, whereas genotyping in the normal tissue may have failed to detect one or two alleles. The majority of VNTR loci detected in both tissues for each patient exhibited no change. The most common genotype change was loss in the tumor of one allele (LOH) or both alleles. Allele mutation was apparently uncommon.

>84%, and detected about 1,500 to 3,500 VNTRs per genome. VNTRseek has high sensitivity when the read length and coverage is high. However at similar read length and coverage, the number of VNTRs depends on the population, with Africans having higher per genome VNTRs. We characterize these VNTR loci by studying their frequencies, number of alleles, and heterozygosity. Although VNTRseek has lower recall on detecting losses, the number of losses detected was almost twice as the number of gains, with loss of one copy being the most frequent VNTR allele. About a third of the VNTR loci had three or more alleles, and in 150 cases the reference allele was never observed, suggesting the reference genome is incorrect at these loci. Further, we found that VNTRs have high heterozygosity, as predicted in the literature, and that Africans and East Asians have higher ratios of heterozygous VNTRs than other populations.

Chapter 3

Population-wide study of VNTRs

3.1 Introduction

As discussed in Chapter 1, detecting variants on a large population of humans is useful to understand the diversity of our genome. This information is crucial to differentiate disease causing mutations from healthy variation in the human population. The variants enriched in one population compared to other populations, could predispose a population to a phenotype or disease, which could be useful to design better treatments. Population-biased variants shed light on the evolutionary path of the human genome and can be used to understand the origin of human life, the diversity in each population, bottlenecks and natural selection that the population has gone under. Also, population-biased variants can be used to predict ancestry. In this chapter we analyze the VNTRs across the population and detect common VNTRs that occur in >5% of the individuals. The potential role of these common VNTRs in gene regulation is investigated by genomic annotation enrichment and gene set enrichment. Then, the expression of proximal genes to VNTRs are examined to find correlations. Finally, common VNTRs with population-biases alleles, i.e. alleles that are more frequently found in one population compared to other populations, are used to predict ancestry of the individuals.

In Section 3.2 the materials and methods used to analyse the VNTRs across populations are discussed. Section 3.3 will present the results. Section 3.4 will summarize this chapter's findings.

3.2 Materials and methods

In the following sections methods for detecting common and private VNTRs are described (Section 3.2.1). The material used for enrichment of common VNTRs by annotation is explained (Section 3.2.2). Then the methods of two population-wide analyses on common VNTRs are discussed: association with gene expression (Section 3.2.3) and population-biased VNTR alleles (Section 3.2.4).

3.2.1 Common and private VNTRs

To classify common and private VNTRs, we used results from the NYGC dataset (2,504 individuals) because the read length and coverage were comparable across all genomes (See Table 2.1). Additionally, these genomes contain no related individuals and represent a wide set of populations (26 populations from five continents).

VNTR loci were classified as common if they were identified as VNTR in at least 5% (126 individuals) of the individuals and classified as private if they were identified as VNTR in less than 1% (25 individuals). For these classifications, VNTRs genotyped on the sex chromosomes were excluded.

3.2.2 Annotation and enrichment

Annotation based on overlap with functional genomic regions was performed for the reference TR loci. Genomic annotations for GRCh38 were obtained from the UCSC Table Browser (Karolchik et al., 2004) in BED format. Known gene transcripts from GENCODE V32 (Hsu et al., 2006) were used along with tracks for introns, coding exons, and 5' and 3' exons. Regulatory annotations included transcription factor binding site (TFBS) clusters (ENCODE Project Consortium, 2012; Davis et al., 2018) and DNAse clusters (Thurman et al., 2012) from ENCODE 3 (Roadmap Epigenomics Consortium et al., 2015), and CpG island tracks (Gardiner-Garden and Frommer, 1987), comprising 25%, 15%, and 1% of the genome, respectively. Bedtools (Quinlan and Hall, 2010) was used to find overlaps between TR loci and the annotation features.

LOLAweb (Nagraj et al., 2018) was used to determine VNTR enrichment for genomic regions in comparison to the background TR annotations, and common and private VNTR enrichment in comparison to all VNTR annotations. TRs on the sex chromosomes were excluded in the background set. To identify gene and pathway functions that could be affected by common VNTR copy number change, genes with exons or introns overlapping with common VNTRs were collected and their enrichment computed using GSEA (Subramanian et al., 2005) for biological process Gene Ontology (GO) terms (Ashburner et al., 2000) and KEGG pathways (Kanehisa and Goto, 2000) with FDR p-value <0.05.

3.2.3 Association of VNTR alleles with gene expression.

To detect expression differences among individuals with different VNTR genotypes, mRNA expression counts from lymphoblastoid cell lines of 660 individuals by the Geuvadis consortium (Accession: E-GEUV-1) were downloaded (Lappalainen et al., 2013). A total of 445 individuals overlapped with the 2,504 NYGC genomes set.

We paired VNTR loci with genes within 10 Kbp, and extracted the genotypes for each individual at those VNTRs. When no genotype was observed for an individual, we classified the genotype as *other*. We did this because we assumed that the alleles were outside the detection range, given that genotypes were observed in other individuals with similar coverage. VNTR loci were retained for analysis if at least two genotypes were detected for that VNTR across all individuals (at least three if *other* was one of the genotypes) and if each genotype was observed in at least 20 individuals. Genes were excluded from analysis if the median TPM (Transcripts Per Kilobase Million) expression value equaled zero.

To control for confounders we used covariates for sex and population structure and detected additional hidden covariates using Iteratively Adjusted Surrogate Variable Analysis (IA-SVA) (Lee et al., 2018) on the log_2 normalized TPM values. For population structure we used the top five principal components determined from a principal components analysis of the informative SNP genotypes from the 445 individuals as reported by the 1000 Genomes project (`http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/`). Using IA-SVA and observing that covariates sixteen and above were over 85% correlated with other covariates (Figure 3·1), we chose fifteen hidden factors to include in our model. Finally, we used a linear regression $expression \sim sex + population\_PCAs + hidden\_factors$ with the log_2 normalized TPM values to extract residuals to be used in the downstream association model.

For each gene-VNTR pair, we used a one-way ANOVA test as $residuals \sim genotype$ to detect if the mean of any genotype class was different from the others. The p-values of the ANOVA tests were adjusted using FDR. Any gene-VNTR pair with FDR<5% was reported. For significant eQTLs, we calculated the maximum mean difference of the residuals for all pairs of genotype classes for reporting purposes.

To associate eQTLs with histone marks or open chromatin, we downloaded narrow peaks data in GRCh38 in bed format from 14 experiments on histone marks and one on DNAse hypersensitive sites from the GM12878 (B-Lymphocyte) cell line from the ENCODE project (ENCODE Project Consortium, 2012) (source IDs are given in Table 3.1). Any overlaps of peaks with the eQTL VNTRs were reported.

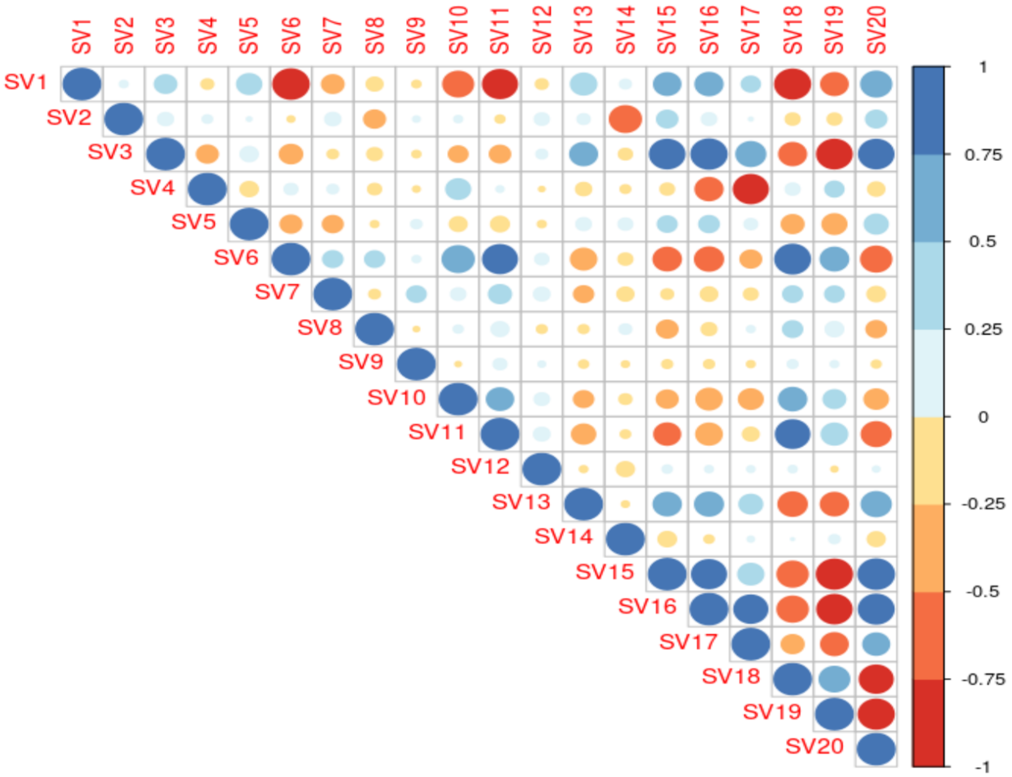


Figure 3·1: Detecting hidden covariates with SV.A To detect unknown confounders, Iteratively Adjusted Surrogate Variable Analysis (iasva) (Lee et al., 2018) was applied on the log_2 normalized TPM values of mRNA expression from the Geuvadis consortium (Accession: E-GEUV-1). We observed the first covariates were independent and the covariates 16–20 were >85% correlated to other covariates. Therefore, I chose five hidden factors to include in our model.

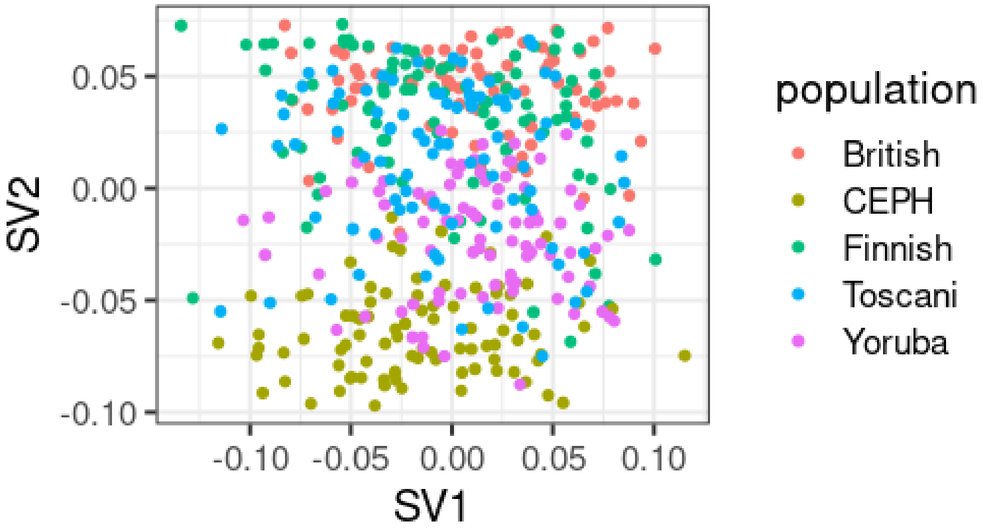


Figure 3·2: First and second hidden factors.

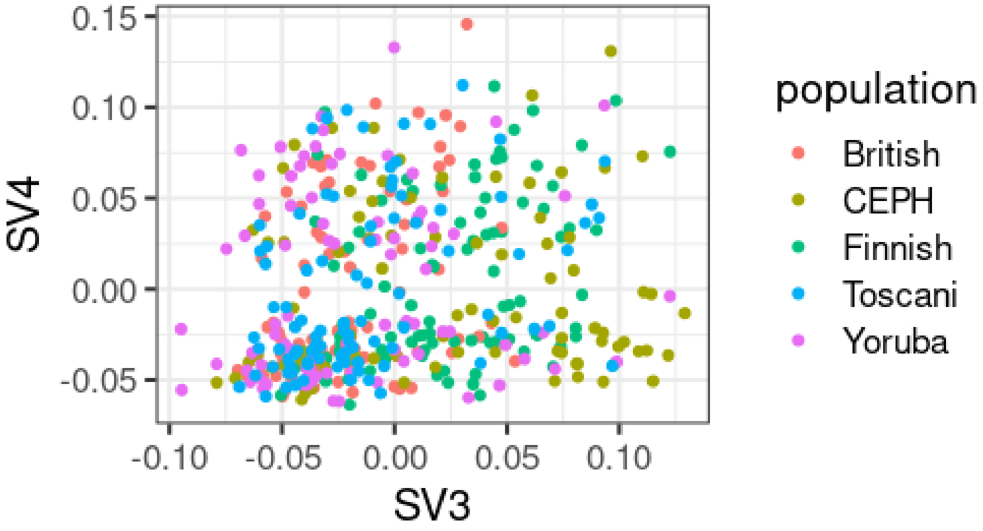


Figure 3·3: Third and fourth hidden factors.

Marker	Source ID	Number of overlap
H3K27me3	ENCFF695TUH	0
H3K27ac	ENCFF835NLI	38
H3K36me3	ENCFF695TUH, ENCFF829MGL	7
H3K4me1	ENCFF710GQV	11
H3K4me2	ENCFF772RNW	48
H3K4me3	ENCFF566VFL, ENCFF920KUR	53
H3K79me2	ENCFF522JVO	24
H3K9ac	ENCFF797IEH	41
H3K9me3	ENCFF874UEV	0
H4K20me1	ENCFF774QTB	0
DNAse	ENCFF588OCA	40
Total		89/138

Table 3.1: Number of overlap of histone markers and open DNAse peaks with the eQTL VNTRs. From the ENCODE, experiments on the GM12878 cell line were selected. Narrow peak calls in bed format on GRCh38 were downloaded. Any intersection of the peak with the VNTR loci was counted as an overlap. The ID of the file from ENCODE is given in the "Source ID" column. Total 89 eQTL VNTRs overlapped with one or more histone marker.

3.2.4 Population-biased alleles and predicting ancestry

The 2,504 genomes in the NYGC dataset consisted of 26 populations of individuals with ancestry from five super-populations: African, American, East Asian, European, and South Asian. To investigate the predictive power of common VNTRs with regard to super-population membership, Principal Component Analysis (PCA) clustering was applied. For each sample, a vector of common loci alleles showing presence/absence (1/0) was produced. Uninformative alleles (that were not present in at least 5% of the samples) were removed and principal components (PCs) were calculated over the resulting vector set. Using a 70% training to 30% testing split of the data, a decision tree based on the first 10 PCs was trained using 10-fold cross-validation and was then validated on the testing data.

In order to find super-population markers among the common VNTRs, a one-sided Fisher's exact test was used to calculate the odds ratio and p-value of each allele being in one super-population versus being collectively in all the others. Only over-represented alleles were considered (rather than both over- and under-represented) because we were interested in identifying alleles that have a phenotypic effect. Odds ratio values were log2 transformed and p-values were adjusted for false discovery rate (FDR). Any allele with FDR¡0.05 and log2(odds ratio) >1 was chosen as a significant marker for that population.

3.3 Results

The number of common and private VNTRs will be discussed in Section 3.3.1 and the enrichment by genomic location and gene set will be presented in Section 3.3.2. Section 3.3.3 investigates the correlation of VNTR alleles with proximal gene expression levels. And Section 3.3.4 presents a model to predict ancestry from the population-biased VNTR alleles.

3.3.1 A total of 19% of minisatellite VNTRs are polymorphic across the human population

To detect common VNTRs, following methodology used with SNPs (Psychiatric GWAS Consortium Coordinating Committee et al., 2009), we classified VNTRs in the 2,504 healthy, unrelated individuals from the NYGC dataset (150 bp and coverage $> 30\times$) as common if they occurred in at least 5% of a population (126 individuals) and private if they occurred in less than 1% (25 individuals). We classified 5,676 VNTRs as common (17% of the 33,403 VNTRs detected in this population) and 68% as private. The number of genomes calling each loci as VNTR is illustrated in Figure 3·4. Each sample averaged 1,783 common VNTRs (median 1,677) and 46 private VNTRs (median 17). Figure 3·5 shows that when the threshold of 5% of the population is increased for common VNTRs, the number of common VNTRs does not change drastically. Widespread occurrence of common VNTRs indicates a fitness for use in Genome Wide Association Studies (GWAS).

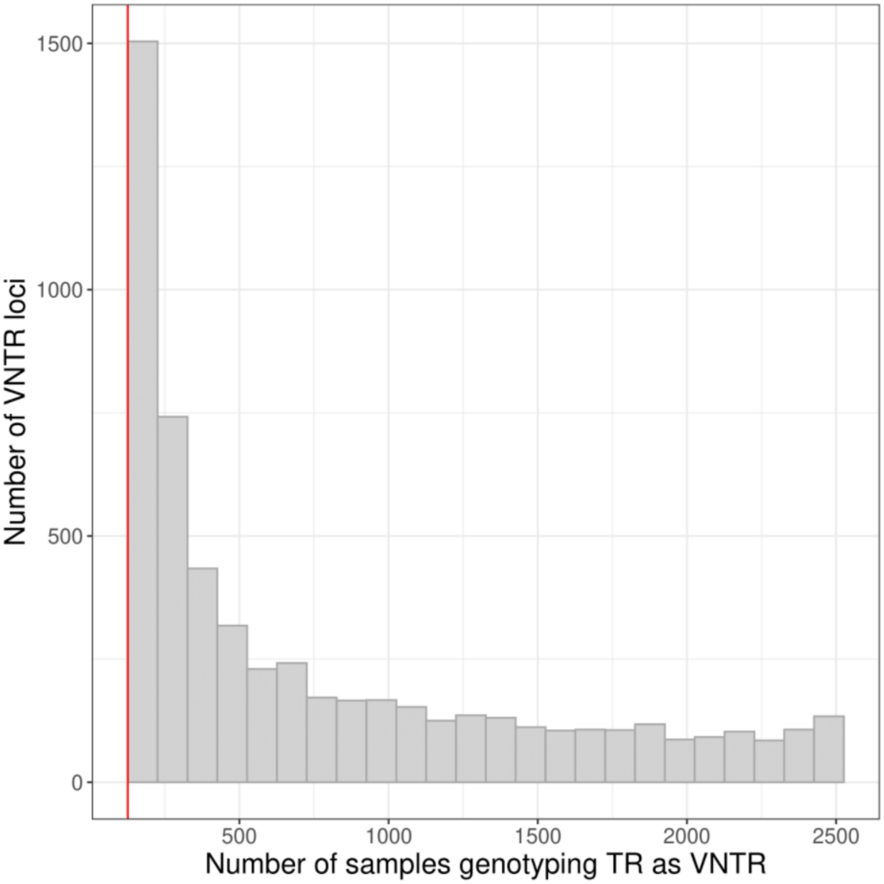


Figure 3·4: The number of genomes calling each loci as VNTR. Data shown are the common loci from the 2,504 sample NYGC dataset. Each bar represents the number of samples calling a locus as a VNTR. Bin size is 100. Bar height is the number of loci with that sample support. Red line indicates the 5% cutoff for common loci (126 samples).

3.3.2 Polymorphic minisatellite VNTRs are enriched in functionally annotated regions

To determine possible functional effects of the common VNTRs, we classified the overlap of reference TRs with various functionally annotated genomic regions: upstream and downstream of genes, 3' UTRs, 5' UTRs, introns, exons, transcription factor binding site (TFBS) clusters, CpG islands, and DNAse clusters.

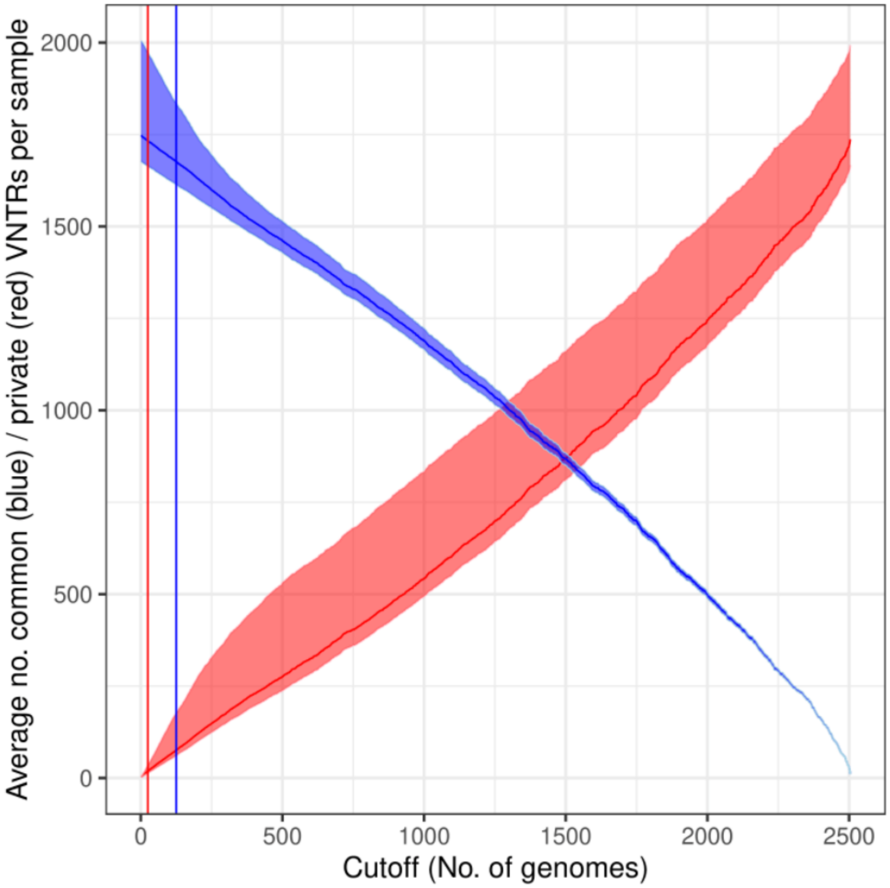


Figure 3·5: Number of common/private VNTRs by cutoff in the 2,504 NYGC samples. Blue area specifies the interquartile range (line is the median) of common VNTR loci counts per sample as the number of samples required to be called common increases from zero to 2,504. For example, at the 5% cutoff (126 - blue vertical line), there were 1,783 common VNTRs on average per genome (median 1,677). The pink area specifies the interquartile range of private VNTRs. At the 1% cutoff (25 - red vertical line), each genome had 46 private VNTRs on average (median 17). The graph shows that common VNTR loci were indeed very common since the numbers do not drop dramatically even if the cutoff were raised to 500 samples.

Our reference TR reference set comprised only 0.52% of the genome, however, 49% of human genes contained at least one TR and 5% of all the TFBS clusters overlapped with TRs. Moreover, high proportions of our TR reference set and common VNTRs intersected with genes (63% and 64% respectively), TFBS clusters (38% and 51%), and DNAse clusters (21% and 28%) (Table 3.2 and Table 3.3).

A total of 3,627 common VNTRs overlapped with 2,173 protein coding genes including 254 exons. In comparison to TRs, VNTR loci were positively enriched in 1 Kbp upstream and downstream regions of genes, 5' and 3' UTRs, coding exons, TFBS clusters, DNAse clusters, and CpG islands (p-values <0.05) (Table 3.2 and Table 3.3). The common VNTRs, on the other hand, compared to all VNTRs, were enriched in 1 Kbp upstream regions of genes, TFBS, and CpG islands, suggesting regulatory function. Private VNTRs were less likely to occur in 1 Kbp regions upstream or downstream of genes, inside TFBS clusters, open DNAse clusters, or CpG islands.

Focusing on the common VNTRs, we used the LOLAweb (Nagraj et al., 2018) online tool to perform enrichment analysis with various curated feature sets, including transcription factor binding sites from ENCODE (ENCODE Project Consortium, 2012), DNase hypersensitive sites clustered by tissue (Sheffield et al., 2013), ChIP-Seq experiments for histone markers from the CODEX database (Sánchez-Castillo et al., 2015), and transcription factor ChIP-Seq peaks from the Cistrome database (Liu et al., 2011). The aim was to identify potential VNTR functional effects through overlap with these experimentally validated regulatory regions. LOLAweb found that common VNTRs are mostly located inside introns and intergenic regions (Figure 3·6) and, compared to all VNTRs, the common VNTRs were enriched in TSS and enhancer segments (Figure 3·7). Enrichment within ChIP-Seq transcription factor peaks included POL1 and POL2 (Odds ratio >3), suggesting that they have effects on gene

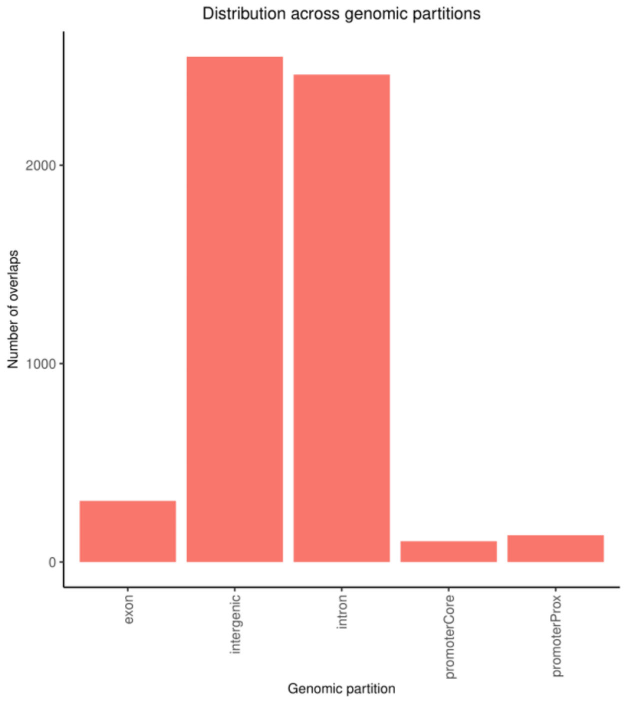


Figure 3·6: Genetic distribution of common VNTRs for different partitions by lolaWeb. The common VNTRs are spread across chromosomes. Most occur inside introns and intergenic regions. Hundreds occur inside promoters and gene exons.

transcription (Figure 3·8). Among the results from LOLAweb, DNAse enrichments by tissue type were enriched in brain, muscle, epithelial, fibroblast, bone, hematopoietic, cervix, skin, and endothelial (Figure 3·9) with brain showing up multiple times, consistent with findings in the literature that associate VNTRs with loss or gain of cognitive function and neuron function (Sonay et al., 2015). These results suggest that VNTR alleles can affect gene regulation in multiple tissues, which is consistent with previous reports (Bakhtiari et al., 2020).

3.3.3 VNTR genotypes are correlated with gene expression differences

To detect association between VNTR genotypes and expression of nearby genes, we paired VNTRs to any gene within 10 Kbp and after removing genes with low expres-

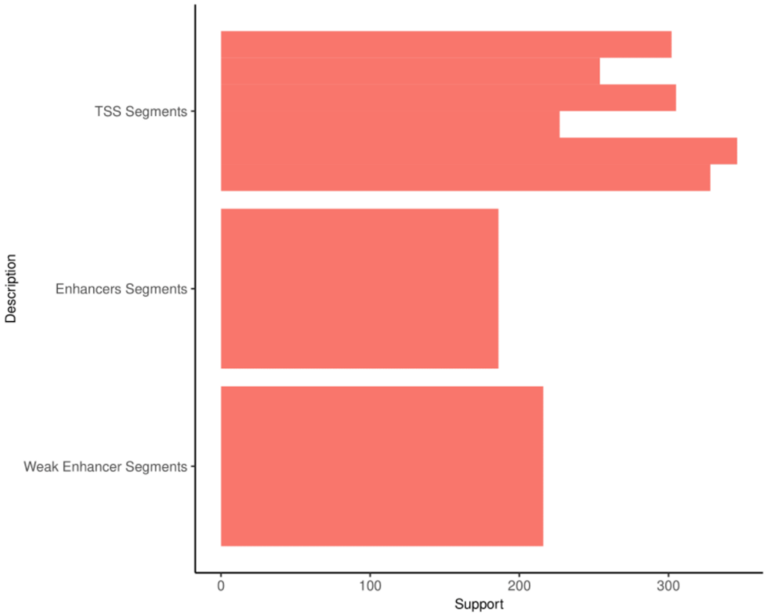


Figure 3·7: Encode segmentation enrichment for common VNTRs by lolaWeb. The common VNTRs are enriched in enhancer segments and TSS segments. All results were filtered by Odds ratio >1 and p-value <5%.

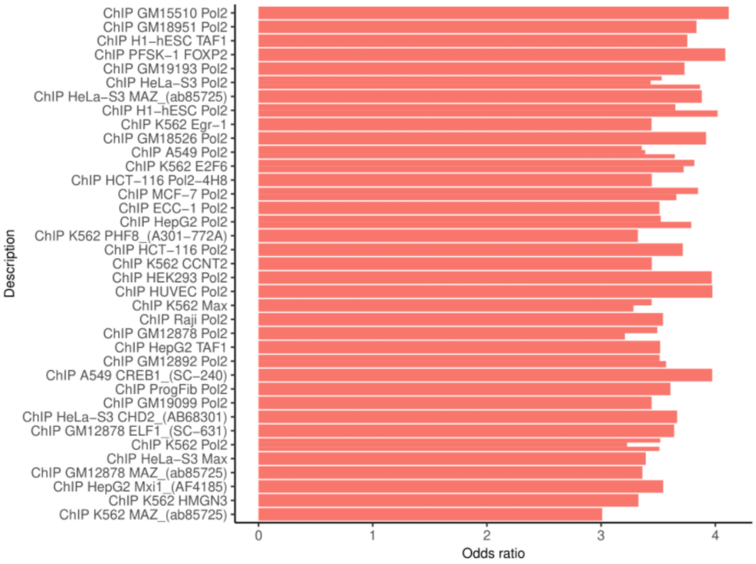


Figure 3·8: Enrichment of common VNTRs using chipSeq data by LolaWeb. Common VNTRs are enriched in binding sites of Pol1 and Pol2 transcription factors suggesting they may play a role in gene regulation. The results were filtered by p-value <5% and Odds ratio >1.

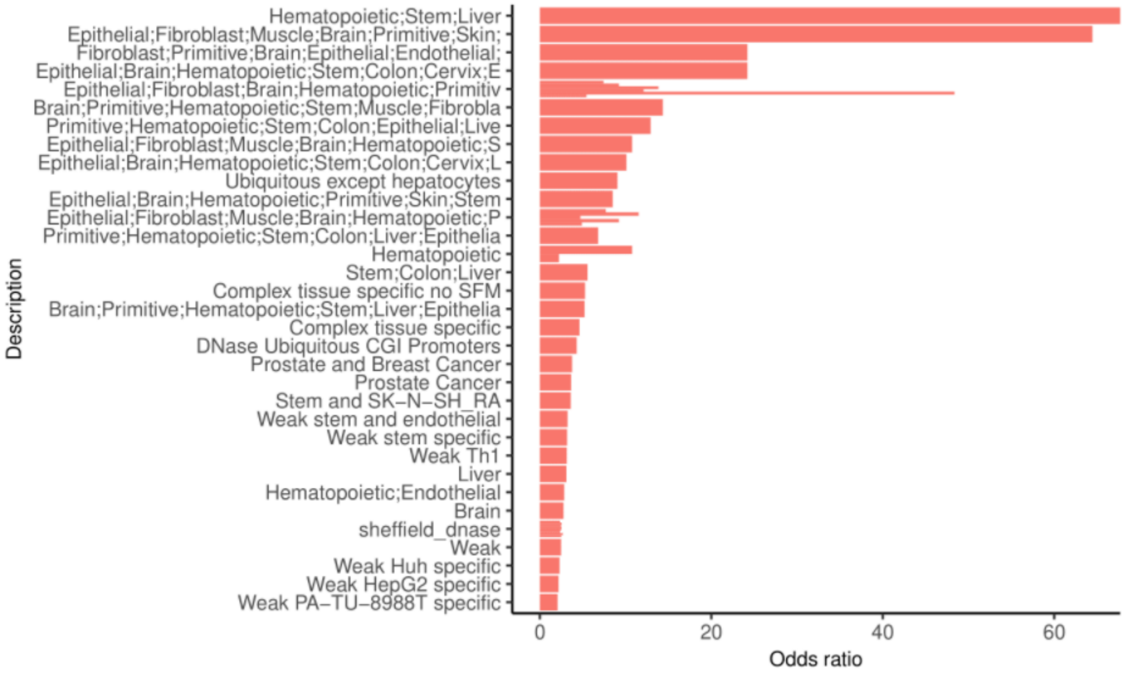


Figure 3·9: Enrichment of common VNTRs with DNAse clusters by LolaWeb. DNAse experiments across 91 tissues are tested for enrichment of common VNTRs. Brain is a recurrent hit. The results were filtered by p-value <5% and Odds ratio >1.

sion and controlling for confounders, applied a one-way ANOVA test to determine if there was a significant difference between the average gene expression levels for the VNTR genotypes. A total of 1,071 gene-VNTR pairs were tested and 197 pairs (190 genes, 192 VNTRs) exhibited significant expression differences at FDR<5% (Figure 3·10). The top 10 genes were FARP1, HEBP1, MXRA7, CD151, THNSL2, DNAJA4, PIP5K1B, B4GALNT3, KLF11, and DPYSL4.

Three top genes are shown in Figure 3·11, Figure 3·12, and Figure 3·13. Gene MXRA7 is associated with a VNTR (id 182606303) in the 5' UTR exon, DPYSL4 is associated with a VNTR (id 182316137) in the first intron, and CSTB is associated with an upstream VNTR (id 182814480). The VNTR region in MXRA7 is a target site for transcription factors METTL23 and JMJD6. METTL23 is known to function

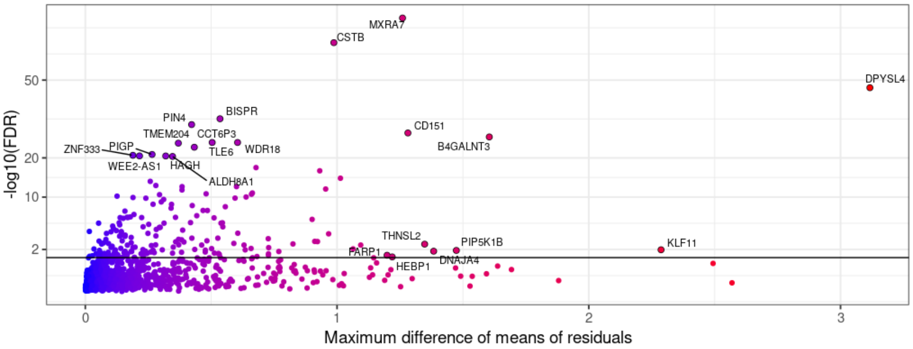


Figure 3·10: The eQTL VNTRs. In 198 gene-VNTR pairs (dots above the dashed black line), a significant difference in gene expression was correlated with VNTR genotype. the y-axis is $-\log_{10}$ of the FDR value and the dashed black line denotes FDR=0.05. The x-axis is the maximum difference between the mean expression for the different genotypes. Genes with the most significant expression differences are labeled.

as a regulator in the transcriptional pathway for human cognition and has been associated with mental retardation and intellectual disability. JMJD6 is associated with congenital myasthenic syndrome associated with AChR deficiency and pancreatitis. Copy number expansions in the VNTR upstream of CSTB have been previously associated with progressive myoclonic epilepsy (EPM1) (Lalioti et al., 2003). For this VNTR, we observed the -1 and 0 alleles (2 and 3 copies, respectively), which are common in healthy individuals. However, 201 individuals had genotypes outside of our detection range which likely represented longer expansions and these individuals showed higher expression of this gene.

Feature	Reference TRs	All VNTRs	Common (>5%)	Private (<1%)
Total	191,286	33,403	5,676	22,538
Upstream (1Kb)	12,415	3,424	671	2,181
5' UTR	4,994	1,294	236	847
Intron	116,002	20,266	3,451	13,526
Coding exon	2,990	699	91	500
3' UTR	6,628	1,295	238	817
Downstream (1Kb)	10,844	2,200	401	1,385
Gene	121,205	21,410	3,624	14,351
TFBS cluster	71,779	16,220	2,913	10,459
DNAse cluster	40,517	9,296	1,613	6,003
CpG Island	6,718	2,989	638	1,757

Table 3.2: Annotation and enrichment of VNTRs. Column *Reference TRs* shows the genomic feature annotations of the reference VNTRs. Numbers do not add to the total due to multiple classifications. We performed a Fisher's Exact Test to find enrichment of all VNTRs against all TRs and common/private VNTRs against all VNTRs. Significant p-values at the 5% threshold are presented in colored font, with blue and red indicating odds ratios less than one and greater than one, respectively. Compared to reference TRs, VNTRs were enriched in genes, gene upstream and downstream regions, TF binding sites, CpG islands, and open DNAse sites. Common VNTRs were more likely to occur in gene upstream regions, TF binding sites, and CpG islands (suggesting possible gene regulation effects); while they were less likely to occur inside exons, possibly due to disruption of protein product function. Private VNTRs, on the other hand, were less likely to occur at gene upstream and downstream regions, TF binding sites, open DNAse sites, and CpG islands (possibly due to the potential to randomly disrupt gene expression).

Feature	Reference TRs	All VNTRs	Common (>5%)	Private (<1%)
Total	191,286	33,403	5,676	22,538
Upstream (1Kb)	6.49	10.25	11.82	9.68
5’ UTR	2.61	3.87	4.16	3.76
Intron	60.64	60.67	60.80	60.01
Coding exon	1.56	2.09	1.60	2.22
3’ UTR	3.46	3.88	4.19	3.62
Downstream (1Kb)	5.67	6.59	7.06	6.15
Gene	63.36	64.10	63.85	63.67
TFBS cluster	37.52	48.56	51.32	46.41
DNAse cluster	21.18	27.83	28.42	26.64
CpG Island	3.51	8.95	11.24	7.80

Table 3.3: Annotation and enrichment of VNTRs as percentages. This table presents the enrichment data in Table 3.2 as percentages. We performed a Fisher’s Exact Test to find enrichment of all VNTRs against all TRs and common/private VNTRs against all VNTRs. Significant p-values at the 5% threshold are presented in colored font, with blue and red indicating odds ratios less than one and greater than one, respectively.

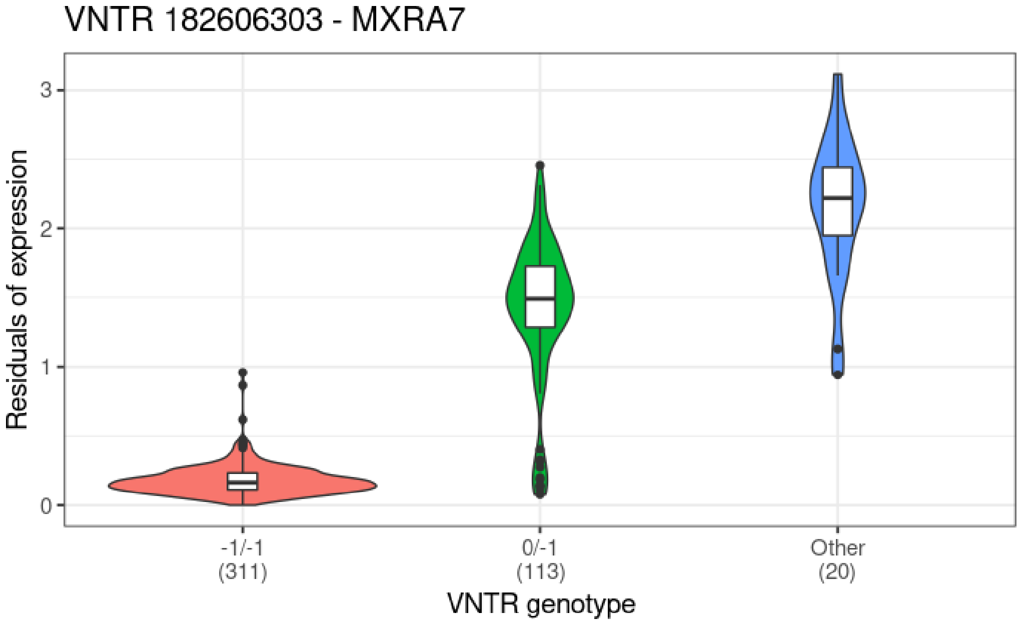


Figure 3·11: Gene expression of MXRA7 with VNTR id 182606303. Shown are violin plots of gene expression values ($log2$ normalized TPM) for the MXRA7 gene which displayed significant differential expression when samples were partitioned by VNTR allele genotype. Genotype is indicated in labels on the x-axis and numbers refer to copies gained or lost relative to the reference allele. ”Other” indicates a partition with undetected alleles presumed outside the range of VNTRseek detection. Number of samples in each partition is shown in parenthesis.

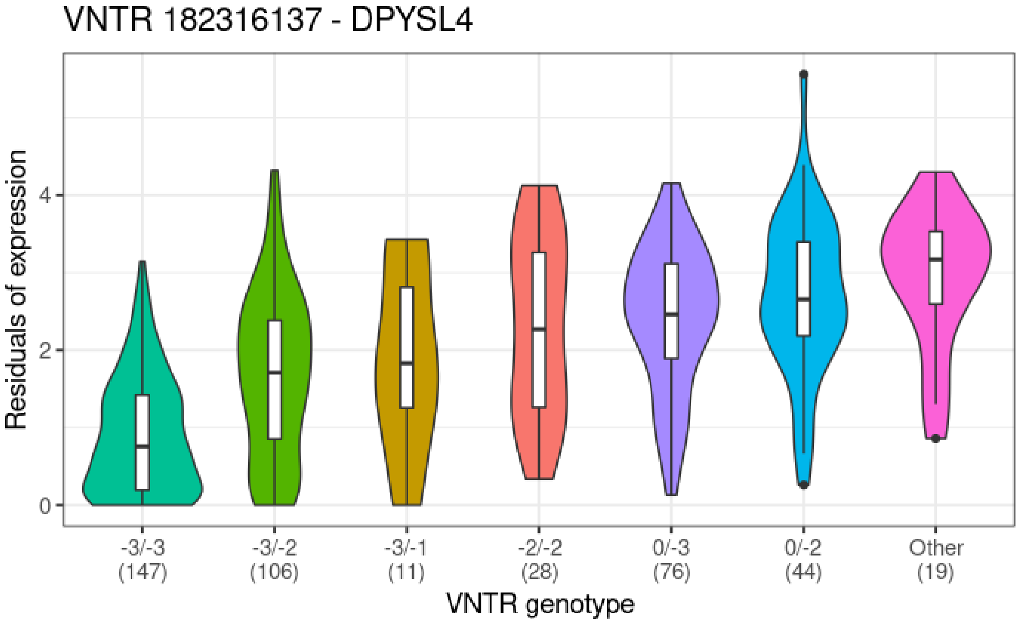


Figure 3·12: Gene expression of DPYSL4 with VNTR id 182316137. Shown are violin plots of gene expression values ($log2$ normalized TPM) for the DPYSL4 gene which displayed significant differential expression when samples were partitioned by VNTR allele genotype. Genotype is indicated in labels on the x-axis and numbers refer to copies gained or lost relative to the reference allele. "Other" indicates a partition with undetected alleles presumed outside the range of VNTRseek detection. Number of samples in each partition is shown in parenthesis.

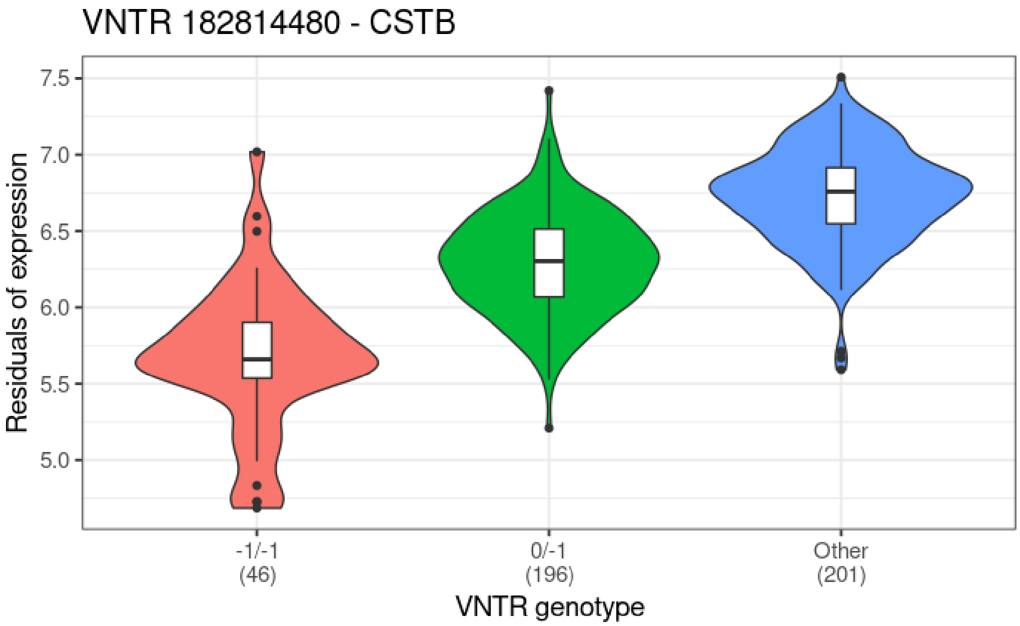


Figure 3·13: Gene expression of CSTB with VNTR id 182814480. Shown are violin plots of gene expression values ($log2$ normalized TPM) for the CSTB gene which displayed significant differential expression when samples were partitioned by VNTR allele genotype. Genotype is indicated in labels on the x-axis and numbers refer to copies gained or lost relative to the reference allele. ”Other” indicates a partition with undetected alleles presumed outside the range of VNTRseek detection. Number of samples in each partition is shown in parenthesis.

Of these eQTL VNTRs 89 overlapped with peaks for histone marks and DNAse hypersensitive sites (Table 3.4 and Figure 3·14).

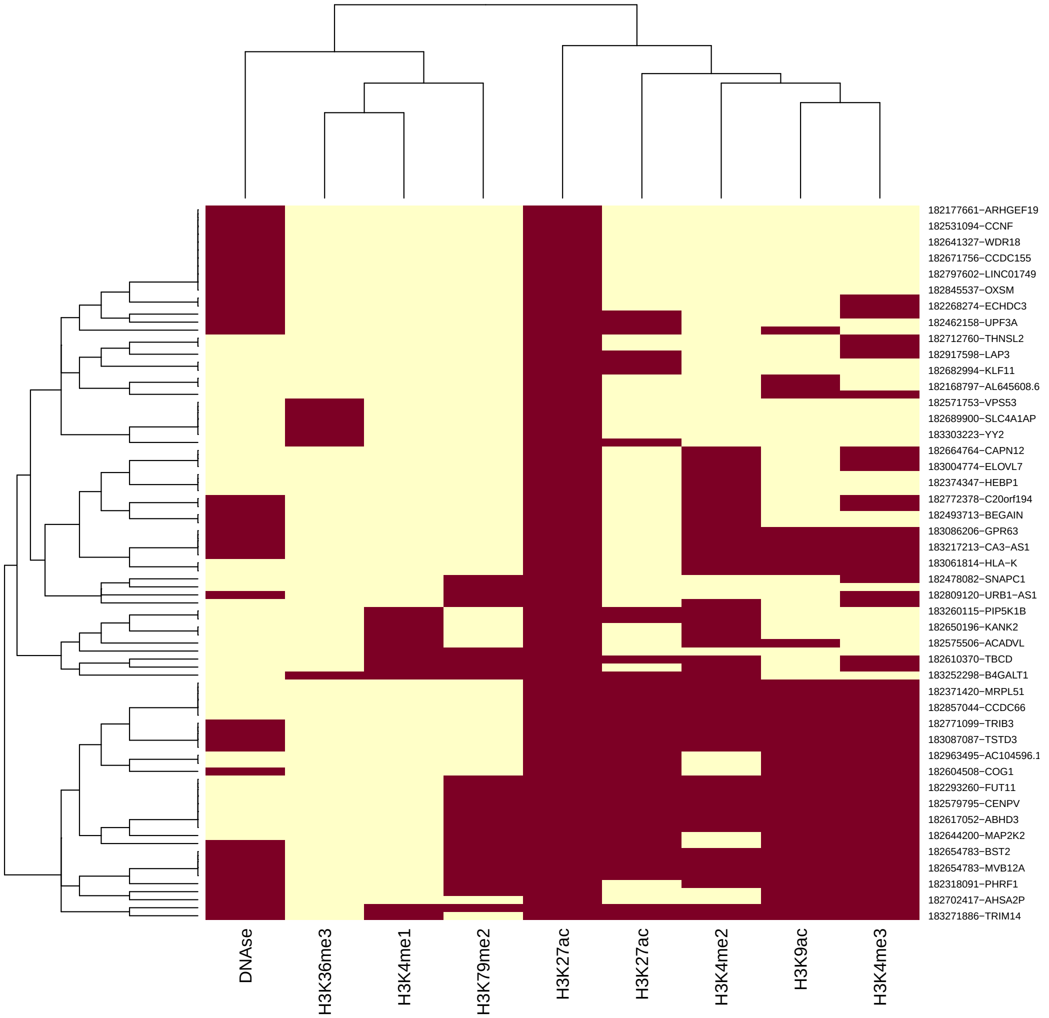


Figure 3·14: Clustering of eQTL VNTRs with histon markers and open DNAse peaks. For source of data see Table 3.1. A total of 98 out of 195 eQTL VNTRs overlapped with histone markers and DNAse peaks. Hierarchical clustering was performed to illustrate the overlaps.

	Overlap with histones	No overlap with histones	Marginal row total
In eQTL	89	106	195
Not in eQTL	187	585	772
Marginal column total	276	691	967 (Grand Total)

Table 3.4: Enrichment of histone markers in eQTL VNTRs. The eQTL VNTRs are more likely to overlap with eQTL VNTRs at p-value <0.00001.

3.3.4 Population-specific VNTR alleles

We next investigated whether VNTR alleles are population-biased and whether they can be used to predict ancestry. Understanding the occurrence of population-biased VNTR alleles will be useful when controlling for population effects in GWAS, and more generally in interpreting gene expression differences among people of different ancestry.

A total of 4,605 alleles from the common VNTR loci were classified as common if they occurred in at least 5% of the population (NYGC). We then constructed a matrix of presence/absence of each allele by sample and clustered the samples using Principal Component Analysis. We found that the first, second, fourth, and fifth principal components (PCs) separated the super-populations as shown in Figure 3·15. Each PC captured a small fraction of the variation in the dataset, suggesting that there was substantial variation between individuals from the same population.

The first PC separated Africans, suggesting that they have the furthest evolutionary distance from the other super-populations analyzed. The second PC separated East Asians. The fourth and fifth PCs separated South Asians and Americans, respectively. The American population had a sub-population of Puerto Ricans that clustered with the Iberian Spanish population, suggesting mixed ancestry (Lalioti et al., 2003; Sudmant et al., 2015).

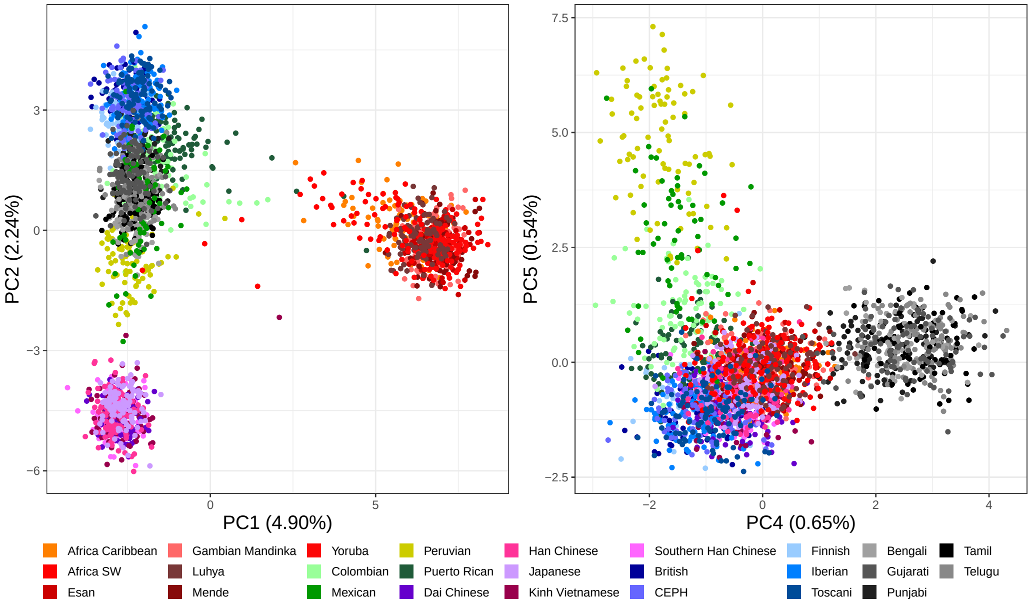


Figure 3·15: Principle Component Analysis (PCA) of common VNTR alleles in the NYGC population (150 bp). PCA was performed to reduce the dimensions of the data. *Left*: PC1 captured ∼5% of the variation and separated Africans from the other super-populations, suggesting that they had the greatest distance from the others. PC2 separated East Asian and European populations but left individuals from the Americas and South Asia mixed. *Right*: PC4 separated the South Asian population and PC5 separated the American populations. PC3 (not shown) captured batch effects due to differences in coverage. Some American sub-populations proved hardest to separate, likely due to ancestry mixing.

To show the power of these alleles to predict ancestry, we next trained a decision tree model (Figure 3·16) using the top 10 PCs (11% of the total variation) and achieved a recall of over 98% on every population when applied to the 30% test partition (Table 3.5).

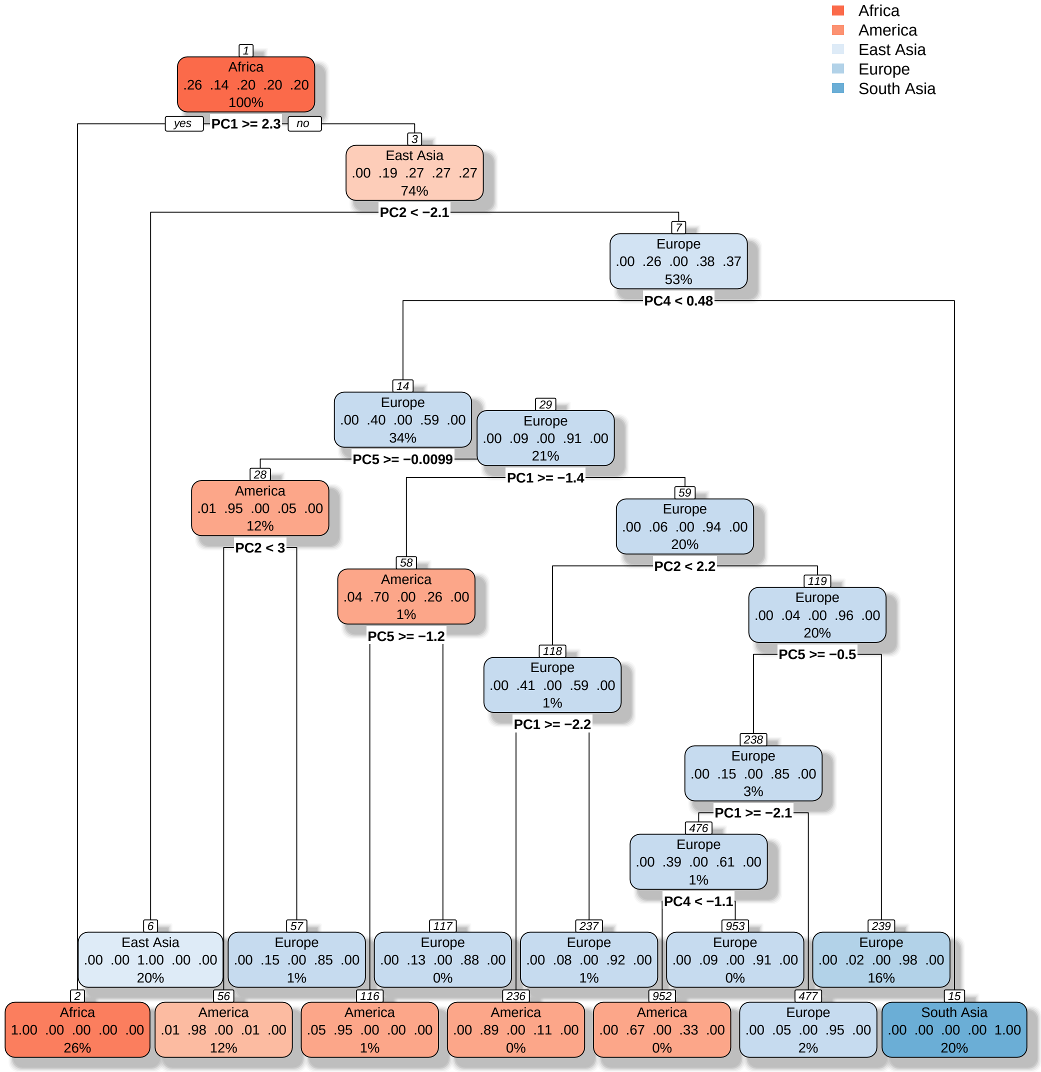


Figure 3·16: Decision tree for prediction of superpopulation ancestry from common VNTRs. Each box is a node in the decision tree. Color and superpopulation name indicate majority label in the training data entering the node. Five decimal values are the fractions of population labels in training data entering the node with values in order as Africa, America, East Asia, Europe, and South Asia. Percentage shown is percent of training data entering the node. Equation with PC number indicates principle component test value to exit the node down ”yes” (left) branch.

N=751	African	American	East Asian	European	South Asian	Precision
African	205	0	0	0	0	100%
American	4	91	0	3	0	93%
East Asian	0	1	146	0	0	99%
European	0	6	0	141	0	96%
South Asian	0	0	0	0	154	100%
Recall	98%	93%	100%	98%	100%	

Table 3.5: Confusion matrix of decision tree results on the test data to predict ancestry. Common VNTR predictions on 2,504 unrelated genomes from NYGC were used to train a model to predict ancestry. Principle component analysis was performed to reduce dimensionality. The first 10 principle components were used to train 70% of the data (train). The model was tested on the remaining test data (30%). The confusion matrix on the test data is presented here. The total number of genomes in the test was 751. Columns indicate the *true* label, rows the *predicted* label. The last column shows the *precision*, and the last row shows the *recall*. Populations of African, East Asian, and South Asian ancestry were the easiest to predict. People with American ancestry, on the other hand, had more admixed genomes and fewer samples (as described by the data source) making them more difficult to predict. Overall accuracy was 98%.

A one-sided Fisher's Exact Test was applied to determine the population-biased VNTR alleles that were over-represented in one population versus all the others. A total of 3,850 VNTR alleles were identified as population-biased in one or more super-populations, corresponding to 1,096 VNTR loci (Figure 3·17). The population-biased VNTR loci overlapped with 689 genes and 51 coding exons. Africans had the highest number of population-biased alleles (266), followed by East Asians (65), while Americans had the lowest (13), suggesting more mixed ancestry in the American super-population. We observed 63 loci that had a population-biased allele in each population (Figure 3·18). Figure 3·19 shows seven of the 1,096 population-biased loci in a “virtual gel” representation, mimicking the appearance of bands on an Agarose gel for easier interpretation. The details of these seven population-biased loci are given in Table 3.6.

A total of 49 genes that displayed expression differences correlated with VNTR genotype were also associated with population-biased VNTR loci (Table 3.7), including the VNTR 182316137 associated with the gene DPYSL4, discussed in the previous section, which exhibited seven different alleles, five of which were population-biased.

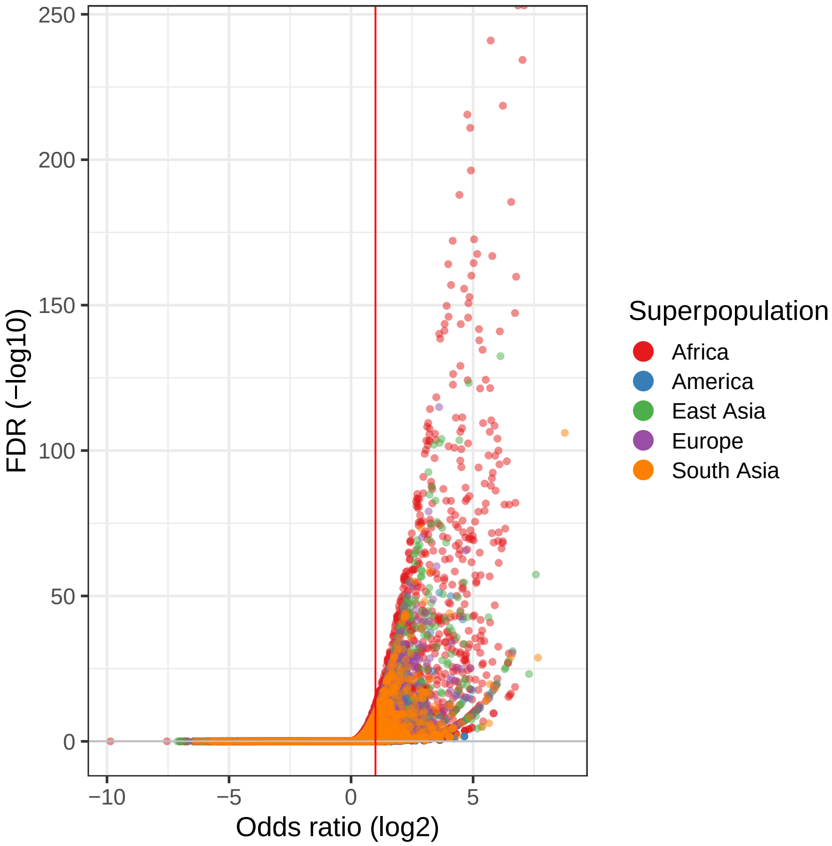


Figure 3·17: Volcano plot for alleles at population-specific VNTR loci. One-way Fisher's Exact Test was used to find common VNTR *alleles* over-represented in each population versus the others. Each dot in the volcano plot represents one allele tested for over-representation in one population (depicted by color). The p-values were adjusted using FDR. Alleles with FDR<5% (above the horizontal gray line nearly coinciding with zero) and with odds ratio > 2 (right of the red line) were selected.

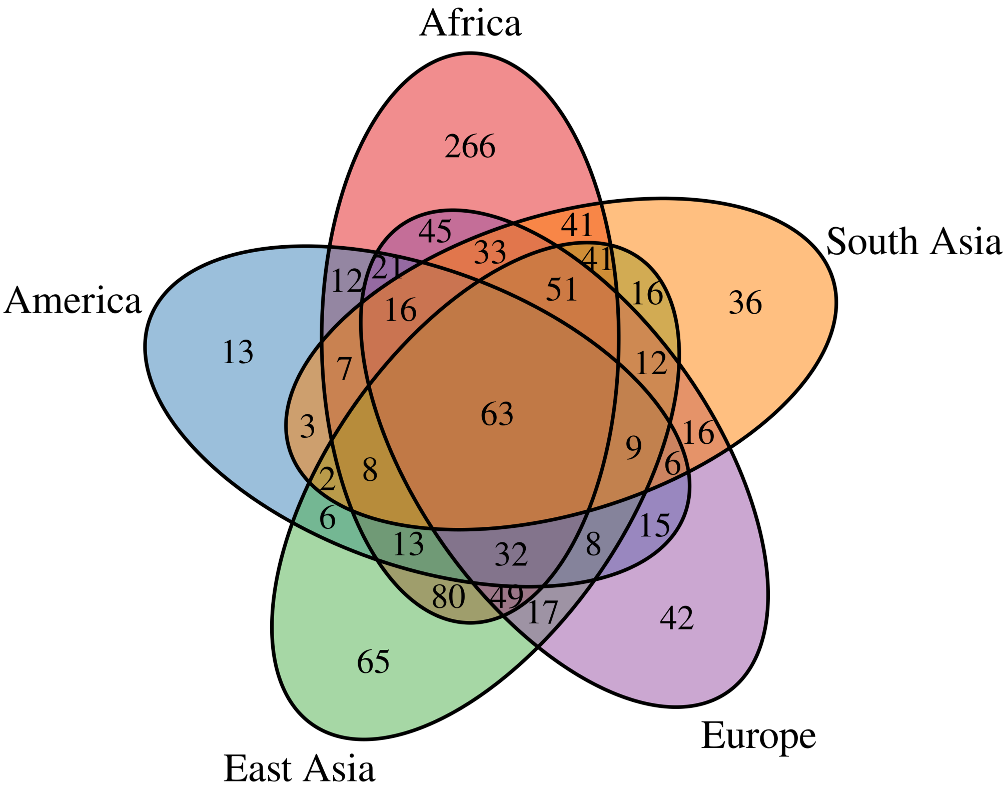


Figure 3·18: Venn diagram of population-specific VNTR loci. Africans have the highest number of loci with population-specific alleles, followed by East Asians. Americans had the least which could be because of the lower number of samples (statistical power) and/or more admixed genomes (Sudmant et al., 2015). Interestingly 63 loci had an allele that is over-represented in each population.

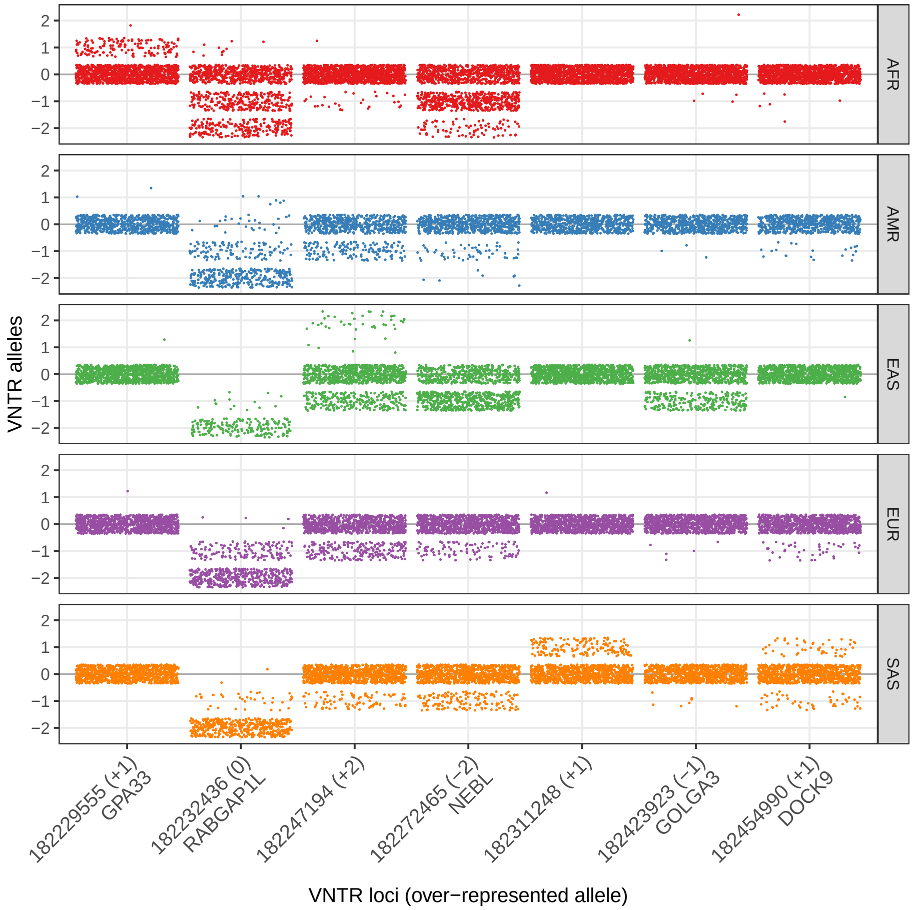


Figure 3·19: "Virtual gel" representation of seven population-specific VNTR alleles. Each dot represents an allele in one sample. Samples are separated vertically by super-population. Dots are jiggered in a rectangular area to reduce overlap. Population-specific alleles show up as bands over-represented in one population. Numbers and labels at bottom are VNTR locus ids with nearby genes indicated and the population-specific allele expressed as copy number change (+1, -2, *etc.*) from the reference. For example, in the leftmost column, the +1 allele was over-represented in the African population. Note that the allele bias towards pattern copy loss relative to the reference allele is apparent and that at one locus (second from left) the reference allele was the population-specific allele since almost no reference alleles were observed in the four other populations. The details of these seven loci are given in Supplementary Table 3.6.

TRid (Allele)	Specific to	Odds Ratio (log2)	FDR	AFR (661)	AMR (347)	EAS (504)	EUR (503)	SAS (489)
182229555 (+1)	African	6.73	1E-82	148	2	1	1	0
182232436 (0)	African	7.03	5E-235	381	15	0	2	1
182247194 (+2)	East Asian	7.29	7E-24	0	0	36	0	0
182272465 (−2)	African	4.85	4E-34	73	7	0	0	0
182311248 (+1)	South Asian	8.76	8E-107	0	0	0	1	147
182423923 (−1)	East Asian	6.12	4E-133	4	3	208	5	7
182454990 (+1)	South Asian	7.63	1E-28	0	0	0	0	43

Table 3.6: Example of population specific alleles. Seven significant alleles were chosen to draw a virtual gel (main text). The details of the test on those seven VNTR alleles and the raw counts in each population for the specified allele are given here. *N* is the total number of genomes in that population. Allele column indicates copy number change relative to the reference allele.

Fifty eQTL VNTRs also had population-biased alleles (Table 3.7), including the VNTR 182316137 associated with the gene DPYSL4, discussed in the previous section, which exhibited seven different alleles, five of which were population-biased.

Table 3.7: Intersection of population-specific VNTRs and VNTRs with genotypes correlated with gene expression. For a total of 51 genes, which had expression levels correlated with proximal VNTR genotype, the VNTRs were also found to be population-specific.

TRid	**Gene**	**Maximum difference in mean**	**FDR**
182187335	AC004865.2	0.64	1E-02
182370381	AC006207.1	0.36	6E-07
182422326	ADGRD1	0.18	1E-02
182168797	AL645608.6	0.09	5E-02
182168889	AL645608.7	0.37	3E-03
182317968	ANO9	0.70	2E-19
182344338	AP003071.5	0.16	4E-02
182318359	AP2A2	0.11	7E-10
182177661	ARHGEF19	0.30	2E-05
182369070	B4GALNT3	1.53	1E-27
182319651	C11orf21	0.96	1E-12
182318295	CD151	1.27	1E-30

Continued on next page

Table 3.7 – *Continued from previous page*

TRid	Gene	Maximum difference in mean	FDR
182405813	CFAP54	0.09	4E-02
182417629	CLIP1	0.09	3E-02
182171124	DFFB	0.31	5E-02
182176034	DHRS3	0.23	4E-02
182387680	DIP2B	0.10	4E-02
182303301	DNMBP	0.32	3E-11
182316137	DPYSL4	2.88	3E-47
182268274	ECHDC3	0.67	5E-11
182318207	EPS8L2	0.47	4E-02
182454752	FARP1	1.19	3E-02
182293260	FUT11	0.15	2E-02
182374347	HEBP1	1.25	4E-02
182374347	HTR7P1	0.07	3E-03
182312799	LHPP	0.18	8E-03
182188225	MEAF6	0.13	5E-05
182314781	MGMT	0.19	5E-02
182371420	MRPL51	0.07	3E-03
182272465	NEBL	0.36	7E-03
182172261	NPHP4	0.27	3E-05
182172324	NPHP4	0.34	3E-07
182225943	NTRK1	0.11	3E-02
182251099	NVL	0.17	9E-03
182278429	PARD3	0.49	3E-08
182331520	PDHX	0.16	1E-04
182317816	PGGHG	0.37	2E-09
182318090	PHRF1	0.41	4E-05
182318091	PHRF1	0.07	3E-04
182242989	PTPRVP	0.27	3E-03
182242992	PTPRVP	0.45	4E-02
182281227	RET	0.02	3E-02
182228303	RGS5	0.09	2E-02
182223480	S100A10	0.47	1E-08
182182434	SELENON	0.82	3E-02
182330802	TCP11L1	0.25	5E-14
182169710	TMEM52	0.19	7E-03
182388680	TNS2	0.03	8E-03
182270754	TRDMT1	0.18	6E-10
182462158	UPF3A	0.39	1E-02
182386450	ZNF641	0.18	5E-07

Next, to identify potential functional roles of the population-biased VNTR loci we performed Gene Set Enrichment Analysis (GSEA) for the associated genes against the Broad Institute MSigDB (Liberzon et al., 2011). Genes overlapping with the population-biased VNTRs were enriched for Endocytosis (hsa04144), Fatty acid metabolism (hsa01212), and Arrhythmogenic right ventricular cardiomyopathy (ARVC) (hsa05412) pathways (Table 3.8). Among the GO biological processes affected by these genes were neurogenesis (GO:0022008; FDR=3.62e-8), neuron differentiation (GO:0030182; FDR=2.52e-7), and neuron development (GO:0048666; FDR=4.31e-7). These processes are potentially related to other findings that have linked VNTRs to neurodegenerative disorders and cognitive abilities (Marinho et al., 2019; Katsumata et al., 2019; Chang et al., 2019; De Roeck et al., 2018; Scott et al., 1991; Hoxha et al., 2019; Šerý et al., 2015; Grünblatt et al., 2019; Van Assche et al., 2016) The GO term Behavior (GO:0007610; FDR=2.22e-4) was also found, which could be related to the association of VNTR loci with aggressive behavior (Schlüter et al., 2014; Zammit et al., 2004; Schlüter et al., 2014).

Other notable GO terms that the population-biased VNTRs were enriched in were regulation of muscle contraction (GO:0006937) and neuromuscular processes related to balancing (GO:0050885) with FDR <1%. The genes were also highly enriched in midbrain neurotype cell gene signatures (FDR=5.49e-25), which might affect movement and emotions (Mill et al., 2002; Diatchenko et al., 2007; Kang et al., 1999).

3.4 Summary

Our research has classified a large subset of VNTRs (5,676) as common (occurring in >5% of the population). When considering the largest dataset in our study (2,504

Gene Set Name	KEGG id	Gene set size (K)	No. of genes in overlap (k)	k/K	p-value	FDR q-value
Endocytosis	hsa04144	181	11	6%	2.51 e-5	4.67 e-3
Fatty acid metabolism	hsa01212	42	5	12%	1.98 e-4	1.84 e-2
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	hsa05412	74	6	8%	3.87 e-4	2.4 e-2
Calcium signaling pathway	hsa04020	178	9	5%	5.29 e-4	2.46 e-2
Vascular smooth muscle contraction	hsa04270	115	7	6%	7.39 e-4	2.75 e-2

Table 3.8: KEGG pathways enriched for population-specific VNTR genes. GSEA (Subramanian et al., 2005) was used to find enrichment of the 560 protein coding genes overlapping with 1,096 population-specific VNTR loci.

individuals), on average, each genome was found to be variable at 1,951 VNTR loci and among those, nearly 1,700 were common VNTRs.

In addition to their widespread occurrence, further evidence of minisatellite VNTR importance can be seen in the enrichment of these loci in genes and gene regulatory regions (promoters, transcription factor binding sites, DNAse hypersensitive sites, and CpG islands). Our entire set of VNTRs overlapped with 7,698 protein coding genes and 3,512 exons. The common VNTRs occurred within or were proximal to over 2,173 protein coding genes, including that overlapped with 254 exons. Biological function enrichment among the genes containing VNTRs includes neuron development and differentiation, and behavior. This finding is consistent with the finding that VNTR expansions in humans compared to primates are associated with gain of cognitive abilities (Sulovari et al., 2019), and possible involvement of VNTRS with many neurodegenerative diseases and behavioral disorders.

The overabundance of VNTR proximity to genes suggests that variability at these loci affects gene expression and indeed, we observed that the expression levels of 118 genes were significantly correlated with the presence of specific VNTR alleles in lymphoblastoid cell lines of 445 individuals.

Common VNTRs were used to predict ancestry of the genomes. PCA clustering separated the genomes by super-population. Using the top 10 PCs, a decision tree was trained that predicted the ancestry with >98% accuracy. In addition, A total of 4,605 VNTR alleles were found to be over-represented in one population compared to others. VNTRs with population-biased alleles were enriched in gene sets related to neuron function and gene signatures of various mid-brain cell types.

Chapter 4

Validating VNTRseek predictions

4.1 Introduction

In the previous chapters we predicted VNTR loci in 2,800 genomes using VNTRseek. In this chapter the reliability of our predictions are evaluated. Detecting minisatellites using short reads has several difficulties. As discussed in Chapter 1, common aligners such as BWA MEM (Li, 2013) and Bowtie (Langmead and Salzberg, 2012) perform very well on a whole genome scale, however, they do not perform ideally on repeat-rich regions. Many reads originating from the TR loci will be mismapped or unmapped. Second, as mentioned in Chapter 1, minisatellite repeats, unlike microsatellites, may contain SNPs and indels, making the mappability ratio inconsistent. Local assembly methods are not applicable to TR loci, because of the low complexity or the repeats. Finally, since many TRs are GC-rich, the library preparation protocol can affect the coverage of TR arrays. For example, PCR amplified datasets can lose complete coverage on minisatellites or create slippage. To overcome these difficulties, VNTRseek uses the basic idea to run TRF on each read and compare the read pattern consensus to the reference TRs (see Section 1.8.5). Aligning the consensus pattern of read TRs to the reference TRs, makes the algorithm more robust to SNPs and indels in each pattern or caused by sequencing errors.

The precision and recall of VNTRseek is assessed on various simulations in Chapter 5. In this chapter, we focus on validating the prediction results from Chapter 2, and show that the predictions are reliable. Our predictions will be validated in several ways, including experimental validation, through confirmation of predicted alleles in long reads, and by comparisons across related genomes and genomes sequenced with different Illumina machines. We investigate the type 1 error of VNTRSeek prediction by providing a discussion on *Multis*. A Multi is a TR that is genotyped with more alleles than logically possible (Section 2.2.3).

Section 4.2 discusses the methods used for in vitro and in silico validations and measuring consistency of the calls. Section 4.3 summarizes the results. The experimental validations are presented in Section 4.3.1 and the validations using long reads are given in Section 4.3.2. The consistency of VNTRseek prediction by Mendelian inheritance and by different sequencing platforms is discussed in Section 4.3.3 and Section 4.3.4, respectively. Section 4.3.5 discusses the type 1 error rate and possible sources of error. Section 4.3.6 evaluates the trade-off between type 1 error rate and sensitivity. Finally, Section 4.4 summarizes the results of this chapter.

4.2 Materials and methods

In vitro validation was performed in Fuxman Bass lab by Samantha D. Drinan and by comparison to validations in the adVNTR paper (Bakhtiari et al., 2018) on the NA12878 genome. In silico validation was performed using long precise reads. Consistency of VNTR predictions is measured across sequencing platforms and across related genomes.

4.2.1 In vitro Validation

Accuracy of VNTRseek genotyping was experimentally tested for 13 predicted VNTR loci in the Ashkenazi Jewish (AJ) trio. The following DNA from B-Lymphocytes were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: NA24385, NA24149, and NA24143 (also identified as GIAB IDs HG002, HG003, and HG004). Selection criteria required that the PCR product was not contained in a repeat region, unique primers could be designed, the primer-defined allele length difference was between 10% and 20% of the longest allele, and the primer-defined GC content was between 40% and 60%. Given these criteria, we prioritized VNTRs in genes and regulatory regions which might be of interest to researchers. List of the selected loci are given in Table 4.1.

Primers were designed with Primer-BLAST (Ye et al., 2012) (Table 4.2). Samantha D. Drinan (Fuxman Bass lab, Boston University) amplified the VNTR loci from the genomic DNA using the specific primers for each individual using the following reagents: 0.2 μL 5 U/μL DreamTaq DNA Polymerase (ThermoFisher Scientific), 4.0 μL 10× DreamTaq Buffer (ThermoFisher Scientific), 0.8 µL 10mM dNTP mix (ThermoFisher Scientific), 3.2 μL primer mix at a final concentration of 0.5 μM, 1.6 μL genomic DNA (40 ng), and 30.2 μL nuclease-free water. PCR cycling conditions were as follows: 30 s at 95°C, 30 s at 56-60°C, 20 s at 72°C, for 30 cycles, with an initial denaturation of 3 min at 95°C and a final extension of 7 min at 72°C. The resulting amplicons were electrophoresed on a 2% agarose gel at 100 V for 2 h and visualized with UV light using ethidium bromide.

#	TR id	Pattern size (bp)	Ref. copy no.	Description
1*	182316181	105	3.8	intron 1 of STK32C
2	182316985	27	5.8	regulatory region targeting LINC01168
3	182453735	30	2.8	intron 1 of DNAJC3
4	182461997	38	7.2	intron 1 of RASA3
5	182493720	70	3.1	intron 1 of BEGAIN
6	182515357	34	8.2	intron 5 of MEGF11
7	182608886	27	6.3	intron 21 of RPTOR
8	182620950	48	3	exon 8 of RNF138
9	182982510	34	4.1	intron 6 of SLC12A7
10#	183046759	38	4.1	intron 3 of ARL10
11†	183081195	15	4.5	exon 2 of TENT5A
12	183117043	17	9.2	regulatory region targeting MRM2, LFNG
13	183169331	15	4.5	exon 6 of IRF5

Table 4.1: TRs selected for experimental validation.

#	VNTR_ID	Forward primer sequence (5' to 3')	Reverse primer sequence (5' to 3')
1	182316181	TACTCCCAATGAGGACAGCAA	TTCTCCAGCTCTTGAGACAGC
2	182316985	GTCACCCAAGGTCCTGTAGC	CTGGGACCAACAGCCAGTAG
3	182453735	CTAGCAATGGAGCTCAGTCTTC	GCAAGGGGTTGTACAATGGAT
4	182461997	GCAGCACAAGAAAAAGAGGCTG	CTCTGACCTTCACTGCTGTTCT
5	182493720	GGGTAGCTGCATGGCTGAAA	TCCCTGACCATCTCCTCTCTG
6	182515357	GCCTGGACTGTCTCAAAGCC	TGCTACGAGGTAGGGATGAGA
7	182608886	GCCGGGAATCTGTTCTCAGT	CAACCTAGTGCCTCATGGCT
8	182620950	GACTTTTGACCATAGTGTTTTCCAG	TCCATCAAGATGACCTCTACTACA
9	182798584	GACCCTCAAGGAGGAATGAGG	GTAAGGAAGTCTGCCTCCCAC
10	183046759	CCAGAGGCTACTTCTGGGAAC	GCTGGCAGCATTTCCTAACAC
11	183081195	GCTTTCGCAATAGTCCAAGCAA	TCGCCATGTCTGAGGACGAG
12	183117043	ACCTGCTTCCCTCATCTACCC	CTAACCTGAGTGCCTTCTGC
13	183169331	TCCACACGCACTCTCTGTAGAT	GGACCTCAGAGAGAAGCTCCC

Table 4.2: Primers designed for the experimental validations. Total 13 VNTRs were chosen where the change was 10–20% of the VNTR array, unique primers could be designed and CG content was about 40–60%. The primers were designed using the primer-BLAST tool (Ye et al., 2012).

As a second validation, VNTRseek predictions in the NA12878 genome were compared to experimental validations in a paper describing the adVNTR software (Bakhtiari et al., 2018). We had three datasets for NA12878; HG001 (148 bp) from GIAB, NA12878 (150 bp) from NYGC, and NA12878 (250 bp) from 1000 Genomes. The adVNTR predictions used GRCh37 coordinates that were converted using the UCSC liftover tool (Kent, 2002) to coordinates to GRCh38.

4.2.2 In silico validation using long reads

Aligned PacBio reads for the AJ child (GIAB ID HG002) were processed to validate VNTRseek predictions (Section 2.2.1). The read sequences were extracted from the BAM file using picard tools (2019) and mapped back to the GRCh38 genome using BWA MEM default settings (Li, 2013). Using bedtools (Quinlan and Hall, 2010), the reads aligning to each TR reference locus were extracted. For each read, a local wraparound dynamic programming alignment (Appendix xxx) was performed using the reference pattern and the same scoring parameters used to generate the reference set (match=+2, mismatch=-5, and gap=-7). The number of copies of the pattern in the resulting alignment was then determined and compared with the VNTRseek predictions. If the difference between a PacBio copy number in at least one read and the VNTRseek copy number was within ±0.25 of a copy, we considered the VNTRseek allele to be validated.

4.2.3 Measuring consistency of Mendelian inheritance

A locus on an autosomal chromosome is consistent with Mendelian inheritance if the genotype of a child can be explained as one allele from the mother and one from the father. Genotype consistency was evaluated for all mother-father-child trios, i.e., the AJ, CEU, HAN, and YRI trios.

We evaluated loci defined by several increasingly stringent criteria:

- both parents heterozygous,
- all members of the trio heterozygous,
- all members of the trio heterozygous and with different genotypes

These criteria were selected to avoid false interpretations of consistency.

TR loci on the X and Y chromosome of male children were also selected for evaluation when both the son and the appropriate parent had a predicted genotype. In these cases, inheritance consistency means a son's X chromosome allele is observed on one of the mother's X chromosomes, and a son's Y chromosome allele is observed on the father's Y chromosome.

4.2.4 Measuring allele consistency across platforms

VNTR calls were compared for each of 27 genomes that were represented twice, once in the 1000 Genomes dataset, sequenced in 2015 on an Illumina HiSeq2500 with 250 bp read length and once in the NYGC dataset, sequenced in 2019 on an Illumina NovaSeq 6000 with 150 bp read length. The two platforms have different error profiles.

Because read length and coverage differed among datasets, for each pairwise comparison, we only considered VNTR loci that were genotyped in both samples. We extracted the *non-reference* VNTR alleles (detected in at least one sample) and computed consistency as the ratio of those alleles detected by both platforms over the total alleles found by both. For alleles detected in 250 bp reads, we only counted those that could have been detected in the shorter 150 bp reads. Reference alleles were excluded to avoid inflating the ratio.

4.3 Results

To show the reliability of our results, we experimentally validated VNTR predictions at 13 loci in the three related AJ genomes, and also compared VNTRseek predictions to alleles experimentally validated in the literature. We additionally used accurate long reads on one genome (HG002) to find evidence of the predicted alleles.

Separately, we determine the consistency of our predictions in two ways: first, we looked at inheritance consistency among four trios (mother, father, child), and second, we compared the results for genomes sequenced on two different platforms.

4.3.1 In vitro validation

All but one of the 66 predicted VNTR alleles were confirmed at 13 loci in the three related AJ genomes (child HG002, father HG003, and mother HG004). In the remaining case, two predicted alleles were separated by only 15 nucleotides and could not be distinguished. At two loci, other bands were also observed. In one, all three family members contained an allele outside the detectable range of VNTRseek (longer than the reads). In the other, one allele that was detectable was missed in two family members. See Table 4.3 for a complete list of experiment results and Figure 4·1, Figure 4·2, Figure 4·3, Figure 4·4, and Figure 4·5 for the gel images and more details.

#	TR id	VNTRseek			Expected bands			Validation		
		Child	Father	Mother	Child	Father	Mother	Child	Father	Mother
1∗	182316181	-2	-2	-2	237/342	237	237	-2/-1	-2/-1	-2/ +1
2	182316985	-3/ 0	-1/0	-3/-2	118/ /199	172/199	118/145	yes	yes	yes
3	182453735	0/ +1	0/+1	0/+1	180/210	180/210	180/210	yes	yes	yes
4	182461997	-5/ -4	-5/-4	-5/-2	193/231	193/231	193/307	yes	yes	yes
5	182493720	0	0	0/-1	266	266	266/196	yes	yes	yes
6	182515357	-5/-5	-5/-6	-5	195	195/161	195	yes	yes	yes
7	182608886	0/ -2	0/-3	-2	208/181	208/154	235/181	yes	yes	yes
8	182620950	0/ -1	0/-1	0/-1	243/195	243/195	243/195	yes	yes	yes
9	182982510	0/ -1	0	0/-1	301/267	301	301/267	yes	yes	yes
10#	183046759	0	0/-1	-1	222/260	222/184	184	0/ +1	yes	-1/ +1
11†	183081195	+2/ -1	2/1	-2	182/122	182/167	182/122	yes	2	yes
12	183117043	-5/ -3	-5/-3	0/-3	180/214	180/214	265/214	yes	yes	yes
13	183169331	-2	-2/0	-2	209	209/239	209	yes	yes	yes

Table 4.3: Experimental validation of 13 loci on the AJ trio. Thirteen VNTR loci were selected for experimental validation in the AJ trio. The experiment was performed by Samantha D. Drinan (Fuxman Bass lab, Boston University). All but one of the 66 bands predicted by VNTRseek were validated. † For the remaining band, the results were questionable because the two predicted alleles for the father were only 15 nucleotides different in length, which was too close to distinguish in the image. ∗ For all three individuals, the gel contained bands (bold font) not predicted (or detectable) by VNTRseek. The extra band for the son corresponded to the -1 allele as found in the PacBio reads. The father's extra band appeared to match with the -1 allele. The mother's extra band appeared to be a +1 allele (552 nucleotides). # An extra band for the mother and son (bold font) was not predicted by VNTRseek, although it seemed to match the +1 allele that was detectable.

#	TR	VNTRseek			Expected bands			Validated		
		son	father	mother	son	father	mother	son	father	mother
1	182316181	-2	-2	-2	237	237	237	-2, **-1**	-2, **-1**	-2, **+1**
2	182316985	-3, 0	-1, 0	-3, -2	118, 199	172, 199	118, 145	yes	yes	yes
3	182453735	0, +1	0, +1	0, +1	180, 210	180, 210	180, 210	yes	yes	yes

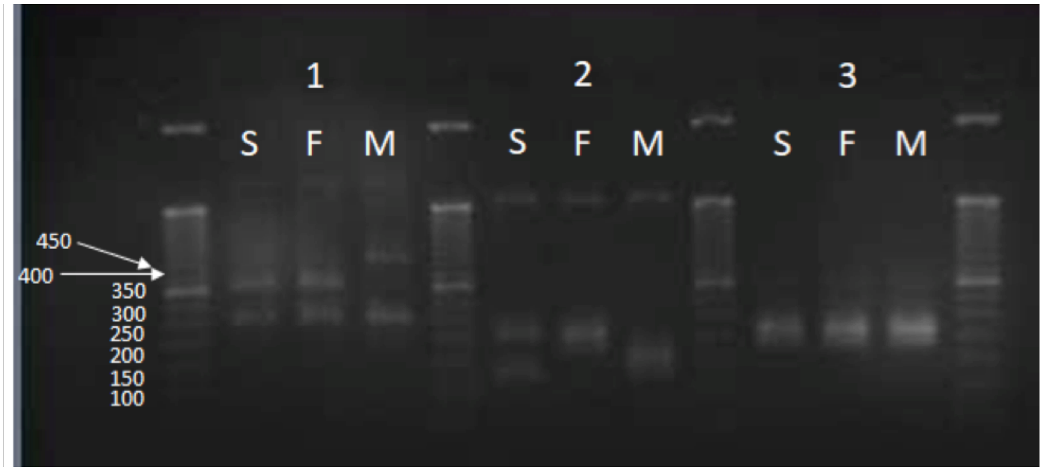


Figure 4·1: Experimental validation of loci 1, 2, and 3. In locus 1, all three individuals share a band at 237bp. The father and son share a band at approximately 342bp and the mother has an additional band above 450bp. The bands at and above 342bp correspond to alleles that are not detectable by VNTRseek and were not predicted. The 342 bp band corresponds to a -1 allele found in the PacBio reads for the son. The mother's extra band appears to be a +1 allele (552 bp). In locus 2, the son and father share a band at 199 bp and the son and mother share a band at 118 bp. The father has an additional band at 172 bp and the mother at 145 bp. In locus 3, all individuals (son, father, mother) share the same bands, 180 and 210 bp). All three loci confirm the VNTRseek predictions.

#	TR	VNTRseek			Expected bands			Validated		
		son	father	mother	son	father	mother	son	father	mother
4	182461997	-5,-4	-5,-4	-5,-2	193, 231	193, 231	193, 307	yes	yes	yes
5	182493720	0	0	0, -1	266	266	266, 196	yes	yes	yes
6	182515357	-5	-5, -6	-5	195	195, 161	195	yes	yes	yes

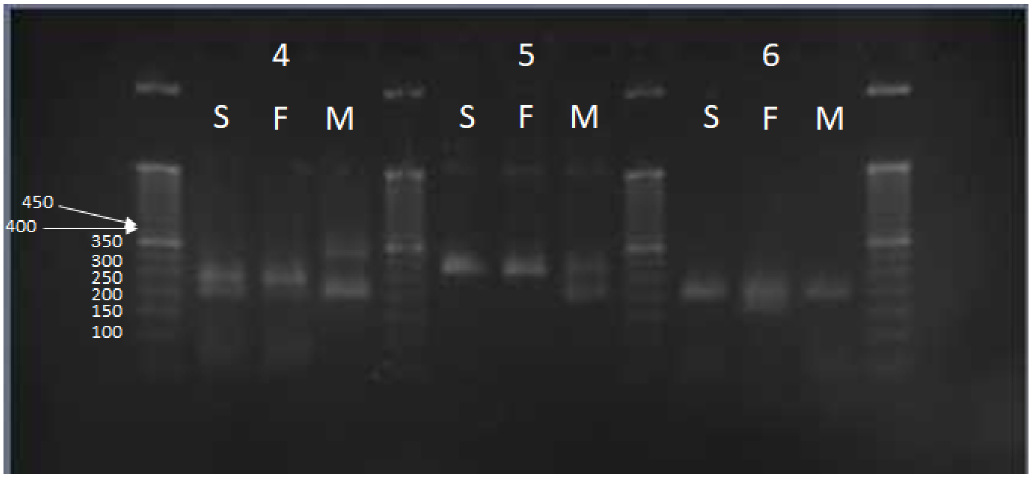


Figure 4·2: Experimental validation of loci 4, 5, and 6. In locus 4, the son shares a larger band with the father (231bp) and a smaller band with both the mother and the father (193bp). The mother seems has a larger band at 307bp. In locus 5, all three individuals share a band at 266bp. The mother has an extra band which is smaller at 196bp. In locus 6, all individuals share a band at 195bp. The father has a smaller band at 161. All three loci confirm the VNTRSeek predictions.

#	TR	VNTRseek			Expected bands			Validated		
		son	father	mother	son	father	mother	son	father	mother
7	182608886	0, -2	0, -3	-2, +1	208, 181	208, 154	235, 181	yes	yes	yes
8	182620950	0, -1	0, -1	0, -1	243, 195	243, 195	243, 195	yes	yes	yes
9	182982510	0, -1	0	0, -1	301, 267	301	301, 267	yes	yes	yes

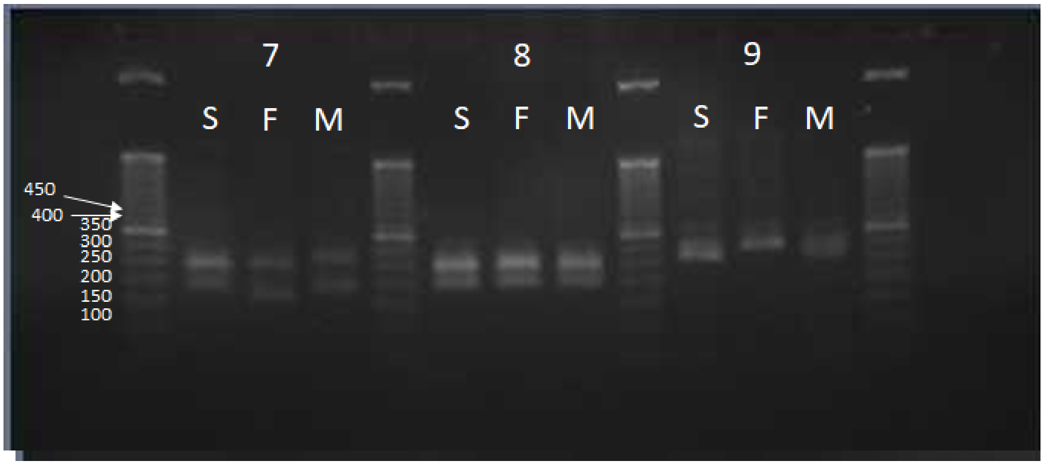


Figure 4·3: Experimental validation of loci 7, 8, and 9. In locus 7, the son and father share a band at 208 bp and the son and mother share a band at 181 bp. The father has another smaller band at 154bp and the mother has a band slightly higher at 235 bp. In locus 8, all individuals share two bands at 195 bp and 243 bp. In locus 9 all individuals share a band at 301 bp. The son and mother share another smaller band at 267 bp. All three loci confirm the VNTRSeek predictions.

#	TR	VNTRseek			Expected bands			Validated		
		son	father	mother	son	father	mother	son	father	mother
10*	183046759	0	0, -1	-1	222	222, 184	184	0, **+1**	0, -1	-1, **+1**
11	183081195	+2, -1	+2, +1	+2, -1	182, 122	182, 167	182, 122	yes	yes*	yes

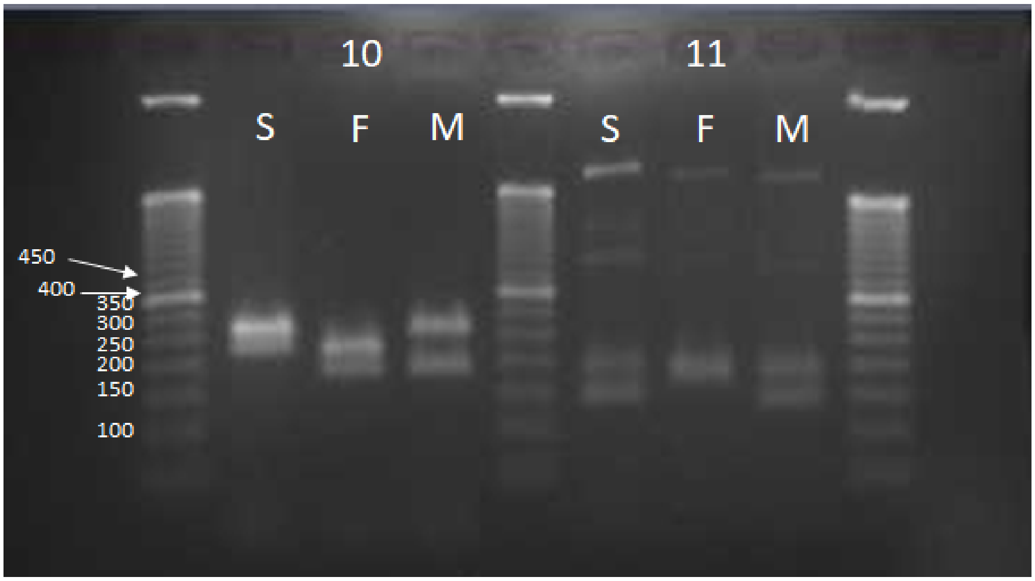


Figure 4·4: In locus 10 the son and father share a band at 222 bp. The father and mother share a band at 184 bp. The son and mother share a larger band which appears to be the +1 allele (260 bp), which was detectable, but not predicted in either sample by VNTRseek. The alleles predicted by VNTRSeek were validated and an extra band that could have been predicted was missed in two samples. In locus 11, all individuals share the larger band at 182bp. The son and mother have a smaller band at 122bp. If the father has two alleles as predicted by VNTRseek, then they are very close together (15 nucleotides different in length). There is a suggestion in the gel that there are two close bands. In any event, all but one of the alleles was validated.

#	TR	VNTRseek			Expected bands			Validated		
		son	father	mother	son	father	mother	son	father	mother
12	183117043	-5, -3	-5, -3	0, -3	180, 214	180, 214	265, 214	yes	yes	yes
13	183169331	-2	-2, 0	-2	209	209, 239	209	yes	yes	yes

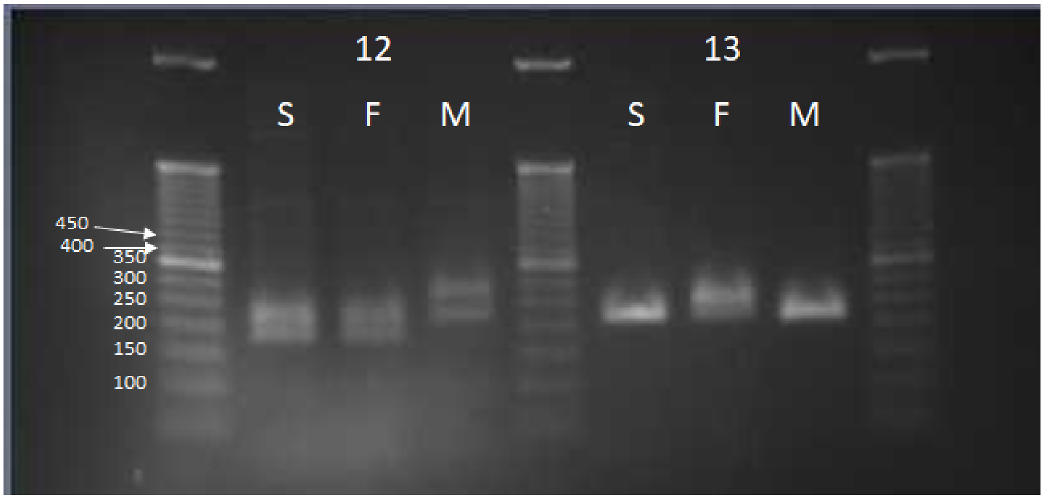


Figure 4·5: Experiment Validation of loci 12 and 13. In locus 12, all individuals share a band at 214 bp. The son and father share a smaller band at 180bp. The mother has a larger band at 265 bp. In locus 13, all individuals share a band at 209 bp. The father has a larger band at 239bp. Both loci confirm the VNTRSeek predictions.

Among the experimentally verified loci, three were located within coding exons (genes IRF5, TENT5A, and RNF138) and might be expected to affect protein function. We therefore examined predicted allele occurrence at these loci within the 150 bp NYGC dataset. In the case of IRF5, 675 individuals were genotyped as reference, 609 had genotype –2/–2 and 1,209 individuals had both 0 and –2 alleles. For TENT5A, only 127 individuals had the reference genotype; the more common genotypes were +1/+1 (854 individuals), -1/-1 (693 individuals), 0/+1 (462 individuals), and 0/-1 (178 individuals). In the case of RNF138, the reference genotype could not be observed because it was 144 bp (¿130 bp), but in 2,158 individuals we observed a loss of one copy. In all three VNTR loci, the repeat pattern length was a multiple of three

(15 bp in IRF5 and TENT5A and 48 bp in RNF138). This suggests that exon translation would not be interrupted, potentially resulting in proteins of different lengths in different individuals.

We also compared VNTRseek predictions in three datasets from the NA12878 genome with VNTRs validated in the adVNTR paper (Bakhtiari et al., 2018). Out of the original 17 VNTR loci experimentally validated in the adVNTR paper, four were not included in our reference set and for one, the matching TR could not be determined.

Overall, 11 of 16 detectable alleles were correctly predicted, four were not found in the NA12878 sample with sufficient read size (250 bp), and one was incorrectly predicted in the HG001 sample and was not found in the other two (Table 4.4).

4.3.2 Validation of predicted VNTRs using long reads

PacBio Circular Consensus Sequencing reads from the HG002 genome (Wenger et al., 2019), with an average length of 13.5 Kbp and an estimated 99.8% sequence accuracy, were computationally tested to determine if they confirmed VNTRseek predicted alleles for the GIAB Illumina reads from the same genome. Overall, more than 97% of predicted alleles were confirmed, and at the predicted VNTR loci, more than 87% of alleles were confirmed (Table 4.5).

Gene	TRid	Pat.	CN	Array	Δbp	Genotype	Predicted	HG001	NA12878	NA12878
								(148 bp)	(150 bp)	(250 bp)
LCE4A	182223827	24	2.21	53	24	1	0	error	not found	not found
DRD4	182318145	48	4.29	206	0	0		N/A	N/A	not found
					-48	-1		N/A	N/A	not found
IL1RN	182721519	86	4	344	0	0		N/A	N/A	N/A
					-172	-2		N/A	N/A	not found
C14orf180	182495840	9	3.11	28	0	0	0	match	match	match
CSTB	182814480	12	3.08	37	0	0	0	match	match	match
					-12	-1	-1	match	match	match
PAOX	182317476	35	4	140	-35	-1	-1	match	match	match
SSTR1	182469843	24	2.42	58	0	0	0	match	match	match
					24	1	1	match	match	match
EIF3G	182649420	21	2.29	48	0	0	0	match	match	match
					-21	-1	N/A			
JAKMIP3	182316063	39	3.13	122	0	0	0	match	not found	match
					-78	-2	N/A			
SRSF8	182354360	21	2.1	44	0	0	0	match	match	match
					-21	-1	N/A			
MAOA	183311386	30	3.5	105	30	1	1	N/A	N/A	match
GP1BA	182574350	39	3.92	153	0	0		N/A	N/A	not found
					-40	-1	-1	not found	not found	match
BRWD1	182812082	22	2.14	45	0	0				
					-22	-1	not in ref.			
CLCA4	184859017	6	5.67	34	-6	-1	not in ref.			
SLC6A4	182584594	63	2.27	358	-66		not in ref.			
STK39	182741803	6	4.67	28			not in ref.			
UBXN11					-18		not in ref.			

Table 4.4: Comparison of VNTRseek allele calls to validations from the adVNTR paper. VNTRseek validation results for three datasets from the same genome, HG001 (148 bp) from GIAB, NA12878 (150 bp) from NYGC, and NA12878 (250 bp) from 1000 Genomes Phase 3 HC for VNTR loci as reported in the adVNTR paper. In one case, a matching TR could not be determined. In total, 11 out of 16 detectable alleles were correctly predicted, four were not found in the NA12878 250 bp sample, and one was incorrectly predicted in the HG001 sample.

Locus Category	All	All VNTR	All Hom	Hom Ref.	Hom VNTR	All Het	Het 0/1	Het 1/2
VNTRseek genotyped	170,481	2597	169,175	1167,884	1,291	1,306	1,139	167
With 0 alleles validated	4,121	226	4,074	3,895	179	47	38	9
With 1 allele validated	165,303	1,314	164,996	165,101	1,112	202	173	29
With 2 alleles validated	1,057	1,057	–	–	–	1,057	928	129
Alleles PPV	97.46%	87.83%	97.59%	97.68%	86.13%	88.67%	89.07%	85.93%

Table 4.5: VNTRseek validation in PacBio long reads. PacBio Circular Consensus Sequencing reads from the HG002 genome (Wenger et al., 2019), with an average length of 13.5 Kbp and an estimated 99.8% sequence accuracy, were computationally tested to determine if they confirmed VNTRseek predicted alleles from the GIAB Illumina reads. If a VNTRseek predicted allele copy number from the GIAB data was matched in at least one PacBio read previously mapped to the VNTR locus, the allele was considered confirmed. Matching copy numbers were defined as differing by at most 0.25 copy. Counts are shown for all loci and separately for homozygous and heterozygous loci. Het 0/1 are heterozygous VNTR loci called with one reference allele and one variant allele. Het 1/2 are heterozygous VNTR loci called with two variant alleles. Alleles PPV (positive predictive value) is the percentage of VNTRseek predicted *alleles* that were validated. Overall, more than 97% of predicted alleles were validated, and at the predicted VNTR loci, more than 87% of alleles were validated.

4.3.3 VNTR predictions are consistent with Mendelian inheritance.

We compared the predicted alleles in four trios (CEU and YRI trios from 1000 Genomes; Chinese HAN and AJ from GIAB), testing loci on autosomes and X and Y chromosomes (see Section 4.2.3). We considered loci defined by several increasingly stringent criteria: both parents heterozygous, all members of the trio heterozygous, all members of the trio heterozygous and with different genotypes. Additionally, loci in the X and Y chromosomes of male children were chosen to test for consistency with direct inheritance from the mother or father, respectively. In all cases, only 15 of loci were inconsistent (Table 4.6).

Trio:	AJ	HAN	CEPH	YRI
Heterozygous in both parents	358	224	302	469
Inconsistent	2	2	2	7
Heterozygous in all three	262	178	215	349
Inconsistent	0	2	1	3
Heterozygous in all three and different	59	32	53	105
Inconsistent	0	2	1	3
All on chrY of son	911	884	-	-
Inconsistent	0	0	-	-
All on chrX of son	6,701	6,493	-	-
Inconsistent	2	0	-	-

Table 4.6: Consistency with Mendelian inheritance of VNTR genotypes in trios from GIAB and the 1000 Genomes Phase 3 HC datasets. Only loci detected in all members of a trio were considered, but with increasingly stringent criteria. Loci were overwhelmingly consistent.

4.3.4 VNTR predictions are consistent across platforms

In 2015, the 1000 Genomes Phase 3 sequenced 30 genomes using Illumina HiSeq2500 at read length 250bp. In 2020, 27 of those 30 genomes were resequenced by NYGC using Illumina Novaseq 6000 at read length 150 bp. Comparing VNTR loci genotyped in both platforms and non-reference alleles detectable at both read lengths, agreement ranged from 76%–91% (Figure 4·6). Note, however, that read coverage was not the same for both datasets, causing variation in statistical power.

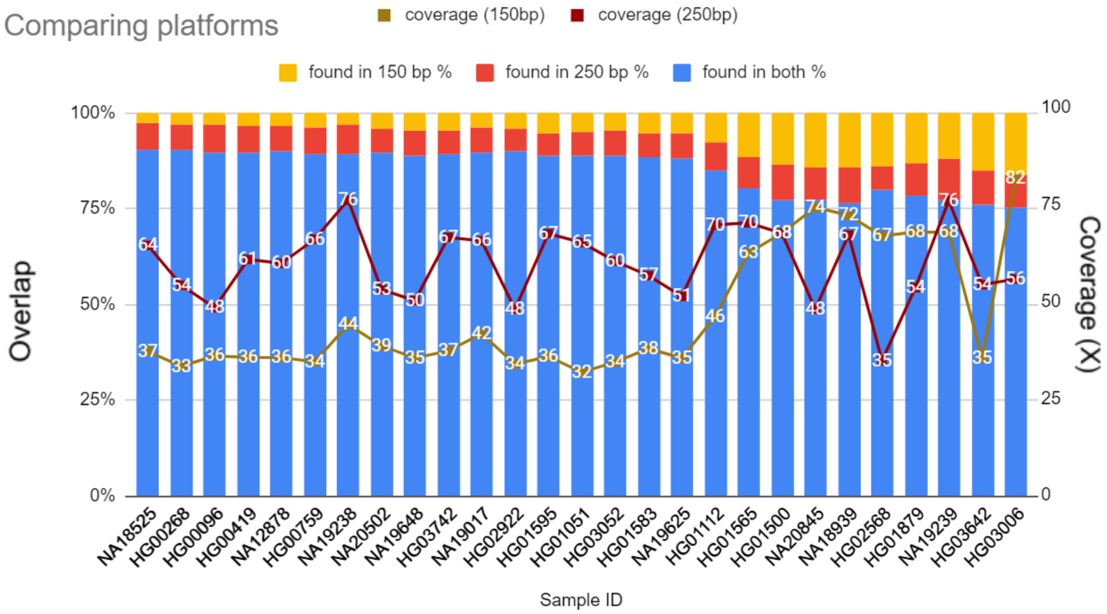


Figure 4·6: Comparison of VNTRseek allele prediction across platforms. Comparison was done on the 27 genomes in common in the 1000 Genomes Phase 3 HC and NYGC datasets. Only non-reference alleles detected in both genomes were included. Agreement of allele calls in both platforms ranged from 76% to 91%. Each bar represents one genome. Colored bar heights indicate percentage of agreement between predicted alleles (blue) or alleles found in one sample only (read and yellow). Lines indicate the read coverage in the two platforms in each sample. Generally, as the read coverage of the 150 bp sample increased above that of the 250 bp sample, more alleles were found in the former that were not detected in the latter, which was expected due to higher statistical power. Conversely, when the coverage was higher in the 250 bp sample, that platform found more alleles not detected in the 150 bp sample.

4.3.5 Controlling for type 1 errors

"Multis" are TR loci genotyped with more alleles than logically possible (Section 2.2.3). These alleles contribute false positive calls or type 1 errors. A total of 98,150 Multis were detected across all genomes corresponding to 3,808 loci. Figure 4.7 shows the histogram of how many times a TR is genotyped as a Multi across all genomes. It can be observed that most Multi loci occur in only a few genomes: A total of 1,384 (36%) of the Multi loci occurred in only one genome. Another 525 (14%) occurred in exactly two genomes. Only 115 (3%) occurred in more than 5% (140) of the genomes.

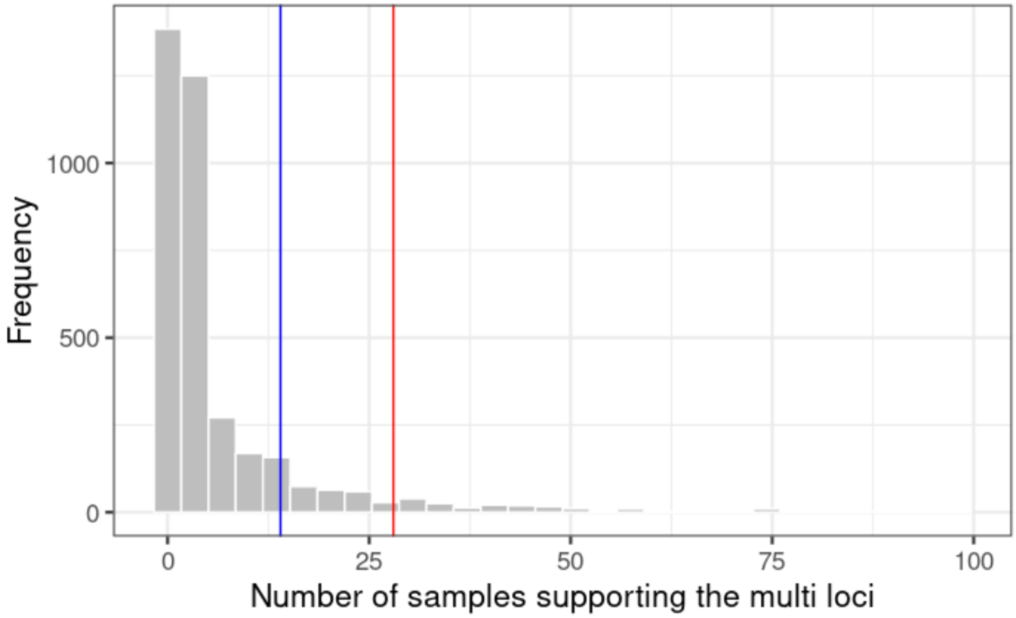


Figure 4·7: Frequency of recurrent Multis. The number of samples calling the same loci as Multi is presented as a histogram. Loci that occurred in more than 100 individuals are not shown, because all occurred in <4 individuals. The blue vertical line shows the 0.5% cutoff (14 samples) and red line shows the 1% (28 samples). Most Multis reoccur in few individuals (<1%), suggesting that the Multis are not related to the sequence or characteristics of the TR, such as the CG content, pattern size, copy number, etc.

We next investigated two sources for type 1 error. First, in case of recurrent Multis, we investigated whether similarity in sequence is causing mapping errors. Using a sliding window approach (Section 2.2.3) the reference TRs that cause ambiguous mapping due to similarity in sequence are labelled as Indistinguishable. Another source of repeats, which might not be captured by the sliding window, is Segmental Duplications. Segmental Duplications are duplications of 1 Kbp or longer with over 90%

similarity across the human genome (Vallente and Eichler, 2005). Segmental duplications are known to reduce mappability and cause false positive CNV and structural variation calls (Bailey et al., 2001; Alkan et al., 2011). Figure 4·8 summarizes the percentage of recurrent Multi loci (in more than 1% of the genomes) that were Indistinguishable or overlapped with Segmental Duplications. Figure 4·8 shows that about half of the recurrent Multis could be explained by these two criteria. About 36% of the recurrent Multi loci were correctly marked as Indistinguishables and another 5–10% were located inside Segmental Duplications. Interestingly, overlapping with Segmental Duplications increased the chances of recurrent Multis. VNTRseek indistinguishable filtering is doing well to identify Multis. Using a linear regression model:

$$N \sim Indistinguishable + Segmental\_Duplication \tag{4.1}$$

the effect size of being indistinguishable was 39% (p-value=1.61e-13) and overlapping Segmental Duplications had effect size 48 (7.89e-4).

Second, in the case of random Multi loci (occurring in <1% of genomes), we propose that the sequencing errors are causing false calls. In this case, we would see a correlation between the read coverage and number of singleton Multis per genome. Indeed, the number of Multis increased as the read coverage increased (Figure 4·9). To reduce the effect of sequencing errors, we would need to increase the minimum read support for an allele. The default minimum read support setting for VNTRseek is two reads, meaning that a TR allele will require two reads supporting the same copy number to be reported. In the next section the effect of the minimum read support parameter on Multi calls is investigated.

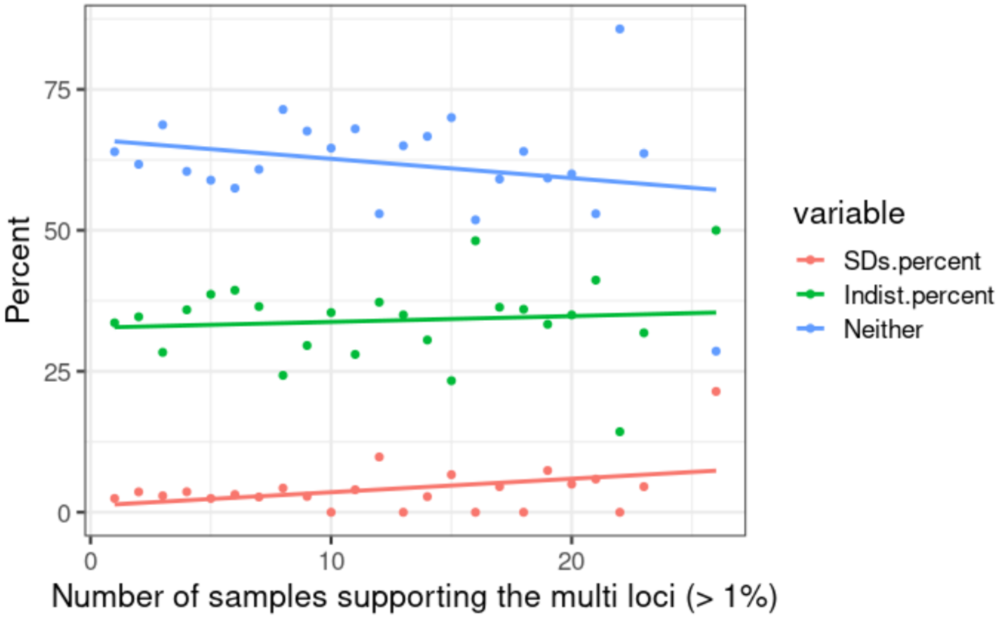


Figure 4·8: Number of Multis explained by Segmental Duplications and Indistinguishable TRs. For each Multi loci detected in at least 1% of the genomes (28 genomes), we measured the percentage that were Indistinguishable and the percentage that overlapped Segmental Duplications (SD). Approximately 30–35% of the Multis were caught by the indistinguishable filter. Another 5–10% could be explained by occurrence inside large Segmental Duplications. Interestingly, overlap with Segmental Duplications increases the change of a Multi TR call.

4.3.6 Type one error reduced by increasing minimum read support cutoff

CHM1 and CHM13 are haploid genomes, meaning we expect all the genotyped TRs to have only one allele. Under such assumption, every genotyped TR with two alleles (mimicking heterozygous VNTR), would be marked as Multi. We used such Multi loci on these haploid genomes to investigate the effect of minimum read support cutoff on Multi errors. We separated the Multi loci with two alleles on CHM1 and CHM13 by read support, the alleles with the less read support were labeled as “Lower” and the other one, “Higher”. Figure 4·10 shows the distribution of the two classes. By

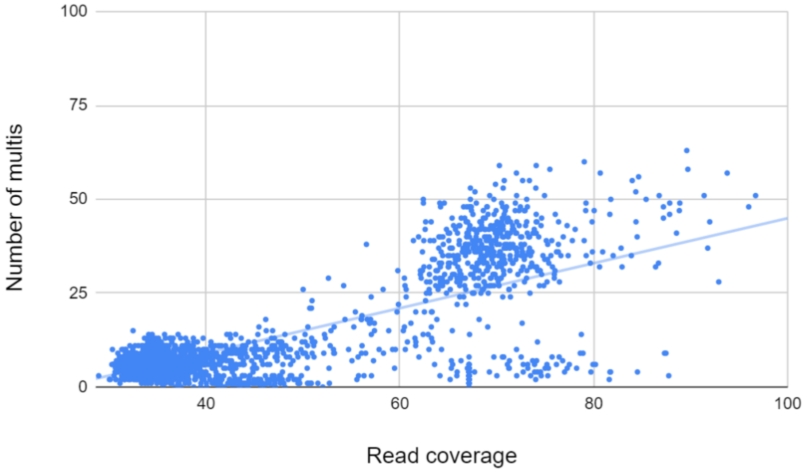


Figure 4·9: Number of Multis per read coverage. Only datasets coverage below 100X are shown to avoid outliers. There is a clear trend between read coverage and number of Multis. As the read coverage increases, sequencing errors accumulate, resulting in higher type 1 errors.

increasing the minimum read support, the erroneous allele would have been filtered and the loci would not have been a Multi.

The percentage of loci categorized as Multi in CHM1 and CHM13 was 0.1% and 0.3% of all TR loci and 18% and 25% of VNTR loci, respectively, at the default requirement for a minimum of two reads supporting an allele call. Increasing the requirement to three reads reduced the percentages to 8% and 15%, respectively (Figure 4·11 and Figure 4·12). This shows the necessity of using more stringent calling criteria in the case of higher coverage, as for CHM13 with a coverage of 137×. However, increasing the minimum required read support could result in lost sensitivity, because longer arrays are expected statistically to have fewer spanning reads. We are currently developing more sophisticated methods to determine the minimum required cutoff dependent on read coverage and array length. Because the expected read support depends on array length, we do not recommend such a coarse-grained correction. We are currently working on a more refined approach based on array length.

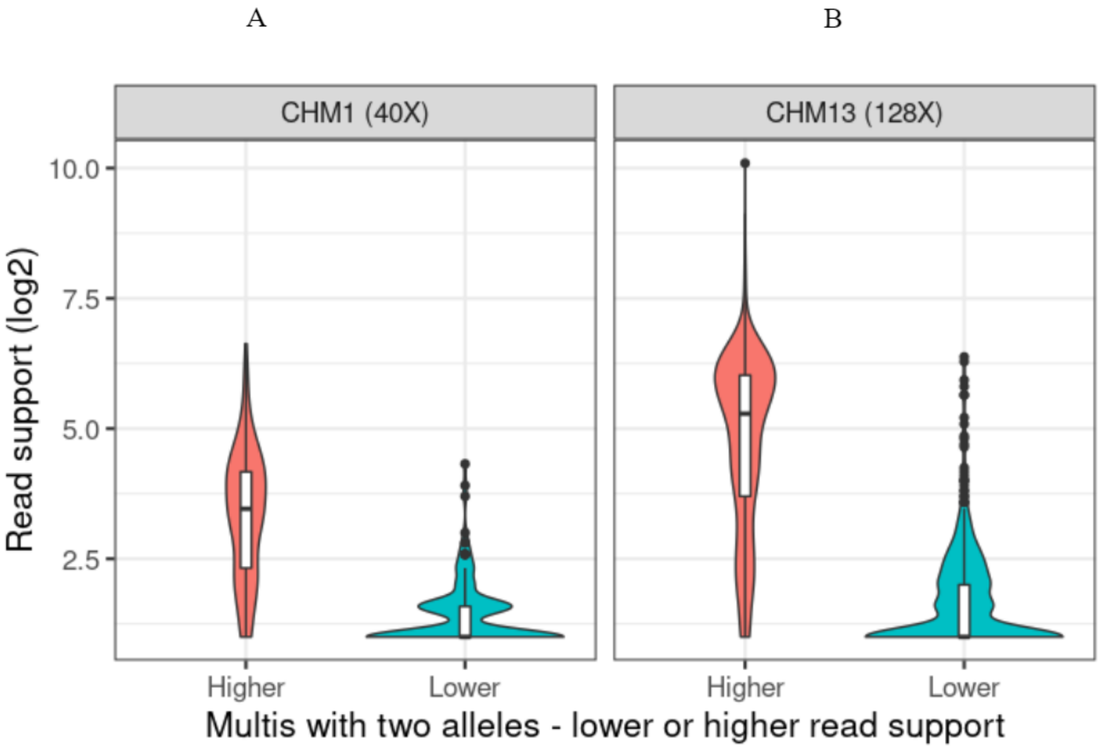


Figure 4·10: The Multi calls on haploid genome by read support. Haploid genomes CHM1 and CHM13, theoretically, should not have any heterozygous calls. Such calls are marked as Multis. We divided two alleles of such Multi calls on these two genomes into two groups: the allele with less read support (Lower), and the allele with more read support (Higher). For each group we plotted the distribution of read support. Most of these Multis had one allele with very few reads (<4). These false positive calls could be resolved by increasing the minimum read support parameter in VNTRseek. The default value for minimum read support is two.

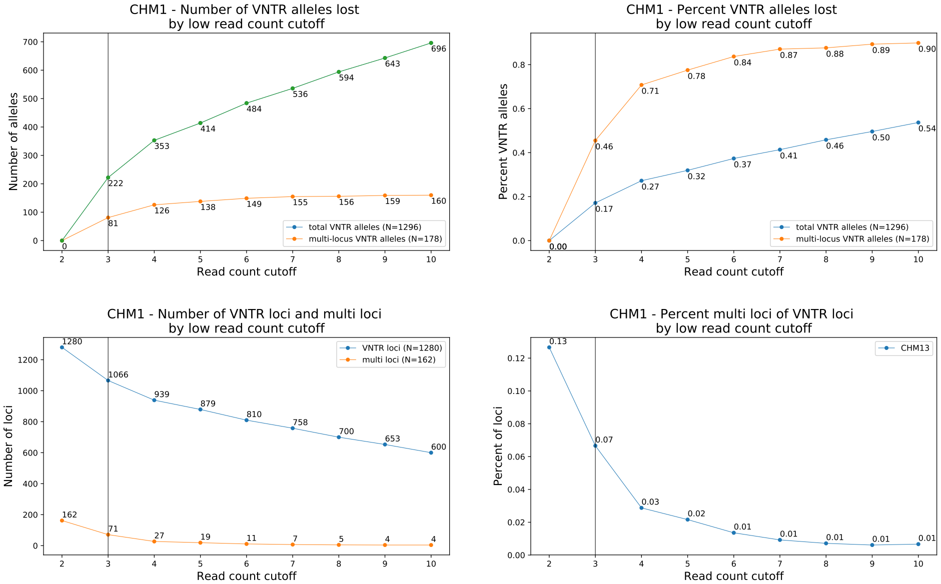


Figure 4·11: False positive VNTR alleles on the haploid genome CHM1. CHM1 is a haploid genome without Y chromosomes and theoretically carries no heterozygous VNTRs. We measure false positive calls by VNTRseek by counting loci with Multi alleles. Here we compared the number of Multi calls (heterozygous calls or any allele call on chromosome Y) as a function of the minimum required read support for an allele. We observed that as the minimum required support increased, the number of false positive calls fell (bottom figures). However if the support was increased too much, the sensitivity suffered. The minimum read support in VNTRseek is, by default, set to two. At this setting for CHM1 (40× coverage), 13% of the VNTR loci were Multis, i.e, had at least one excess allele (lower right graph in the first quartet of figures). Raising the minimum support to three reads, decreased this to 7%.

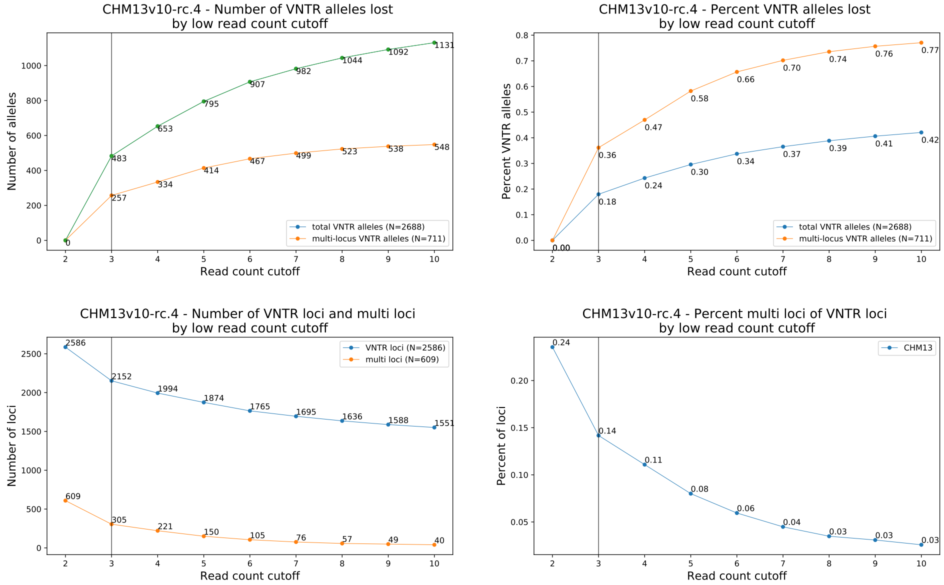


Figure 4·12: False positive VNTR alleles on the haploid genome CHM13. CHM13 is a haploid genome without Y chromosomes and theoretically carries no heterozygous VNTRs. We measure false positive calls by VNTRseek by counting loci with Multi alleles. Here we compared the number of Multi calls (heterozygous calls or any allele call on chromosome Y) as a function of the minimum required read support for an allele. We observed that as the minimum required support increased, the number of false positive calls fell (bottom figures). However if the support was increased too much, the sensitivity suffered. The minimum read support in VNTRseek is, by default, set to two. At this setting for CHM13 (128× coverage), at the default setting, 24% of the loci were Multis, and the support threshold had to be raised to five to decrease this to 8% (lower right graph).

4.4 Summary

In this chapter, the robustness of our VNTR predictions are confirmed in several ways, including the following:

1. successful experimental validation of 66 alleles at 13 VNTR loci in the AJ family trio,
2. comparison of allele genotyping prediction results from Illumina reads to allele occurrence in PacBio reads from the AJ child showing 97% agreement of all alleles and 87% agreement at VNTR loci,
3. comparison of our predictions in the NA12878 genome with experimentally validated results from the adVNTR prediction tool (Bakhtiari et al., 2018) confirming 11 correct VNTR allele predictions and one incorrect prediction, and
4. comparison of results from two Illumina sequencing platforms on 27 genomes that shows 76–91% consistency of the VNTR allele calls.

In this chapter, we also investigated type 1 errors, marked as Multis. These loci are genotyped with more alleles than logically possible on the genome (more than two alleles on a diploid chromosome, or more than one allele on a haploid chromosome such as sex chromosomes in male individuals). We investigated two sources of error: sequence similarity of the TRs and DNA sequencing errors. Sequence similarity would cause ambiguous mapping of reads and should increase the likelihood of a loci becoming a recurrent Multi. We see that Indistinguishable TRs are the source of the Multi about 30% of the time. Segmental Duplication increased the likelihood of recurrent Multi loci.

Another source of Multis is DNA sequencing error. As the read coverage increases, the effect of sequencing errors rises. Genomes with higher read coverage had higher counts of Multi calls. On CHM1 and CHM13 we show that most Multi calls have one allele with an abnormal read count. We show that increasing the minimum read support cutoff would reduce these erroneous alleles. However, longer arrays would suffer due to this filtering. In Chapter 5, a computational method is proposed to dynamically set the cutoff for each array length and detect such errors.

Chapter 5

Inferring heterozygous VNTRs using read support

5.1 Introduction

As discussed in Chapter 1, genomic variant detection tools report the “read support” as an indicator of confidence. The idea is that when a large number of reads support the same variant, that variant is less likely to have been called by mistake, i.e. less likely to be a false positive. Similarly, VNTRseek reports the read support for each detected allele. By default VNTRseek requires two reads to support any allele in order for it to be reported.

As we saw in Chapter 4, VNTRseek's allele-finding capability is limited in two ways: 1) the allele array length must fit in the read, and 2) the allele must have at least two copies in order to be detected by TRF.

For example, consider genotyping a reference TR with the following attributes: pattern length 40 bp and copy number 3, in an individual with read length of 150 bp and coverage 100X. The following three error scenarios could occur:

Scenario 1) *The individual has genotype 0/-2.* Let n(0) and n(-2) be the number of reads spanning the 0 and -2 allele, respectively. VNTRseek will run TRF on the n(0) reads and detect the reference allele each time. However, TRF will not detect the -2 allele in the remaining n(-2) reads because the copy number is below two. As a result VNTRseek will output this genotype as 0/0 with read support n(0).

Scenario 2) *The individual has genotype 0/+2.* Let n(0) and n(+2) be the number of reads spanning the 0 and +2 allele, respectively. Similar to scenario 1, n(0) reads will be reported with allele 0. However, in the case of the allele +2, the array size has become reference array length + (2 pattern size), which is equal to 200 bp. A reads of 150 bp cannot span an array of 200 bp, thus, the +2 allele will not be detected. VNTRseek will report 0/0 with read support n(0).

Scenario 3) *The individual has genotype 0/0.* Let n(0) be the number of reads spanning the loci from both chromosomes. In two reads, the right flank has sequencing errors such that the flank becomes similar to the repeat unit. Then, TRF could possibly align the flank as another repeat unit and report a read TR with four copies. Under these circumstances VNTRseek would incorrectly report 0/+1.

The first two scenarios illustrate the detectable range of VNTRseek. The third scenario shows how “Multis” occur (see Section 4.3.5). This chapter presents a computation tool named mlZ (machine learning on Z-scores) to statistically infer missing alleles (scenario 1 and 2) and “Multi” errors (scenario 3). mlZ uses the read support to predict whether VNTRSeek allele calls are truly heterozygous or homozygous.

The first two scenarios illustrate the detectable range of VNTRseek. The third scenario shows how “Multis” occur (see Section 4.3.5). This chapter presents a computation tool named mlZ (machine learning on Z-scores) to statistically infer missing alleles (scenario 1 and 2) and “Multi” errors (scenario 3). mlZ uses the read support to predict whether VNTRSeek allele calls are truly heterozygous or homozygous.

5.2 Materials and methods

Section 5.2.1 describes the simulated datasets that were used to test VNTRseek and mlZ. The real data used to illustrate mlZ’s application is presented in Section 5.2.2. Then, Section 5.2.3 lists the abbreviations used in the rest of the chapter. Section 5.2.4

summarizes the mlZ algorithm and Section 5.2.5 describes how the Decision Tree model was trained.

5.2.1 Simulated datasets

To evaluate the mlZ algorithm, three diploid genomes with random SNPs and indels were simulated using varsim (Mu et al., 2015) with default parameters on the GRCh38 reference genome. Out of the total reference set, a VCF file with random VNTRs was designed. The reference TR set was randomly stratified into six subsets of equal size. The first, second, third, and fourth subset were designed to have the following genotypes: heterozygous gain (0/+1), homozygous gain (+1/+1), heterozygous loss (0/-1), and homozygous loss (-1/-1), respectively. The remaining two-sixth were left with the reference genotype (0/0). In all cases, the loss and gain of one copy number were implanted.

varsim does not allow overlapping variants, e.g. an SNP inside a VNTR array. To ensure that the designed VNTRs were implanted, all SNPs and indels overlapping the VNTRs from the VCF file were removed. Then, Illumina reads at different read lengths and fragment distributions were simulated using ART read simulator (Huang et al., 2012) to create three testing datasets comparable to real datasets. Table 5.1 summarized the characteristics of the simulated datasets. The three simulated datasets are referred to as Sim1, Sim2, and Sim3 with respective read lengths 250 bp, 148 bp, and 101 bp, and physical coverage 100×.

5.2.2 Real datasets

High coverage PCR-free WGS datasets of seven individuals from the Personal Genome Project (Church, 2005) sequenced by the Genome In A Bottle (GIAB) consortium (Zook et al., 2016) were downloaded. The seven genomes are labelled HG001 to HG007:

ID:	Sim3	Sim2	Sim1
Fragment mean	350	550	550
Fragment standard deviation	100	150	150
Read length	101bp	148bp	250bp
Error profile (ART)∗	HS10 - HiSeq 1000	HSXn - HiSeqX PCR free	MSv3 - MiSeq v3
Physical (fragment) coverage∗∗	100X	100X	100X
Number of reads simulated	2,818,459,600	1,923,354,266	2,277,177,120
Read coverage	∼95X	∼95X	∼190X

Table 5.1: The simulated datasets. ∗ The error profile is provided by ART (Huang et al., 2012) based on real runs. ∗∗ Three datasets were simulated for testing purposes. The characteristics of simulations were designed to represent the real data. Read coverage is calculated as the product of the read length and number of reads over the total size of the genome.

- HG001 is the pilot NA12878 genome from the CEPH/UTAH family (mother) which is commonly used in variation studies,
- HG002, HG003, and HG003 are the Ashkenazi Jew (AJ) trio,
- and HG005, HG006, HG007 are the Chinese trio.

The HG001 and Chinese parents were sequenced at read length 148 bp (using the reference set of 150 bp), and the Chinese son and AJ trio have 250 bp reads. The Chinese son and HG001 were sequenced at 300X coverage and the others were sequenced at 100X. The high coverage and read length allows for more VNTRs to be detected by VNTRseek, thus, more input data for mlZ.

A dataset on the NA12878 genome from the 1000 Genomes Project Phase 3 was compared to the HG001 results, in order to measure the consistency of mlZ across sequencing platforms. The NA12878 dataset was sequenced with read length 250 bp. In addition, long precise CCS PacBio reads were obtained on the HG002 genome (Wenger et al., 2019) with read length ∼13,500 bp. This data set was to evaluate the mlZ predictions on that genome. These datasets are described in Chapter 2 (Section 2.2.1).

5.2.3 Definitions

- *VRS:* read support for each allele reported by VNTRSeek

- ERS: expected read support for a given array length

- *SRS:* simulated read support on bins of size 10, gives medians and standard deviations based on real data characteristics i.e, the read coverage, fragment mean and standard deviation, and read length

- *ORS:* the mean and standard deviation of fitted bimodal normal distribution on the binned VRS

- *lmRS:* the smoothed ORS with a simple linear regression to filter out variation due to noise

- *normal-VRS:* for each allele, VRS is divided by (ERS+1); 1 is added to the ERS to avoid division by zero

5.2.4 The mlZ algorithm

The mlZ algorithm compares the read support of each allele detected by VNTRseek to the expected read support distributions for that allele. The steps of this algorithm are depicted in Figure 5·1.

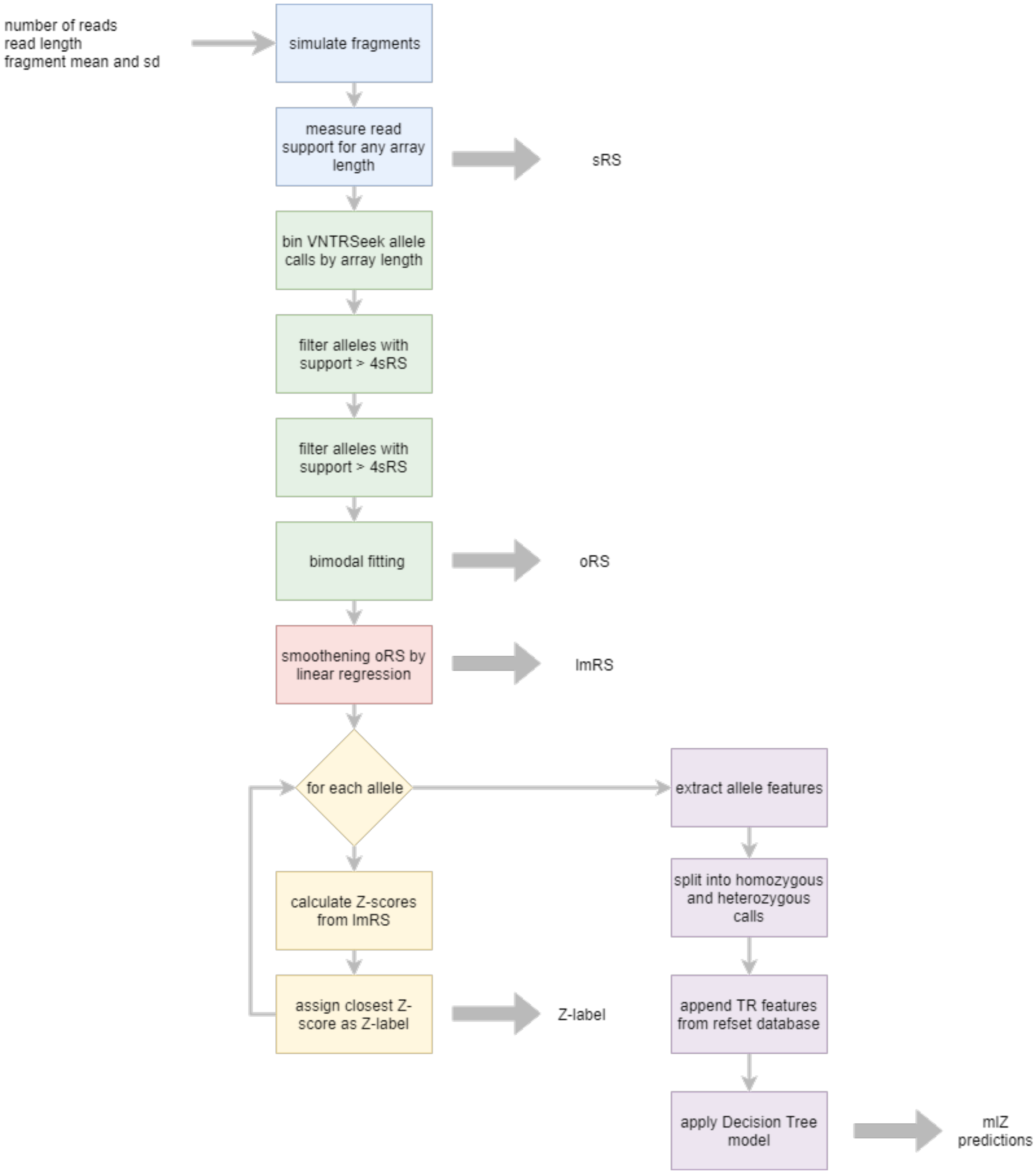

number of reads
read length
fragment mean and sd
simulate fragments
measure read support for any array length
sRS
bin VNTRSeek allele calls by array length
filter alleles with support > 4sRS
filter alleles with support > 4sRS
bimodal fitting
oRS
smoothening oRS by linear regression
lmRS
for each allele
extract allele features
calculate Z-scores from lmRS
split into homozygous and heterozygous calls
assign closest Z-score as Z-label
Z-label
append TR features from refset database
apply Decision Tree model
mlZ predictions

Figure 5·1: Flowchart of the mlZ algorithm. mlZ extracts VRS for each VNTR and compares the VRS to SRS to find the most likely distribution: heterozygous or homozygous.

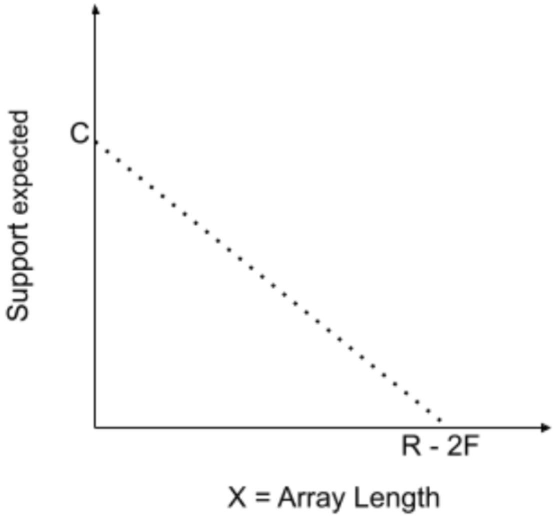


Figure 5·2: Relation between expected read support and array length. The X-axis is the array length. The Y-axis is the read support, i.e. the number of reads spanning the array with given flank length. R and C are the read length and coverage extracted from the real data. F is the minimum flank required by VNTRseek. The expected read support (ERS) is linearly proportional to the array size. At array length $X = 1$, C reads span a loci (array length $X = 1$). No read could possibly span an array with length $R - 2F$.

Assuming read coverage follows a Poisson distribution, expected read support increases linearly with read length and decreases linearly with allele length (Figure 5·2). Assume VNTRseek detected an allele A with observed read support VRS. The distribution of expected read support can be estimated computationally. Random reads are simulated with the same characteristics of the real data, and the number of times they span a given array size of A is counted. This is repeated many times via bootstrapping, resulting in the expected read support distribution, SRS. Since A can be homozygous or heterozygous, this simulation is done twice: first assuming A is homozygous, $SRS^{(hom)}$, and again assuming A is heterozygous, $SRS^{(het)}$. Respectively, two Z-scores are calculated for VRS. The first calculation is done using the $ERS^{(hom)}$ distribution, hereon referred to as $Zscore(VRS, ERS^{(hom)})$. The second calculation uses the $ERS^{(het)}$ distribution, referred to as $Zscore(VRS, ERS^{(het)})$. The allele is given a label according to the distribution it was more likely to come from, i.e the distribution that resulted in the Z-score with the smallest absolute value. This is described below:

1. if $Zscore(VRS, ERS^{(het)}) < -3$ or $Zscore(VRS, ERS^{(hom)}) > +3$, the allele A is marked as ERROR;

2. if the $|Zscore(VRS, ERS^{(het)})| \leq |Zscore(VRS, ERS^{(hom)})|$, allele A is labelled heterozygous (HET);

3. otherwise, allele A is labelled homozygous (HOM).

This label classifies the alleles predicted by VNTRseek into homozygous and heterozygous categories. However, this prediction is done on alleles, and not the locus. To predict whether the locus is homozygous or heterozygous, another layer of machine learning was added that trains a model based on both TR and allele features. TR features include: the reference array length, the pattern length, the reference copy number, and whether gain/loss of one copy could be observed. Allele features include: the Z-score values and label for each allele, the normalized VRS, and whether the allele was an VNTR or not. The features are explained in detail in the following paragraphs. RapidMiner [1] was used to train the decision tree model.

Extracting and binning VRS: mlZ extracts the allele genotypes and their read support (VRS) from the VNTRseek output. Multis (TRs with more than two alleles on autosomal chromosomes or chromosome X of females) and TR alleles on the sex chromosomes of male individuals are excluded. The alleles are grouped according to their array lengths into bins of size 10 bp, starting from $2 \times flank$ to $read\_length -$

[1] Mierswa, Ingo, and Ralf Klinkenberg. "RapidMiner Studio." RapidMiner Account, 9.1.000 (rev: ef0090, platform OSX), RapidMiner, Inc., 12 Dec. 2018, rapidminer.com

$2 \times flank$, where flank is the minimum flank required by VNTRseek (by default set to 10 bp).

On the VRSs of each bin, two normal distributions are fitted using mixtools (Benaglia et al., 2009), one for the heterozygous and one for the homozygous alleles. To avoid errors due to different bin sizes and the nature of noisy data, i.e. VRS with extremely low or extremely high values, outliers (beyond three standard deviations) are not considered in the model fitting. The means and standard deviations of these two normal distributions for each bin are saved as the $ORS^{(het)}$ and $ORS^{(hom})$ distributions.

Estimating the $SRS^{(het)}$ and $SRS^{(hom)}$ for each bin: The expected distribution of the read support, SRS, is estimated by bootstrapping. First, the characteristics of the real data, i.e. read length, fragment mean and standard deviation, and the total number of reads (or coverage) are extracted from the bam file. Then, using these characteristics, random fragments are drawn from a normal distribution of $\mathcal{N}(\mu_{fragment}, \sigma_{fragment})$, with the DNA strand drawn from a Boolean distribution with p=0.5. For each fragment, two reads are set at the ends with a fixed length of read length. Random arrays of given length are simulated and the number of reads overlapping the arrays is counted for homozygous (on both strands) and heterozygous (on the first strand only). The array sizes are the middle of the VRS bins created in the previous step. This should theoretically be done for the same number of reads as exists in the real dataset. However, in order to speed up the runtime and reduce memory requirements, a speedup parameter is used that scales down the number of fragments and the genome size. This gives us the SRS distribution for the homozygous and heterozygous alleles for each bin of array lengths.

Inferring the expected read support from real data: Given that the read support and the array size correlate linearly, that is $read\_support \propto array\_size$, and the fact that the error across the bins from real data is not uniform, a linear model:

$$ORS[array\_size] \sim arraysize \tag{5.1}$$

is applied on the ORS estimates to smooth the variation. The final estimates are called the linear model read supports, $lmRS$. A plot is drawn to compare the SRS, ORS, and $lmRS$ on each data set, for both heterozygous and homozygous VRS. This plot can guide the user to detect incorrect parameter settings or unexpected errors.

Normalizing the observed read supports: The expected read support, ERS, decreases linearly with the array length. This relation can be formulated as:

$$ERS = C \times (1 - \frac{Array}{R - 2F}) \tag{5.2}$$

where Array is the observed array length, R is the read length, F is flank size, and C is the read coverage (Figure 5·2).

In order to make the ORS comparable across alleles with different array size, we need to normalize the VRS using the ERS. To normalize the read count we divide it by the expected read support.

$$normal\text{-}VRS = \frac{VRS}{ERS + 1} \tag{5.3}$$

The 1 is added to avoid division by zero. The $normal\text{-}VRS$ will be about 0.5 and 1.0 for heterozygous and homozygous calls, respectively. However, due to VNTRSeek errors and noise, some reads are expected to be unmapped, causing the $normal\text{-}VRS$ to be slightly less than ERS ($ORS \leq ERS$). In other words, the observed real support in real data is lower than the theoretical calculated values.

Computing Z-scores of heterozygous and homozygous lmRS distributions: For each allele genotyped, given the corresponding bin size and VRS, the Z-scores of the homozygous and heterozygous lmRS distributions are calculated. These Z-score values are called $lm\text{-}Z^{(het)}$ and $lm\text{-}Z^{(hom)}$. $lm\text{-}Z$ values are used to find the more likely class of the allele:

$$Z\text{-}label = \begin{cases} ERR, & lm\text{-}Z^{(het)} < -3 \text{ or } lm\text{-}Z^{(hom)} > +3 \\ HET, & |lm\text{-}Z^{(het)}| \leq |lm\text{-}Z^{(hom)}| \\ HOM, & \text{otherwise} \end{cases} \tag{5.4}$$

The Z-label improves the prediction of heterozygosity (discussed in results). However, *Z-label* provides information on the allele and not the locus, raising ambiguity when the Z-labels of two alleles from the same locus do not agree.

5.2.5 Learning the Decision Tree model

After estimating the read support distributions, a decision tree model is trained on simulated data to predict the heterozygous or homozygous class using the Z-label and Z-score measurements calculated above. This decision tree is calculated once and saved for future use on real data. Two sets of features are collected for the training dataset:

- *The reference TR attributes:* pattern size, array size, copy number, and the ability of detecting loss /gain of one copy

- *The allele features:* normalized read support (normal-VRS), lm-Z(het), lm-Z(hom), and the Z-label.

The alleles are separated into two datasets, one for homozygous calls and one for heterozygous calls. Each dataset will include the reference TR attributes. The homozygous dataset will have one set of allele features, since only one allele was detected. The heterozygous dataset will have two sets of allele features, one per detected allele, ordered by array size. Features of the homozygous dataset are given in Table 5.2 and those of the heterozygous dataset are given in Table 5.3. All machine learning procedures were done in RapidMiner Studio (Mierswa and Klinkenberg, 2018) Educational License edition.

Feature	Type	Distribution in data	
TR id	text	—	
Ability of detect gain of one copy	binomial	TRUE	91.7%
		FALSE	8.3%
Ability of detect loss of one copy	binomial	TRUE	87.5%
		FALSE	12.5%
lmZ(hom)	real	minimum	-8.8
		maximum	122.0
		mu	-1.0
		σ	2.0
lmZ(het)	real	minimum	-5.8
		maximum	182.9
		mu	3.1
		σ	2.9
normal-VRS	real	minimum	-49.7
		maximum	89.4
		mu	0.9
		σ	0.4
Z-label	categorical	HOM	66.9%
		HET	31.0%
		ERROR	2.1%

Table 5.2: Features for the homozygous dataset. A total of six features were selected for the homozygous dataset. Type is the type of data. Distribution is the distribution of that feature in the data.

Feature	Type	Distribution in data	
TRID	text	–	
Ability of detect gain of one copy	binomial	TRUE	99.0%
		FALSE	1.0%
Ability of detect loss of one copy	binomial	TRUE	82.4%
		FALSE	17.6%
lmZ(hom) of allele 1	real	minimum	-8.8
		maximum	115.6
		μ	-4.0
		σ	2.1
lmZ(hom) of allele 2	real	minimum	-8.6
		maximum	27.5
		μ	-2.7
		σ	1.9
lmZ(het) of allele 1	real	minimum	-5.8
		maximum	161.9
		μ	-0.7
		σ	2.3
lmZ(het) of allele 2	real	minimum	-5.7
		maximum	54.2
		μ	-0.6
		σ	2.4
normal-VRS of allele 1	real	minimum	0.0
		maximum	13.9
		μ	0.4
		σ	0.2
normal-VRS of allele 2	real	minimum	-7.0
		maximum	36.2
		μ	0.6
		σ	0.3
Z-label of allele 1	categorical	HET	81.1%
		ERROR	14.9%
		HOM	4.0%
Z-label of allele 2	categorical	HET	75.7%
		HOM	19.1%
		ERROR	5.3%

Table 5.3: Features for the heterozygous dataset. A total of ten features were selected for the heterozygous dataset. Type is the type of data. Distribution is the distribution of that feature in the data. Allele 1 is the allele with the smaller array length and allele 2 is the allele with the longer array length.

To optimize the decision tree parameters and find the best model, an optimization grid is applied. A grid runs through a set of parameters and tries every combination. Due to memory and time constraints, numeric features are discretized with a step size. The optimization grid for each model was:

- criterion $\in$ {gini_index, gain_ratio, information_gain, accuracy}
- minimal gain $\in$ range(min=1.0E-7, max=0.3, step=0.01)
- confidence $\in$ range(min=1.0E-7, max=0.3, step=0.01)
- minimal depth $\in$ range(min=5, max=15, step=1)

The optimization objective was to maximize the AUC measure. A ten-fold cross-validation was used to avoid over-fitting. For the homozygous model, the minimum split size was set to 200 and the minimum leaf size was set to 10. For the heterozygous model, the minimum split size was set to 20 and the minimum leaf size was set to 10, because there are less heterozygous loci. The decision tree was trained on a random split of 75% of the training data and tested on the remaining 25%.

To ensure that the models are not biased according to the dataset used in training, we tested each decision tree model on the other two simulation datasets. The homozygous models are called Hom1, Hom2, and Hom3, corresponding to the models trained on the homozygous genotypes of the Sim1, Sim2, and Sim3 datasets, respectively. Similarly, the heterozygous models are called Het1, Het2, and Het3, corresponding to the models trained on the heterozygous genotypes of the Sim1, Sim2, and Sim3 datasets, respectively.

After making sure models are representative of every dataset and there is no bias for sequencing features such as read length, fragment size, and coverage (i.e. the features never showed up in the models), all three simulated datasets were combined to obtain higher sample size. The final heterozygous and homozygous models were saved and implemented in the mlZ tool. The user can choose to use the Z-labels or the mlZ predictions.

5.3 Results

Table 5.4 summarizes the number of detectable reference TRs for each read length. In over 80% of the reference TRs a loss of one copy could not be detected, while loss of copy was shown to be the most common allele in humans in Section 2.3.4. As the read length increases, more gain can be observed; the number of reference TRs that gain of one copy could not be detected is about 32%, 16%, and 8% for read lengths 100 bp, 150 bp, and 250 bp, respectively. In the case of heterozygous VNTRs, there is a good probability that at least one allele will be detected. The performance of VNTRseek and mlZ on simulated data is presented in Section 5.3.1 and Section 5.3.2, respectively. The accuracy of the normalRS measurement is shown in Section 5.3.3. The Decision Tree performance is discussed in Section 5.3.4 and the best confidence cutoff is presented in Section 5.3.5. In Section 5.3.6 the result of mlZ on real data is presented. Section 5.3.7 shows the consistency of mlZ across sequencing platforms. Section 5.3.8 and Section 5.3.9 are dedicated to the erroneous predictions.

5.3.1 Performance of VNTRseek on simulated data

To measure the performance of VNTRseek on simulated data, we extracted the detectable genotypes on each simulated dataset (Table 5.1) and evaluated the precision and recall of each of the five classes: 0/0. 0/-1, 0/+1, -1/-1, and +1/+1. Table 5.5 shows the results of VNTRseek on the simulated datasets. In general, the number of TRs genotyped increases with read length and coverage, as previously shown in Section 2.3.2.

Reference set	**100 bp** (%)	**125 bp** (%)	**150 bp** (%)	**250 bp** (%)
Total singleton TRs	190,080	189,475	189,772	187,541
Loss of ≥1 copy cannot be detected	153,603 (81%)	153,504 (81%)	153,607 (81%)	153,207 (82%)
Gain of ≥1 copy cannot be detected	61,070 (32%)	40,306 (21%)	30,261 (16%)	13,916 (7%)
Union (either loss or gain of one copy cannot be detected)	173,599 (91%)	169,616 (90%)	167,315 (88%)	160,901 (86%)
Intersect (both loss or gain of one copy cannot be detected)	41,074 (22%)	24,194 (13%)	16,553 (9%)	6,222 (3%)

Table 5.4: Percentage of reference TRs not detectable by VNTRSeek. This table summarizes singletons TRs that cannot be detected by VNTRSeek when there is a loss of at least 1 copy (TRF will not find it) or gain of at least one copy (the TR $array + 2flank$ will become larger than the read length). The flank considered here is 10bp on each side. The numbers are presented for each reference set at 100 bp, 125 bp, 150 bp and 250 bp. The number of total TRs changes based on the read length due to the sliding window approach to detect indistinguishable TRs. In most loci, at least one allele could be detected.

VNTRSeek results	**Sim1**	**Sim2**	**Sim3**
Read length	250 bp	148 bp	100 bp
Read coverage	190X	95X	95X
No. of reference set TRs	223,674	226,510	226,782
No. of indistinguishables	36,133	36,738	36,702
No of read TRs	262,001,823	135,346,720	127,080,013
Total TRs found (singleton)	132,344	147,093	133,739
Total VNTRs	70,544	58,395	46,112
Homozygous TRs (singleton)	110,009	111,500)	101,274
Heterozygous TRs (singleton)	37,850	24,893	18,930

Table 5.5: VNTRSeek results on simulated data. VNTRseek was run with default parameters on the Sim1, Sim2, and Sim3 datasets. Read length and coverage are the simulation characteristics. At longer read lengths, more heterozygous TRs were genotyped.

In order to not penalize VNTRSeek on what it cannot find, we calculated the performance on the detectable genotypes, i.e. if the given genotype could be detected by VNTRseek (as discussed in Section 5.1). Table 5.3.1, Table 5.3.1, and Table 5.3.1 summarize the performance of VNTRseek in each dataset Sim1, Sim2, and Sim3, respectively. Overall the accuracy of VNTRseek to find the alleles it could detect was 84%, 80%, and 79% on Sim1 (250 bp), Sim2 (150 bp), and Sim3 (101 bp), respectively.

Sim1	Truth 0/0	Truth 0/1	Truth 0/-1	Truth -1/-1	Truth 1/1	Unde- tectable	Preci- sion
Pred. 0/0	84,140	1,452	1,817	60	210	321	96%
Pred. 0/1	135	26,113	7	0	9,037	2	74%
Pred. 0/-1	213	15	1,949	166	0	11	83%
Pred. -1/-1	22	2	173	2,202	2	92	88%
Pred. 1/1	25	1,761	0	1	17,458	2	91%
Pred. other	75	111	30	39	128	260	0%
Not detected	5,370	524	396	2,107	1,014	34,911	79%
Total	89,980	29,978	4,372	4,575	27,849	35,599	
Recall	94%	87%	45%	48%	63%	98%	
Recall in detected	99%	89%	49%	89.22%	65%	NA	

Table 5.6: Confusion matrix of VNTRSeek genotypes on the Sim1 dataset. Homozygous and heterozygous gain/losses of one copy were simulated for testing purposes. Expected genotype is what VNTRseek would have observed according to its detectable range. VNTRseek cannot detect alleles with array size larger than the read length or alleles with copy number less than two. Read length for the Sim1 dataset was 250 bp and physical coverage was 100X. The confusion matrix of the genotypes predicted by VNTRseek. undetectable: implanted TR alleles that could be detected by VNTRseek. Recall in found: the recall of that class among all the genotyped TRs. Note that VNTRseek does not find 79% of the TRs. Overall accuracy was 84.12% and accuracy of detected genotypes was 89.49%.

Sim2	Truth 0/0	Truth 0/1	Truth 0/-1	Truth -1/-1	Truth 1/1	Undetectable	Precision
Pred. 0/0	72,278	6,295	1,767	70	226	154	89%
Pred. 0/1	40	17,367	3	1	643	0	96%
Pred. 0/-1	41	14	758	12	2	0	92%
Pred. -1/-1	22	8	391	1,068	15	37	69%
Pred. 1/1	20	1,489	0	0	17,393	1	92%
Pred. other	18	33	6	4	35	7	0%
Not found	17,380	4,685	1,403	3,358	9,477	35,263	49%
Total	89,799	29,891	4,328	4,513	27,791	35,462	
Recall	80%	58%	18%	24%	63%	99%	
Recall in found	100%	69%	26%	92%	95%	NA	

Table 5.7: Confusion matrix of VNTRSeek genotypes on the Sim1 dataset. Homozygous and heterozygous gain/losses of one copy were simulated for testing purposes. Expected genotype is what VNTRseek would have observed according to its detectable range. VNTRseek cannot detect alleles with array size larger than the read length or alleles with copy number less than two. Read length for the Sim1 dataset was 250 bp and physical coverage was 100X. The confusion matrix of the genotypes predicted by VNTRseek. undetectable: implanted TR alleles that could be detected by VNTRseek. Recall in found: the recall of that class among all the genotyped TRs. Note that VNTRseek does not find 79% of the TRs. Overall accuracy was 84.12% and accuracy of detected genotypes was 89.49%.

Sim3	Truth 0/0	Truth 0/1	Truth 0/-1	Truth -1/-1	Truth 1/1	Undetectable	Precision
Pred. 0/0	79,390	4,298	1,903	89	300	185	92%
Pred. 0/1	101	21,830	4	0	1,298	0	94%
Pred. 0/-1	236	16	1,244	30	1	0	81%
Pred. -1/-1	26	11	405	1,532	8	40	76%
Pred. 1/1	28	1,850	0	1	21,350	2	92%
Pred. other	60	55	19	15	70	36	0%
Not found	9,775	1,865	772	2,871	4,786	35,246	64%
Total	89,616	29,925	4,347	4,538	27,813	35,509	
Recall	89%	73%	29%	34%	77%	99%	
Recall in found	99%	78%	35%	92%	93%	NA	

Table 5.8: Confusion matrix of VNTRSeek genotypes on the Sim3 dataset. Homozygous and heterozygous gain/losses of one copy were simulated for testing purposes. Expected genotype is what VNTRseek would have observed according to its detectable range. VNTRseek cannot detect alleles with array size larger than the read length or alleles with copy number less than two. Read length was 101 bp and physical coverage was 100X. The confusion matrix of the genotypes predicted by VNTRseek. undetectable: implanted TR alleles that could be detected by VNTRseek. Recall/Accuracy in found: the recall/accuracy of that class/VNTRseek among all the genotyped TRs. Note that VNTRseek does not find 79% of the TRs. Overall, while the per class precision is high, the recall of losses is lower. The overall accuracy was 69.64% and accuracy of detected genotypes was 90.71%.

While VNTRseek can find reference alleles with >95% precision, the -1/-1 and 0/-1 genotypes have recall <50%. When a VNTR has a copy loss such that the copy number falls less than 1.8 (1.9 for patterns smaller than 50 bp), TRF will not be able to discover it and the VNTR allele will not be genotyped by VNTRSeek. In another scenario, when the VNTR has gained too many copies such that the array length becomes larger than the read length -2flanks, no read can span it, and thus, that allele cannot be detected by VNTRseek. In general, about 80% of the reference TRs have a copy number ≤ 2.8, meaning a loss of one copy or more cannot be discovered in them. The number of TRs where a gain of at least one copy cannot be detected depends on the read length. Table 5.9 shows the precision of VNTRseek for the predicted homozygous genotypes. On simulated datasets Sim1, Sim2, and Sim3, VNTRseek homozygous predictions had precision 73%, 72%, 69%, respectively. On the predicted heterozygous genotypes, the precision was 74%, 93%, and 96% for Sim1, Sim2 and Sim3, respectively (Table 5.10).

Homozygous	Sim1	Sim2	Sim3
True HOM	76,173	76,507	66,973
True HET	28,294	29,618	29,608
Total	104,467	106,125	96,581
VNTRseek accuracy	72.92%	72.09%	69.34%

Table 5.9: VNTRseek precision on simulated data for the predicted homozygous genotypes. The precision of VNTRseek was measured for homozygous genotypes in the detectable range for the simulated datasets Sim1, Sim2, and Sim3. Note that most homozygous calls are reference genotypes. The shorter reads in the Sim3 dataset miss many heterozygous alleles, causing false positive homozygous calls.

Heterozygous	Sim1	Sim2	Sim3
True HOM	9,616	1,698	741
True HET	27,867	23,138	18,164
Total	37,482	24,835	18,904
Precision	74.35%	93.17%	96.09%

Table 5.10: VNTRseek precision on simulated data for the predicted heterozygous genotypes. The precision of VNTRseek was measured for heterozygous genotypes in the detectable range for the simulated datasets Sim1, Sim2, and Sim3. VNTRseek has high precision on heterozygous class. The Sim1 dataset, which has the longest read length, causes more Multis (false positive heterozygous calls).

5.3.2 Evaluating the accuracy of normal-VRS

To show the correctness of our assumptions about the heterozygous and homozygous read supports, especially the final normal-VRS, after running VNTRseek on the simulated data, the VRS values were extracted. The distribution of normal-VRS of each allele was plotted (Figure 5·3 A). As it can be seen, we have three Poisson (binomial) distributions: one centered slightly less than 0.5, corresponding to the heterozygous alleles, another centered slightly under 1.0, corresponding to homozygous calls, and another at 0, indicating erroneous calls. We separated heterozygous and homozygous genotypes (as reported by VNTRseek) and plotted the VRS again, this time filtering outliers (beyond 3 standard deviations). The heterozygous normal-VRS values peaked at 0.5 (Figure 5·3 B) and the homozygous normal-VRS values peaked at 1.0 (Figure 5·3 C). As it can be observed, normal VRS is a good feature to predict heterozygous alleles. In Figure 5·3 C, a small bump around 0.5 is observed. These are the genotypes that are missing an allele.

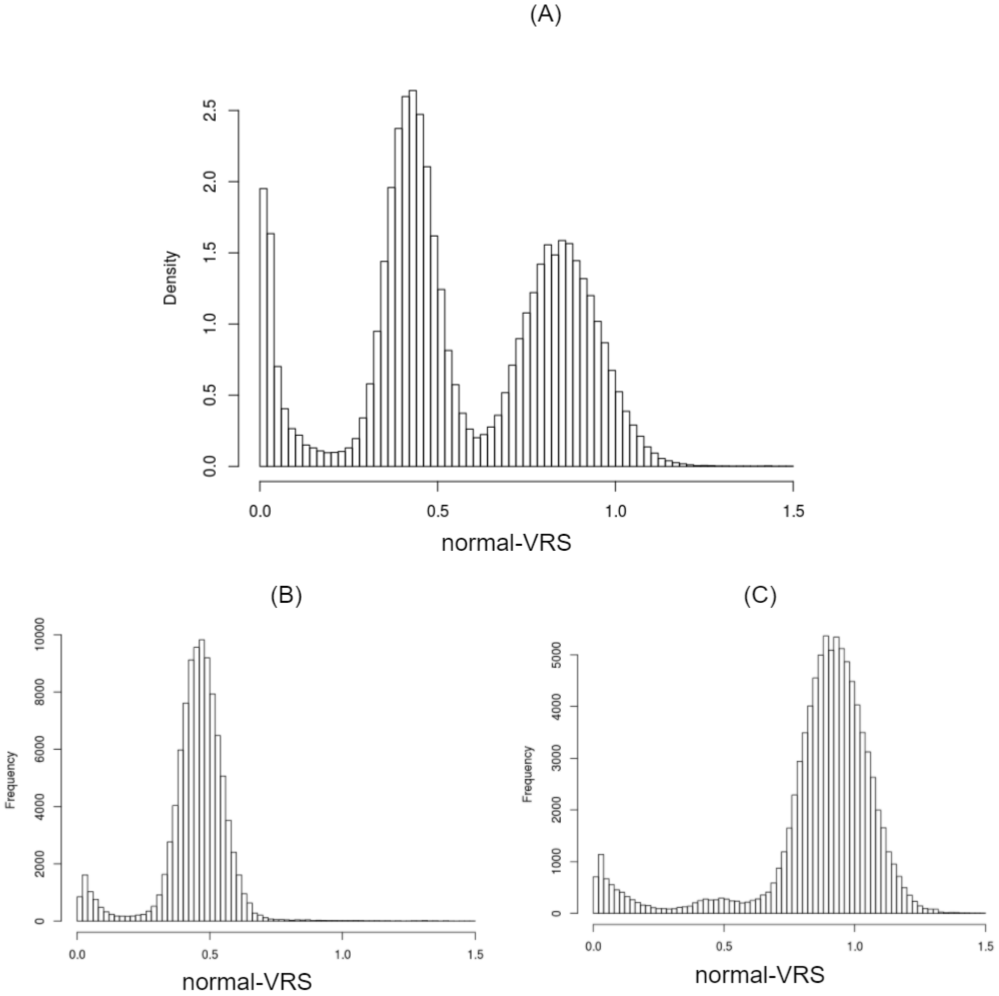


Figure 5·3: normal-VRS distributions for alleles. On the simulated dataset Sim1, Sim2, and Sim3, the normal VRS was plotted for all the alleles (A), the alleles detected as by VNTRseek (B) and alleles detected as homozygous by VNTRseek (C). Alleles with normal-VRS ¿ 1.5 have been excluded because there were very few and caused a very long tail. The homozygous alleles have mean about 1.0 and the heterozygous alleles have mean about 0.5. The homozygous alleles have a small peak at 0.5, suggesting in some cases the true genotype was heterozygous. The alleles that peak around 0 are the Multis.

5.3.3 Robustness of the decision tree across sequencing platforms

The homozygous model: Homozygous VNTRseek genotype prediction on the simulated datasets was used to train the homozygous datasets. To ensure the accuracy of the mlZ decision trees, we test a tree model that was trained on each simulated dataset and tested on the other two independent datasets. The models and performances are compared to ensure the decisions trees are not biased by dataset characteristics such as read length, fragment size, and coverage.

Table 5.11 shows the learning performance (five-fold cross-validation) of the Hom1 model that was trained on the Sim1 dataset, and Table 5.12 and Table 5.13 show the results of the Hom1 model on the Sim2 and Sim3 datasets, respectively. Similarly, Table 5.14 presents the learning performance of the Hom2 model on the cross-validations, and Table 5.15 and Table 5.16 present the testing performance of the Hom2 model on the Sim1 and Sim3 datasets; and Table 5.17 is the learning performance on the Hom3 model, and Table 5.18 and Table 5.19 are the testing performance of the Hom3 model on the Sim1 and Sim2 datasets. While all the models had similar performance, the Hom3 model that was trained on the Sim3 dataset has the lowest performance because the Sim3 dataset had less training data (lowest read length results in less TRs being genotyped). The Hom1, Hom2, and Hom3 models had 96%, 95%, and 93% accuracy, respectively. They had similar performance across datasets. The recall on the heterozygous class was the lowest, but the precision was high. All models performed with >90% precision on homozygous and heterozygous classes, except for the heterozygous recall of the Hom1 model on Sim2 and Sim3 datasets which was <80%. The possible explanation is that the Hom1 model was trained on the Sim1 dataset that was simulated with the MiSeq Illumina error profile with higher error rate including indels. In general, the three models, Hom1, Hom2, and Hom3, were comparable across simulation datasets.

Learning Hom1	True HOM	True HET	Class Precision
Pred. HOM	73,382	1,518	97.97%
Pred. HET	2,791	26,776	90.56%
class recall	96.34%	94.63%	

Table 5.11: Learning performance of Hom1 model on Sim1 (N=127,449). The Hom1 model was trained on homozygous calls of the Sim1 dataset (read length 250 bp) with stratified 5× cross-validation. The optimal parameters used were: Information Gain criterion, minimal gain and confidence set to minimum (1.0E-7) and maximum depth to 9. The accuracy was 95.88%±0.15 (micro average: 95.88%).

Hom1 on Sim2	True HOM	True HET	Class Precision
Pred. HOM	87,317	7,842	91.76%
Pred. HET	2,084	27,422	92.94%
class recall	97.67%	77.76%	

Table 5.12: Testing performance of Hom1 on Sim2. The Hom1 model was tested on the Sim2 dataset (read length 148 bp).The overall accuracy was 92%. The heterozygous (HET) recall was 78%.

Hom1 on Sim3	True HOM	True HET	Class Precision
Pred. HOM	77,156	13,798	84.83%
Pred. HET	1,699	21,613	92.71%
class recall	97.85%	61.03%	

Table 5.13: Testing performance of Hom1 on Sim3. The Hom1 model was tested on the Sim3 dataset (read length 101 bp). The overall accuracy was 86%. The heterozygous (HET) recall was 61%.

Learning Hom2	True HOM	True HET	Class Precision
Pred. HOM	72,699	2,050	97.26%
Pred. HET	3,808	27,568	87.86%
class recall	95.02%	93.08%	

Table 5.14: Learning performance of the Hom2 model on Sim2 (N=124,665). The Hom2 model was trained on homozygous calls of the Sim2 dataset (read length 148 bp) with stratified 5× cross-validation. The optimal parameters used were: information gain criterion, minimal gain and confidence set to minimum (1.0E-7) and maximum depth to 9. The accuracy was accuracy: 94.48%±0.17 (micro average: 94.48%).

Hom2 on Sim1	True HOM	True HET	Class Precision
Pred. HOM	83,456	1,409	98.34%
Pred. HET	4,963	31,749	86.48%
class recall	94.39%	95.75%	

Table 5.15: Testing performance of Hom2 on Sim1. The Hom2 model was tested on the Sim1 dataset (read length 250 bp). The overall accuracy was 95%. The heterozygous (HET) recall was 96%.

Hom2 on Sim3	True HOM	True HET	Class Precision
Pred. HOM	74,024	4,852	93.85%
Pred. HET	4,831	30,559	86.35%
class recall	93.87%	86.30%	

Table 5.16: Testing performance of Hom2 on Sim3. The Hom2 model was tested on the Sim3 dataset (read length 101 bp.) The overall accuracy was 92%. The heterozygous (HET) recall was 86%.

Learning on Hom3	True HOM	True HET	Class Precision
Pred. HOM	63,064	3,040	95.40%
Pred. HET	3,909	26,568	87.17%
class recall	94.16%	89.73%	

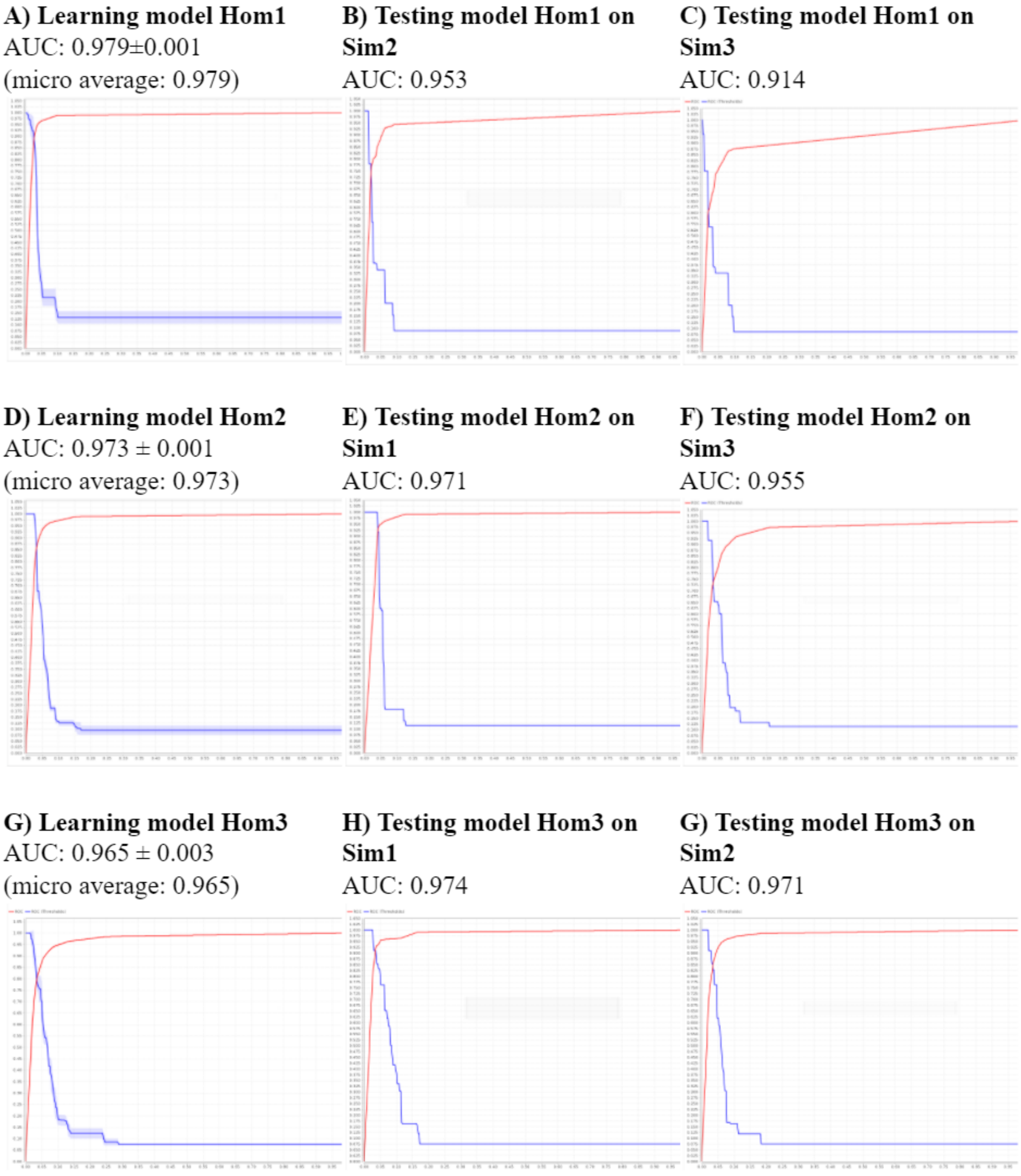
Table 5.17: Learning performance of the Hom3 model on Sim3 (N=114,267). The Hom3 model was trained on homozygous calls of the Sim3 dataset (read length 101 bp) with stratified 5× cross-validation. The optimal parameters used were: information gain criterion, minimal gain and confidence set to minimum (1.0E-7) and maximum depth to 9. accuracy: 92.81%±0.07% (micro average: 92.81%).

Hom3 on Sim1	True HOM	True HET	Class Precision
Pred. HOM	81,195	1,259	98.47%
Pred. HET	7,224	31,899	81.54%
class recall	91.83%	96.20%	

Table 5.18: Testing performance of Hom3 on Sim1. The Hom3 model was tested on the Sim1 dataset (read length 250 bp.) The overall accuracy was 93%. The heterozygous (HET) recall was 96%.

Hom3 on Sim2	True HOM	True HET	Class Precision
Pred. HOM	84,109	2,060	97.61%
Pred. HET	5,292	33,204	86.25%
class recall	94.08%	94.16%	

Table 5.19: Testing performance of Hom3 on Sim2. The Hom3 model was tested on the Sim2 dataset (read length 148 bp.) The overall accuracy was 94%. The heterozygous (HET) recall was 94%.



A) Learning model Hom1
AUC: 0.979±0.001
(micro average: 0.979)

B) Testing model Hom1 on Sim2
AUC: 0.953

C) Testing model Hom1 on Sim3
AUC: 0.914

D) Learning model Hom2
AUC: 0.973 ± 0.001
(micro average: 0.973)

E) Testing model Hom2 on Sim1
AUC: 0.971

F) Testing model Hom2 on Sim3
AUC: 0.955

G) Learning model Hom3
AUC: 0.965 ± 0.003
(micro average: 0.965)

H) Testing model Hom3 on Sim1
AUC: 0.974

G) Testing model Hom3 on Sim2
AUC: 0.971

Figure 5·4: AUC of the homozygous models: Hom1, Hom2, and Hom3. Positive class is HET

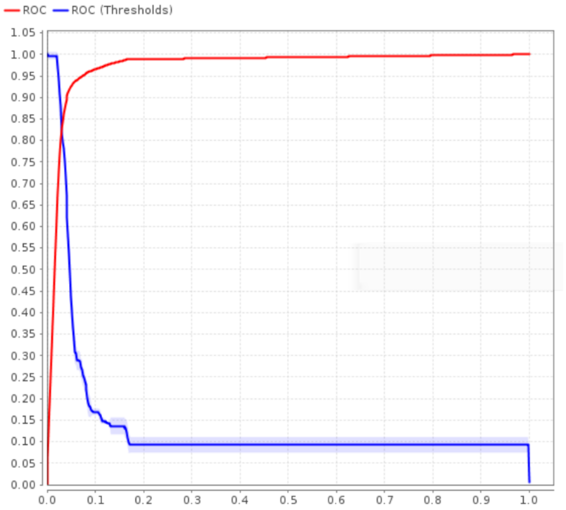


Figure 5·5: AUC of the final homozygous model. The ROC curve for the final homozygous model is shown. The AUC was 0.973±0.001 (micro average: 0.973) on ten-fold cross-validation.

Figure 5·4 shows all AUCs corresponding to the testing done above. Note that learning AUC is calculated on the cross-validations and shows the 95% confidence interval as shadows (Figure 5·4 A for the Hom1 model, Figure 5·4 D for the Hom2 model, and Figure 5·4 G for the Hom3 model). The AUC on the heterozygous class was >95% in all models, except for the Hom1 model on the Sim3 dataset.

After ensuring the decision trees were robust across simulation datasets, the simulated data was combined and the final homozygous tree was trained and incorporated into mlZ. The final homozygous model had 94.31%±0.15 accuracy and F-measure 90.20%±0.24 on the ten-fold cross-validation (Table 5.20). The AUC was 97.3%±0.001 (Figure 5·5).

In addition I compared the decision tree to a Naive Bayes model. The decision tree performed slightly better than the Naive Bayes model (Figure 5·6). Random Forest did not improve the results (not shown), so we decided to use the decision tree for simplicity. In the final tree (not shown), the first depth node was the Z-label and the second depth node was normal-VRS. This shows the importance of these features.

Final homozygous model	True HOM	True HET	Class Precision
Pred. HOM	209,378	7,189	96.68%
Pred. HET	10,275	80,331	88.66%
Class recall	95.32%	91.79%	

Table 5.20: Learning performance of the final homozygous model (N=366,380). The final homozygous model was trained on the homozygous calls of the Sim1, Sim2, and Sim3 datasets combined. The optimal parameters used were: information gain criterion, minimal gain and confidence set to minimum (1.0E-7) and maximum depth to 9 with ten-fold cross-validations. Overall accuracy was accuracy: 94.31%±0.15 (micro average: 94.31%) and the F1 measure was 90.20%±0.24 (micro average: 90.20).

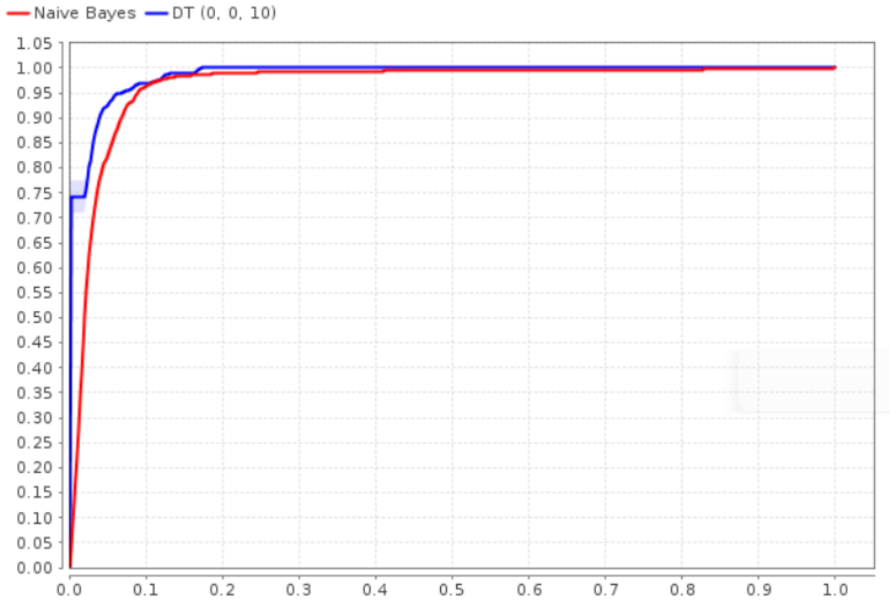


Figure 5·6: Comparison of the ROC curves of the final homozygous model and the candidate model. Naive bayes is the default. Random tree model was made with the same parameters of DT. ROC curves of all models drawn with 75%/25% split ratio and 10 folds validation. (positive class: HET). The 95% confidence interval on the ten-fold cross-validation is presented as shadows.

The heterozygous model: The same testing and training was performed on the heterozygous genotypes as the homozygous genotypes. Three tree models, named Het1, Het2, and Het3, were trained on 75% of the Sim1, Sim2, and Sim3 datasets with five-fold cross-validation and tested on the corresponding remaining 25%, respectively to get the learning performance. The three models were then tested on the other two simulated datasets. Table 5.21, Table 5.24, and Table 5.27 show the learning performance of the Het1, Het2, and Het3 models, respectively. The learning accuracy was >98% in all three models. Table 5.22 and Table 5.23 show the result of the Het1 model on the Sim2 and Sim3 datasets, respectively. Table 5.25 and Table 5.26 show the performance of the Het2 model on the Sim1 and Sim3 datasets, respectively. And Table 5.28 and Table 5.29 show the performance of the Het3 model on the Sim1 and Sim2 datasets. The AUC for the learning models and tests are shown in Figure 5·7. While all models had accuracy >95%, the results did not significantly improve beyond the overall performance of VNTRseek. Table 5.5 shows that VNTRSeek accuracy on heterozygous calls is around 97%. As a result there are very few TRUE homozygous calls in the training data. This results in a big bias in the training data, and reduces statistical power for the heterozygous model. This problem can be solved by simulating more and more data.

Learning Het1	True HOM	True HET	Class Precision
Pred. HOM	8,965	280	97.69%
Pred. HET	651	27,587	96.97%
class recall	93.23%	99.00%	

Table 5.21: Learning performance of the Het1 model on Sim1 (N=366,380). Model Het1 was trained on the heterozygous calls of the Sim1 dataset (read length 250 bp) with stratified 5× cross-validation. The optimal parameters used were: information gain criterion, minimal gain and confidence set to minimum (1.0E-7) and maximum depth to 7. F1 measure was 95.06%±0.25 (micro average: 95.06%) (positive class: HOM).

Het1 on Sim2	True HOM	True HET	Class Precision
Pred. HOM	1,747	197	89.87%
Pred. HET	563	26,421	97.91%
class recall	75.63%	99.26%	

Table 5.22: Testing performance of Het1 on Sim2. The Het1 model was tested on the heterozygous calls of the Sim2 dataset (read length 148 bp). The overall accuracy was 97.37%.

Het1 on Sim3	True HOM	True HET	Class Precision
Pred. HOM	725	300	70.73%
Pred. HET	305	20,780	98.55%
class recall	70.39%	98.58%	

Table 5.23: Testing performance of Het1 on Sim3. The Het1 model was tested on the heterozygous calls of the Sim3 dataset (read length 101 bp). The overall accuracy was 97.26%.

Learning Het2	True HOM	True HET	Class Precision
Pred. HOM	1,307	109	98.33%
Pred. HET	391	23,029	92.30%
class recall	76.97%	99.53%	

Table 5.24: Learning performance of the Het2 model on Sim2 (N=24,836). The Het2 model was trained on heterozygous calls of Sim2 dataset (read length 148 bp) with stratified 5× cross-validation. The optimal parameters used were: information gain criterion, minimal gain and confidence set to minimum (1.0E-7) and maximum depth set to 8. Overall accuracy was 97.99%±0.23 (micro average: 97.99%) and F1 measure was 83.92%±1.99 (micro average: 83.94%). (positive class: HOM)

Het2 on Sim1	True HOM	True HET	Class Precision
Pred. HOM	11,169	536	95.42%
Pred. HET	871	31,839	97.34%
class recall	92.77%	98.34%	

Table 5.25: Testing performance of Het2 on Sim1. The Het2 model was tested on the heterozygous calls of the Sim1 dataset (read length 250 bp). The overall accuracy was 96.83%.

Het2 on Sim3	True HOM	True HET	Class Precision
Pred. HOM	788	219	78.85%
Pred. HET	242	20861	98.85%
class recall	76.50%	98.96%	

Table 5.26: Testing performance of Het2 on Sim3. The Het2 model was tested on the heterozygous calls of the Sim3 dataset (read length 101 bp). The overall accuracy was 97.91%.

Learning Het3	True HOM	True HET	Class Precision
Pred. HOM	575	67	89.56%
Pred. HET	166	18,097	99.09%
class recall	77.60%	99.63%	

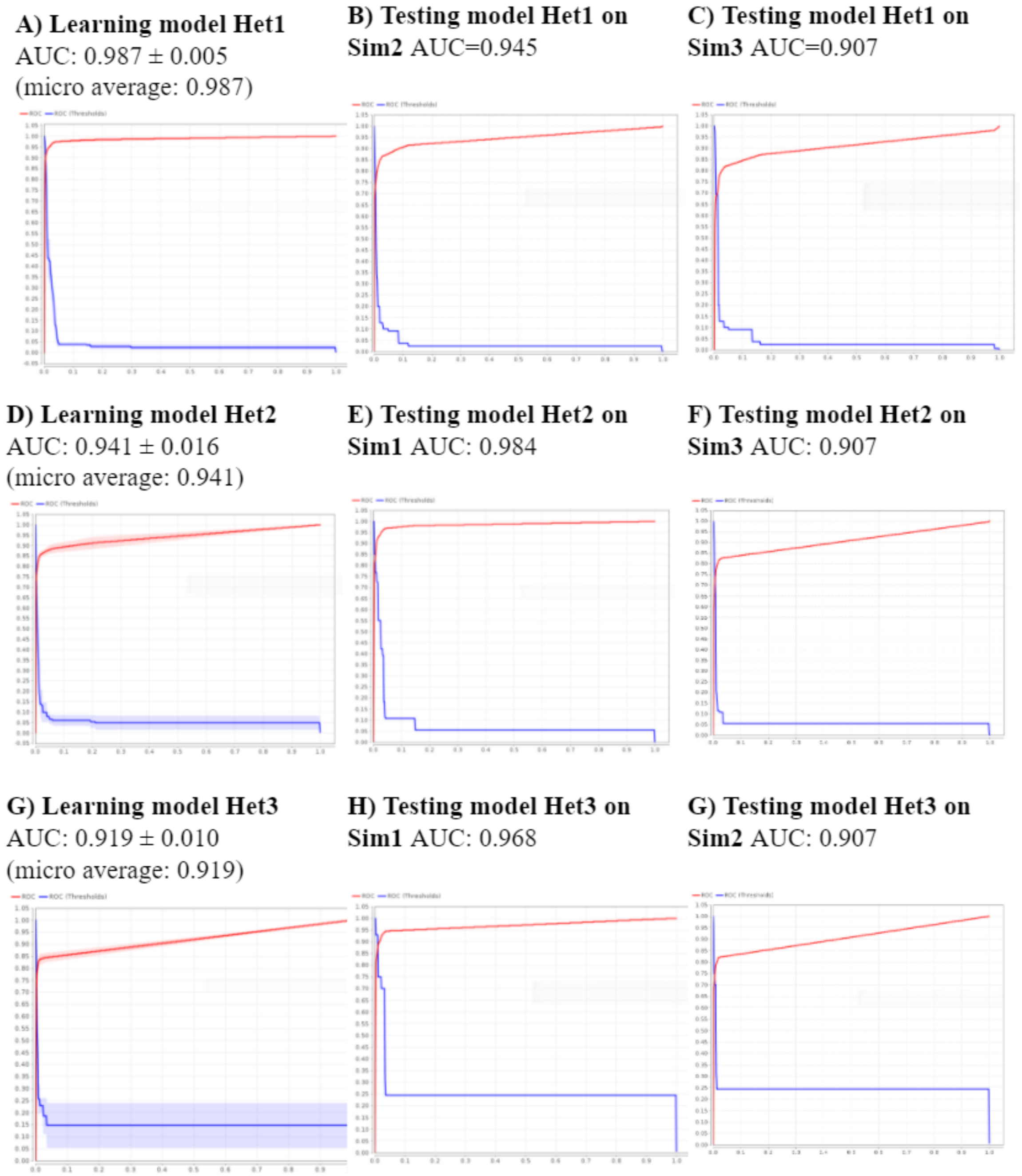
Table 5.27: Learning performance of the Het3 model on Sim3 (N=18,905). The Het3 model was trained on the heterozygous calls of the Sim3 dataset (read length 101 bp) with stratified 5× cross-validation. The optimal parameters used were: information gain criterion, minimal gain and confidence set to minimum (1.0E-7) and maximum depth set to 5. Overall accuracy was 98.77%±0.08 (micro average: 98.77%) and F1 measure was 83.13%±1.33 (micro average: 83.15%). (positive class: HOM)

Het3 on Sim1	True HOM	True HET	Class Precision
Pred. HOM	11,127	700	94.08%
Pred. HET	913	31,675	97.20%
class recall	92.42%	97.84%	

Table 5.28: Testing performance of Het3 on Sim1. The Het3 model was tested on the heterozygous calls of the Sim1 dataset (read length 250 bp). The overall accuracy was 96.37%.

Het3 on Sim2	True HOM	True HET	Class Precision
Pred. HOM	1,753	193	90.08%
Pred. HET	557	26,425	97.94%
class recall	75.89%	99.27%	

Table 5.29: Testing performance of Het3 on Sim2. The Het3 model was tested on the heterozygous calls of the Sim2 dataset (read length 148 bp). The overall accuracy was 97.41%.



A) Learning model Het1 AUC: 0.987 ± 0.005 (micro average: 0.987)

B) Testing model Het1 on Sim2 AUC=0.945

C) Testing model Het1 on Sim3 AUC=0.907

D) Learning model Het2 AUC: 0.941 ± 0.016 (micro average: 0.941)

E) Testing model Het2 on Sim1 AUC: 0.984

F) Testing model Het2 on Sim3 AUC: 0.907

G) Learning model Het3 AUC: 0.919 ± 0.010 (micro average: 0.919)

H) Testing model Het3 on Sim1 AUC: 0.968

G) Testing model Het3 on Sim2 AUC: 0.907

Figure 5·7: AUC of the heterozygous models: Het1, Het2, and Het3. Positive class is HOM

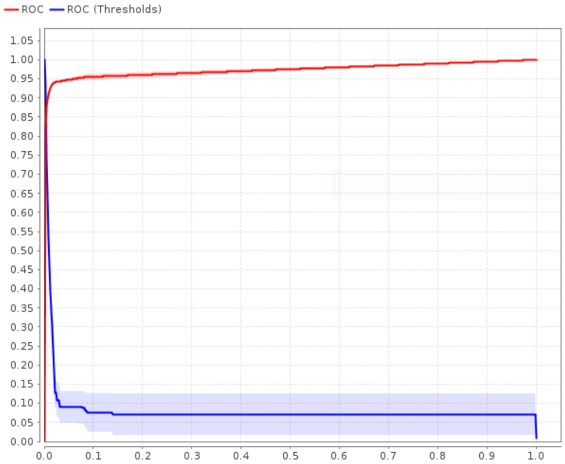


Figure 5·8: AUC of the final heterozygous model. The ROC curve for the final heterozygous model is shown. The AUC was 0.973±0.004 (micro average: 0.973) on ten-fold cross-validation.

Ensuring the decision trees for the Het1, Het2, and Het3 models were robust across simulated datasets, the three datasets were combined and a final heterozygous model was trained. The overall learning accuracy on ten-fold cross-validation was 98.00%±0.15 and the F-measure was 90.20%±0.24. The homozygous precision was 95.63% and the heterozygous precision was 98.39% (Table 5.30). The AUC was 0.973±0.004 (Figure 5·8).

Final heterozygous model	True HOM	True HET	Class Precision
Pred. HOM	10,928	499	95.63%
Pred. HET	1,127	68,670	98.39%
class recall	90.65%	99.28%	

Table 5.30: Learning performance of the final homozygous model (N=81,224). The final homozygous model was trained on the heterozygous calls of the Sim1, Sim2, and Sim3 datasets combined. The optimal parameters used were: information gain criterion, minimal gain and confidence set to minimum (1.0E-7) and maximum depth to 10 with ten-fold cross-validations. Overall accuracy was 98.00%±0.15 (micro average: 98.00%) and the F1 measure was 90.20%±0.24 (micro average: 90.20%).

In the final tree, the first and second node is the normal read count for the first and second allele followed by the Z-label. Figure 5·9 compares the final heterozygous tree to the Naive Bayes model which performed similarly.

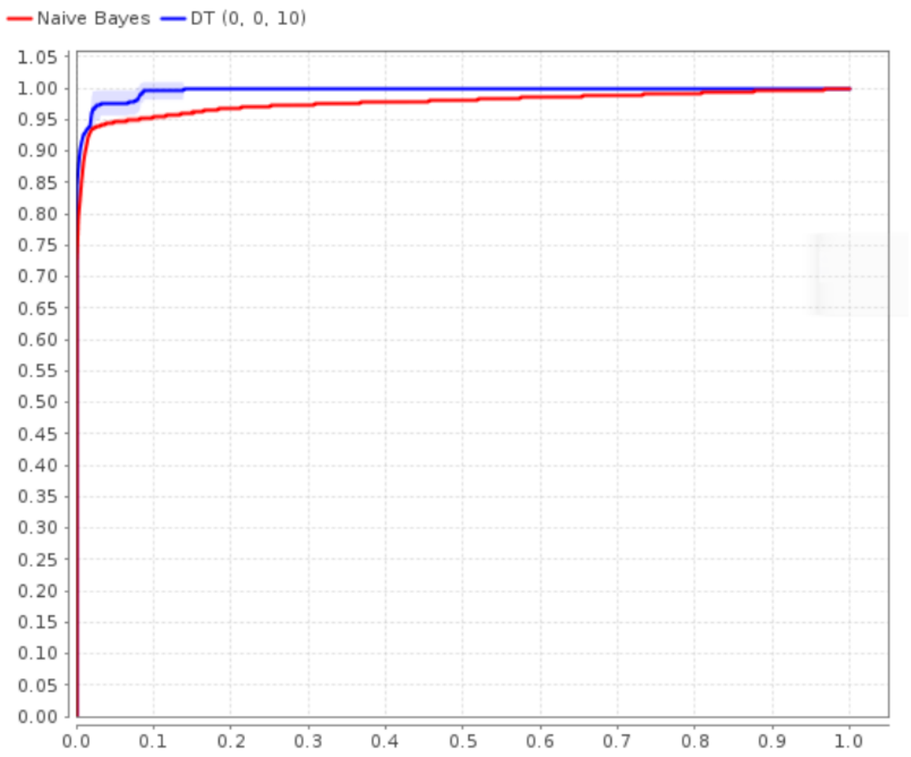


Figure 5·9: Comparison of the ROC curves of the final heterozygous model and the candidate model. Naive Bayes is the default. Random tree model was made with the same parameters of DT. ROC curves of all models drawn with 75% / 25% split ratio and ten-fold cross-validation.

Overall performance of the decision tree on simulated data: The overall accuracy of mlZ on simulated data was >95%. On the homozygous and heterozygous the precision was 96% and 93% and the recall was 95% and 94%, respectively (Table 5.31).

Final mlZ performance	True HOM	True HET	Class Precision
Pred. HOM	19,228+324	715+37	96.30%
Pred. HET	839+103	7,093+5,759	93.17%
Class recall	95.40%	94.47%	

Table 5.31: Final mlZ performance (heterozygous and homozygous model combined). The predictions from the final heterozygous model and the final homozygous model were combined and the confusion table was calculated. Overall accuracy was 95.03%.

5.3.4 Comparison of VNTRSeek , Z-label, and mlZ

The predictions of mlZ on VNTRseek can be presented with the annotation "VNTRseek class"→"mlZ class". HOM→HOM is used when a locus was predicted homozygous by both VNTRseek and mlZ, HET→HET is used when both VNTRseek and mlZ predicted heterzoygous, HOM→HET is used when VNTRseek predicted a homozygous locus that mlZ predicts should have been heterozygous, and HET→HOM is when VNTRseek predicted heterozygous and mlZ predicts is was homozygous. We are particularly interested in the HOM→HET class where more VNTRs can be inferred.

Comparison of performances by VNTRseek, Z-label, and mlZ are given in Table 5.32. The overall accuracy of VNTRseek was 79%, which by using the Z-labels alone, improved to 90%. mlZ could further improve this accuracy to 98%. The precision on homozygous class improved from 72% to 94% using Z-labels and 98% by mlZ. This is because mlZ can infer many missing alleles that would make a homozygous prediction heterozygous (HOM→HET).

5.3.5 The confidence threshold

The decision tree model used in mlZ reports a confidence value that can be used to filter the predictions to a more reliable subset. Figure 5·10, Figure 5·11, and Fig-

Comparison	True homozygous		True heterozygous		All
Method	Precision	Recall	Precision	Recall	Accuracy
VNTRSeek	71.99%	95.91%	93.16%	59.75%	78.52%
Z-label	94.29%	86.42%	86.56%	94.34%	90.23%
mlZ	97.63%	97.25%	89.90%	91.21%	95.97%

Table 5.32: Comparison of VNTRseek, Z-label, and mlZ. Alleles were extracted. If the allele was from a homozygous VNTRseek prediction, it was labelled as homozygous and otherwise, it was labelled as heterozygous. The Z-label can be homozygous, heterozygous, or error. The errors were excluded in the precision calculation. mlZ predicts whether the allele was heterozygous or homozygous. The precisions and overall accuracy were compared on simulated data. mlZ does not improve the precision of VNTRseek on heterozygous alleles, because VNTRseek can find 2,000–3,000 heterozygous calls per genome, which is not enough to train the model.

ure 5·12 show the effect of increasing the confidence threshold on performance on the Sim1, Sim2, and Sim3 datasets, respectively. At a threshold of 90% confidence, the performance improves significantly, while losing <10% of the predictions. Table 5.33 shows that, on the Sim1 dataset, a confidence threshold of 90% improves the overall accuracy from 96% to 98% and improves the heterozygous recall from 97% to 99%, while maintaining >90% of the predictions. Table 5.34 shows that >86% of the predictions on the Sim2 dataset had confidence >90%, and that the accuracy improved from 95% to 98% and the heterozygous recall improved from 95% to 99%. Table 5.35 shows that >77% of the predictions in the Sim3 dataset have confidence >90%, the accuracy improved from 93% to 98%, and the heterozygous confidence improved from 92% to 98%. Using a threshold of 90% confidence, overcomes the lower recall of mlZ on heterozygous compared to using Z-labels alone. Using this threshold when applying mlZ is recommended.

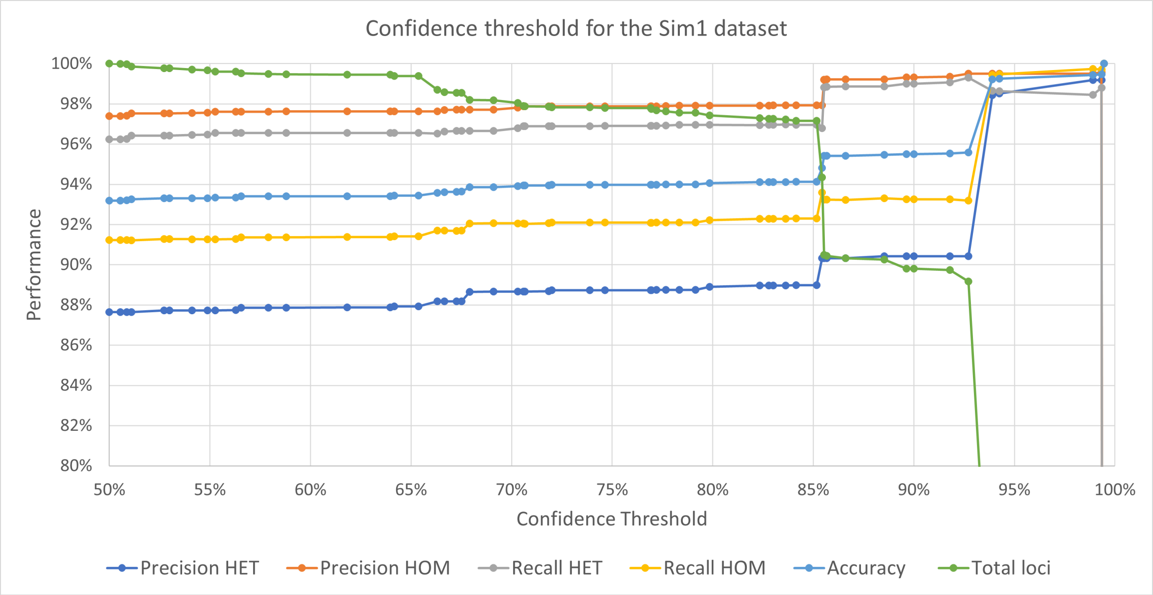


Figure 5·10: The confidence threshold effect on the Sim1 dataset performance. The confidence reported by the decision tree can be used to filter uncertain predictions. As the confidence threshold increases, the performance improves. About 90% of the test samples had confidence >90%. Note that the Y-axis starts from 80 to 100%.

Sim1		True HET	True HOM	Precision
Conf.≥0.9 Accuracy=98%	Pred. HET	58,890	3,091	95%
	Pred. HOM	540	87,764	99%
	Recall	99%	97%	
Conf.<0.9 Accuracy=76%	Pred. HET	4,552	2,146	68%
	Pred. HOM	1,551	7,458	83%
	Recall	75%	78%	
Combined Accuracy=96%	Pred. HET	63,442	5,237	92%
	Pred. HOM	2,091	95,222	98%
	Recall	97%	95%	
VNTRSeek	Not found	12,129	50,365	
	Percent	16%	33%	

	Conf.≥0.9	Conf.<0.9	Total	Percent
HET	61,981	6,698	68,679	90%
HOM	88,304	9,009	97,313	91%
Total	150,285	15,707	165,992	
Percent	91%	9%		

Table 5.33: Performance of mlZ on the Sim1 dataset at confidence threshold 90%. The table above compares the mlZ predictions with confidence ≥90% to the predictions with confidence <90%. The bottom table shows the number of samples with confidence ≥90% compared to the number of samples with confidence <90%. Percent in real data is the number of samples with the same criteria in the real data results.

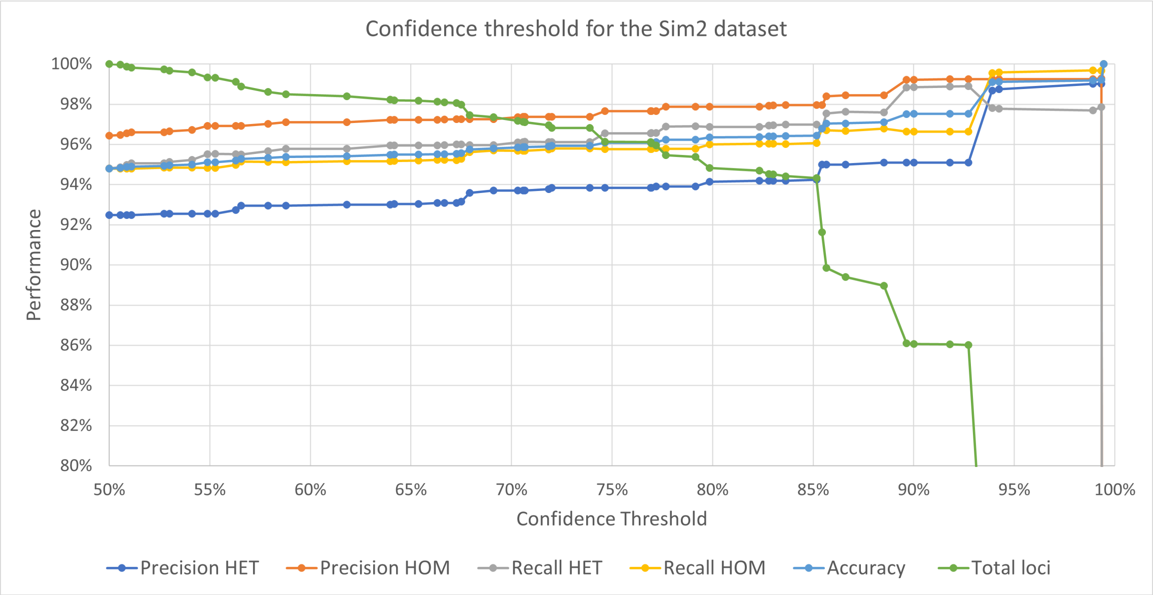


Figure 5·11: The confidence threshold effect on the Sim2 dataset performance. The confidence reported by the decision tree can be used to filter uncertain predictions. As the confidence threshold increases, the performance improves. About 86% of the test samples had confidence >90%. Note that the Y-axis starts from 80 to 100%.

Sim2		True HET	True HOM	Precision
Conf.≥0.9 Accuracy = 98%	Pred. HET	52,013	2,680	0.95
	Pred. HOM	604	76,891	99%
	Recall	99%	97%	
Conf.<0.9 Accuracy=78%	Pred. HET	6,654	2,090	76%
	Pred. HOM	2,611	10,050	79%
	Recall	72%	83%	
Combined Accuracy=95%	Pred. HET	58,667	4,770	92%
	Pred. HOM	3,215	86,941	96%
	Recall	95%	95%	
VNTRSeek	Not found	15,780	59,113	
	Percent	20%	39%	

	Conf.≥0.9	Conf.<0.9	Total	Percent
HET	54,693	8,744	63,437	86%
HOM	77,495	12,661	90,156	86%
Total	132,188	21,405	153,593	
Percent	86%	14%		

Table 5.34: Performance of mlZ on the Sim2 dataset at confidence threshold 90%. The table above compares the mlZ predictions with confidence ≥90% to the predictions with confidence <90%. The bottom table shows the number of samples with confidence ≥90% compared to the number of samples with confidence <90%. Percent in real data is the number of samples with the same criteria in the real data results.

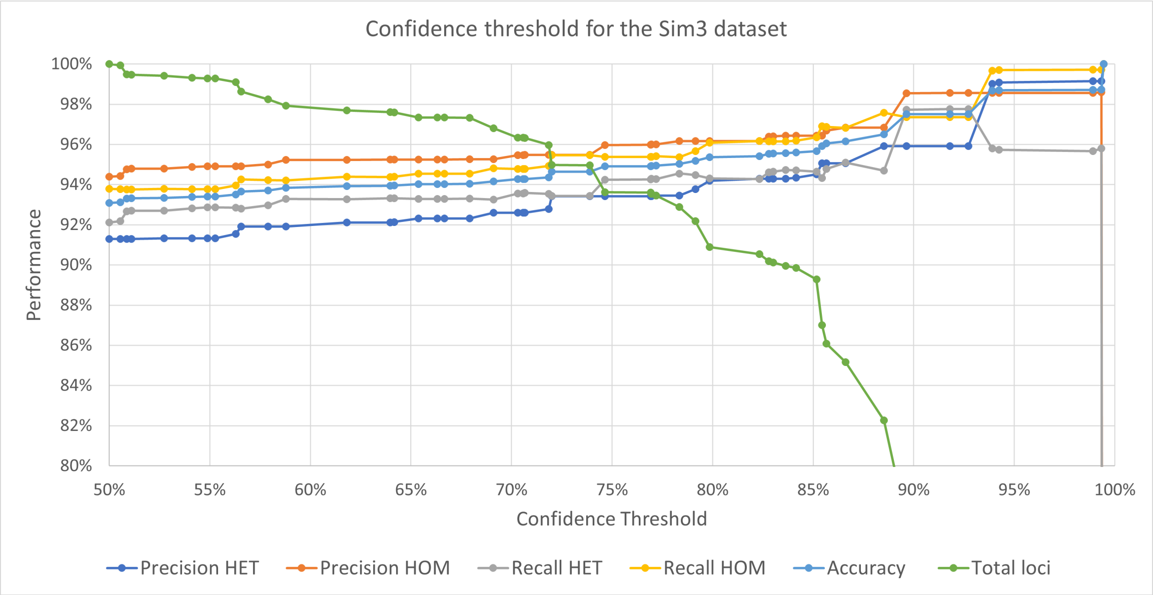


Figure 5·12: The confidence threshold effect on the Sim3 dataset performance. The confidence reported by the decision tree can be used to filter uncertain predictions. As the confidence threshold increases, the performance improves. About 78% of the test samples had confidence >90%. Note that the Y-axis starts from 80 to 100%.

Sim3		True HET	True HOM	Precision
Conf.≥0.9 Accuracy=98%	Pred. HET	39,865	1,700	96%
	Pred. HOM	912	62,603	99%
	Recall	98%	97%	
Conf.<0.9 Accuracy=78%	Pred. HET	12,173	3,265	79%
	Pred. HOM	3,541	12,317	78%
	Recall	77%	79%	
Combined Accuracy=93%	Pred. HET	52,038	4,965	91%
	Pred. HOM	4,453	74,920	94%
	Recall	92%	94%	
VNTRSeek	Not found	21,171	70,939	
	Percent	27%	47%	

	Conf.≥0.9	Conf.<0.9	Total	Percent
HET	41,565	15,438	57,003	73%
HOM	63,515	15,858	79,373	80%
Total	105,080	31,296	136,376	
Total	132,188	21,405	153,593	
Percent	77%	23%		

Table 5.35: Performance of mlZ on the Sim3 dataset at confidence threshold 90%. The table above compares the mlZ predictions with confidence ≥90% to the predictions with confidence <90%. The bottom table shows the number of samples with confidence ≥90% compared to the number of samples with confidence <90%. Percent in real data is the number of samples with the same criteria in the real data results.

5.3.6 Results of mlZ on the GIAB dataset

To test mlZ on real data, TRs were genotyped using VNTRseek on the seven genomes from GIAB (previously discussed in Chapter 2). Table 5.36 summarizes the VNTRseek runs on the GIAB datasets. First, the normal-VRS distribution on the HG002 (AJ son) was examined to see if the assumptions of mlZ apply on the real data (Figure 5·13). The homozygous calls by VNTRSeek had a thick left tail compared to a Poisson distribution suggesting some homozygous calls were heterozygous with undetected allele pairs. Also, similar to simulated datasets (shown in Figure 5·3), the heterozygous and homozygous normal-VRS is slightly less than the expected values of 0.5 for heterozygous and 1.0 for homozygous. This means that due to some errors, VNTRSeek missed some reads. This problem is yet to be investigated by the authors of VNTRSeek.

Genome	Sex	Read length (bp)	Fragment (μ, σ)	Physical coverage (X)	Total no. of reads	Total TRs genotyped	Total VNTRs
HG001	F	148	(550,150)	300	62.0E8	193,185	3,788
HG002	M	250	(400,100)	100	8.8E8	200,446	3,437
HG003	M	250	(400,100)	100	7.8E8	200,181	3,294
HG004	F	250	(400,100)	100	8.7E8	199,258	3,394
HG005	M	250	(550,150)	300	42.6E8	203,836	5,424
HG006	M	148	(550,150)	100	23.7E8	192,143	2,895
HG007	F	148	(550,150)	100	24.4E8	190,999	2,933

Table 5.36: VNTRseek results on the real datasets. VNTRseek was used to genotype the reference TRs in seven individuals from the GIAB consortium. The number of genotyped TRs increased with read length and coverage.

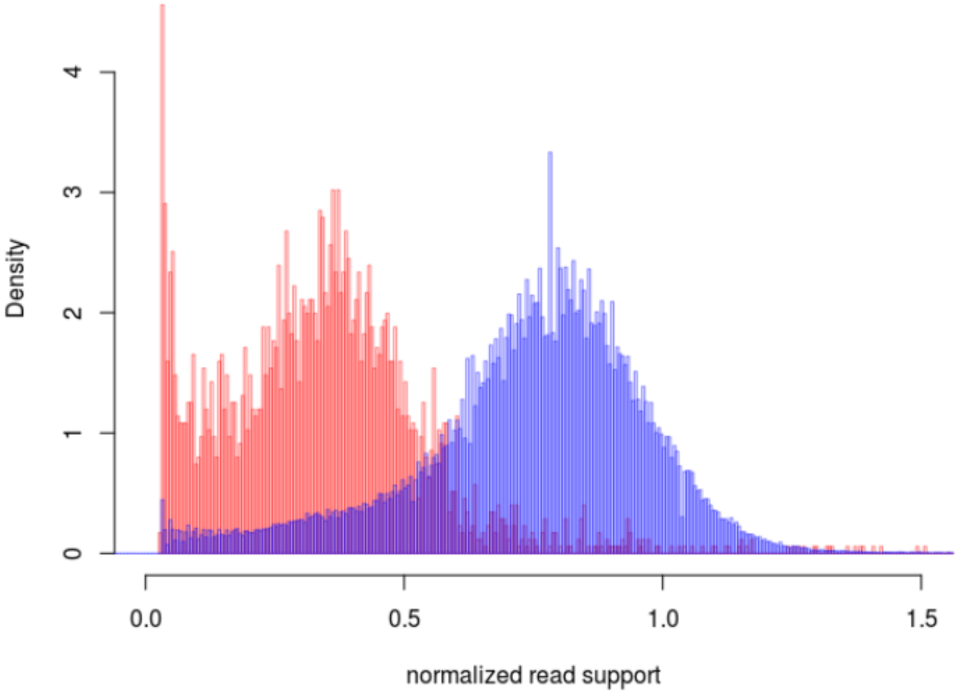


Figure 5·13: Histogram of normal-VRS values in the HG002 dataset. The number of reads supporting homozygous calls (blue) show a thicker left tail, indicating a mixture with heterozygous calls (red). The read support is normalized by the expected (theoretical) read support to be comparable across different array sizes and coverages.

In order to investigate how close our simulated read support estimates (SRS) are to the real data and how representative our model is, mlZ plots the three values SRS, ORS, and lmRS for each run. For HG001 (Figure 5·14), the Ashkenazi Jew trio (Figure 5·15, Figure 5·16, and Figure 5·17), and the Chinese Han trio (Figure 5·18, Figure 5·19, and Figure 5·20) the lmRS and SRS are in almost perfect agreement and lmRS successfully reduces the variability of ORS. These plots are automatically drawn by mlZ and are useful to detect incorrect parameter settings, i.e. if the user had a wrong assumption and incorrectly set the parameter.

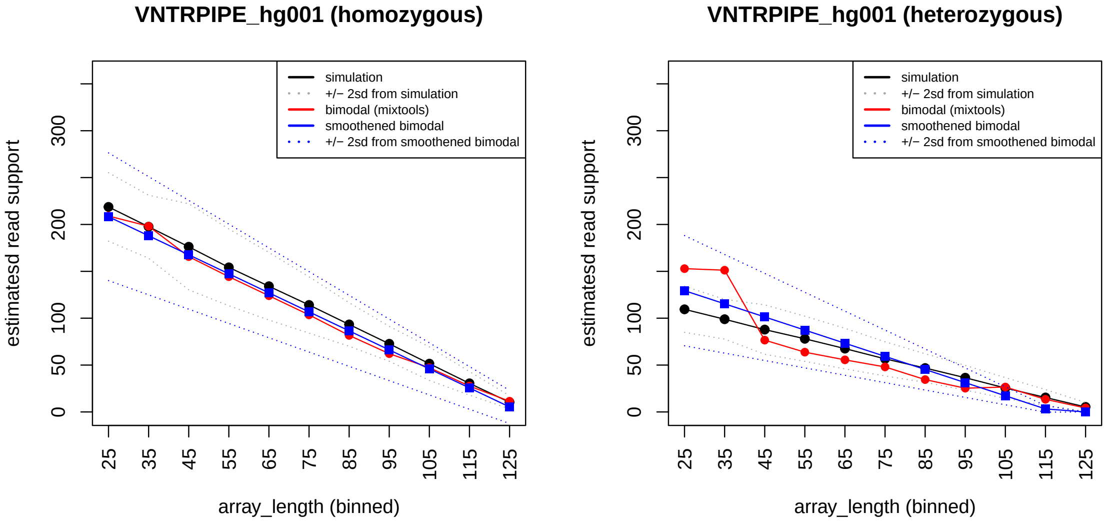


Figure 5·14: Comparison of the SRS, ORS, and lmRS values on HG001. mlZ plots the SRS (simulated read support) in black, ORS (observed read support inferred by fitting a normal distribution on the real data) in red, and lmRS (smooth ORS) in blue. The homozygous alleles are plotted in the left and the heterozygous alleles are plotted in the right. The gray dashed lines are the 2σ from the SRS mean and the dashed blue lines are 2σ from the lmRS lines.

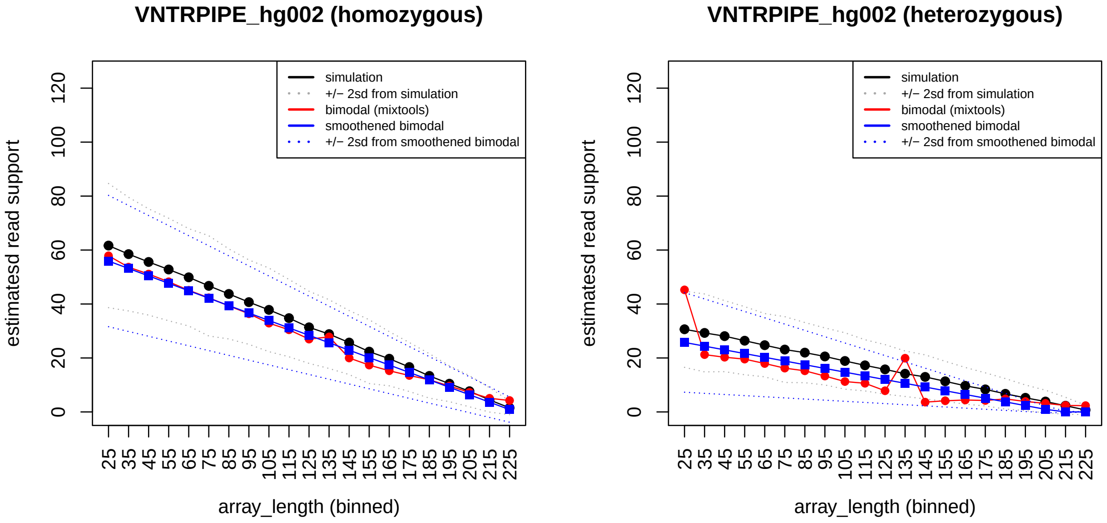


Figure 5·15: Comparison of the SRS, ORS, and lmRS values on HG002. mlZ plots the SRS (simulated read support) in black, ORS (observed read support inferred by fitting a normal distribution on the real data) in red, and lmRS (smooth ORS) in blue. The homozygous alleles are plotted in the left and the heterozygous alleles are plotted in the right. The gray dashed lines are the 2σ from the SRS mean and the dashed blue lines are 2σ from the lmRS lines.

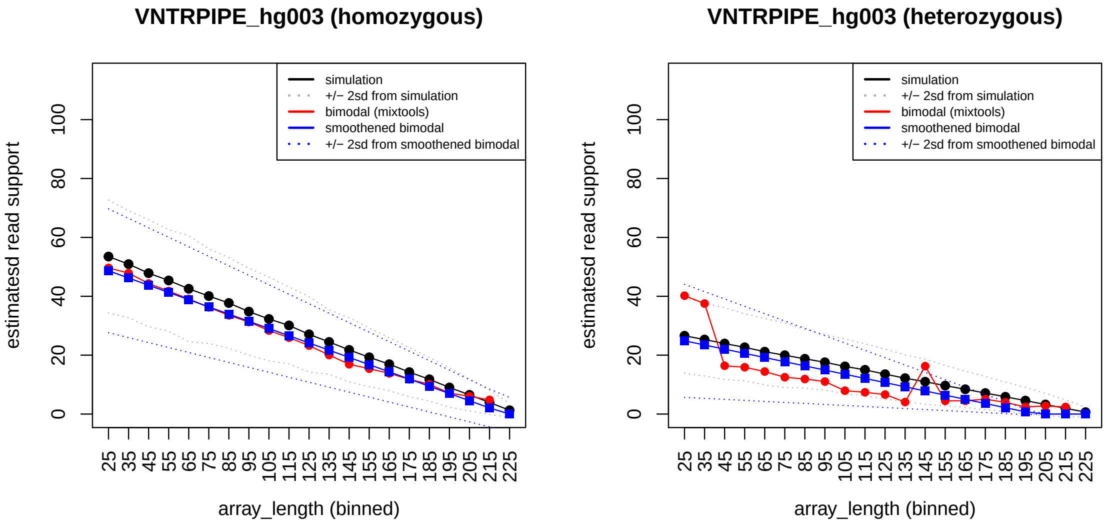


Figure 5·16: Comparison of the SRS, ORS, and lmRS values on HG003. mlZ plots the SRS (simulated read support) in black, ORS (observed read support inferred by fitting a normal distribution on the real data) in red, and lmRS (smooth ORS) in blue. The homozygous alleles are plotted in the left and the heterozygous alleles are plotted in the right. The gray dashed lines are the 2σ from the SRS mean and the dashed blue lines are 2σ from the lmRS lines.

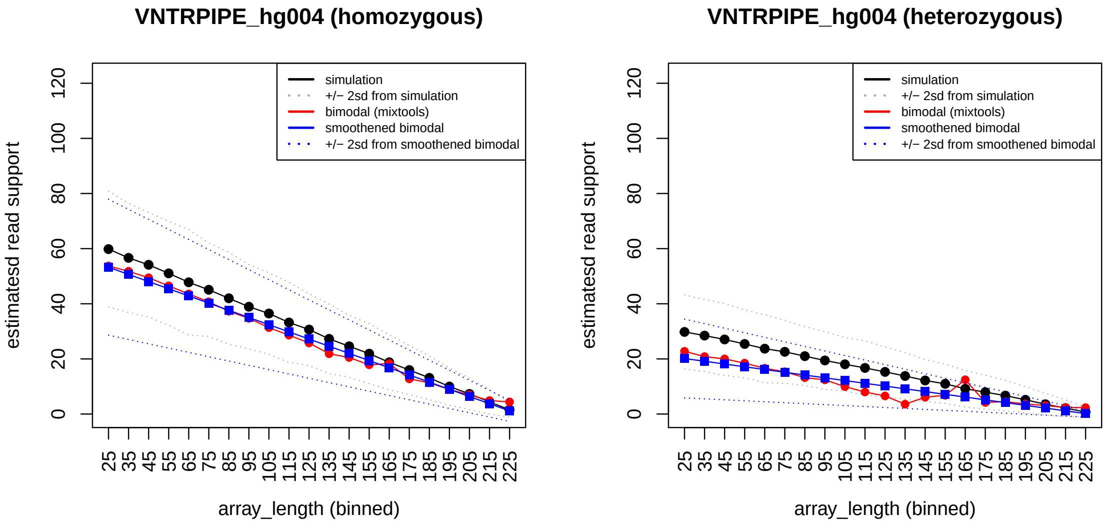


Figure 5·17: Comparison of the SRS, ORS, and lmRS values on HG004. mlZ plots the SRS (simulated read support) in black, ORS (observed read support inferred by fitting a normal distribution on the real data) in red, and lmRS (smooth ORS) in blue. The homozygous alleles are plotted in the left and the heterozygous alleles are plotted in the right. The gray dashed lines are the 2σ from the SRS mean and the dashed blue lines are 2σ from the lmRS lines.

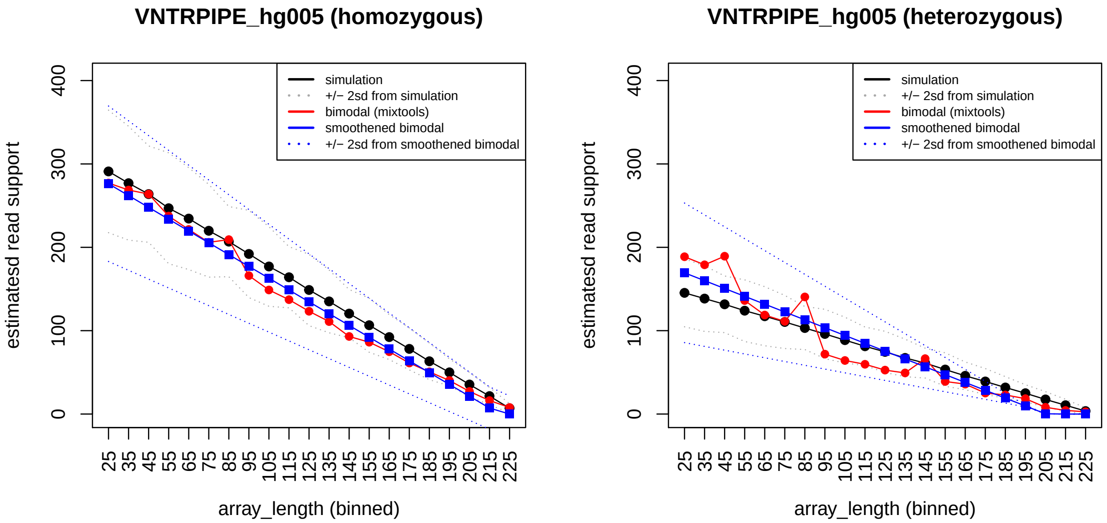


Figure 5·18: Comparison of the SRS, ORS, and lmRS values on HG005. mlZ plots the SRS (simulated read support) in black, ORS (observed read support inferred by fitting a normal distribution on the real data) in red, and lmRS (smooth ORS) in blue. The homozygous alleles are plotted in the left and the heterozygous alleles are plotted in the right. The gray dashed lines are the 2σ from the SRS mean and the dashed blue lines are 2σ from the lmRS lines.

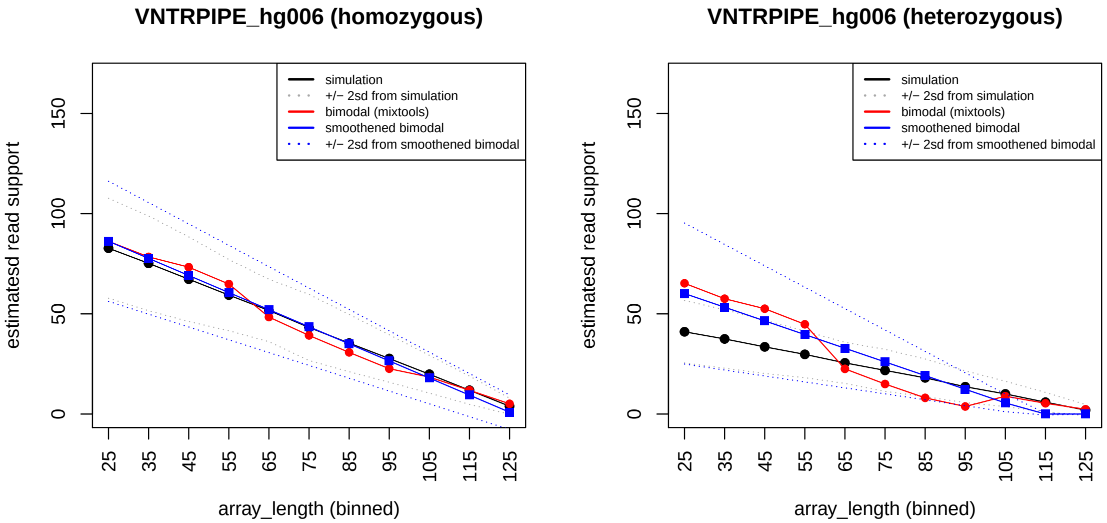


Figure 5·19: Comparison of the SRS, ORS, and lmRS values on HG006. mlZ plots the SRS (simulated read support) in black, ORS (observed read support inferred by fitting a normal distribution on the real data) in red, and lmRS (smooth ORS) in blue. The homozygous alleles are plotted in the left and the heterozygous alleles are plotted in the right. The gray dashed lines are the 2σ from the SRS mean and the dashed blue lines are 2σ from the lmRS lines.

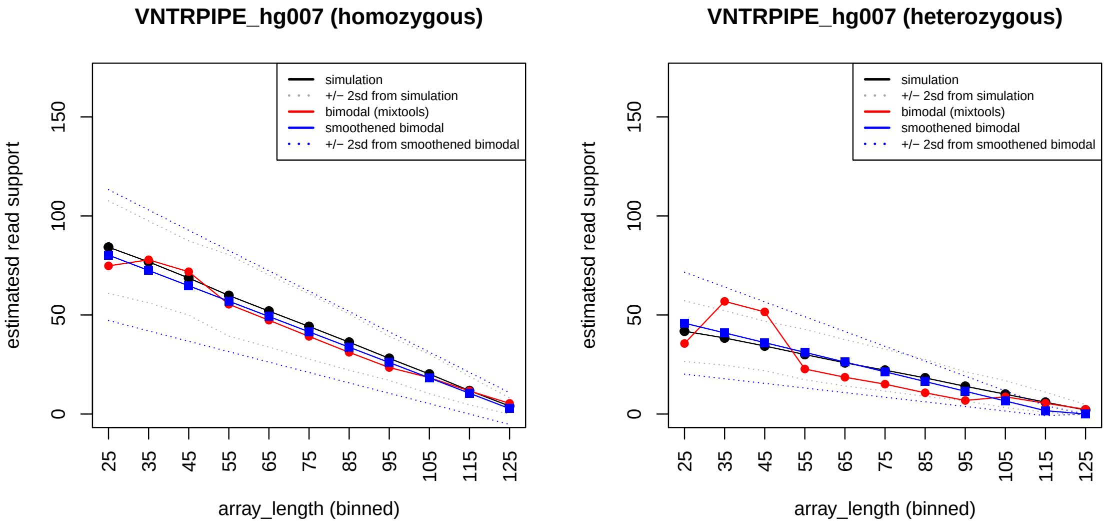


Figure 5·20: Comparison of the SRS, ORS, and lmRS values on HG007. mlZ plots the SRS (simulated read support) in black, ORS (observed read support inferred by fitting a normal distribution on the real data) in red, and lmRS (smooth ORS) in blue. The homozygous alleles are plotted in the left and the heterozygous alleles are plotted in the right. The gray dashed lines are the 2σ from the SRS mean and the dashed blue lines are 2σ from the lmRS lines.

The results of mlZ on real data are given in Table 5.37 for the TRs genotyped by VNTRseek. Table 5.38 compares the number of VNTRs per genome before and after applying mlZ. In each case about 19–20 thousand new heterozygous calls could be inferred, increasing the VNTR count per genome from thousands to tens of thousands. This new number agrees with the predictions in the literature (Näslund et al., 2005).

Data	**HOM→HOM**	**HOM→HET**	**HET→HOM**	**HET→HET**
HG001	142,084	21,376	495	971
HG002	142,038	21,334	109	1,204
HG003	143,343	19,857	83	1,144
HG004	147,922	22,168	101	1,221
HG005	144,064	20,730	645	1,451
HG006	134,630	21,069	317	739
HG007	141,151	21,127	313	747

Table 5.37: Results of mlZ on real data. The predictions are shown as "VNTRseek prediction" →"mlZ predictions". Around 20,000–22,000 homozygous calls were predicted to be heterozygous (HOM→HET). The genomes: HG001 (NA12878) is from the CEU family; HG002, HG003, and HG004 are the Ashkenazi son, father, and mother, respectively; and HG005, HG006, and HG007 are the Chinese son, father, and mother, respectively.

Sample	Description	Read length	VNTR Count pre-mlZ	VNTR Count w/ mlZ inferred
HG001	NA12878	148 bp	2,582	23,589
HG002	AJ son	250 bp	2,571	23,466
HG003	AJ father	250 bp	2,477	21,893
HG004	AJ mother	250 bp	2,558	24,312
HG005	Chinese son	148 bp	3,829	24,058
HG006	Chinese father	148 bp	2,039	22,827
HG007	Chinese mother	148 bp	2,052	22,787

Table 5.38: Inferring VNTRs using mlZ. The results of mlZ on real data were used to infer VNTRs. The mlZ predictions were labelled as VNTRs for all alleles with HOM→HOM predictions that had a non-reference allele or mlZ heterozygous (HET) predictions. The HOM→HET predictions are the VNTRs inferred by mlZ that VNTRseek did not predict as VNTR.

5.3.7 Consistency of mlZ across platforms

In Section 4.3.4 of Chapter 4, we showed that VNTRseek predictions are consistent across sequencing platforms. Here, we evaluate the consistency of mlZ across sequencing platforms. We used two datasets on the NA12878 genome. One was from GIAB, with read length 148 bp and 300× physical coverage, and another was from 1000 Genomes Project Phase 3, sequenced as 250 bp and >40× physical coverage. Table 5.39 shows that at 90% confidence threshold, the mlZ predictions were the same in both datasets 93% of the time. However, when we allowed missing genotypes (not penalizing if one dataset did not detect the genotype), the consistency went up to 99%. The consistency on the HOM→HET predictions was 87% and, when allowing missing genotypes, the consistency was 96%.

Table 5.40 presents a detailed comparison between the two datasets. The most common inconsistency is when the missing allele was detected in one, but not in the other. The rest of the inconsistent cases were negligible. As we increase the confidence threshold for mlZ predictions, the consistency increases, meaning the more precise calls are more consistent/reliable (Figure 5·21).

ID	HG001	NA12878
Source	GIAB	1000 Genomes Project
Fragment size	N(550,150)	N(400,100)
Read length (bp)	148	250
Read coverage	306×	60×
Percent of TRs Genotyped	87%	91%
VNTRs	2,583	2,181

Confidence≥90%	Without NA (found in both)	Allowing NA (found in either)
Consistency in HOM→HET	87%	96%
Consistency in all	93%	99%

Table 5.39: Consistency of mlZ predictions on NA12878 from two platforms. Two datasets from the same genome were used to compare the consistency of mlZ on different sequencing platforms. Results of mlZ on HG001 from GIAB with read length 148 bp and 300× physical coverage was compared to the results of mlZ from NA12878 from the 1000 Genomes Project with read length 250 bp and ∼40X physical coverage. Only mlZ predictions with >90% confidence were used. Consistency was measured in two ways: 1) only in the loci genotyped in both datasets (calculated a lower bound on consistency), and 2) allowing missing data (calculating an upper bound on consistency).

GIAB	1000 Genomes Project	Count
HOM→HET	HOM→HET	5,019
HOM→HOM	HOM→HOM	5,019
HOM→HOM	HOM→HET	606
HET→HET	HET→HET	541
HOM→HET	HOM→HOM	165
HOM→HET	HET→HET	29
HOM→HOM	HET→HET	29
HET→HET	HOM→HET	27
HET→HOM	HOM→HOM	6
HOM→HOM	HET→HOM	6
HET→ERR	HET→HET	3
HET→HOM	HET→HET	3
HET→HOM	HET→HOM	2
HOM→HET	HET→HOM	2
HET→HET	HET→HOM	1
HET→HOM	HOM→HET	1

Table 5.40: Detail comparison of results from VNTRseek and mlZ on NA12878 from two platforms. The predictions from mlZ on two datasets from the same genome (NA12878) were compared. Total 16 different cases can occur. HOM→HET predictions in any dataset are highlighted as the most interesting class which results in inferring new VNTRs. Rows highlighted in green are the consistent predictions (HOM→HET in both dataset) and highlighted in orange are the inconsistent predictions where a least in one dataset a HOM→HET prediction was made.

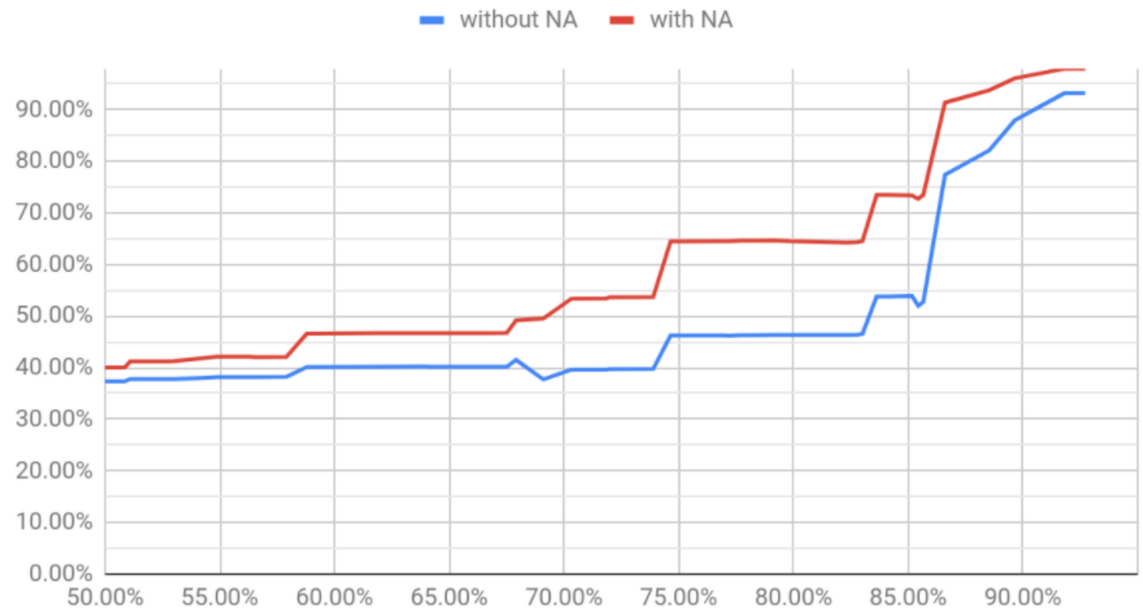


Figure 5·21: The effect of the confidence threshold on the consistency of mlZ on sequencing platforms. The consistency of mlZ predictions on the same genome from two datasets with different sequencing platforms was calculated in two ways: 1) without NAs: only on loci which was genotyped in both datasets by VNTRseek, and 2) with NAs: Considering the loci that was not genotyped on one of the datasets as a correct prediction. The first method calculated a lower bound for consistency and the second method calculated an upper bound for consistency.

5.3.8 Investigating the predicted errors

In Section 4.3.5 of Chapter 4, the error type 1 of VNTRseek was addressed. To investigate the power of mlZ to detect such erroneous alleles, we counted the percentage of alleles that were labeled as ERROR by mlZ (using the Z-label). Table 5.41 shows that all of the datasets had $\leq 1\%$ error rate except for HG001 and HG005 that were sequenced using MiSeq Illumina machines at 300× physical coverage. Sim1 was simulated using an error profile mimicking a MiSeq machine and had much higher Multi ratio, too. Figure 5·22 shows that, as the coverage increases, the number of ERROR predictions by mlZ also increases. This was also seen in Section 4.3.5 where the number of Multis increased by read coverage.

Genome	**VN-TRs**	**Any allele was error**	**All alleles were error**	**Error (%)**	**Read coverage (×)**	**Read length (bp)**
HG001	2,583	5,183	3,845	**1.99**	306	148
HG002	2,522	1,627	1,575	0.82	65	250
HG003	2,559	1,609	1,566	0.82	72	250
HG004	2,621	1,726	1,682	0.84	74	250
HG005	2,081	9,441	7,071	**3.65**	355	250
HG006	2,053	2,233	1,851	1.01	117	148
HG007	3,897	1,837	1,576	0.83	119	148
Sim1	70,540	23,321	4,239	**2.46**	201	250
Sim2	58,395	5,796	1,452	0.95	112	148
Sim3	46,112	2,242	821	0.60	65	101

Table 5.41: Error predictions by mlZ. The Z-labels were used to count the ERROR labels. "Any error" refers to loci that at least one allele had Z-label ERROR. "All error" is the number of loci that all alleles were ERROR. The error percentage is the percentage of loci that had error labels for all the alleles. In **bold** is the error rate of HG001, HG005, and sim1. These datasets had the MiSeq Illumina machine error profile that has higher errors including indels.

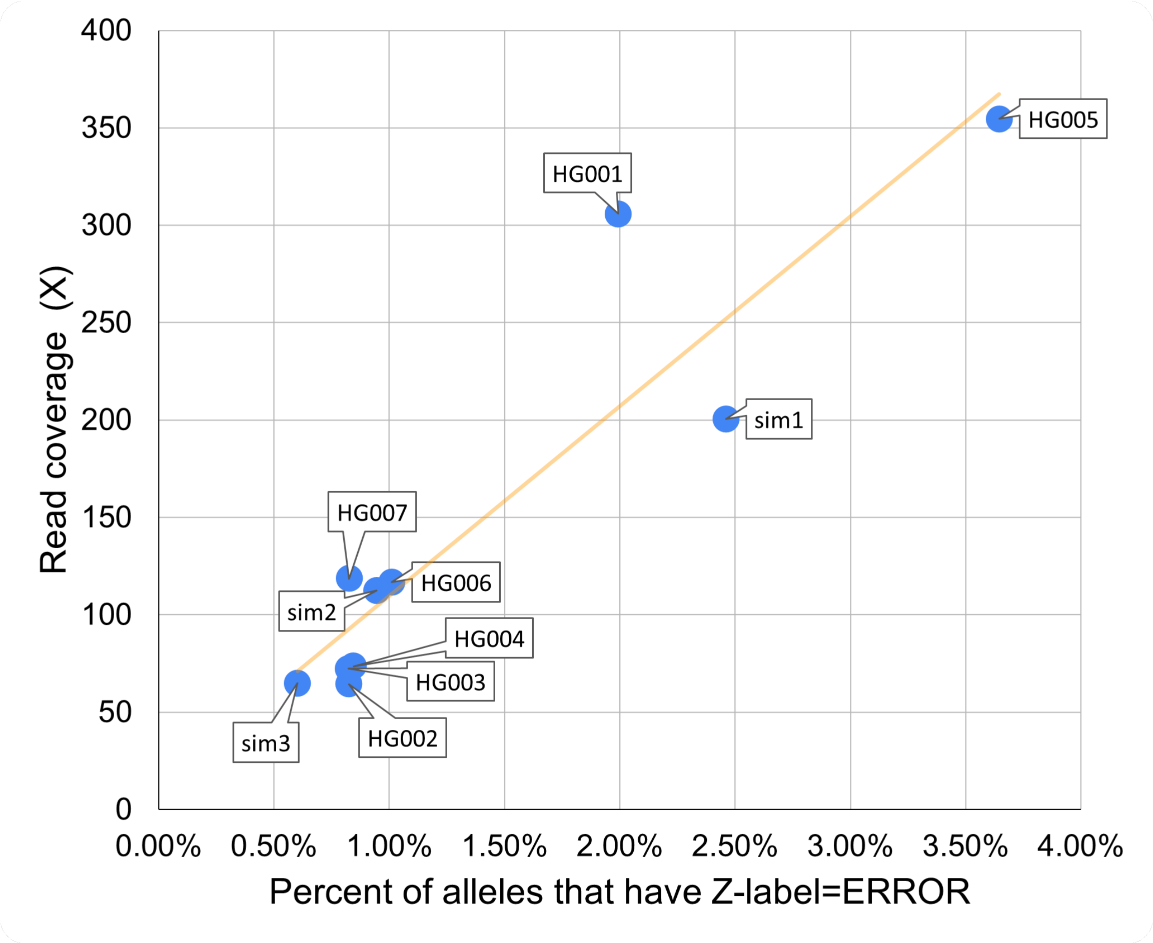


Figure 5·22: Predicted ERRORs as a function of the read coverage. The percent of loci that had Z-label ERROR for all their alleles is plotted against the read coverage. At higher read coverage more ERRORs occur. This is consistent with the finding that the number of Multis increases with the read coverage (Section 4.3.5).

5.3.9 Ability of mlZ to remove Multis (error type 1)

To investigate the effectiveness of mlZ on removing false positive alleles, i.e. Mutis, the number of heterozygous calls on the sex chromosomes of male individuals were used. Sex chromosomes on male individuals are haploid and should not have any heterozygous VNTR. We counted the number of such multis on the chromosome X and Y of HG002, HG003, HG005, and HG006. Given that there 8,312 TRs on the sex chromosomes and 181,461 TRs on autosomal chromosomes, we can estimate the number of multis expected on the whole genome. mlZ does not run on sex chromosomes of male individuals, so the number of mlZ predictions of HET→HOM could be used as an estimate of the multi loci. Table 5.42 shows that the number of HET→HOM predictions by mlZ agrees well with the estimated values. Figure 5·23 plots this correlation with $r^2 = 99\%$.

Genome	Multis with two alleles on chrX or chrY	Estimated no. of Multis on the whole genome	mlZ HET→HOM	Difference
Sim1	284	6,200	9,287	-3,087
Sim2	59	1,288	1,405	-117
Sim3	27	589	614	-25
HG002	5	109	109	0
HG003	3	65	83	-18
HG005	22	480	645	-165
HG006	10	218	317	-99

Table 5.42: The mlZ HOM→HET predictions as an estimate for Multi loci. The heterozygous Multi genotypes on the sex chromosome of male individuals were used to estimate the total number of multis on the whole genomes. Total TRs on chromosome X and Y was 8,312, and the total number of TRs on the other chromosomes is 181,461. The HET→HOM predictions by mlZ was compared to the estimated number of Multis on the whole genome. The column Difference shows the difference between the estimated Multis and the HET→HOM predictions. The differences were small. Note that mlZ does not perform predictions on the sex chromosomes of male individuals.

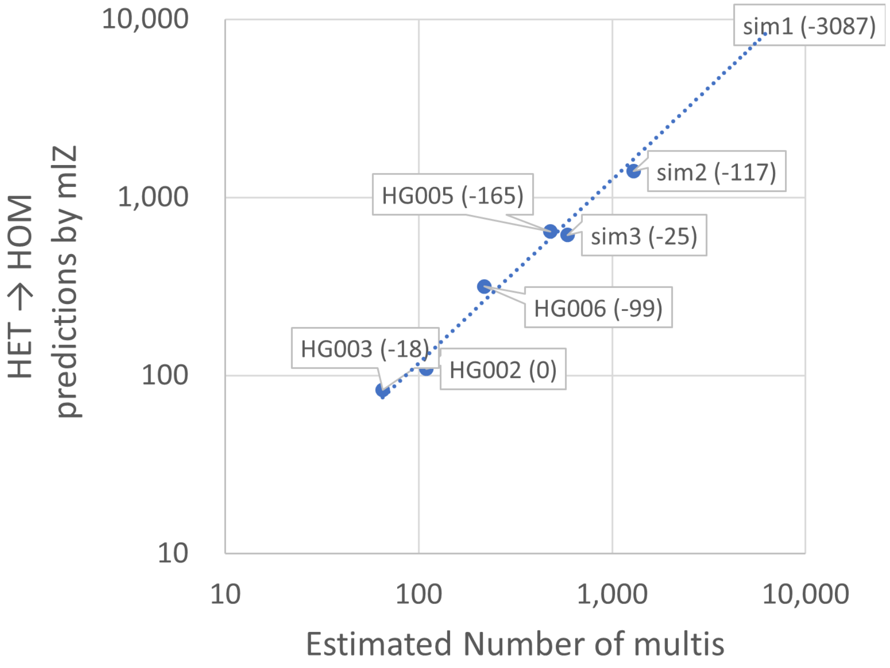


Figure 5·23: The expected number of multis compared to the HET→HOM predictions by mlZ. The heterozygous Multi genotypes on the sex chromosome of male individuals were used to estimate the total number of multis on the whole genomes (see Table 5.42). The HET→HOM predictions by mlZ correlated with the estimated number of multis (r2 = 99%).

5.4 Summary

In this chapter, a computational tool named mlZ (machine learning on Z-scores) was introduced to reduce type 1 and type 2 errors in VNTR predictions. On the genotyped TRs by VNTRseek, mlZ compares the observed read support to the theoretically expected value to predict whether the genotype is heterozygous or homozygous. On simulated datasets, mlZ had an accuracy of >95% and the precision and recall on the heterozygous class was 94% and 93%, respectively. By applying mlZ to VNTRseek predictions, the accuracy increased from ∼70% to >95%.

On real data, mlZ inferred an additional 20,000 heterozygous VNTRs increasing the number of VNTRs to ∼21,000–24,000 per genome. mlZ prediction and performance was consistent between datasets from the same genome from two different sequencing platforms with ∼96% consistency on the heterozygous class.

In addition, the number of predicted ERROR labels increases with the read coverage. Using the number of heterozygous VNTRs predicted by VNTRseek on the sex chrosomomes on male individuals, the total number of multis on the rest of the genome is estimated. A close correlation ($r^2 = 99\%$) between the number of HET→HOM predictions by mlZ and the estimated number of multis was found.

Chapter 6

Genotyping Macrosatellites Using Read Depth (MaSUD)

6.1 Introduction

Macrosatellites are tandem repeats with patterns of 100 bp or larger. Currently no high-throughput methods exist for genotyping macrosatellites. However, few studies have examined macrosatellites in vitro and concluded that macrosatellite copy numbers are highly polymorphic across the human population (Schaap et al., 2013; Brahmachary et al., 2014). Macrosatellites may affect gene expression by repeat-induced gene silencing (RIGS) (Garrick et al., 1998; Ye and Signer, 1996) (see Chapter 1). In this chapter I present a novel computational method, `MaSUD` (genotyping **Ma**cro**S**atellite **U**sing **d**epth) to detect total copy number gain or loss using short WGS datasets. Section 6.2 described the data used in this chapter and the methodology of `MaSUD`. Section 6.3 presents the performance of `MaSUD` and an analysis of the characteristics of macrosatellites in a large cohort of unrelated individuals. This chapter concludes in Section 6.4 with a brief summary.

6.2 Materials and methods

The data used in this analysis is presented in Sections 6.2.1, 6.2.2 and 6.2.3. The `MaSUD` algorithm is described in Section 6.2.4. Sections 6.2.5 and 6.2.6 describe the validation of `MaSUD` on real data. Section 6.2.7 provides the methods to annotate macrosatellites. Sections 6.2.8 and 6.2.9 describe the population-biases of macrosatellite genotypes and Section 6.2.10 presents the methods on finding eQTL macrosatellites.

6.2.1 Tandem repeat reference set

Tandem repeats on the GRCh38 reference genome were downloaded from the Tandem Repeat Database (TRDB) (Gelfand et al., 2007) and filtered using the methodology explained in (Gelfand et al., 2014). We further filtered the TRs to include only macrosatellites, i.e. those with pattern length ≥100 bp. Additionally, macrosatellites were removed if they overlapped with segmental duplications, i.e.,regions >1 Kbp long and with >90% similarity to another region (Vallente and Eichler, 2005). For this step, a list of segmental duplications database provided by the Eichler lab was used (Bailey et al., 2001; Alkan et al., 2011).

6.2.2 Simulated datasets

Five simulated haploid genomes were generated using simuG (Yue and Liti, 2019), which allows genome simulation with random or targeted variants. The simulated genomes were designed to test `MaSUD` under different conditions: small copy number changes, large copy number changes, and no changes. For each of the genomes, the changes described below were applied to all the macrosatellites from the filtered reference set:

– *Ref*: No change in macrosatellite copy number;

– *Simple*: macrosatellite copy number modified by creating a copy number change in the range $[-rcn, +rcn]$ where rcn = reference copy number. The range includes only integer values. For example if $rcn = 4$, the new copy number could be any integer value from 0 to 8. Since most macrosatellites have copy number <5, the changes here were small;

- *Large*: changes of at least three copy number when possible, if rcn≥3 the copy change for gain would be in the range $[3, rcn]$ and for loss would be $[-rnc, -3]$, otherwise, if rnc<3, no change is implemented;
- *Gain*: only copy number gains in the range $+3 + [0, rcn]$; and
- *Loss*: only copy number losses in the range $-3 - [0, rcn]$ when rnc≥3, otherwise no change is implemented.

All randomized change values were rounded to the closest integer. Table 6.1 summarizes the copy number gain and loss implanted on each simulated genome.

Name	No. of ref.	No. of gains	No. of loss	Sequencing
Ref	4,292	0	0	A, C
Simple	1,948	1,166	1,178	A, C
Large	4,038	130	124	A, B, D
Gain	0	4,292	0	A, B, D
Loss	3,615	0	677	A, B, D

ID	Read length	Fragment size	Coverage
A	100 bp	$\mathcal{N}(\mu = 350, \sigma = 100)$	40×
B	150 bp	$\mathcal{N}(\mu = 550, \sigma = 150)$	40×
C	150 bp	$\mathcal{N}(\mu = 550, \sigma = 150)$	100×
D	250 bp	$\mathcal{N}(\mu = 550, \sigma = 150)$	100×

Table 6.1: Sequencing simulations.

The copy number changes were saved as VCF files and given as CNV targets to simuG. An additionally random 3,000,000 random SNPs were also inserted into each genome using simuG. As a result, a simulated haploid genome was produced for each of the five conditions and saved in FASTA format.

We simulated paired-end Illumina reads from the FASTA files using ART (Huang et al., 2012). Four different settings were used to represent publicly available real data, with read lengths of 100 bp, 150 bp, and 250 bp and physical coverages of 100× and 40× (Table 6.1). The reads were mapped back to the reference genome using BWA-MEM (Li, 2013) and the BAM files were sorted and indexed using samtools (Li et al., 2009).

6.2.3 Whole genome sequencing data

GIAB: BAM files for seven genomes sequenced at high coverage were downloaded from the Genome In A Bottle (GIAB) consortium (Zook et al., 2016) (see Data Availability Section) including: the NA12878 genome (HG001) from the 1000 genomes CEU family, the Ashkenazi Jewish (AJ) trio (HG002, HG003, HG004) and the Chinese (HAN) trio (HG005, HG006, HG007). The Ashkenazi Jewish trio and Chinese Han trio genomes are from the Personal Genome Project (Church, 2005). Table 6.2 summarizes the characteristics of these datasets.

GIAB ID	**NIST ID**	**Description**	**Sex**	**Read length (bp)**	**Coverage**
HG001	NA12878	Western European	Female	148	300
HG002	NA24385	Ashkenazi Jewish son	Male	250	100
HG003	NA24149	Ashkenazi Jewish father	Male	250	100
HG004	NA24143	Ashkenazi Jewish mother	Female	250	100
HG005	NA24631	Chinese son	Male	250	300
HG006	NA24694	Chinese father	Male	148	100
HG007	NA24695	Chinese mother	Female	148	100

Table 6.2: GIAB genomes.

NYGC: In 2019, the New York Genome Center (NYGC), sequenced 2,504 unrelated individuals from five super-populations and 26 sub-populations[1]. Most of these genomes overlap with the 1000 Genomes Project. The read length was 150 bp and the coverage was >30X. These genomes were downloaded as CRAM files.

[1] These data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S.

6.2.4 The MaSUD algorithm

MaSUD uses read depth to estimate the total change in pattern copy number at a TR locus in comparison to a reference copy number. As a preliminary, we consider reads mapping inside a TR array.

The expected number of reads aligning inside a macrosatellite array of length A is positively correlated to the physical coverage and inversely? correlated to the read length. This relation can be formulated as:

$$Inside \propto \frac{Physical\ coverage}{Read\ length} \times A, \tag{6.1}$$

where *Inside* is the number of reads aligning inside the array of size A.

The accuracy of this equation is shown using simulated data in Figure 6·1. Using simulated data, we plotted the $Inside$ counts against the array size and a corresponding linear regression line for each read length and coverage combination. In each case, the slope is equal to $Physical\ coverage/Read\ length$.

Consider now the example of genotyping the same macrosatellite locus in two genomes, i and j, with the same sequencing characteristics, *i.e.*, read length and coverage. If the locus is homozygous in both genomes, but they have two different alleles, with array sizes A_i and A_j, respectively, then by counting $Inside_i$ and $Inside_j$, we can estimate the difference between the array sizes as:

$$A_i - A_j = \frac{Read\ length}{Coverage}(Inside_i - Inside_j), \tag{6.2}$$

and the fold change (ratio change in the sizes) as:

$$\frac{A_i}{A_j} = \frac{Inside_i}{Inside_j}. \tag{6.3}$$

Because the pattern lengths are the same in the two alleles, we can explain the

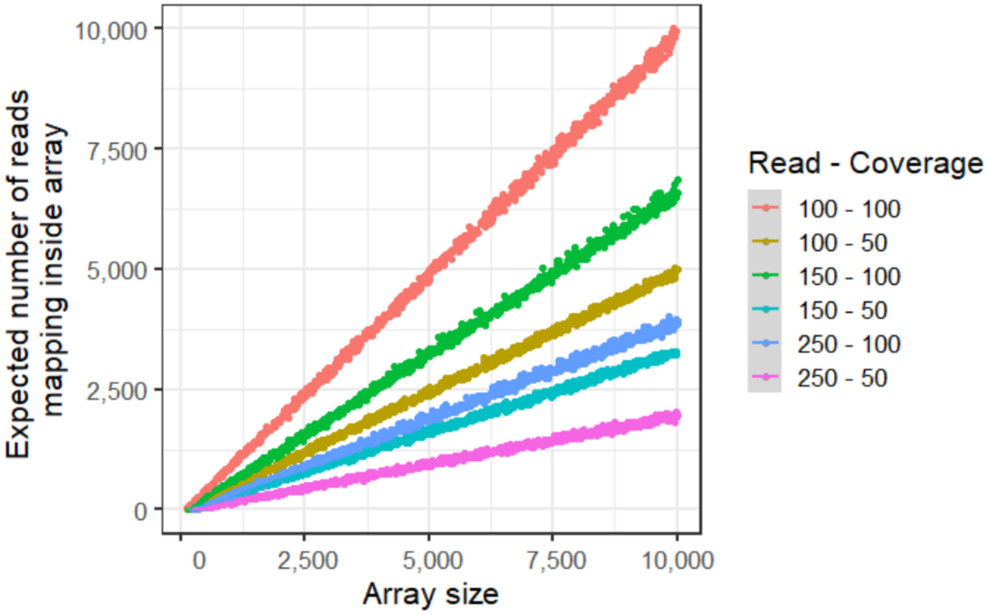


Figure 6·1: The relation of the *Inside* measurement with read length and physical coverage. *Legend:* the numbers for each test are given as “read length”-“coverage”. The expected slope for read length 100 bp and coverage 100X (red line) is one.

difference in array sizes as a difference in their copy numbers. (Here we assume that the major differences in the allele lengths are confined to whole pattern copy number differences. This restriction will be relaxed below?) In other words:

$$A_i = pattern\ length \times copy_i \tag{6.4}$$

and

$$A_j = pattern\ length \times copy_j, \tag{6.5}$$

where *pattern length* is the pattern size of the macrosatellite and $copy_i$ and $copy_j$ are the pattern copy numbers in alleles A_i and A_j, respectively. Then, the copy number change can be estimated as:

$$copy_i - copy_j = (A_i - A_j) \ / \ pattern \ length, \tag{6.6}$$

or:

$$copy_i - copy_j = (\frac{A_i}{A_j} - 1) \times copy_j. \tag{6.7}$$

We use genome i and thus allele A_i as the reference allele in what follows. For the sample (diploid) genome, the locus may have one (homozygous) or two (heterozygous) alleles. The homozygous case is as presented above. For the heterozygous case, the value of $copy_i$ will be the average copy number of the two alleles. Because MaSUD cannot detect the difference between a homozygous and heterozygous locus, this means that the estimated copy number difference from Equations 6.6 or 6.7 will also be an average of the difference between the reference and the two heterozygous alleles.

Figure 6·2 shows the algorithm for MaSUD. Equations 6.6 and 6.7 are used to estimate the copy number change of a macrosatellite locus in an individual compared to the reference genome in the following steps:

1. Given a sorted BAM file, for each macrosatellite, reads aligning inside the array ($Inside_i$) are counted.

2. Simulated reads are generated with the same characteristics as the real data from a genome with reference alleles. These characteristics are the physical coverage, read length, mean fragment size and standard deviation. The simulated reads are mapped back to the same reference genome and the reads mapping inside each array ($Inside_j$) are counted. The change in copies is calculated according to Formula 6.6 (ΔC) and Formula 6.7 (FC).

3. Step 2 is repeated several times (by default 10 times) to account for random noise. The mean and standard deviation of ΔC and FC are calculated are reported for each reference macrosatellite.

These steps are described in more detail in the following paragraphs.

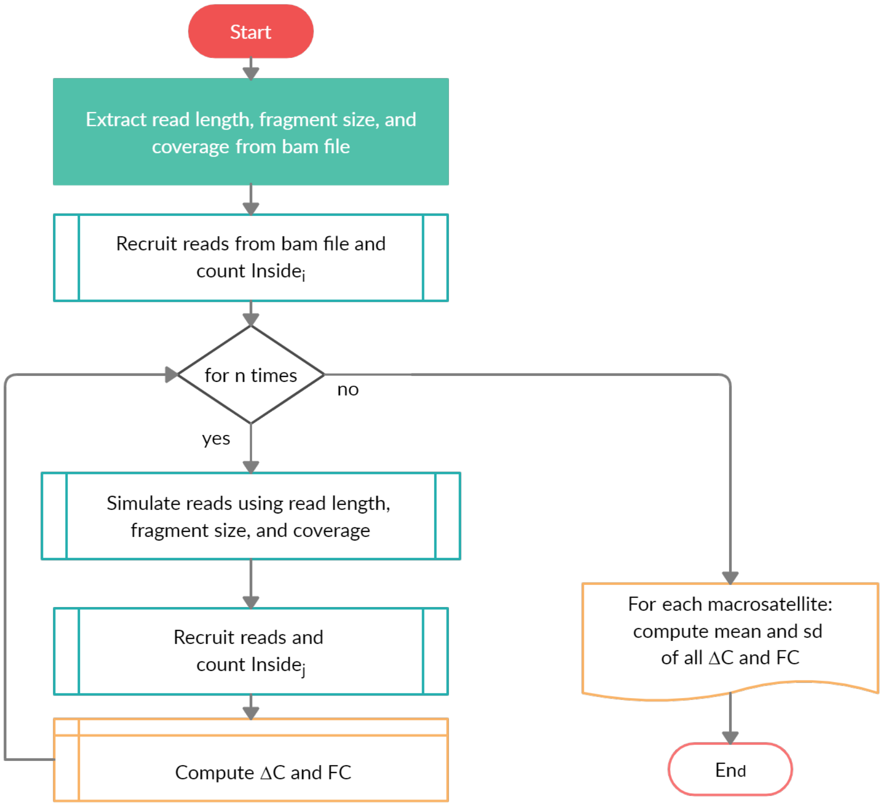

Start
Extract read length, fragment size, and coverage from bam file
Recruit reads from bam file and count Inside$_i$
for n times
no
yes
Simulate reads using read length, fragment size, and coverage
Recruit reads and count Inside$_j$
Compute ΔC and FC
For each macrosatellite: compute mean and sd of all ΔC and FC
End

Figure 6·2: The `MaSUD` algorithm

Recruiting reads Given a sample genome BAM file consisting of reads aligned to a given reference genome, `MaSUD` extracts reads mapping completely inside the macrosatellite array. This observed number of reads is termed $Inside_{obs}$ and represents the combined effect of the two (heterozygous or homozygous) alleles in the sample genome. This requires the array length to be greater than the read length.

Creating background counts Using the characteristics of the input BAM file, *i.e.*, the read length, fragment length mean and standard deviation, and physical coverage, reads are randomly generated from the reference genome in the region(s) of interest. The simulated reads are mapped back to the reference genome and those reads aligning completely inside the macrosatellite array are counted. This background number of reads is termed $Inside_{bg}$ for the reference copy number, $copy_{ref}$.

The copy number change estimate Copy number change at the macrosatellite locus is estimated in two ways. First, using the $Inside$ differences (formula 6.6):

$$\Delta C = (\frac{Read\ length}{Coverage}) \times (Inside_{obs} - Inside_{bg})\ /\ pattern\ length, \tag{6.8}$$

and, second, using the $Inside$ fold change (formula 6.7):

$$FC = (\frac{Inside_{obs}}{Inside_{back.}} - 1) \times copy_{ref}. \tag{6.9}$$

Theoretically ΔC and FC should return the same number. Since ΔC is a linear function of two normal distributions, the standard deviation of this distribution will be the summation of the standard deviation of the two normal distributions. However, FC is the *ratio* of two normal distributions and the standard deviation will be undefined. In the case of gain ($Inside_i > inside_j$), the standard deviation will be lower than the standard deviation of each distribution. However, ΔC and FC are calculated using the reference pattern length and copy number, respectively, which are independent values. It is interesting to report both measurements to compare the robustness and how they perform under different characteristics of the reference macrosatellite.

The refining step The background simulation is repeated ten times and the ΔC and FC for each simulation is calculated. The *mean* and the *standard deviation* of both statistics, ΔC and FC, are computed. The standard deviation is used as a measure of certainty. Observing that ΔC performs better on loss of copies and FC performs better on gain, (see Section 6.3.2) we define the final estimate, $MaSUD$ as:

$$MaSUD = \begin{cases} \Delta C, & \text{if } \Delta C < 0 \\ FC, & \text{otherwise} \end{cases} \tag{6.10}$$

One limitation of `MaSUD` is in regard to short array sizes. When the loss of copies is such that the $array_j$ becomes smaller than the *Read length*, `MaSUD` returns -rcn, because no read could span inside the array ($A_i = 0$). Any copy change of -rcn should be treated as a loss of at least $rcn - \frac{read}{pattern}$, rnc is the reference copy number.

6.2.5 In vitro validation

DNA samples for the AJ trio, NA24385, NA24149, and NA24143 (also identified as GIAB IDs HG002, HG003, and HG004), were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. We chose 10 macrosatellite loci predictions for experimental validation using the following criteria:

- product length ∼1,000–1,500 bp,
- expected length change within 10-20% of the reference array length,
- GC content of 40-60%, and
- unique primers detectable.

We prioritize those macrosatellite loci which were nearby to genes or regulatory regions. The primers were designed using primer-BLAST (Ye et al., 2012) (Table 6.3).

#	TR id	Position on GRCh38	Left primer	Right primer
1	182520406	chr15:78945370-78946281	TCTGGGAGGAAGGGCTTATCT	GAGTGTGGGGGAAAGTCTGAG
2	183350417	chrX:151913368-151919994	AACCCATCTACACTCCCATCC	AAACGGAATGCGATTCAACC
3	182265155	chr10:4815436-4816095	ACAGCTCTGTCTGCATTTTGAG	AGGATGTACTTCACTTGAGCTT
4	182319087	chr11:1684453-1685020	CTGTGATTGTCCCCAGACACC	TCAAAGGTGGAGGCTGTCAGA
5	182496331	chr14:105346463-105347033	GGTCCCATGCAGAATCACACA	AAATGGAGACCACAGAGACCC
6	182641232	chr19:879444-879969	CGCACATGCCCCAGCA	CAGTGCCTGTGCCCCAC
7	182673031	chr19:50974778-50975726	CTTCAGGGGAATGACTGGTCG	TCTGCAGCCGCTAATTCTGAT
8	182674254	chr19:53211484-53212591	ATTGAGAAGTCTGTCACCCGT	TTGAGCGCACGCTAAAGGC
9	182766559	chr2:233396872-233397623	CACCACGTGGAGTGTGCAA	CCACCCTACAGCCAGAGTAGA
10	182831502	chr22:43280700-43281599	TCTCAGCCATGCTCCTACCA	ATATGGGGACAGCAAGGGGA
11	182908814	chr3:197517200-197519140	TAGCATTCCCTTGGGTCCTC	CACTCATGCCTGGATCAGACC
12	182917919	chr4:18611524-18612759	TTGGTAGTGTTTACTGGGGTGT	AAATCCCGGGCAGTAGCAGA
13	183113244	chr6:168316239-168318178	CACCCCCTTGACAGAACAGT	GGCCGTCTGTTTAGGTCGTT
14	183184241	chr8:1388079-1389137	TTAGCGATCAAGGCGTGGAG	CCCTTAGTCTGTGGGACCAG
15	183184316	chr8:1461012-1462594	TGGGTCGGTAGGGATCAGAA	CACCCAAGGTAGCCACAGTC
16	183288865	chr9:136535433-136536036	CAATCTGGAGTGTAGACGGC	ACCTGTGGAATTCTGGGAGGA
17	182910398	chr4:1580570-1581279	AAGCTAGGGAGAAAGCATCCC	AAGTCTGGTCTGAGCAACTGTG
18	182263726	chr10:2290344-2291180	GCATTGGAGTGAAAGCTGGTG	CACAGGGGTGCAAGGATTGTA
19	183364583	chrY:19481992-19482841	GTTGGCTCCAGGAATGATGTTG	TGTTGATAACACACAGGCCCA
20	183183624	chr8:617460-618467	TTTGGTCCTCATCTGACCCCA	TGGTAATGTTGCCCAAGGTCT

Table 6.3: Experimental primers for macrosatellites.

6.2.6 Confirmation on the Y chromosome in related genomes

The copy number on the Y chromosome of son and fathers should be identical.

6.2.7 Gene set enrichment of genes overlapping macrosatellites

All genes proximal to macrosatellites (1 Kbp upstream to 1 Kbp downstream) were tested with GSEA (Subramanian et al., 2005) against the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) for enrichment of KEGG pathways (Kanehisa and Goto, 2000), cell type signatures (C8), and biological process GO terms (GO:BP) (Ashburner et al., 2000; The Gene Ontology Consortium, 2019). Results with adjusted p-value (FDR) <1% were considered significant.

6.2.8 Population ancestry prediction

The NYGC dataset consists of 2,504 genomes from individuals with ancestry in five super-populations (African, American, East Asian, European, and South Asian) and 26 sub-populations. Principal Component Analysis (PCA) was used to investigate the predictive power of `MaSUD` genotyping results with regard to super-population membership. For each sample, a vector of loci with `MaSUD` prediction values was produced. Macrosatellite loci on the sex chromosomes, were excluded, resulting in xxx loci per vector. The top 30 Principal Components (PCs) were selected to train a decision tree to predict super-population ancestry using a 75% to 25% training to testing split of the data. Training parameters were a maximum depth of 30, minimum split size of 40, and minimum bucket size of 20 (default value). Performance of the model, in terms of precision, recall, and overall accuracy, was evaluated on the testing split.

6.2.9 Population-biased macrosatellites

We applied two approaches to find macrosatellite loci with different `MaSUD` genotyping distributions among the super-populations:

1. **ANOVA model**. An ANOVA model was used to detect differences of average copy number changes predicted by `MaSUD` across super-populations. We used the Tukey's honestly significant difference test (Tukey's HSD) (Tukey, 1949) for multiple comparisons among the super-population groups. Tukey's HSD reports a maximum difference and the adjusted p-value for the null hypothesis that the means in all the groups are the same. The Tukey's HSD p-values were FDR corrected and loci with FDR<5% and with predicted copy number difference of at least one were reported as population-biased.

2. **SHAP value.** SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), a method to measure feature importance. was used to find macrosatellites that are most predictive of super-population. We used the R package SHAPforxgboost V0.0.4 (Liu and Just, 2020) to calculate SHAP values. The top 1, 10, 20, 50, and 100 macrosatellites by SHAP value were used train a predictive model (similar to Section 6.2.8).

6.2.10 Association of macrosatellite copy number with nearby genes

To detect expression differences among individuals with different VNTR genotypes, mRNA expression counts from lymphoblastoid cell lines of 660 individuals by the Geuvadis consortium (Accession: E-GEUV-1) were downloaded (Lappalainen et al., 2013). A total of 445 individuals overlapped with the 2,504 NYGC genomes set. To exclude genes that are not expressed in blood, we excluded genes with median expression value equal to 0. Finally, we used log2 normalized TPM values to ensure comparability across genomes.

To control for confounders we used known covariates of sex and population and detected additional hidden covariates. To detect hidden factors such as unreported batch effects, we applied "Iteratively Adjusted Surrogate Variable Analysis" (Lee et al., 2018) on the log2 normalized TPM values. Observing that covariates six and above are over 85% correlated with other covariates (Figure 3·1), we chose five hidden factors to include in our model. Finally, we used a linear regression $expression \sim sex + population + hidden_factors$ to extract residuals to be used in the downstream association model.

Next, for each gene, we extracted the genotypes of VNTRs within 10 Kbp. When no genotype was observed for an individual, we classified the genotype as *other* (assuming that the actual alleles were outside the detection range of VNTRseek because genotypes were observed in other individuals with similar coverage). VNTR loci were retained for analysis if more than one genotype was detected for that VNTR across all individuals (at least three if *other* was one of the genotypes) and if each genotype was observed in at least 20 individuals.

For each gene-VNTR pair, we used an linear regression model as $residuals \sim genotype$ to detect if the mean of any genotype class is different from the others. The p-values of the linear models were adjusted using FDR. Any gene-VNTR pair with FDR<5% was reported.

Finally, we downloaded experiments on the GM12878 cell line from the ENCODE project (ENCODE Project Consortium, 2012) with histone markers or DNAse peaks. Narrow peaks in bed format on GRCh38 were downloaded (source IDs are given in Table 3.1). Any overlap of the peaks with the eqtl VNTRs were reported.

6.3 Results

The characteristics reference macrosatellites used in this study are described in Section 6.3.1. was tested vigorously datasets and then validated on real dataset. The performance of `MaSUD` was tested on simulate data (Section 6.3.2) and experimentally proven on real data (Section 6.3.3). An overview of macrosatellites genotyped in a large cohort of 2,504 unrelated individuals from NYGC is given in Section 6.3.4. Sections 6.3.5 and 6.3.6 present an analysis of population-biased macrosatellites. Finally, Section 6.3.7 presents the macrosatellites associated with gene expression.

6.3.1 Characteristics of the macrosatellite reference set

Tandem repeats on the human reference genome assembly GRCh38 were downloaded from TRDB (Gelfand et al., 2007). A total of 4,292 (3,877 with array length $\geq$300 bp) macrosatellite loci were extracted from TRDB that comprised 2,524,991 base pairs (0.1%) of the reference genome. Figure 6·3 summarizes the characteristics of the reference macrosatellites. The pattern length varied from 100 bp to 1,994 bp (Figure 6·4a) and the copy number ranged from two to 120 (Figure 6·4c). Overall the array sizes were between 178 bp and 57,978 bp (Figure 6·4b).

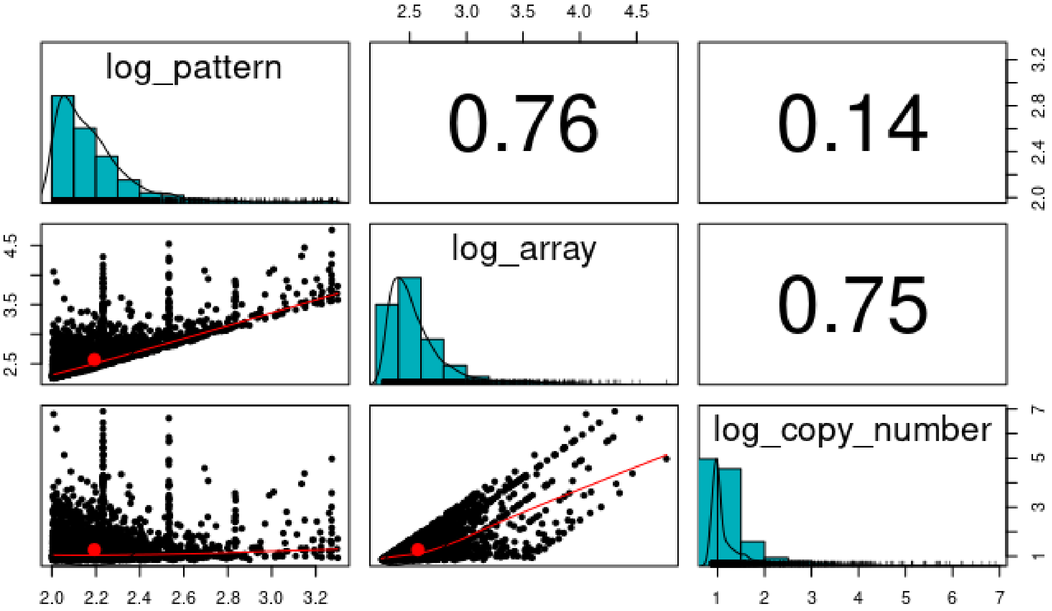
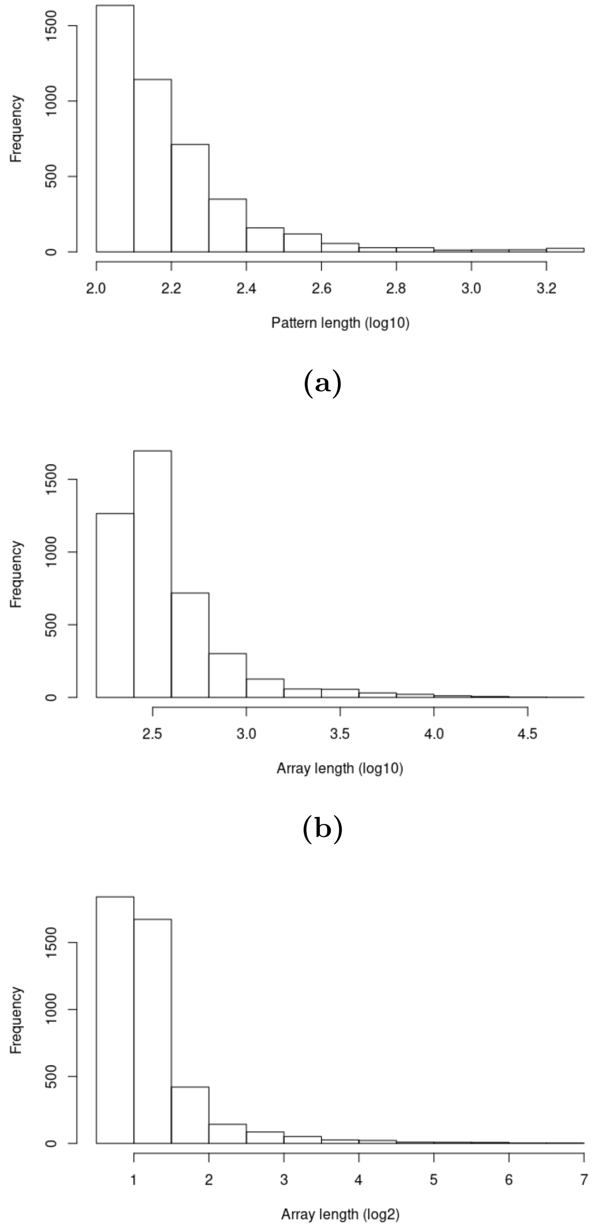


Figure 6·3: Correlation of macrosatellites characteristics.



(a)

(b)

(c)

Figure 6·4: Distribution of pattern length, array length and copy number of macrosatellites used in this study.

The macrosatellite loci were located mainly in intergenic regions and introns (Figure 6·5). They overlap with 2,114 genes (1,492 protein coding), including 677 exons. The protein coding genes overlapping with the macrosatellites were enriched in pathways related to immunity and neuron function. The genes were enriched in gene signatures of various midbrain cell types, adult kidney cells, and pancreas cells (all FDR$< 10^{-10}$). Among the biological functions were neurogenesis, neuron development, neuron differentiation, synaptic signaling, ion transport, cell morphosis and organization, and transportation of small molecules in the membrane. These enrichment results suggest macrosatellites play a role in brain and neuron function.

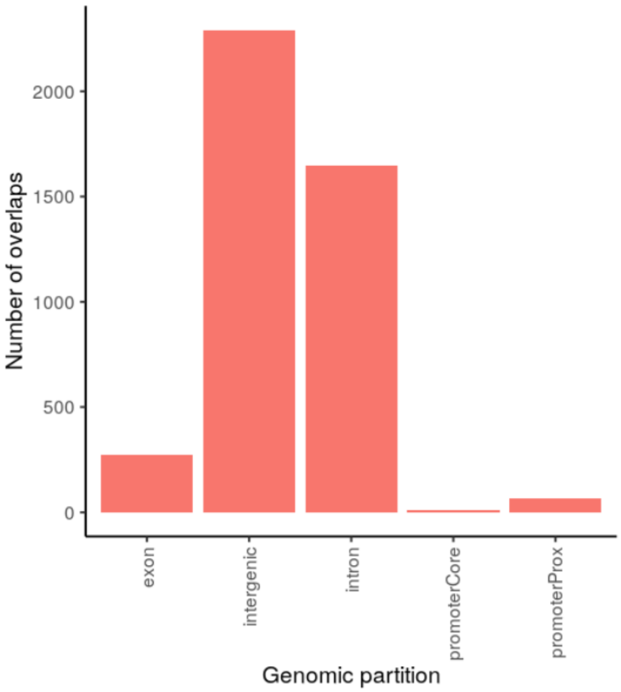


Figure 6·5: Annotation of macrosatellite loci on the human reference genome.

6.3.2 Performance of MaSUD on simulation results

Total 34,428 macrosatellites were simulated in thirteen datasets (see Section 6.2.2). The prediction results of MaSUD on simulated datasets compared to the true copy number change is plotted in Figure 6·6 (Figure 6·7 and Figure 6·8 for ΔC and FC, respectively). Overall, the ΔC and FC estimates were 99% correlated (Figure 6·9). The performance of MaSUD on simulated data was measured in three ways:

1. Using a linear regression (Truth $\sim$ MaSUD): the r-squared value was 0.90 for MaSUD (0.88 and 0.91 for ΔC and FC measurements, respectively). The RMSE was 1.37 (1.66 and 1.34 for ΔC and FC, respectively). The correlation between the MaSUD prediction and the true copy number changes was 95% (94% and 95% for ΔC and FC measurements, respectively). The first and third quartile of the residual values was -0.48 and -0.01, respectively with median -0.22 (median was -0.47 for ΔC and -0.28 for FC).

2. By examining whether the true copy change is inside the ΔC and FC distributions (within three standard deviations): Overall, for 85% of the predictions, the true change was within the normal range of the MaSUD predictions (75% and 78% for ΔC and FC, respectively). The standard deviation of the predictions are shown in Figure 6·10.

3. By converting the problem to a classification problem to predict loss or gain: If zero was within three standard deviations of the prediction, we considered that *NA*, meaning no loss or gain. Otherwise, if the prediction was positive, we consider it a gain and otherwise, a loss. MaSUD precision was 96% and 92% for gain and loss, respectively and the recall was 95% and 96%. Overall accuracy was 96%. FC had lower precision for losses in comparison to ΔC, and that is why we use ΔC values for predicting losses (See Table 6.6).

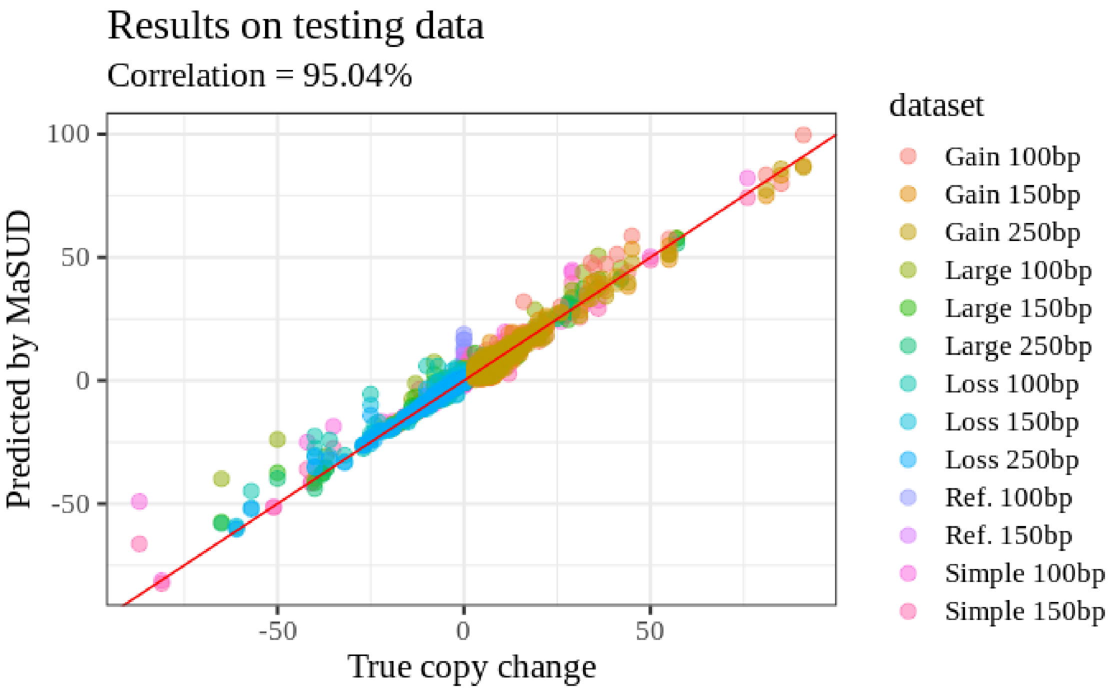


Figure 6·6: `MaSUD` predictions on simulated data. The copy number changes (Y-axis) as a function of the true copy number change implanted (X-axis) are plotted. Total 34,428 predictions are shown for 13 datasets (different colors). Red line is the y=x line.

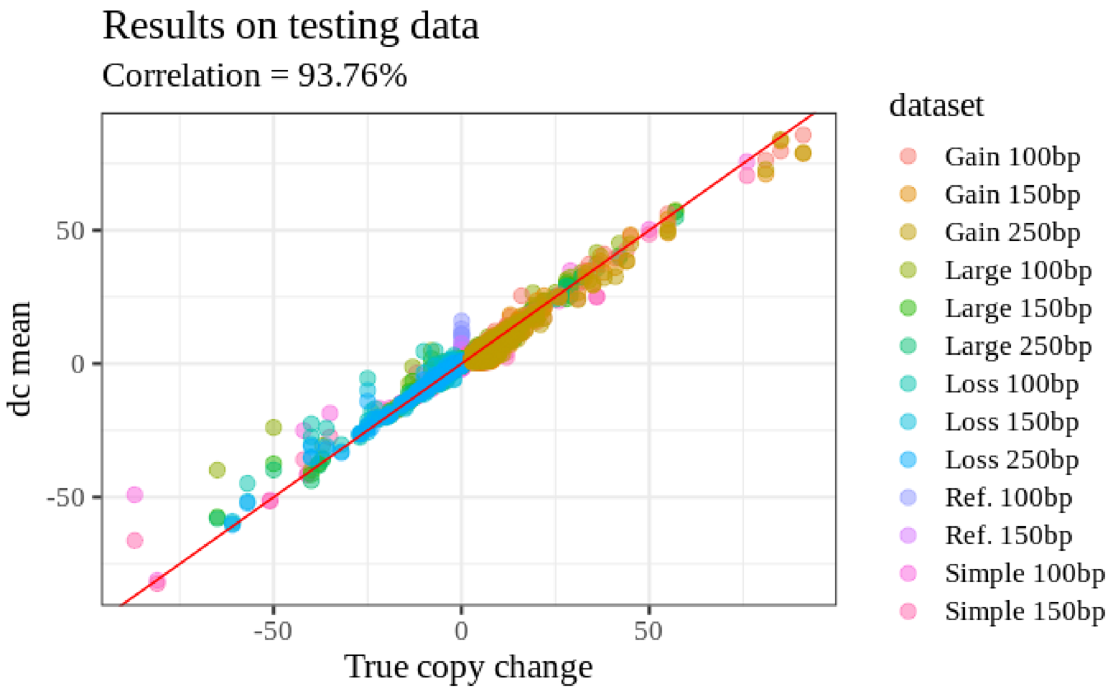


Figure 6·7: Correlation of the ΔC estimate and the true copy change.

						Precision			Recall			
Dataset	N	R^2	RMSE	Cor.	Correct (3sd)	Gain	Loss	NA	Gain	Loss	Ref.	Accuracy
Gain 100bp	4,119	0.85	3.42	92%	5%	100%	N/A	N/A	100%	N/A	N/A	100%
Gain 150bp	2,244	0.89	3.37	94%	8%	100%	N/A	N/A	100%	N/A	N/A	100%
Gain 250bp	871	0.94	3.20	97%	16%	100%	N/A	N/A	100%	N/A	N/A	100%
Large 100bp	4,119	0.93	0.74	97%	94%	50%	66%	100%	100%	96%	95%	95%
Large 150bp	2,244	0.98	0.51	99%	94%	77%	72%	100%	100%	99%	96%	96%
Large 250bp	871	0.99	0.70	99%	92%	89%	96%	100%	99%	100%	97%	97%
Loss 100bp	4,119	0.93	0.71	97%	95%	N/A	90%	100%	N/A	97%	96%	96%
Loss 150bp	2,244	0.97	0.61	99%	95%	N/A	98%	99%	N/A	98%	97%	97%
Loss 250bp	871	0.98	0.80	99%	81%	N/A	98%	96%	N/A	97%	94%	96%
Ref. 100bp	4,119	N/A	0.58	N/A	95%	N/A	N/A	95%	N/A	N/A	95%	95%
Ref. 150bp	2,244	N/A	0.32	N/A	95%	N/A	N/A	95%	N/A	N/A	95%	95%
Simple 100bp	4,119	0.95	0.82	97%	90%	96%	97%	93%	94%	93%	96%	95%
Simple 150bp	2,244	0.98	0.66	99%	82%	97%	97%	94%	98%	94%	96%	96%

Table 6.4: Performance of ΔC for each simulated dataset

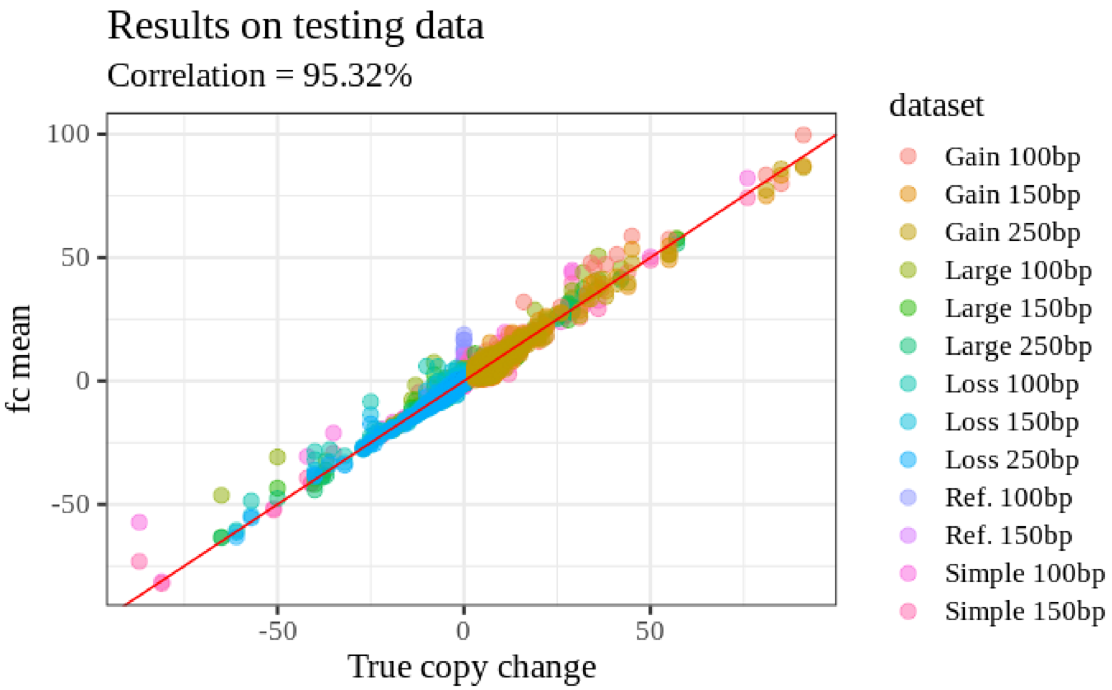


Figure 6·8: Correlation of the FC estimate and the true copy change.

						Precision			Recall			
Dataset	N	R^2	RMSE	Cor.	Correct (3sd)	Gain	Loss	NA	Gain	Loss	Ref.	Accuracy
Gain 100bp	4,119	0.83	2.68	91%	34%	100%	N/A	N/A	100%	N/A	N/A	100%
Gain 150bp	2,244	0.87	2.47	94%	47%	100%	N/A	N/A	100%	N/A	N/A	100%
Gain 250bp	871	0.93	2.27	96%	50%	100%	N/A	N/A	100%	N/A	N/A	100%
Large 100bp	4,119	0.93	0.78	96%	94%	68%	53%	100%	99%	96%	96%	96%
Large 150bp	2,244	0.98	0.53	99%	92%	80%	61%	100%	100%	99%	95%	95%
Large 250bp	871	0.99	0.80	99%	88%	91%	96%	96%	91%	91%	97%	95%
Loss 100bp	4,119	0.93	0.67	97%	88%	N/A	83%	100%	N/A	97%	95%	96%
Loss 150bp	2,244	0.97	0.6	98%	79%	N/A	93%	100%	N/A	99%	96%	97%
Loss 250bp	871	0.98	0.75	99%	61%	N/A	97%	97%	N/A	98%	93%	96%
Ref. 100bp	4,119	N/A	0.77	N/A	96%	N/A	N/A	96%	N/A	N/A	96%	96%
Ref. 150bp	2,244	N/A	0.45	N/A	95%	N/A	N/A	95%	N/A	N/A	95%	95%
Simple 100bp	4,119	0.94	0.87	97%	85%	97%	93%	86%	79%	96%	95%	91%
Simple 150bp	2,244	0.97	0.73	99%	76%	97%	95%	95%	96%	97%	94%	95%

Table 6.5: Performance of FC for each simulated dataset.

ΔC	Gain	Loss	Ref	Precision
Gain	8,943	18	336	96%
Loss	0	3,880	342	92%
NA	459	156	20,294	97%
Recall	95%	96%	97%	

(a) MaSUD

ΔC	Gain	Loss	Ref	Precision
Gain	9,318	19	569	99%
Loss	0	3,880	342	96%
NA	84	155	20,061	96%
Recall	99%	96%	96%	

(b) ΔC

FC	Gain	Loss	Ref	Precision
Gain	8,943	18	336	96%
Loss	0	3,945	653	86%
NA	459	91	19,983	97%
Recall	95%	97%	95%	

(c) FC

Table 6.6: Confusion matrix for MaSUD, ΔC and FC measurements on simulated data. For three class of *Ref* (no change), *Gain* (gains of copies), and *Loss* (loss of copies) we calculated the precision and recall. Columns are labeled with true labels and rows are labeled with predictions. If the $\mu \pm 3\sigma$ interval of ΔC (or FC) included zero, the prediction was considered *NA*. If the interval was strictly >0 it was considered Gain, and otherwise Loss. Total number of predictions was 34,428. The accuracy by ΔC and FC was 97% and 95%, respectively.

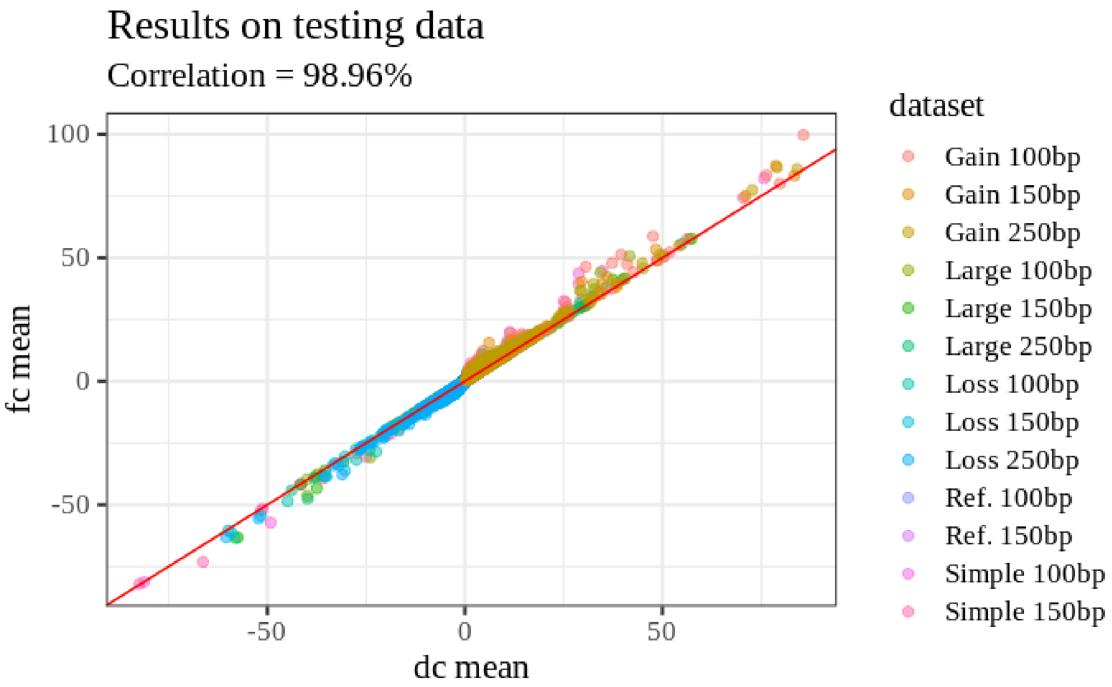


Figure 6·9: Correlation of ΔC and FC on simulated data.

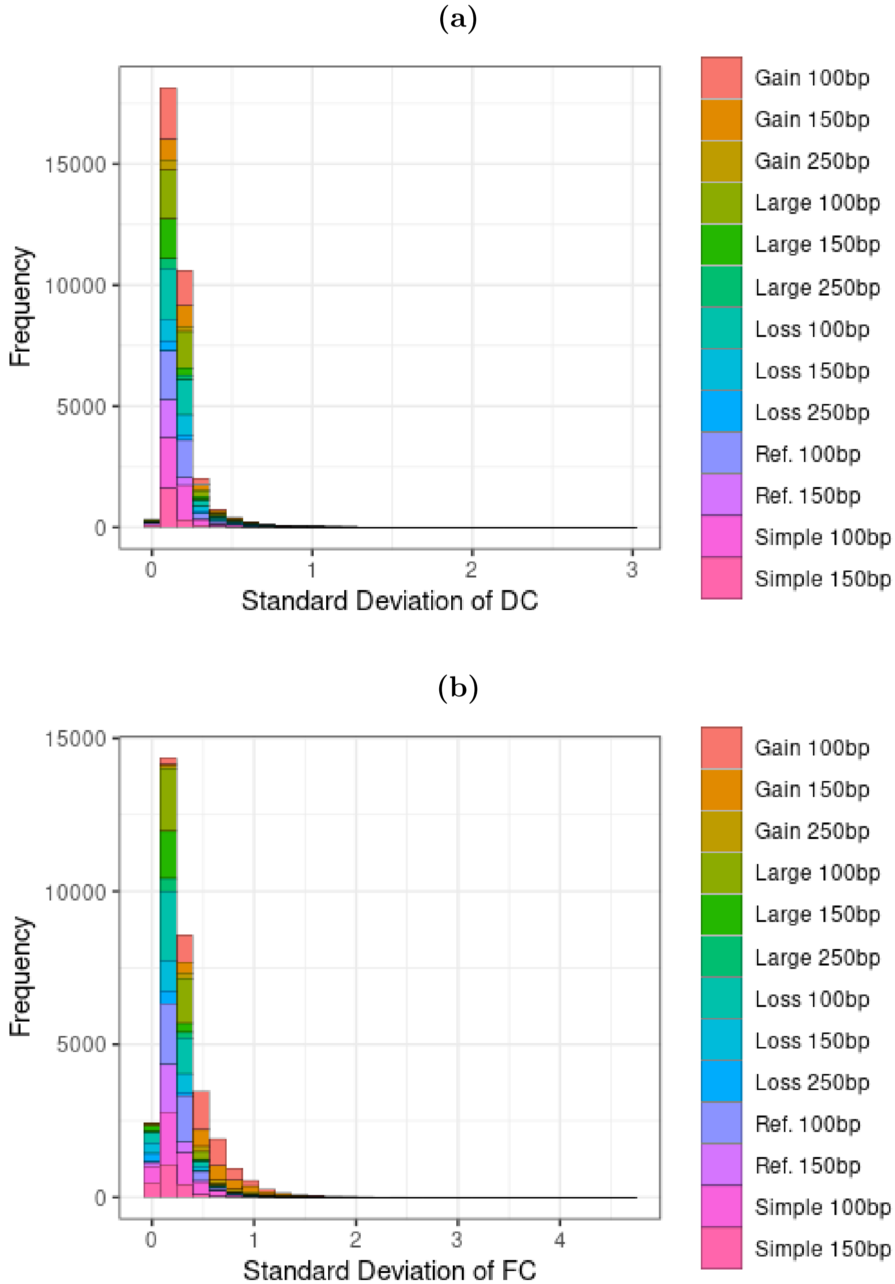
Table 6.8 shows the effect of dataset characteristics of `MaSUD` error. Higher coverage reduced the error because the statistical power increases with the coverage. Similarly, as the pattern length increased the error of `MaSUD` decreased, because the change is more dramatic and easier to detect. The read size had a slight negative effect on the performance. Smaller reads provide higher resolution and allow for smaller changes to be detected with higher precision.

Dataset	Estimate Std.	Error	t value	$\mathrm{Pr}(>\lvert t\rvert)$	
Gain 100bp	2.34	0.01	172	<2e-16	***
Gain 150bp	2.11	0.02	-10	<2e-16	***
Gain 250bp	2.76	0.03	-13	<2e-16	***
Large 100bp	0.29	0.02	-106	<2e-16	***
Large 150bp	0.22	0.02	-92	<2e-16	***
Large 250bp	0.46	0.03	-58	<2e-16	***
Loss 100bp	0.26	0.02	-108	<2e-16	***
Loss 150bp	0.30	0.02	-89	<2e-16	***
Loss 250bp	0.44	0.03	-58	<2e-16	***
Ref. 100bp	0.28	0.02	-107	<2e-16	***
Ref. 150bp	0.19	0.02	-94	<2e-16	***
Simple 100bp	0.37	0.02	-103	<2e-16	***
Simple 150bp	0.38	0.02	-86	<2e-16	***

Table 6.7: Error by dataset.

	Estimate	Std. Error	t value	$\mathrm{Pr}(>\lvert t\rvert)$	
(Intercept)	5.53E-01	1.94E-02	29	<2e-16	***
read	6.82E-03	2.10E-04	32	<2e-16	***
coverage	-1.46E-02	3.24E-04	-45	<2e-16	***
copy_number	2.11E-02	1.45E-03	15	<2e-16	***
pattern	-3.17E-04	3.56E-05	-9	<2e-16	***
copy_number:pattern	1.22E-04	4.88E-06	25	<2e-16	***

Table 6.8: Effect of dataset characteristics on error.
model used was: $abs(truth - MaSUD) \sim read + coverage + pattern \times copy\_number$


(a)

(b)

Figure 6·10: Standard deviation of ΔC and FC for each dataset.

6.3.3 Confirmation on real data

Confirmation on the Y chromosomes of related genomes: In the GIAB datasets, the genomes of the sons should have the same copy number as the father. Using a linear model *son father* had an estimate of 94% (p-value <2e-16).

Experimental validation of `MaSUD` predictions. To test `MaSUD` prediction, ten loci were experimentally validated on the Ashkenazi Jew trio from GIAB by Samantha D. Drinan (Fuxman lab). For each loci and individual, we interpolated the band size using the ladder. To compare the observed bands to the predicted values, When two bands were present for a given loci in an individual, the average of both band sizes were used (Table 6.9). The predictions were plotted against the observed band size or the average of two bands when present (Figure 6·11). The R^2 was 0.91 with slope 0.93 and error of 0.05.

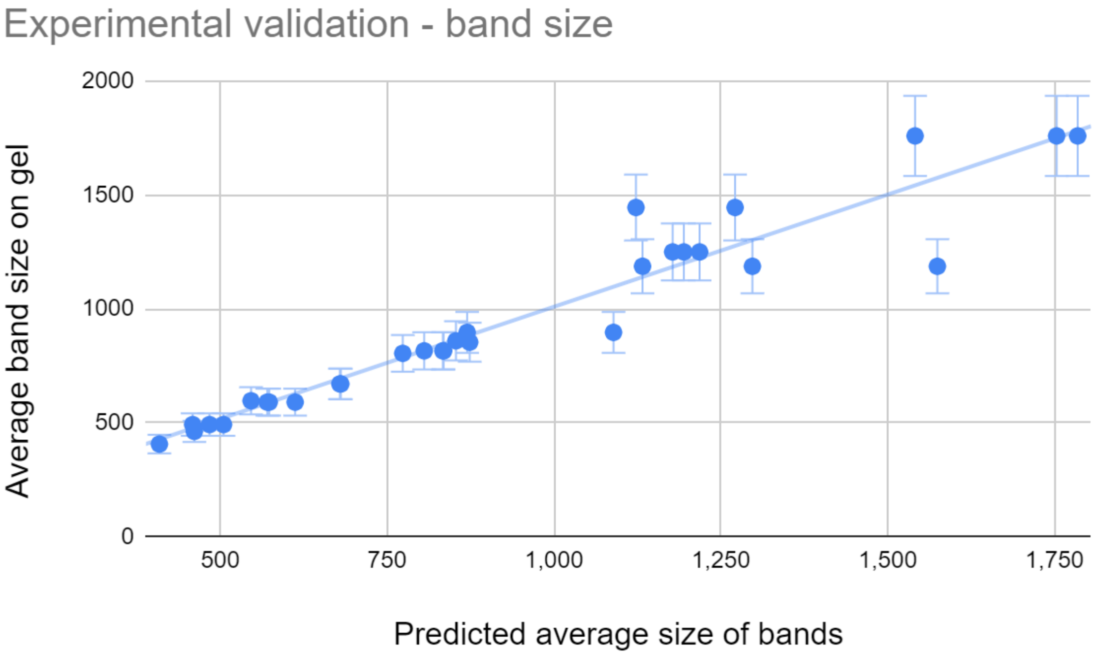


Figure 6·11: Predicted vs. observed band sizes of macrosatellites. $r^2 = 0.907$

#	Ref. copy no.	Ref. array	Ref. pattern	Prediction Child	Prediction Father	Prediction Mother
MS1	6.03	837	138	-2.14	-2.15	-0.89
MS2	6.00	660	110	-2.41	-2.60	-2.83
MS3	5.07	568	112	-2.78	-2.52	-2.53
MS4	4.01	571	144	-2.08	-2.06	-1.79
MS5	6.60	1,108	168	2.20	0.55	-0.43
MS6	2.35	1214	518	0.32	0.26	-0.15
MS7	2.50	668	267	-0.13	1.06	0.24
MS8	6.00	1,940	327	-4.66	-4.82	-4.40
MS9	3.28	1,008	308	0.14	0.01	0.06
MS10	9.60	1,910	165	-3.87	-6.28	-4.77

#	Band size Child	Child genotype	Band size Father	Father genotype	Band size Mother	Mother genotype
MS1	671	-2/-2	671	-2/-2	671/1,049	-2/0
MS2	429/556	-3/-2	429/556	-3/-2	429/556	-3/-3
MS3	794/838	-3/-2	794/838	-3/-2	794/838	-2.69
MS4	443/739	-3/0	443/739	-3/0	443/739	-1.94
MS5	1,187	0	1,187	0	1,187	-0.11
MS6	1,759	0	1,759	0	1,759	0.27
MS7	805	0	897	0	897	0.34
MS8	407/518	-5/-4	407	-5/-5	518/676	-4.24
MS9	1,250	0	1,250	0	1,250	0.24
MS10	774/2,116	-7/+1	774/934	-7/-6	774/2,116	-6/+1

Table 6.9: Validation results of MaSUD.

Then using the estimated average band sizes, the average change in copy number was calculated and compared to the MaSUD predictions (Figure 6·12). Similar to Section 6.3.2, we evaluate the accuracy of the validated prediction in three ways:

1. Using a linear regression ($Truth \sim MaSUD$), the estimate was 1.06, the error was 0.06, the R-squared was 0.92, and the RMSE was 0.62. The correlation between the predictions and the validations was 0.95. The first, second, and third quartile of the residuals were -0.21, -0.03, and 0.22, respectively and the mean of residuals was 0.00.
2. By classifying if the prediction range with *three* standard deviation included the true copy number change, accuracy was 100%. A total of 28/30 predictions were within *two* standard deviations.
3. By considering the problem a classification of three categories: gain, loss, and reference, all predictions were correct and accuracy was 100%.

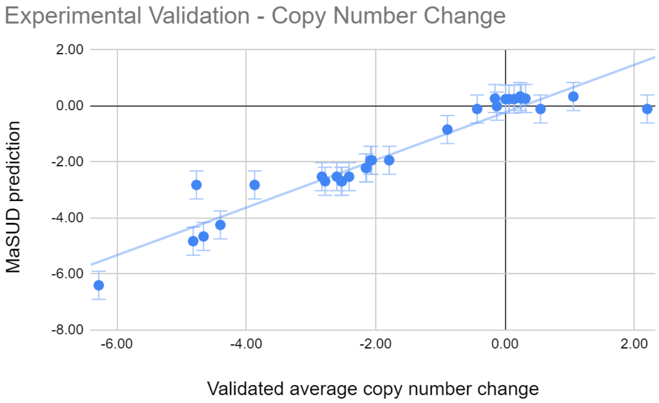
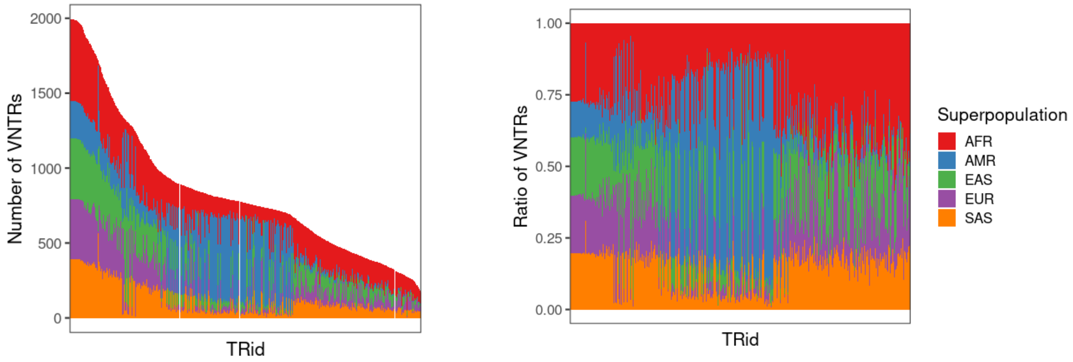


Figure 6·12: Validation results of macrosatellites. The X-axis is the observed copy number change estimated from the agerose gel. This was performed by dividing the average difference of bands from the expected reference band by the pattern size. Y-axis is the MaSUD prediction. The error bars are 10% to represent the agerose gel error rate. The line is the Y∼X regression line with $R^2 = 0.917$.

6.3.4 Results of MaSUD on NYGC

On 2,137 reference macrosatellites in the NYGC samples, a total of 1,084,869 (19%) had gain of copies, 1,521,923 (27%) has loss of copies, and 3,012,184 (54%) were unchanged (reference genotype). Figure 6·13 shows the number of samples that were non-reference per macrosatellite locus. Some loci are more variable in certain populations, e.g. Europeans (shown in purple) have the least number of variable loci, probably because of the reference being mostly European. In 21 loci, the reference was never seen. Minimum number of VNTRs per macrosatellite loci was 474.



(a) Frequency

(b) Proportion

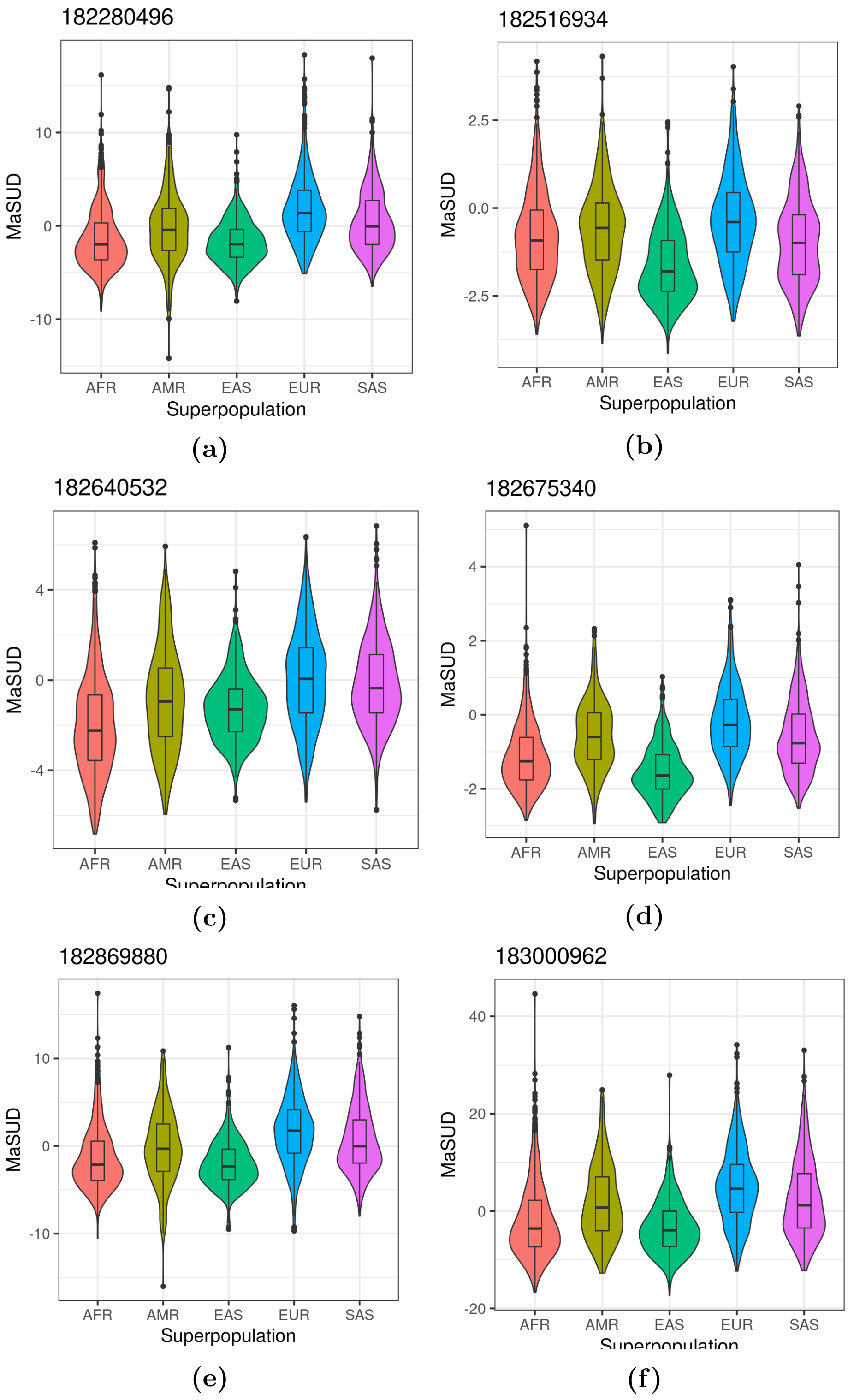
Figure 6·13: Number of VNTRs per macrosatellite locus. In the NYGC genomes, the number of samples with non-reference alleles were counted for each macrosatellite locus. The X-axis are the macrosatellite loci sorted by the total number of samples with VNTRs. The different populations are shown in different colors. The height of each bar represents the number of samples with non-reference genotype. *Left* plot shows the frequency of VNTRs per locus. *Right* plot shows the relative proportions of VNTRs per locus.

6.3.5 Prediction of Super-population using MaSUD

To examine whether the macrosatellite copy numbers are predictive of the super-population, I applied unsupervised clustering using Principal Component Analysis (PCA) on the NYGC genomes. The first 30 principal components were used (capturing 61% of the variation) to train a decision tree on 75% of the data and tested it on the remaining 25%. The decision tree could predict the super-population with 93% accuracy (Table 6.10).

	AFR	AMR	EAS	EUR	SAS	Precision
AFR	658	6	1	1	4	98%
AMR	2	294	10	28	19	83%
EAS	1	9	483	3	10	95%
EUR	0	23	1	453	16	92%
SAS	0	15	9	18	440	91%
Recall	100%	85%	96%	90%	90%	

Table 6.10: Confusion table of prediction model. The precision of each class (super-population) in shown in the last column. The recall is calculated in the last row. The overall accuracy was 93%.



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 6·14: Population specific loci by ANOVA.

6.3.6 Population-biased VNTRs

We detected population specific macrosatellite loci in two ways: *first*, using an ANOVA model to compare the mean `MaSUD` values of each super-population ($MaSUD \sim Superpopulation$), and *second* by SHAP values (see methods). Using the first approach, a total of 397 loci had population-bias with a minimum difference of at least one copy change. These loci could predict the super-population with 87% accuracy. Figure 6·14 shows the loci with the highest difference in mean compared to the overall mean.

In the *second* approach, we used SHAP values to find population-speific macrosatellites. ShAP values methodology to determine feature importance in a predictive model usin Gradient Boosted Trees (Liu and Just, 2020). The top macrosatellite loci by SHAP value was id=182571530, with reference pattern length 105 bp and copy number 3.07 located at chr17: 401846-402168 (hg38) (Figure 6·16).

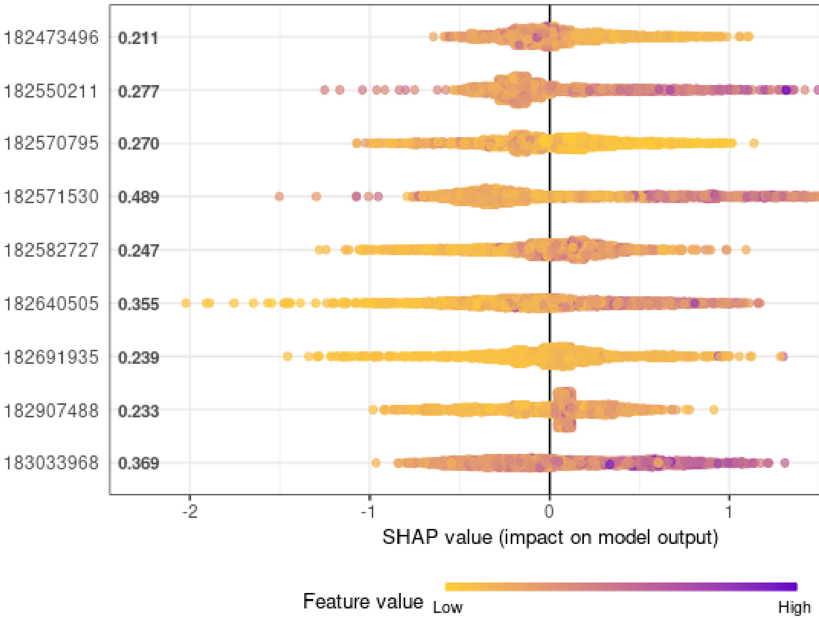


Figure 6·15: Top features by SHAP values.

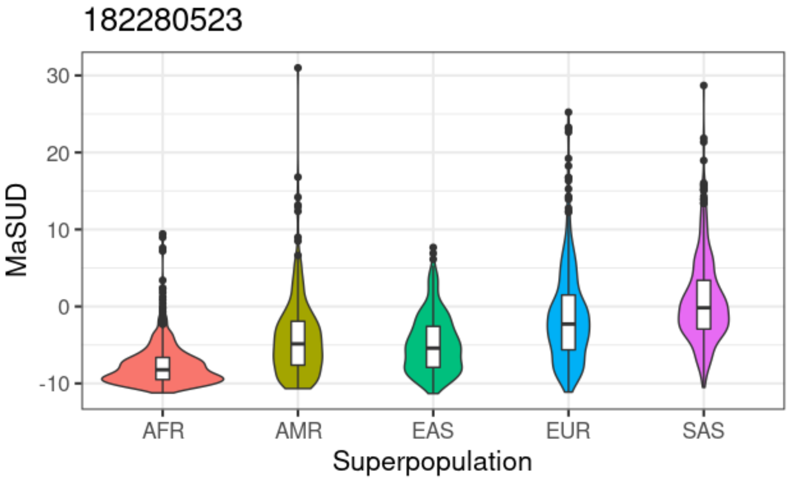


Figure 6·16: Macrosatellite 182218798 has the highest predictive value. The violin plot shows the distribution of `MaSUD` prediction for each super-population. The macrosatellite with the highest SHAP value (0.489) could predict super-population ancestry with 63% accuracy. The macrosatellite (id 182218798) had reference pattern length 105 bp and copy number 3 and is located in at chr17:400,531-411,263.

There were 2,245 macrosatellites with SHAP values with mean impact score of at least 0.1. Using the MaSUD predictions of these loci, I could predict the superpopulation class on the test dataset with 95% accuracy (Table 6.11). The top 10, 20, 50, and 100 macrosatellite loci by SHAP value, the super-population prediction accuracy was 82%, 86%, 87%, and 88%, respectively (not shown).

6.3.7 Association of macrosatellite genotypes with gene expression

To find the association of macrosatellites on gene expression, we used gene expression in blood tissue of 445 indivuals from the Geuvadis 1 dataset. A regression model was used to remove the effect of known covariates (i.e. sex and common SNPs) and unknown covariates which were calculated using SVA. Then, a regression model was used to examine association between the residuals of the gene expression with MaSUD predictions. A total of 979 macrosatellites-gene pairs were tested, out of which 23 macrosatellites were in eQTL with at FDR <5%. Figure 6·17 shows the eQTL

	True AFR	True EAS	True EUR	True SAS	Class precision
Pred. AFR	179	1	3	0	98%
Pred. EAS	0	124	0	2	98%
Pred. EUR	0	1	113	13	89%
Pred. SAS	0	1	6	105	94%
Class recall	98%	98%	89%	94%	

Table 6.11: Prediction of superpopulation using features with SHAP value >0.1. Using the macrosatellites with SHAP value >0.1, an SVM model was trained on 75% of the samples. The model was tested on the remaining 25% and the class precision and recall was calculated. The overall performance was 95%. Note that the individuals with American ancestry were excluded from the analysis, because they have admixed genomes.

macrosatellite genes. These genes included CDK11A, NBPF3, NBPF26, NBPF11, MUC1, LRRC27, TUBGCP2, AL442125.2, GAS6-AS1, NEO1, OSGIN1, ACSF3, CPNE7, P2RX5, RBFADN, INSIG2, AP001056.2, CEP63, SLC9A3-AS1, SLC9A3, DUSP22, and KIF25-AS1. CDK11A have been associated to neuroblastoma (Duan et al., 2012) as well as Neuroblastoma Breakpoint Family Members NBPF3, NBPF11 and NBPF26. The MUC1 macrosatellite is a known marker for cancer therapy (Singh and Bandyopadhyay, 2007).

The top two macrosatellite eQTLs by FDR are shown in Figures 6·18 and 6·19. Figure 6·18 shows the association of NBF26 with macrosatellite 182180200. This macrosatellite spans the NBF26 gene and has pattern length 1,559 bp and reference copy number 2.4. `MaSUD` predicted changes from loss of one copy to gain of 10 copies. Most individuals from the Yoruban population had loss of one copy. In another example, Figure 6·19 shows the association of CDK11A with macrosatellite 182169588. This macrosatellite is located in the fourth intron of CDK11A and has a pattern length 483 bp and reference copy number 2.2. `MaSUD` predicted up to two copies of gain or loss across the 445 individuals.

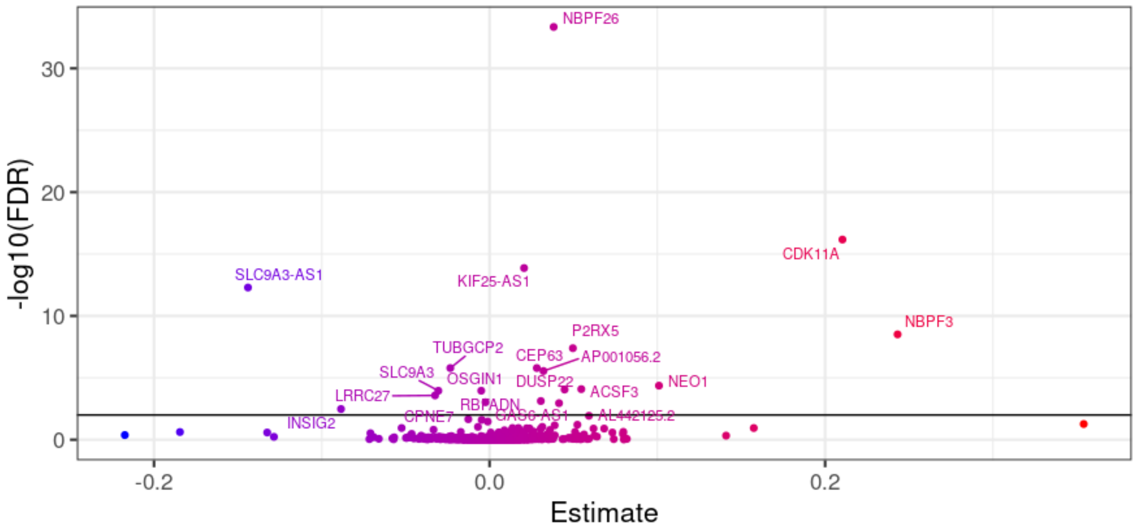


Figure 6·17: Macrosatellite eQTLs. A regression model was used to determine the association of macrosatellite genotypes with the gene expression. The *Estimate* of this model is shown on the X-axis, and the -log10 FDR value on the Y-axis. The horizontal line is the 5% cutoff for FDR. A total of 22 macrosatellite eQTLs were deteced.

6.4 Summary

In this chapter a novel computational tool, `MaSUD`, was presented that genotypes macrosatellite copy number changes using short WGS reads. `MaSUD` was tested vigorously on simulated data and shown to have 90% accuracy with an average error of half a copy. To further evaluate the accuracy of `MaSUD`, it was applied to real data of related genomes. The macrosatellite predictions on the Y chromosome of father and sons agreed with 94% accuracy. A random set of 10 loci were also tested on three related genomes and the accuracy of copy number change was 92% with an error of 0.6 copies. However, agerose gels also have about 10% error, which was not considered in our validation. Overall, `MaSUD` performs with high accuracy.

To characterize 4,292 macrosatellites in human populations, `MaSUD` was applied to a cohort of 2,504 unrelated individuals with ancestries from five super-populations. On average 474 VNTRs per sample were observed, which consisted of 19% gains,

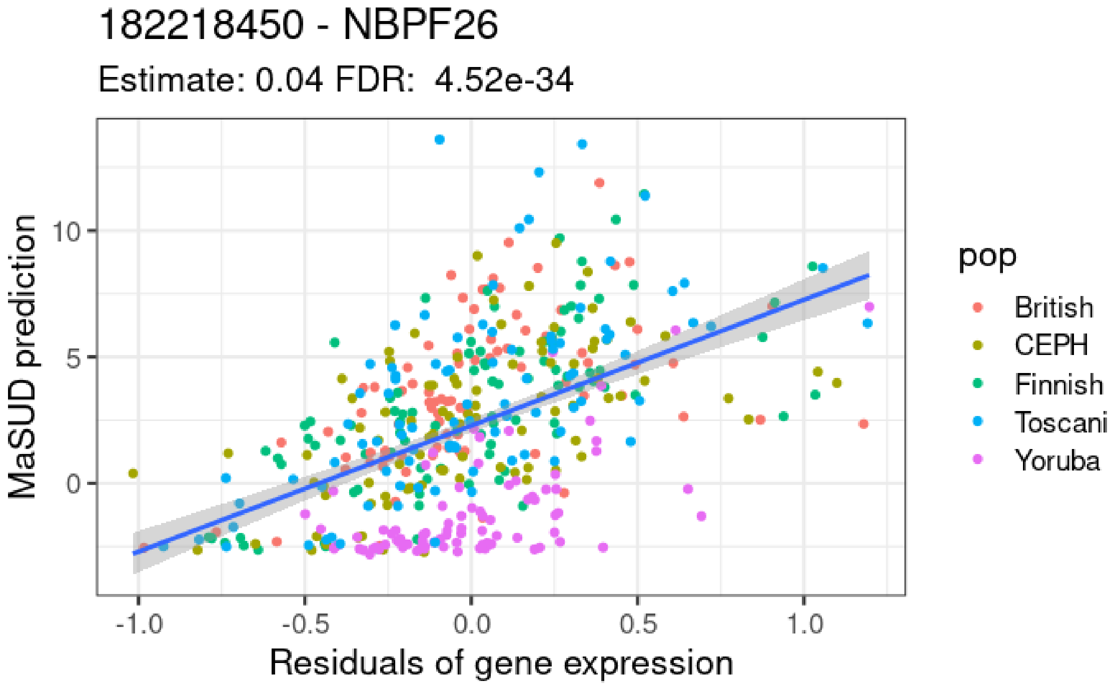


Figure 6·18: Association of macrosatellite 182180200 with NBPF26.

27% loss and 54$ reference alleles. Similar to minisatellites and microsatellites, in macrosatellite VNTRs loss of copied is more common than gain.

A total of 2,245 macrosatellites genotypes were population-specific and using these loci, the ancestry of the individuals could be predicted with 95% accuracy. In blood tissue, a total of 21 macrosatellites were in eQTL with 22 genes. These results show that macrosatellites are highly polymorphic and can be used in GWAS studies to further the knowledge of germline variation.

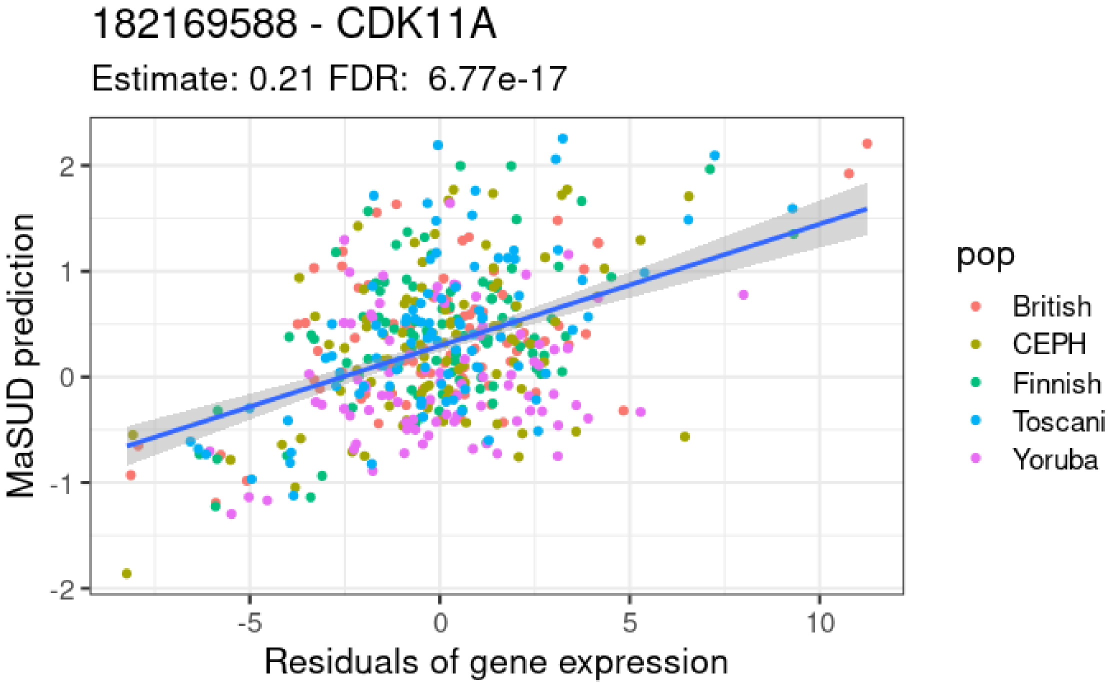


Figure 6·19: Association of macrosatellite 182169588 with CDK11A. This macrosatellite is located in the intron 4 of CDK11A, has pattern length 483 bp and reference copy 2.2.

Chapter 7

Conclusions and future work

In this thesis I present the most comprehensive analysis of human copy number variable tandem repeats (VNTRs), based on currently available whole genome sequencing data., This research greatly expands our knowledge of this class of genetic variation, whose important role in biology and human disease is becoming increasingly evident.

In Chapter 2, approximately191,000 minisatellites were genotyped. The pattern sizes ranged from seven to 100 bp. Our results reveal that nearly 20% (35,828) of the minisatellite loci in the human genome are variable. This percentage is in agreement with previous findings in the literature (Näslund et al., 2005). In comparison to other forms of variation, such as SNPs, minisatellite VNTRs have high heterozygosity (40%–60%) and many exhibit multiple alleles, which is consistent with previous predictions (Denoeud et al., 2003). The number of polymorphic minisatellite detected per genome depends on the statistical power (read length and coverage); however, at similar statistical power, the specific genetic population was the deterministic factor with Africans and East Asians exhibiting the highest number of VNTRs per individual, suggesting a greater evolutionary distance from the reference genome. On average, the African genomes had the highest ratio of heterozygous calls, which suggests higher diversity in the African population.

In Chapter 3, population-wide analyses of the minisatellite VNTRs in 2,504 unrelated genomes from five super-populations were presented. A total of 5,676 commonly polymorphic minisatellite VNTRs (VNTRs in at least 5% of the population) were identified. Each genome had, on average, 1,783 polymorphic minisatellites. These minisatellite loci overlap with 2,173 protein coding genes including 254 exons, and are generally enriched in upstream and downstream regions of genes, 3' UTRs, 5' UTRs, introns, exons, transcription factor binding site (TFBS) clusters, CpG islands, and DNAse clusters. Association of VNTR alleles with proximal gene expression in blood, showed that 193 minisatellite VNTRs had alleles that were correlated with nearby gene expression. Previous studies have also shown that minisatellite VNTRs affect gene expression in a tissue-specific fashion (Bakhtiari et al., 2020).

We also showed that minisatellite VNTRs can be used to predict ancestry with >98% accuracy (Table 3.5). One third of the minisatellite VNTRs had at least one population-biased allele (Figure 3·17). The VNTRs with population-biased alleles were enriched in gene sets with functions related to neuron function, and were to be gene markers of various mid-brain cell types.

In Chapter 4, the precision of VNTRseek predictions was evaluated. Minisatellite VNTR predictions were validated in vitro and in silico. A total of 66 alleles predicted by VNTRseek on three related genomes were tested experimentally using B-Lymphocyte DNA, and all predictions were confirmed. Twelve previously described experimental validations on the NA12878 genome (Bakhtiari et al., 2018) were compared against our predictions and 11 were predicted correctly and only one was incorrectly predicted by VNTRseek. The consistency of VNTRseek across genomes sequenced by two different sequencing platforms was measured: overall 76%-–91% of the predicted VNTR alleles were detected in both platforms. Consistency was also evaluated in four trios (child-father-mother) and only 15 inconsistent alleles out of 16,040 were found.

Chapter 5 presents a statistical tool named mlZ (machine learning on Z-scores) to predict whether genotyped variants are heterozygous or homozygous. Using mlZ, we were able to predict tens of thousands of new VNTR loci. Also, mlZ can identify erroneous genotypes. The performance of mlZ on simulated data is >95%, and the results are consistent across different sequencing platforms.

Chapter 6 focuses on macrosatellite TRs. Macrosatellites have pattern lengths of 100 bp or longer and are significantly longer than minisatellites, which makes them more difficult to genotype. A novel computation tool, MaSUD, to genotype macrosatellites was presented. At this time, MaSUD is the only computational tool capable of precisely genotyping macrosatellites in a high-throughput fashion using short WGS reads. The performance of MaSUD was demonstrated on various simulated datasets and validated in vitro and in silico on real data from related genomes. A genome-wide analysis of macrosatellites showed that over 2,000 macrosatellites have population-specific genotypes, which could be used to predict ancestry with 95% accuracy. In DNA from human blood cells, 22 macrosatellites were found to have genotypes associated with nearby genes. However, more studies are required to find eQTL macrosatellites in other cell and tissue types.

7.1 Comparison of minisatellite VNTR predictions to the literature

As discussed in Section 1.8.4, different characteristics of minisatellites have been used to predict their variability (Denoeud et al., 2003; Näslund et al., 2005). These two previous studies showed that the reference copy number and the similarity of repeat units are predictive of TR variability. In our study, of the 35,638 VNTRs detected in 2,504 unrelated genomes from NYGC, a regression model was applied to predict the variability of minisatellites using pattern length, reference copy number, array length, and annotation of the minisatellite loci. Variability was determined as the number of alleles observed for any minisatellite. We found that the reference copy number was the most predictive of the number of alleles (Table 7.1) with effect size 0.52 (p-value $< 10^{-16}$). TRs inside TFBS, CpG islands, and 1 Kbp upstream of genes were slightly more likely to have higher number of alleles. Figure 7·1 shows the distribution of the number of alleles per VNTR.

	Estimate	**Std. error**	**t value**	**p value**	**Significance**
(Intercept)	-3.79	0.04	-88.75	<2e-16	***
pattern	-0.01	0.00	-9.58	<2e-16	***
copy_number	**0.52**	0.00	107.53	<2e-16	***
TFBS	0.06	0.00	22.41	<2e-16	***
Upstream	0.07	0.01	12.53	<2e-16	***
Exon	-0.01	0.01	-1.17	2E-01	
Intron	-0.01	0.00	-2.33	2E-02	*
CpG	0.35	0.01	44.11	<2e-16	***
Conservation	**3.66**	0.04	87.88	<2e-16	***
pattern:copy_number	0.01	0.00	28.12	<2e-16	***

Table 7.1: The effect of reference TR characteristics on the TR variability.

Another way to examine variability is to consider the number of individuals that have non-reference alleles at a given minisatellite locus. We used a binary label to determine if a minisatellite is commonly polymorphic, i.e., polymorphic in >5% of the population, or private (<5% of the population). TRs with higher reference copy number are more likely to be commonly polymorphic (Table 7.2). Minisatellites overlapping TFBS, upstream of genes, and in CpG islands are more likely to be common VNTRs. In contrast, TRs inside exons are less likely to be common VNTRs.

7.2 Limitations of VNTRseek

At high read length (250 bp) and coverage (>30×), over 90% of the reference TRs were genotyped. Despite the high sensitivity of VNTRseek, our curation of VNTR loci has almost certainly produced an undercount or the total number VNTRs. This is true because VNTRseek requires that a given tandem array fit within a read of

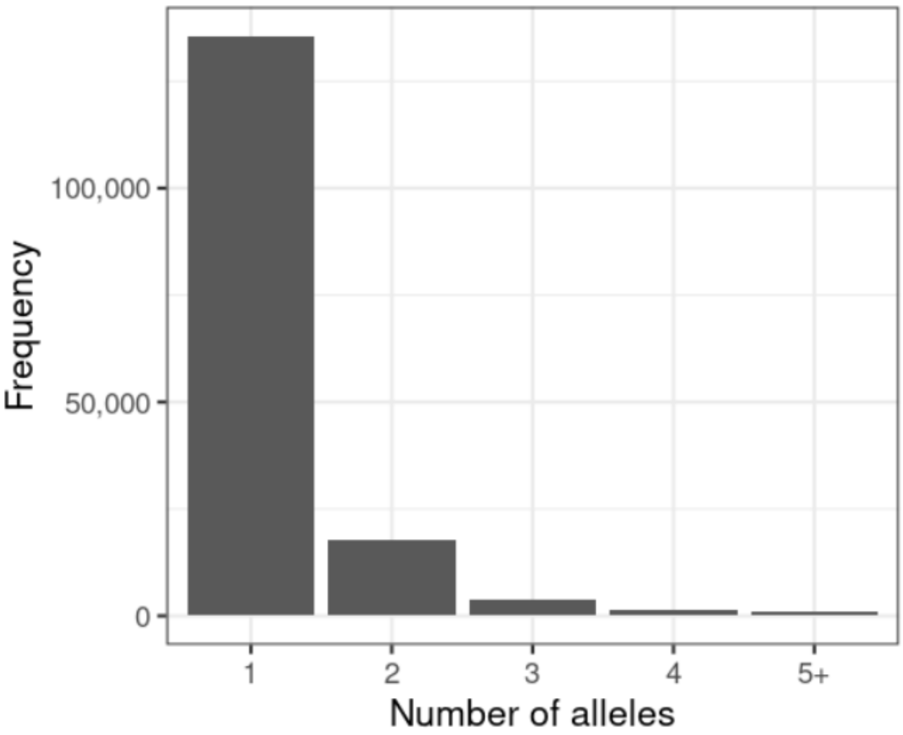


Figure 7·1: Distribution of number of alleles per VNTR loci.

250 bp. Longer reads can span longer TR arrays, but long reads are noisy with high indel rates (Korlach, 2013).

Another limitation of VNTRseek, which results in not detecting TR losses, comes from using the Tandem Repeat Finder (TRF) that is used to find TRs inside reads. TRF requires that the array contains *at least* 1.9 copies to be detected. At a read length of 150 bp, gain of one copy compared to the reference genome could be detected in 82% of the TR loci, whereas loss of one copy could be detected in only 16% of the reference TRs (Table 5.4). Despite this imbalance, loss of one copy was observed nearly 40% more often than gain of one copy (Figure 7·1). This suggests that loss of one copy is the most common allele in minisatellite VNTRs.

The data in Figure 7·2 suggest that loss of one copy is the most common allele, and is not an effect of bias in the VNTRseek detection range. Graph (a) in Figure 7·2

	Estimate	**Std. error**	**t value**	**p value**	**Significance**
	Estimate	Std. error	t value	p value	Significance
(Intercept)	-31.30	0.69	-45.09	<2e-16	***
pattern	-0.16	0.01	-19.76	<2e-16	***
copy_number	0.46	0.04	11.25	<2e-16	***
TFBS	0.30	0.04	7.40	1E-13	***
Upstream	0.34	0.07	4.91	9E-07	***
Exon	-0.59	0.18	-3.34	8E-04	***
Intron	-0.10	0.04	-2.61	9E-03	**
CpG	0.93	0.08	11.17	<2e-16	***
Conservation	25.90	0.68	38.06	<2e-16	***
pattern:copy_number	0.09	0.00	32.76	<2e-16	***

Table 7.2: The effect of characteristics of reference TRs on the likelihood of common VNTRs.

shows information on the entire NYGC dataset with read length 150 bp. The upper subgraph is a histogram showing the number of non-reference alleles detected, grouped by the reference array length. Bin sizes are 1 bp. The data have been filtered so that only those reference loci are shown in which both gain and loss of one copy could be observed. Copy number change of ±1 is by far the most common as seen in Figure 2·4.) The lower subgraph counts, for each reference array length, the number of gain alleles (increase in copy number) minus the number of loss alleles (decrease in copy number). Negative numbers (aqua) indicate that more losses than gains were observed for loci at that reference array length, and the more negative, the greater the bias towards loss. As can be seen, loss with respect to the reference is much more common than gain, except for the shortest reference array lengths. Graph (b) shows this same trend 250 bp reads from the 1000 Genomes data. The same can be observed in individual data as shown in graphs(c) and (d). The bar graph (e) shows that this trend persists in the entire NYGC dataset if one considers loci for which both gain and loss of 1, 2, 3, etc. copies is observed.

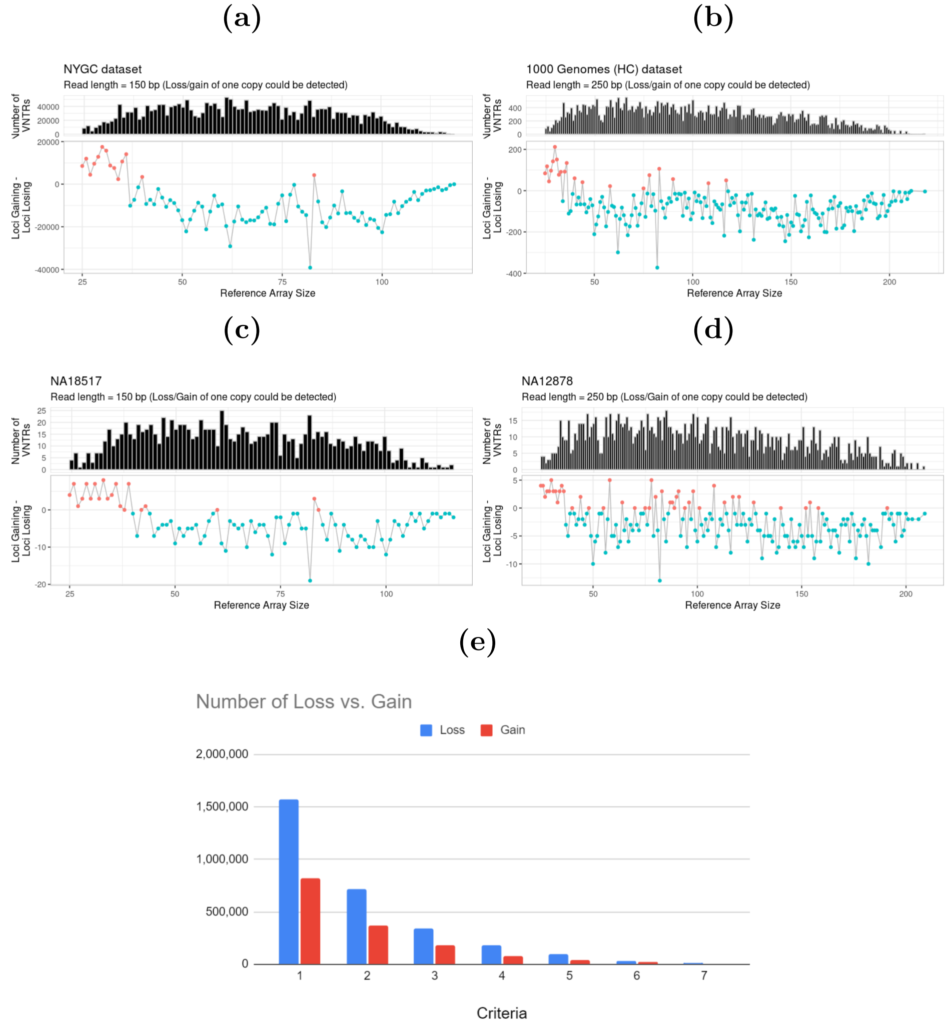


Figure 7·2: Copy gain versus loss relative to the reference. In the upper quartet of graphs, the upper subgraph is a histogram showing the number of non-reference alleles detected for those loci where both a gain and loss of one copy could be observed. Bin sizes are 1 bp. The lower subgraph counts, for each reference array length, the number of gain alleles (increase in copy number) minus the number of loss alleles (decrease in copy number). Negative numbers (aqua) indicate more loss than gain. (a) the entire NYGC dataset, (b) the entire 1000 Genomes dataset, (c) NA18517 (150 bp), (d) NA12878 (250 bp), which is labeled HG001 in GIAB. Graph (e) shows, the aggregate excess of loss over gain for the entire NYGC dataset, when considering loci for which both gain and loss of 1, 2, 3, etc. copies could be observed.

Observing more loss than gain has also been observed in previous studies. Bakhtiari et. al. (2018) genotyped approximately 10,000 loci in the NYGC dataset using adVNTR, which does not have the same array length limitations as VNTRseek, and they also reported finding more losses than gains (see Figure S4 of Bakhtiari et al., 2018). Loss has also been reported to be more common in STR variants using the lobSTR computational tool (Willems et al., 2014). Observing more loss of copies than gain is likely due to a bias during genome assembly whereby a higher number of copies is preferentially selected in repetitive regions (Kent and Haussler, 2001). In Chapter 5 a method to infer missing alleles using read support was proposed.

7.2.1 Error types in VNTRseek

Computational tools that use short WGS reads for detecting genomic variation are evaluated for their sensitivity and type 1 error rate. The main sources of erroneous calls are ambiguous mapping due to reads originating from repeat-rich regions of the genome and sequencing errors. Setting a requirement of minimum read support for each prediction reduces false-positive calls at the cost of reducing sensitivity. The sensitivity of VNTRseek increases as the read length and coverage increase (Figure 2·1). However, as the read coverage increases, sequencing errors accumulate, and cause type 1 errors (Figure 4·9). The type 1 error of VNTRseek was measured by counting the Multi loci, i.e., the loci that were genotyped with more alleles than logically possible. About 1% of the VNTR loci were multi per genome and about one third occurred in only one genome, suggesting type 1 errors are occurring at random. Similarity in sequence of reference TRs accounted for 46% of the Multi loci (Figure 4·8).

Another source of type 1 error was due to sequencing errors. Filtering alleles with low read support reduces the type 1 error with the trade-off of sensitivity. On the haploid genomes, increasing the minimum read support requirement reduced the number of multis by almost two-fold. Requiring higher read support for predicted alleles reduces the number of multis, controlling the type 1 error. However, this would result in less sensitivity as larger TR arrays with lower coverage are filtered.

7.2.2 Effect of read coverage of VNTRseek precision and recall

In order to do determine the effect of read coverage on VNTRseek, the performance of VNTRseek was tested on a simulated genome. Using all the reference set TRs, random gain or loss of one pattern copy was inserted such that:

- one-sixth of the TR loci had *heterozygous gain* of one copy (0/+1),
- one-sixth of the TR loci had *homozygous gain* of one copy (+1/+1),
- one-sixth of the TR loci had *heterozygous loss* of one copy (0/-1),
- one-sixth of the TR loci had *homozygous loss* of one copy (-1/-1), and
- two-sixths of the TR loci had no change (0/0).

The ratios were chosen to balance the different classes.

Two genotypes, paternal and maternal, were simulated using simuG (Yue and Liti, 2019) with 3,000,000 random SNPs and the TR changes in copy as VCF files. Paired-end Illumina reads were simulated using the ART read simulator (Huang et al., 2012) to represent the same profile as the NYGC data: HiSeq2500 error profile, read length of 150 bp, fragment length mean of 550 bp, and fragment length standard deviation of 150 bp. The fragment coverage was calculated to give read coverage of 30X, 50X, 70X, and 100X. In each case, half the reads were from the paternal genome and half from the maternal genome.

VNTRseek was run on each dataset at the different coverages, and the precision and recall were determined for alleles overall and within the detectable range of VNTRseek, *i.e.*, allele copy number of at least 1.9 (TRF limitation) and array size of

130 bp or less (allowing for flanking sequence of 10 bp on each side of the array). Results are presented in Figure 7·3 and Table 7.3 to Table 7.6. The runtimes are given in Table 7.7.

The precision, or positive predictive value, reduced slightly as the coverage increased, but in all cases was greater than 93%. The recall for the detectable alleles ranged from 41% to 44% for the -1 alleles, from 85% to 90% for the +1 alleles, and above 98% for the reference alleles. The very low level of detection for -1 alleles in the “all alleles” graph is due to the TRF limitation. Because a very high percentage of the reference TRs had copy number under 2.9 copies. loss of one copy put them outside the detectable range.

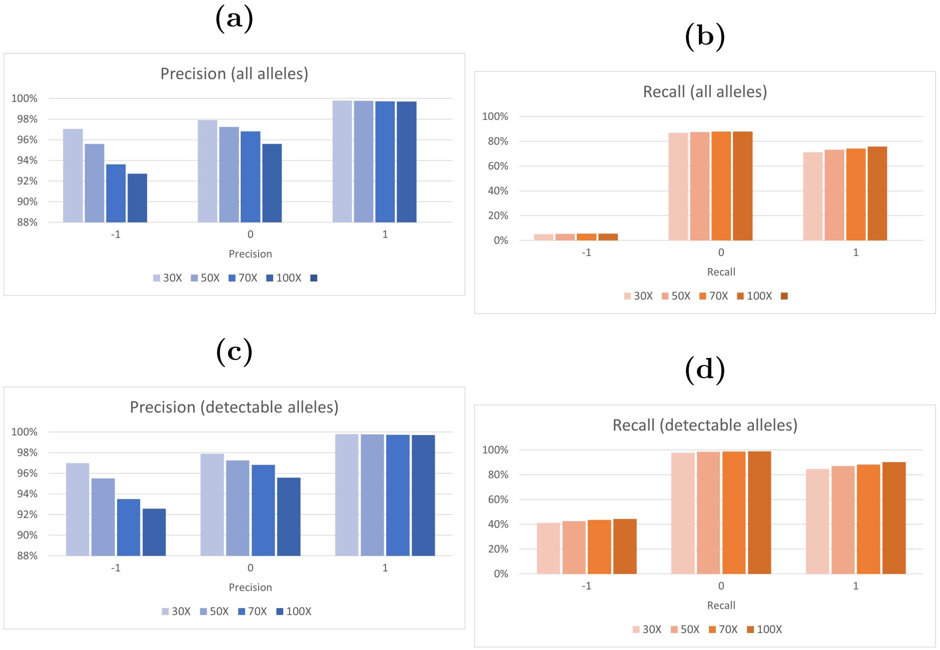


Figure 7·3: VNTRseek performance by coverage. Precision and recall for the entire collection of simulated TR alleles (sub-figures a and b) and the alleles in the detectable range of VNTRseek (sub-figures c and d) are shown. Precision, or positive predictive value, is the fraction of the predicted alleles that are correct (TP/(TP+FP)). Recall is the fraction correctly detected out of the total of that type (TP/(TP + FN)).

30X	True -1	True 0	True 1	Other
TP	3,056	106,065	49,005	0
FP	95	2,275	96	106
FN	4,380	2,463	8,942	0
Precision	97%	98%	100%	NA
Recall	41%	98%	85%	NA

Table 7.3: Detectable alleles confusion matrix of simulation at 30X coverage.

50X	True -1	True 0	True 1	Other
TP	3,161	106,833	50,405	0
FP	149	3,037	118	154
FN	4,275	1,695	7,542	0
Precision	95%	97%	100%	NA
Recall	43%	98%	87%	NA

Table 7.4: Detectable alleles confusion matrix of simulation at 50X coverage.

70X	True -1	True 0	True 1	Other
TP	3,246	107,084	51,076	0
FP	226	3,524	143	184
FN	4,190	1,444	6,871	0
Precision	93%	97%	100%	NA
Recall	44%	99%	88%	NA

Table 7.5: Detectable alleles confusion matrix of simulation at 70X coverage.

100X	True -1	True 0	True 1	Other
TP	3,297	107,294	52,225	0
FP	265	4,954	161	220
FN	4,139	1,234	5,722	0
Precision	93%	96%	100%	NA
Recall	44%	99%	90%	NA

Table 7.6: Detectable alleles confusion matrix of simulation at 100X coverage.

Coverage	User time	System time	Wall clock time	CPU time	Max memory
30X	6:16:26:45	5:49:42	14:44:06	6:22:16:28	37G
50X	11:01:02:20	8:24:16	1:00:32:53	11:09:26:37	38G
70X	15:03:35:49	9:10:04	1:08:58:30	15:12:45:53	39G
100X	22:02:53:22	10:39:42	2:07:48:25	22:13:33:04	40G

Table 7.7: Run time of VNTRseek on simulated datasets.

7.3 Limitations of mlZ

In Chapter 5, I describe mlZ, which can be used to infer missing alleles or incorrect alleles using read support. mlZ requires simulated data to train its machine learning model. However, due to limitation of VNTRseek, few heterozygous calls (less than 1%) are detected per simulation. Thus, at higher array length, the heterozygous machine learning model of mlZ is underfitted and lacks precision. More simulation data should be created for mlZ to overcome this problem.

7.4 Limitations of `MaSUD`

In Chapter 6, I describe MaSUD for genotyping macrosatellites using short WGS reads. MaSUD can only report the total copy number change from all chromosomes and cannot report the exact genotype at a given locus. For example, a MaSUD prediction of zero is found for genotypes of both -1/+1 and 0/0. While the first is a VNTR, while the second is not. Thus, for small MaSUD predictions the true copy number change is unreliable and this is why the results for values close to 0 are reported at N/A (e.g., see Table 6.6 and Figure 6·12). MaSUD can only genotype macrosatellites with arrays longer than the read length. When array sizes are small, the number of reads covering the array is low and MaSUD reports higher standard deviation. This standard deviation can be used to eliminate noisy predictions.

Finally, MaSUD requires a genome sequence coverage of at least 40× to precisely genotype macrosatellites. Shorter reads and higher coverage increase the performance of MaSUD. While these requirements are a limitation for prior data, such as the 1000 Genomes, recent datasets (e.g., NYGC and SGDP) meet this requirement.

7.5 Future work

7.5.1 Correcting reference TRs

In a small number of VNTRs (150 minisatellites and 21 macrosatellites), the reference allele was never observed (e.g., Table 2.4). This suggests that the reference genome is inaccurate at these loci. In addition, for many VNTRs the major allele was not the reference allele. The findings of this dissertation could be used to further improve the reference genome assembly.

7.5.2 Effects of VNTR alleles on proximal gene expression

Many of the VNTRs what we and others (some refs) have identified are found within or near to genes; however, a limited number of studies have analyzed the effects of VNTR changes on gene expression or protein function. As such, further studies are required to provide more evidence of changes in gene expression or function associated with VNTR genotype (Sulovari et al., 2019; Bakhtiari et al., 2020). For example, additional studies are required to determine if correlations between VNTR number and gene expression levels can be directly attributed to specific VNTR alleles, such as altering the binding of transcriptional regulatory proteins. The genotypes presented in this dissertation could be used to detect VNTR eQTLs.

7.5.3 Ancestry DNA

Population-biased alleles have the potential for use in tracing early human migration. We have shown that super-populations can be predicted using the VNTR genotypes (Section 3.3.3). Based on common VNTR alleles, I have constructed a decision tree that obtains nearly perfect classification of individuals at the super-population level (Figure 3·16). It will be interesting to determine whether, with more information, classification can be further refined to encompass specific sub-populations, whether a minimal minisatellite VNTR set can be established for high accuracy population classification, and whether VNTR alleles can be used to estimate mixed ancestry as is done now with SNP haplotyping (Bulbul and Filoglu, 2018; Pritchard et al., 2000).

7.5.4 GWAS

The frequency of VNTR occurrence and the possible effects of VNTRs on gene expression suggest that minisatellite VNTR loci can be useful in genome-wide association studies (GWAS). Relevant to this, we have determined that 1,096 of the common VNTR loci contain alleles show significant population specificity and that these loci intersect with 689 genes. Including VNTRs in GWAS models will be useful in future studies should. For example, VNTRs could be incorporated into GWAS on publicly available datasets on Alzheimer's disease, autism, and centenarians. Further investigate is required to determine whether there are haplotype linkages between specific VNTR alleles and nearby SNP alleles.

7.5.5 VNTR alleles under selection

Commercial cancer diagnosis kits using minisatellite VNTRs have been introduced (Leem et al., 2011) and it has been proposed that VNTRs associated with cancers be used for targeted sequencing in personalized therapies (Singh and Bandyopadhyay, 2007; Yoon et al., 2016; Rose, 2015). VNTR genotypes presented in this dissertation could be studied under the Hardy-Weinberg equilibrium to find deleterious alleles. This would be beneficial to find disease causing VNTR loci which could be included in targeted sequencing panel for clinical purposes.

7.5.6 Visualization of VNTR genotypes

Currently I am developing an Rshiny application to allow users to visualize VNTR genotypes and their characteristics.

7.5.7 Improvement and generalization of mlZ

mlZ can be generalized to be applicable to outcomes from other CNV tools that report the genotype and read support.

7.6 Overall summary

The results presented in this thesis have described several computational tools for identifying and characterizing tandem repeats in the human genome. Although for a long time, such repeats were overlooked as having biological functions, it is becoming increasingly clear that tandem repeats do play roles in biology and disease, in many cases probably by affecting gene expression and/or function. As methods, such as long read sequencing, improve for reading these highly repetitive sequences in the human genome, computational tools for identifying and predicting tandem repeat variations will also become increasingly useful. Applications of computational tools, such as those described in this thesis, may be in areas of predicting disease susceptibility, biological function, forensics, and ancestry.

References

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376.

Antwi-Boasiako, C., Dzudzor, B., Kudzi, W., Doku, A., Dale, C. A., Sey, F., Otu, K. H., Boatemaa, G. D., Ekem, I., Ahenkorah, J., Achel, D. G., Aboagye, E. T., and Donkor, E. S. (2018). Association between eNOS gene polymorphism (T786C and VNTR) and sickle cell disease patients in ghana. *Diseases*, 6(4).

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29.

Asif, H., Alliey-Rodriguez, N., Keedy, S., Tamminga, C. A., Sweeney, J. A., Pearlson, G., Clementz, B. A., Keshavan, M. S., Buckley, P., Liu, C., Neale, B., and Gershon, E. S. (2020). GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Molecular Psychiatry*.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Research*, 11(6):1005–1017.

Bakhtiari, M., Park, J., Ding, Y.-C., Shleizer-Burko, S., Neuhausen, S. L., Halldórsson, B. V., Stefansson, K., Gymrek, M., and Bafna, V. (2020). Variable number tandem repeats mediate the expression of proximal genes. *bioRxiv, doi: 10. 1101/ 2020. 05. 25. 114082*.

Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V., and Bafna, V. (2018). Targeted genotyping of variable number tandem repeats with adVNTR.

Balko, J. M., Cook, R. S., Vaught, D. B., Kuba, M. G., Miller, T. W., Bhola, N. E., Sanders, M. E., Granja-Ingram, N. M., Smith, J. J., Meszoely, I. M., et al. (2012). Profiling of residual breast cancers after neoadjuvant chemotherapy identifies dusp4 deficiency as a mechanism of drug resistance. *Nature Medicine*, 18(7):1052–1059.

Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biology*, 20(1):1–9.

Balog, J., Miller, D., Sanchez-Curtailles, E., Carbo-Marques, J., Block, G., Potman, M., de Knijff, P., Lemmers, R. J. L. F., Tapscott, S. J., and van der Maarel, S. M. (2012). Epigenetic regulation of the x-chromosomal macrosatellite repeat encoding for the cancer/testis gene CT47. *European Journal of Human Genetics*, 20(2):185–191.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009). mixtools: AnR-Package for analyzing finite mixture models.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences.

Benson, G. and Dong, L. (1999). Reconstructing the duplication history of a tandem repeat. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 44–53.

Biémont, C. and Vieira, C. (2006). Junk DNA as an evolutionary force.

Billingsley, K. J., Lättekivi, F., Planken, A., Reimann, E., Kurvits, L., Kadastik-Eerme, L., Kasterpalu, K. M., Bubb, V. J., Quinn, J. P., Kõks, S., and Taba, P. (2019). Analysis of repetitive element expression in the blood and skin of patients with parkinson's disease identifies differential expression of satellite elements.

Brahmachary, M., Guilmatre, A., Quilez, J., Hasson, D., Borel, C., Warburton, P., and Sharp, A. J. (2014). Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. *PLoS Genetics*, 10(6):e1004418.

Bray, S. M., Mulle, J. G., Dodd, A. F., Pulver, A. E., Wooding, S., and Warren, S. T. (2010). Signatures of founder effects, admixture, and selection in the ashkenazi jewish population. *Proceedings of the National Academy of Sciences of the United States of America*, 107(37):16222–16227.

Bulbul, O. and Filoglu, G. (2018). Development of a snp panel for predicting biogeographical ancestry and phenotype using massively parallel sequencing. *Electrophoresis*, 39(21):2743–2751.

Burgner, D., Rockett, K., Ackerman, H., Hull, J., Usen, S., Pinder, M., and Kwiatkowski, D. (2003). Haplotypic relationship between snp and microsatellite markers at the nos2a locus in two populations. *Genes & Immunity*, 4(7):506–514.

Bustamante, A. V., Sanso, A. M., Segura, D. O., Parma, A. E., and Lucchesi, P. M. A. (2013). Dynamic of mutational events in variable number tandem repeats of escherichia coli O157:H7. *Biomed Research International*, 2013:390354.

Campbell, H., Carothers, A. D., Rudan, I., Hayward, C., Biloglav, Z., Barac, L., Pericic, M., Janicijevic, B., Smolej-Narancic, N., Polasek, O., Kolcic, I., Weber, J. L., Hastie, N. D., Rudan, P., and Wright, A. F. (2007). Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Human Molecular Genetics*, 16(2):233–241.

Carey, N. (2015). *Junk DNA: A Journey Through the Dark Matter of the Genome.* Icon Books.

Cavenee, W. K. (1991). Stages of tumor progression: Loss of genetic heterozygosity.

Chadwick, B. P. (2009). Macrosatellite epigenetics: the two faces of DXZ4 and D4Z4. *Chromosoma*, 118(6):675–681.

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. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–611.

Chambon, M., Orsetti, B., Berthe, M.-L., Bascoul-Mollevi, C., Rodriguez, C., Duong, V., Gleizes, M., Thénot, S., Bibeau, F., Theillet, C., et al. (2011). Prognostic significance of trim24/tif-1α gene expression in breast cancer. *The American Journal of Pathology*, 178(4):1461–1469.

Chang, H.-I., Chang, Y.-T., Tsai, S.-J., Huang, C.-W., Hsu, S.-W., Liu, M.-E., Chang, W.-N., Lien, C.-Y., Huang, S.-H., Lee, C.-C., and Chang, C.-C. (2019). MAOA-VNTR genotype effects on ventral Striatum-Hippocampus network in alzheimer's disease: Analysis using structural covariance network and correlation with neurobehavior performance.

Church, G. M. (2005). The personal genome project. *Molecular Systems Biology*, 1:2005.0030.

Cong, L., Tu, G., and Liang, D. (2018). A systematic review of the relationship between the distributions of aggrecan gene VNTR polymorphism and degenerative disc disease/osteoarthritis.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., and Cherry, J. M. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, 46(D1):D794–D801.

de Greef, J. C., Lemmers, R. J. L. F., van Engelen, B. G. M., Sacconi, S., Venance, S. L., Frants, R. R., Tawil, R., and van der Maarel, S. M. (2009). Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Human Mutation*, 30(10):1449–1459.

De Roeck, A., De Coster, W., Bossaerts, L., Cacace, R., De Pooter, T., Van Dongen, J., D'Hert, S., De Rijk, P., Strazisar, M., Van Broeckhoven, C., and Sleegers, K. (2019). NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biology*, 20(1):239.

De Roeck, A., Duchateau, L., Van Dongen, J., Cacace, R., Bjerke, M., Van den Bossche, T., Cras, P., Vandenberghe, R., De Deyn, P. P., Engelborghs, S., Van Broeckhoven, C., Sleegers, K., and BELNEU Consortium (2018). An intronic VNTR affects splicing of ABCA7 and increases risk of alzheimer's disease. *Acta Neuropathologica*, 135(6):827–837.

Debrauwère, H., Buard, J., Tessier, J., Aubert, D., Vergnaud, G., and Nicolas, A. (1999). Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. *Nature Genetics*, 23(3):367–371.

Deka, R., Chakroborty, R., and Ferrell, R. E. (1991). A population genetic study of six VNTR loci in three ethnically defined populations. *Genomics*, 11(1):83–92.

Deka, R., DeCroo, S., Yu, L. M., and Ferrell, R. (1992). Variable number of tandem repeat (VNTR) polymorphism at locus D17S5 (YNZ22) in four ethnically defined human populations.

Deng, T., Liu, J. C., Chung, P. E., Uehling, D., Aman, A., Joseph, B., Ketela, T., Jiang, Z., Schachter, N. F., Rottapel, R., et al. (2014). shrna kinome screen identifies tbk1 as a therapeutic target for her2+ breast cancer. *Cancer Research*, 74(7):2119–2130.

Denoeud, F., Vergnaud, G., and Benson, G. (2003). Predicting human minisatellite polymorphism. *Genome Research*, 13(5):856–867.

Diatchenko, L., Nackley, A. G., Tchivileva, I. E., Shabalina, S. A., and Maixner, W. (2007). Genetic architecture of human pain perception. *Trends Genetics*, 23(12):605–613.

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.

Duan, Z., Zhang, J., Choy, E., Harmon, D., Liu, X., Nielsen, P., Mankin, H., Gray, N. S., and Hornicek, F. J. (2012). Systematic kinome shrna screening identifies cdk11 (pitslre) kinase expression is critical for osteosarcoma cell growth and proliferation. *Clinical Cancer Research*, 18(17):4580–4588.

Duitama, J., McEwen, G. K., Huebsch, T., Palczewski, S., Schulz, S., Verstrepen, K., Suk, E.-K., and Hoehe, M. R. (2012). Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. *Nucleic Acids Research*, 40(5):2041–2053.

Duitama, J., Zablotskaya, A., Gemayel, R., Jansen, A., Belet, S., Vermeesch, J. R., Verstrepen, K. J., and Froyen, G. (2014). Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Research*, 42(9):5728–5741.

Dumbovic, G., Biayna, J., Banús, J., Samuelsson, J., Roth, A., Diederichs, S., Alonso, S., Buschbeck, M., Perucho, M., and Forcales, S.-V. (2018). A novel long noncoding RNA from NBL2 pericentromeric macrosatellite forms a perinucleolar aggregate structure in colon cancer. *Nucleic Acids Research*, 46(11):5504–5524.

Edea, Z., Bhuiyan, M. S. A., Dessie, T., Rothschild, M. F., Dadi, H., and Kim, K. S. (2015). Genome-wide genetic diversity, population structure and admixture analysis in african and asian cattle breeds. *Animal*, 9(2):218–226.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The international genome sample resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1):D941–D947.

Fondon, 3rd, J. W. and Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(52):18058–18063.

Fu, S., Octavia, S., Wang, Q., Tanaka, M. M., Tay, C. Y., Sintchenko, V., and Lan, R. (2016). Evolution of variable number tandem repeats and its relationship with genomic diversity in salmonella typhimurium.

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes.

Garrick, D., Fiering, S., Martin, D. I., and Whitelaw, E. (1998). Repeat-induced gene silencing in mammals. *Nature Genetics*, 18(1):56–59.

Gelfand, Y., Hernandez, Y., Loving, J., and Benson, G. (2014). VNTRseek—a computational tool to detect tandem repeat variants in high-throughput sequencing data.

Gelfand, Y., Rodriguez, A., and Benson, G. (2007). TRDB—The tandem repeats database. *Nucleic Acids Research*, 35(suppl_1):D80–D87.

Geng, L. N., Yao, Z., Snider, L., Fong, A. P., Cech, J. N., Young, J. M., van der Maarel, S. M., Ruzzo, W. L., Gentleman, R. C., Tawil, R., and Tapscott, S. J. (2012). DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Developmental Cell*, 22(1):38–51.

Giacalone, J., Friedes, J., and Francke, U. (1992). A novel GC-rich human macrosatellite VNTR in xq24 is differentially methylated on active and inactive X chromosomes. *Nature Genetics*, 1(2):137–143.

Grünblatt, E., Werling, A. M., Roth, A., Romanos, M., and Walitza, S. (2019). Association study and a systematic meta-analysis of the VNTR polymorphism in the 3'-UTR of dopamine transporter gene and attention-deficit hyperactivity disorder.

Günther, T. and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genetics*, 15(7):e1008302.

Gymrek, M. (2017). A genomic view of short tandem repeats.

Hancock, J. M. and Santibáñez-Koref, M. F. (1998). Trinucleotide expansion diseases in the context of micro- and minisatellite evolution, hammersmith hospital, april 1-3, 1998. *EMBO Journal*, 17(19):5521–5524.

Hoxha, B., Goçi Uka, A., Agani, F., Haxhibeqiri, S., Haxhibeqiri, V., Sabic Dzananovic, E., Kucukalic, S., Bravo Mehmedbasic, A., Kucukalic, A., Dzubur Kulenovic, A., Feric Bojic, E., Marjanovic, D., Kravic, N., Avdibegovic, E., Muminovic Umihanić, M., Jaksic, N., Cima Franc, A., Rudan, D., Jakovljevic, M., Babic, R., Pavlovic, M., Babic, D., Aukst Margetic, B., Bozina, N., Sinanovic, O., Ziegler, C., Warrings, B., Domschke, K., Deckert, J., Wolf, C., and Vyshka, G. (2019). The role of TaqI DRD2 (rs1800497) and DRD4 VNTR polymorphisms in posttraumatic stress disorder (PTSD). *Psychiatria Danubina.*, 31(2):263–268.

Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M., and Haussler, D. (2006). The UCSC known genes.

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

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.

Huichalaf, C., Micheloni, S., Ferri, G., Caccia, R., and Gabellini, D. (2014). DNA methylation analysis of the macrosatellite repeat associated with FSHD muscular dystrophy at single nucleotide level. *PLoS One*, 9(12):e115278.

Ibrahimi, M., Moossavi, M., Mojarad, E. N., Musavi, M., Mohammadoo-khorasani, M., and Shahsavari, Z. (2019). Positive correlation between interleukin-1 receptor antagonist gene 86bp VNTR polymorphism and colorectal cancer susceptibility: a case-control study.

Imam, J., Reyaz, R., Rana, A. K., and Yadav, V. K. (2018). DNA fingerprinting: Discovery, advancements, and milestones.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J., and Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345.

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):239.

Jeffreys, A. J., Neil, D. L., and Neumann, R. (1998). Repeat instability at human minisatellites arising from meiotic recombination. *EMBO Journal*, 17(14):4147–4157.

Jones, T. I., Chen, J. C. J., Rahimov, F., Homma, S., Arashiro, P., Beermann, M. L., King, O. D., Miller, J. B., Kunkel, L. M., Emerson, Jr, C. P., Wagner, K. R., and Jones, P. L. (2012). Facioscapulohumeral muscular dystrophy family studies of DUX4 expression: evidence for disease modifiers and a quantitative model of pathogenesis. *Human Molecular Genetics*, 21(20):4419–4430.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Kang, A. M., Palmatier, M. A., and Kidd, K. K. (1999). Global variation of a 40-bp vntr in the 3'-untranslated region of the dopamine transporter gene (slc6a3). *Biological Psychiatry*, 46(2):151–160.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue):D493–6.

Katsumata, Y., Fardo, D. W., Bachstetter, A. D., Artiushin, S. C., Wang, W.-X., Wei, A., Brzezinski, L. J., Nelson, B. G., Huang, Q., Abner, E. L., Anderson, S., Patel, I., Shaw, B. C., Price, D. A., Niedowicz, D. M., Wilcock, D. W., Jicha, G. A., Neltner, J. H., Van Eldik, L. J., Estus, S., and Nelson, P. T. (2019). Alzheimer disease Pathology-Associated polymorphism in a complex variable number of tandem repeat region within the MUC6 gene, near the AP2A2 gene.

Kent, W. J. (2002). The human genome browser at UCSC.

Kent, W. J. and Haussler, D. (2001). Assembly of the working draft of the human genome with GigAssembler. *Genome Research*, 11(9):1541–1548.

Kondo, T., Bobek, M. P., Kuick, R., Lamb, B., Zhu, X., Narayan, A., Bourc'his, D., Viegas-Péquignot, E., Ehrlich, M., and Hanash, S. M. (2000). Whole-genome methylation scan in ICF syndrome: hypomethylation of non-satellite DNA repeats D4Z4 and NBL2. *Human Molecular Genetics*, 9(4):597–604.

Korlach, J. (2013). Understanding accuracy in smrt® sequencing. *Pacific Biosciences*, pages 1–9.

Krontiris, T. G., Devlin, B., Karp, D. D., Robert, N. J., and Risch, N. (1993). An association between the risk of cancer and mutations in the HRAS1 minisatellite locus. *New England Journal of Medicine*, 329(8):517–523.

Ksiazek, K., Blaszczak, J., and Buraczynska, M. (2019). IL4 gene VNTR polymorphism in chronic periodontitis in end-stage renal disease patients. *Oral Diseases*, 25(1):258–264.

Kulski, J. (2016). *Next Generation Sequencing: Advances, Applications and Challenges*. BoD – Books on Demand.

Laidlaw, J., Gelfand, Y., Ng, K.-W., Garner, H. R., Ranganathan, R., Benson, G., and Fondon, 3rd, J. W. (2007). Elevated basal slippage mutation rates among the canidae. *Journal of Heredity*, 98(5):452–460.

Lalioti, M. D., Antonarakis, S. E., and Scott, H. S. (2003). The epilepsy, the protease inhibitor and the dodecamer: progressive myoclonus epilepsy, cystatin b and a 12-mer repeat expansion.

Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., and Gendler, S. J. (1990). Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit.

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

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359.

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Häsler, R., Syvänen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I. G., Estivill, X., and Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84.

Lee, D., Cheng, A., Lawlor, N., Bolisetty, M., and Ucar, D. (2018). Detection of correlated hidden factors from single cell transcriptomes using iteratively adjusted-sva (ia-sva). *Scientific Reports*, 8(1):1–13.

Leem, S. H., Jeong, Y. H., Yoon, S. L., and Seol, S.-Y. (2011). Diagnosis kits and method for detecting cancer using polymorphic minisatellite. US Patent 7,981,613.

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.

Lemmers, R. J. L. F., Tawil, R., Petek, L. M., Balog, J., Block, G. J., Santen, G. W. E., Amell, A. M., van der Vliet, P. J., Almomani, R., Straasheijm, K. R., Krom, Y. D., Klooster, R., Sun, Y., den Dunnen, J. T., Helmer, Q., Donlin-Smith, C. M., Padberg, G. W., van Engelen, B. G. M., de Greef, J. C., Aartsma-Rus, A. M., Frants, R. R., de Visser, M., Desnuelle, C., Sacconi, S., Filippova, G. N., Bakker, B., Bamshad, M. J., Tapscott, S. J., Miller, D. G., and van der Maarel, S. M. (2012). Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nature Genetics*, 44(12):1370–1374.

Levinson, G. and Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for dna sequence evolution. *Molecular Biology and Evolution*, 4(3):203–221.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv Preprint, doi: 10. 6084/M9. FIGSHARE. 963153. V1*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., Zhou, G., Zhu, X., Wu, H., Qin, J., Jin, X., Li, D., Cao, H., Hu, X., Blanche, H., Cann, H., Zhang, X., Li, S., Bolund, L., Kristiansen, K., Yang, H., Wang, J., and Wang, J. (2010). Building the sequence map of the human pan-genome. *Nature Biotechnology*, 28(1):57–63.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740.

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., Pape, U. J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., and Liu, X. S. (2011). Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biology*, 12(8):R83.

Liu, Y. and Just, A. (2020). SHAPforxgboost: SHAP plots for'XGBoost'. *R package version 0. 0*, 4.

Long, X., Shi, Y., Ye, P., Guo, J., Zhou, Q., and Tang, Y. (2020). Microrna-99a suppresses breast cancer progression by targeting fgfr3. *Frontiers in Oncology*, 9:1473.

López Herráez, D., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M. R., Gross, A., Scholz, M., and Stoneking, M. (2009). Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One*, 4(11):e7888.

LU, G. M., Wang, S., Zhao, L., Zhou, X. Q., Yan, H., Wang, Q. C., Huang, J. F., et al. (2017). Expression of cd44, trim24, tagln-2, er and pr in breast invasive ductal carcinoma and their clinicopathologic significance. *Chinese Journal of Clinical and Experimental Pathology*, 33(7):724–727.

Lu, T.-Y. T., Chaisson, M. J., Consortium, H. G. S. V., et al. (2020). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *bioRxiv, `doi: 10. 1101/2020. 08. 13. 249839`.*

Ludwig, M. Z. (2016). Noncoding DNA evolution: Junk DNA revisited.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc.

Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J., Li, Y., Li, B., Shen, S., Wang, Z., Shurtleff, S., Robison, L. L., Levy, S., Easton, J., and Zhang, J. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biology*, 20(1):50.

Madsen, C. S., Ghivizzani, S. C., and Hauswirth, W. W. (1993). In vivo and in vitro evidence for slipped mispairing in mammalian mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, 90(16):7671–7675.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., Spence, J. P., Song, Y. S., Poletti, G., Balloux, F., van Driem, G., de Knijff, P., Romero, I. G., Jha, A. R., Behar, D. M., Bravi, C. M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O. L., Balanovska, E., Balanovsky, O., Karachanak-Yankova, S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y., Abdullah, M. S., Ruiz-Linares, A., Beall, C. M., Di Rienzo, A., Jeong, C., Starikovskaya, E. B., Metspalu, E., Parik, J., Villems, R., Henn, B. M., Hodoglugil, U., Mahley, R., Sajantila, A., Stamatoyannopoulos, G., Wee, J. T. S., Khusainova, R., Khusnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M. F., Kivisild, T., Klitz, W., Winkler, C. A., Labuda, D., Bamshad, M., Jorde, L. B., Tishkoff, S. A., Watkins, W. S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L., Thangaraj, K., Pääbo, S., Kelso, J., Patterson, N., and Reich, D. (2016). The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.

Marinho, F. V. C., Pinto, G. R., Oliveira, T., Gomes, A., Lima, V., Ferreira-Fernandes, H., Rocha, K., Magalhães, F., Velasques, B., Ribeiro, P., Cagy, M., Gupta, D., Bastos, V. H., and Teixeira, S. (2019). The SLC6A3 3'-UTR VNTR and intron 8 VNTR polymorphisms association in the time estimation. *Brain Structural Function*, 224(1):253–262.

Marks, P., Garcia, S., Barrio, A. M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2019). Resolving the full spectrum of human genome variation using linked-reads. *Genome Research*, 29(4):635–645.

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.

Menyhart, O., Budczies, J., Munkácsy, G., Esteva, F. J., Szabó, A., Miquel, T. P., and Győrffy, B. (2017). Dusp4 is associated with increased resistance against anti-her2 therapy in breast cancer. *Oncotarget*, 8(44):77207.

Mierswa, I. and Klinkenberg, R. (2018). RapidMiner studio (9.2)[data science, machine learning, predictive analytics].

Mill, J., Asherson, P., Browes, C., D'Souza, U., and Craig, I. (2002). Expression of the dopamine transporter gene is regulated by the 3' UTR VNTR: Evidence from brain and lymphocytes using quantitative RT-PCR. *American Journal of Medical Genetics*, 114(8):975–979.

Mu, J. C., Mohiyuddin, M., Li, J., Bani Asadi, N., Gerstein, M. B., Abyzov, A., Wong, W. H., and Lam, H. Y. K. (2015). VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, 31(9):1469–1471.

Nagraj, V. P., Magee, N. E., and Sheffield, N. C. (2018). LOLAweb: a containerized web server for interactive genomic locus overlap enrichment analysis. *Nucleic Acids Research*, 46(W1):W194–W199.

Näslund, K., Saetre, P., von Salomé, J., Bergström, T. F., Jareborg, N., and Jazin, E. (2005). Genome-wide prediction of human VNTRs. *Genomics*, 85(1):24–35.

Pacheco, A., Berger, R., Freedman, R., and Law, A. J. (2019). A VNTR regulates mir-137 expression through novel alternative splicing and contributes to risk for schizophrenia. *Scientific Reports*, 9(1):11793.

Panagiotou, O. A., Evangelou, E., and Ioannidis, J. P. A. (2010). Genome-wide significant associations for variants with minor allele frequency of 5% or Less—An overview: A HuGE review.

Panigrahi, I. (2018). Genetic fingerprinting for human diseases: Applications and implications.

Pâques, F., Richard, G. F., and Haber, J. E. (2001). Expansions and contractions in 36-bp minisatellites by gene conversion in yeast. *Genetics*, 158(1):155–166.

Pathiraja, T. N., Thakkar, K. N., Jiang, S., Stratton, S., Liu, Z., Gagea, M., Shi, X., Shah, P. K., Phan, L., Lee, M.-H., et al. (2015). Trim24 links glucose metabolism with transformation of human mammary epithelial cells. *Oncogene*, 34(22):2836–2845.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Psychiatric GWAS Consortium Coordinating Committee, Cichon, S., Craddock, N., Daly, M., Faraone, S. V., Gejman, P. V., Kelsoe, J., Lehner, T., Levinson, D. F., Moran, A., Sklar, P., and Sullivan, P. F. (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry*, 166(5):540–556.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Raczy, C., Petrovski, R., Saunders, C. T., Chorny, I., Kruglyak, S., Margulies, E. H., Chuang, H.-Y., Källberg, M., Kumar, S. A., Liao, A., Little, K. M., Strömberg, M. P., and Tanner, S. W. (2013). Isaac: ultra-fast whole-genome secondary analysis on illumina sequencing platforms. *Bioinformatics*, 29(16):2041–2043.

Ramírez-Patiño, R., Figuera, L. E., Puebla-Pérez, A. M., Delgado-Saucedo, J. I., Legazpí-Macias, M. M., Mariaud-Schmidt, R. P., Ramos-Silva, A., Gutiérrez-Hurtado, I. A., Flores-Ramos, L. G., Zúñiga-González, G. M., and Gallegos-Arreola, M. P. (2013). Intron 4 VNTR (4a/b) polymorphism of the endothelial nitric oxide synthase gene is associated with breast cancer in mexican women.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339.

Rhoads, A. and Au, K. F. (2015). PacBio sequencing and its applications.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Rose, A. M. (2015). Therapeutics and diagnostics based on minisatellite repeat element 1 (msr1). US Patent App. 14/761,952.

Safarinejad, M. R., Safarinejad, S., Shafiei, N., and Safarinejad, S. (2013). Effects of the T-786C, G894T, and intron 4 VNTR (4a/b) polymorphisms of the endothelial nitric oxide synthase gene on the risk of prostate cancer.

Sánchez-Castillo, M., Ruau, D., Wilkinson, A. C., Ng, F. S. L., Hannah, R., Diamanti, E., Lombard, P., Wilson, N. K., and Gottgens, B. (2015). CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Research*, 43(Database issue):D1117–23.

Schaap, M., Lemmers, R. J. L. F., Maassen, R., van der Vliet, P. J., Hoogerheide, L. F., van Dijk, H. K., Baştürk, N., de Knijff, P., and van der Maarel, S. M. (2013). Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. *BMC Genomics*, 14:143.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17:125.

Schlüter, T., Winz, O., Mohammadkhani-Shali, S., Eggermann, T., Henkel, K. H., Mottaghy, F. M., and Vernaleken, I. (2014). P.8.b.024 MAOA-VNTR polymorphism modulates context-dependent dopamine release and aggressive behaviour.

Schwarzenbach, H., Eichelser, C., Kropidlowski, J., Janni, W., Rack, B., and Pantel, K. (2012). Loss of heterozygosity at tumor suppressor genes detectable on fractionated circulating cell-free tumor DNA as indicator of breast cancer progression. *Clinical Cancer Research*, 18(20):5719–5730.

Scott, H. S., Nelson, P. V., Hopwood, J. J., and Morris, C. P. (1991). PCR of a VNTR linked to mucopolysaccharidosis type I and huntington disease. *Nucleic Acids Research*, 19(22):6348.

Šerý, O., Paclt, I., Drtílková, I., Theiner, P., Kopečková, M., Zvolský, P., and Balcar, V. J. (2015). A 40-bp VNTR polymorphism in the 3'-untranslated region of DAT1/SLC6A3 is associated with ADHD but not with alcoholism.

Sheffield, N. C., Thurman, R. E., Song, L., Safi, A., Stamatoyannopoulos, J. A., Lenhard, B., Crawford, G. E., and Furey, T. S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research*, 23(5):777–788.

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., Levin, A. M., Eng, C., Yazdanbakhsh, M., Wilson, J. G., Marrugo, J., Lange, L. A., Williams, L. K., Watson, H., Ware, L. B., Olopade, C. O., Olopade, O., Oliveira, R. R., Ober, C., Nicolae, D. L., Meyers, D. A., Mayorga, A., Knight-Madden, J., Hartert, T., Hansel, N. N., Foreman, M. G., Ford, J. G., Faruque, M. U., Dunston, G. M., Caraballo, L., Burchard, E. G., Bleecker, E. R., Araujo, M. I., Herrera-Paz, E. F., Campbell, M., Foster, C., Taub, M. A., Beaty, T. H., Ruczinski, I., Mathias, R. A., Barnes, K. C., and Salzberg, S. L. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nature Genetics*, 51(1):30–35.

Singh, R. and Bandyopadhyay, D. (2007). Muc1: a target molecule for cancer therapy. *Cancer Biology & Therapy*, 6(4):481–486.

Sinha, M., Arjun Rao, I., and Mitra, M. (2018). Molecular basis of identification through DNA fingerprinting in humans.

Sonay, T. B., Carvalho, T., Robinson, M. D., Greminger, M. P., Krützen, M., Comas, D., Highnam, G., Mittelman, D., Sharp, A., Marques-Bonet, T., and Wagner, A. (2015). Tandem repeat variation in human and great ape populations and its impact on gene expression divergence.

Stolf, A. R., Cupertino, R. B., Müller, D., Sanvicente-Vieira, B., Roman, T., Vitola, E. S., Grevet, E. H., von Diemen, L., Kessler, F. H. P., Grassi-Oliveira, R., Bau, C. H. D., Rovaris, D. L., Pechansky, F., and Schuch, J. B. (2019). Effects of DRD2 splicing-regulatory polymorphism and DRD4 48 bp VNTR on crack cocaine addiction.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., and Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.

Sulovari, A., Li, R., Audano, P. A., Porubsky, D., Vollger, M. R., Logsdon, G. A., Human Genome Structural Variation Consortium, Warren, W. C., Pollen, A. A., Chaisson, M. J. P., and Eichler, E. E. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 116(46):23243–23253.

Taylor, J. S. and Breden, F. (2000). Slipped-strand mispairing at noncontiguous repeats in poecilia reticulata: a model for minisatellite birth. *Genetics*, 155(3):1313–1320.

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

The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., and Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases.

Tremblay, D. C., Alexander, Jr, G., Moseley, S., and Chadwick, B. P. (2010). Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC Genomics*, 11:632.

Trepicchio, W. L. and Krontiris, T. G. (1992). Members of therel/NF-χB family of transcriptional regulatory proteins bind theHRAS1minisatellite DNA sequence.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Vairaktaris, E., Serefoglou, Z. C., Yapijakis, C., Vassiliou, S., Nkenke, E., Avgoustidis, D., Vylliotis, A., Stathopoulos, P., Neukam, F. W., and Patsouris, E. (2007). The platelet glycoprotein ibalpha VNTR polymorphism is associated with risk for oral cancer. *AntiCancer Research*, 27(6B):4121–4125.

Vallente, R. U. and Eichler, E. E. (2005). Segmental duplications and the human genome.

Van Assche, E., Moons, T., Van Leeuwen, K., Colpin, H., Verschueren, K., Van Den Noortgate, W., Goossens, L., and Claes, S. (2016). Depressive symptoms in adolescence: The role of perceived parental support, psychological control, and proactive control in interaction with 5-HTTLPR. *European Psychiatry*, 35:55–63.

Van Tol, H. H., Wu, C. M., Guan, H. C., Ohara, K., Bunzow, J. R., Civelli, O., Kennedy, J., Seeman, P., Niznik, H. B., and Jovanovic, V. (1992). Multiple dopamine D4 receptor variants in the human population. *Nature*, 358(6382):149–152.

Verstrepen, K. J., Jansen, A., Lewitter, F., and Fink, G. R. (2005). Intragenic tandem repeats generate functional variability. *Nature Genetics*, 37(9):986–990.

Vilas, A., Pérez-Figueroa, A., Quesada, H., and Caballero, A. (2015). Allelic diversity for neutral markers retains a higher adaptive potential for quantitative traits than expected heterozygosity. *Molecular Ecology*, 24(17):4419–4432.

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K. J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, 324(5931):1213–1216.

Vogler, A. J., Keys, C., Nemoto, Y., Colman, R. E., Jay, Z., and Keim, P. (2006). Effect of repeat copy number on Variable-Number tandem repeat mutations in escherichia coli O157:H7.

Weber, J. L. and Myers, E. W. (1997). Human Whole-Genome shotgun sequencing.

Wei, Y.-C. and Huang, G.-H. (2020). CONY: A bayesian procedure for detecting copy number variations from sequencing read depths. *Scientific Reports*, 10(1):10493.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Ruan, J., Marschall, T., Sedlazeck, F. J., Zook, J. M., Li, H., Koren, S., Carroll, A., Rank, D. R., and Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162.

Willems, T., Gymrek, M., Highnam, G., Mittelman, D., Erlich, Y., Consortium, . G. P., et al. (2014). The landscape of human str variation. *Genome Research*, 24(11):1894–1904.

Xie, C. and Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 10:80.

Ye, F. and Signer, E. R. (1996). RIGS (repeat-induced gene silencing) in arabidopsis is transcriptional and alters chromatin configuration. *Proceedings of the National Academy of Sciences of the United States of America*, 93(20):10881–10886.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-blast: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13(1):1–11.

Yoon, S.-L., Roh, Y.-G., Chu, I.-S., Heo, J., Kim, S. I., Chang, H., Kang, T.-H., Chung, J. W., Koh, S. S., Larionov, V., et al. (2016). A polymorphic minisatellite region of boris regulates gene expression and its rare variants correlate with lung cancer susceptibility. *Experimental & Molecular Medicine*, 48(7):e246–e246.

Yue, J.-X. and Liti, G. (2019). simug: a general-purpose genome simulator. *Bioinformatics*, 35(21):4442–4444.

Zammit, S., Jones, G., Jones, S. J., Norton, N., Sanders, R. D., Milham, C., McCarthy, G. M., Jones, L. A., Cardno, A. G., Gray, M., Murphy, K. C., O'Donovan, M. C., and Owen, M. J. (2004). Polymorphisms in the MAOA, MAOB, and COMT genes and aggressive behavior in schizophrenia. *American Journal of Medical Genetics*, 128B(1):19–20.

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*, 3:160025.

Zuo, H., Gandhi, M., Edreira, M. M., Hochbaum, D., Nimgaonkar, V. L., Zhang, P., Dipaola, J., Evdokimova, V., Altschuler, D. L., and Nikiforov, Y. E. (2010). Downregulation of Rap1GAP through epigenetic silencing and loss of heterozygosity promotes invasion and progression of thyroid tumors. *Cancer Research*, 70(4):1389–1397.

CURRICULUM VITAE

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]