UNIVERSITY OF CALIFORNIA SAN DIEGO

Computational Methods to Study Tandem Repeats in Human Genome and Complex Diseases

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Mehrdad Bakhtiari

Committee in charge:

Professor Vineet Bafna, Chair
Professor Vikas Bansal
Professor Kelly Frazer
Professor Melissa Gymrek
Professor Pavel Pevzner

2021

The dissertation of Mehrdad Bakhtiari is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents, for your limitless love and support.

To the passengers of PS752 and 52 graduate students on board, who did not finish their dissertation.

EPIGRAPH

Came here for school, graduated to the high life.

—Shawn Corey Carter

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor and supportive mentor Vineet Bafna. Vineet has always been the source of wisdom that influenced both academic and personal aspects of my life. I am grateful to him for giving me the perfect balance of guidance and freedom in pursuing my broad interests, and helping me to clarify my ideas and explanation with his constant enthusiasm and boundless intellectual curiosity. I am grateful for numerous opportunities that came my way as a result of his hard work and dedication to nurturing his students. I also express my gratitude to all my thesis committee members: Vikas Bansal, Kelly Frazer, Melissa Gymrek, and Pavel Pevzner. Their insightful comments and constructive criticisms improved this dissertation in many ways.

I was fortunate to work alongside colleagues in Computer Science and Bioinformatics programs. In particular, I would like to thank Ali Akbari, Viraj Deshpande, Shahab Sarmashghi, Andrey Bzikadze, Jens Luebeck, Jonghun Park, and Sara Javadzadeh for the many helpful discussions we had, their encouragement, and the assistance they gave.

I have to thank Bashir Sadjad, who presented me the existence of Bioinformatics as a focus-area in Computer Science. His vast knowledge in combinatorial and algorithm has motivated me to solve interesting biological problems using computational techniques and I am thankful that his guidance led me down this path.

I have benefited from countless friends and mentors throughout my career and my graduate studies. While it is not possible to name them all, I am grateful for their company and the influence they had on my thinking and decisions. I owe much of my success on their mentorship.

Last but not least, I would like to thank my parents, my sisters, and my brother for their relentless devotion and encouragement. Finally, I would like to thank my wife, Sara, for her constant support and encouragement and also for her invaluable opinions on various scientific problems.

Chapter 2, in full, contains material from Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, Vineet Bafna. "Targeted Genotyping of Variable Number Tandem Repeats with adVNTR.", *Genome Research*, 2018. The dissertation author was the primary author of this paper[10].

Chapter 3, in full, contains material from Mehrdad Bakhtiari, Jonghun Park, Yuan-Chun Ding, Sharona Shleizer-Burko, Susan L. Neuhausen, Bjarni V. Halldórsson, Kári Stefánsson, Melissa Gymrek, Vineet Bafna. "Variable Number Tandem Repeats mediate the expression of proximal genes." *Nature Communications*, 2021[9]. The dissertation author was the primary author of this paper.

Chapter 4, in part, is based on the ongoing research from Mehrdad Bakhtiari and Vineet Bafna, which is currently being prepared for submission for publication of the material. The dissertation author was a joint primary investigator and author of this material.

VITA

2016	B. Sc. in Computer Engineering , University of Tehran, Iran
2018	M. Sc. in Computer Science, University of California San Diego
2021	Ph. D. in Computer Science, University of California San Diego

PUBLICATIONS

Mehrdad Bakhtiari, Jonghun Park, Yuan-Chun Ding, Sharona Shleizer-Burko, Susan L. Neuhausen, Bjarni V. Halldórsson, Kári Stefánsson, Melissa Gymrek, Vineet Bafna. "Variable Number Tandem Repeats mediate the expression of proximal genes." *Nature Communications*, 2021.

Mehrdad Bakhtiari, Vineet Bafna. "A read classification method to improve identification of tandem repeat variations in human genome.", *ICML WCB*, 2019.

Viraj Deshpande, Jens Luebeck, Nam-Phuong Nguyen, **Mehrdad Bakhtiari**, Kristen Turner, Richard Schwab , Hannah Carter, Paul Mischel, Vineet Bafna. "Exploring the landscape of focal amplifications in cancer using AmpliconArchitect.", *Nature Communication*, 2019.

Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, Vineet Bafna. "Targeted Genotyping of Variable Number Tandem Repeats with adVNTR.", *Genome Research*, 2018.

Ali Akbari, Joseph J. Vitti, Arya Iranmehr, **Mehrdad Bakhtiari**, Pardis C Sabeti, Siavash Mirarab, Vineet Bafna. "Identifying the favored mutation in a positive selective sweep.", *Nature Methods*, 2018.

ABSTRACT OF THE DISSERTATION

Computational Methods to Study Tandem Repeats in Human Genome and Complex Diseases

by

Mehrdad Bakhtiari

Doctor of Philosophy in Computer Science

University of California San Diego, 2021

Professor Vineet Bafna, Chair

A central goal in genomics is to identify genetic variations and their impact on underlying molecular changes that lead to disease. With the advances in whole genome sequencing, many studies have been able to identify thousands of genetic loci associated with human traits. These studies mainly focus on single-nucleotide variants (SNVs) and novel insertion and deletions in the genome, while ignoring more complex variants. Here, I consider the problem of genotyping Variable Number Tandem Repeats (VNTRs), composed of inexact tandem duplications of short (6-100 bp) repeating units that span 3% of the human genome.

While some VNTRs are known to play a role in complex disorders (e.g. Alzheimer's,

Myoclonus epilepsy, and Diabetes), the majority of them have not been studied well due to computational difficulty in genotyping VNTRs on a large scale. Here, I will present our progress on developing efficient computational algorithms to profile VNTRs from high throughput sequencing data and identify possible variations within them. I applied our method to generate the largest catalog of VNTR genotypes to this date, which provides insights into the landscape of VNTR variations in different populations. I show the contribution of tandem repeats in mediating expression levels of key genes with known associations to neurological disorders and familial cancers, and argue the causality of this relation. Finally, I will describe our efforts to directly understand the impact of these variations on human phenotypes, which improves our understanding of genetic architecture of complex diseases.

Chapter 1

Introduction

1.1 Overview

A central goal in genomics is to identify genetic variations and their impact on underlying molecular changes that lead to disease. With the advances in whole genome sequencing, many studies have been able to identify thousands of genetic loci associated with human traits. These studies mainly focus on single-nucleotide variants (SNVs) and novel insertion and deletions in the genome, while ignoring more complex variants. Here, I consider the problem of genotyping Variable Number Tandem Repeats (VNTRs), composed of inexact tandem duplications of short (6-100 bp) repeating units, and study their impact on cell mechanism and human health.

1.2 Tandem Repeats

The human genome consists of millions of tandem repeats (TRs) of short nucleotide sequences. These are often termed as Short Tandem Repeats (STRs) if the repeating unit is < 6bp, and Variable Number Tandem Repeats (VNTRs) otherwise. Together, they represent one of the largest sources of polymorphisms in humans[134, 51]. VNTRs span 3% of the human genome

and since they can be located in coding regions[107], untranslated regions[83], and regulatory regions proximal to a gene[43, 129], the variation in length can have a significant functional impact. Not surprisingly, VNTRs have been implicated in a large number of Mendelian diseases that affect millions of people world-wide[17, 19, 72].

We define VNTR genotyping in the narrower sense of determining VNTR length (number of repeating units). Most VNTRs are highly multi-allelic due to their variable nature. VNTRs gain and lose repeat units at high rates due to polymerase slippage during DNA replication [27]. Due to this error prone replication process, VNTRs have been reported to have a genome-wide average mutation rate of 10^{-5} [10], which is orders of magnitude higher than most other types of *de novo* variations.

Despite the advent of sequence based genotyping, repetitive sequences continue to be challenging for genomic analysis. For example, 'stutter errors' due to polymerase slippage during PCR amplification change VNTR length and reduce genotyping accuracy[134]. For these reasons, VNTRs have been largely missing from genome-wide studies due to technical challenges of genotyping and the computational expense and the majority of the VNTRs have not been studied well.

1.3 Contributions of this thesis

In this thesis, I present our work toward identifying VNTR variations in the human genome and assessing their effect on human disease. In the following chapters, I will present our progress on developing efficient computational algorithms to profile VNTRs from high throughput sequencing data and identify possible variations within them. I applied our method to generate the largest catalog of VNTR genotypes to this date, which provides insights into the landscape of VNTR variations in different populations. I show the contribution of tandem repeats in mediating expression levels of key genes with known associations to neurological disorders

and familial cancers, and argue the causality of this relation. Finally, I will describe our efforts to directly understand the impact of these variations on human phenotypes, which improves our understanding of genetic architecture of complex diseases.

1.3.1 Computational tools to genotype VNTR variations

Traditionally, VNTR genotyping required labor intensive gel-based screens which limited the size of large population based studies of VNTRs [101]. Whole genome sequencing has the potential to detect and genotype all types of genetic variation, including VNTRs. However, computational identification of variation in VNTRs from sequence remains challenging. Existing variant calling methods have been developed primarily to identify short sequence variants in unique DNA sequences that fall into a reference versus alternate allele framework, which is not well suited for detecting variation in VNTR sequences.

Chapter 2 describes the method we developed, adVNTR, an algorithm to identify VNTR variations from high throughput sequencing data [10]. It utilizes a unique alignment strategy by training locus-specific Hidden Markov Models (HMMs) for VNTR loci. Then, it finds the number of repeats and possible point mutations within each VNTR locus using statistical learning methods to account for possible sequencing noises.

1.3.2 Contribution of VNTR variations to gene expression mediation

With the lack of an efficient method for genotyping VNTRs, large-scale studies of VNTRs and their association with gene expression have been limited when compared to other sources of human variation such as SNPs and CNVs[11, 75, 22]. Therefore, 'missing heritability'–the gap between estimates of heritability, measured for example by twin studies[47, 138], and phenotypic variation explained by genomic variation– remains a limitation for eQTL studies[85]. It has been speculated that the inclusion of tandem repeats in association analyses may reduce this heritability

gap[56, 85, 17].

Chapter 3 assesses the impact of VNTR variations on gene expression levels and describes our efforts to reduce heritability gap for genetic variations. With a robust method to leverage high throughput sequencing data for studying VNTRs, we were in a position to do large scale VNTR genotyping in more than 2000 individuals for the first time [9]. Subsequently, we identify contributions of VNTRs on cell mechanisms and specifically mediating gene expression levels in 46 different cell tissues.

1.3.3 Case study of VNTR variation effects on Breast Cancer

For carriers of pathogenic BRCA1 or BRCA2 mutations (BRCA), the lifetime risk of developing breast cancer (up to an 80% lifetime risk) is a six-fold increase over that of average risk women and ovarian cancer risk (up to a 44% lifetime risk) is up to a 30-fold increase [68]. Despite higher average risk, penetrance is incomplete (not all carriers will develop cancer) and age at cancer diagnosis varies. The variation in risk, even in identical mutation carriers, suggests that modifier factors, both genetic and environmental, affect cancer risks[79].

Through genome-wide association studies (GWAS), single nucleotide polymorphisms (SNPs) have been identified to better define those at higher and lower risk of developing breast cancer (e.g., [24, 69, 89]). However, these modifier variants explain only a portion of the variation in risk, particularly for women carrying BRCA1 mutations [93]. Identifying additional genetic modifiers will facilitate better risk estimates for clinical decision-making on timing and options for prevention. Chapter 4 presents a systematic, genome-wide investigation of the role of VNTRs as the causal modifiers of breast cancer risk in BRCA1 and BRCA2 pathogenic mutation carriers.

Chapter 2

Computational tools to genotype VNTR variations

Whole Genome Sequencing is increasingly used to identify Mendelian variants in clinical pipelines. These pipelines focus on single nucleotide variants (SNVs) and also structural variants, while ignoring more complex repeat sequence variants. We consider the problem of genotyping *Variable Number Tandem Repeats* (VNTRs), composed of inexact tandem duplications of short (6-100bp) repeating units. VNTRs span 3% of the human genome, are frequently present in coding regions, and have been implicated in multiple Mendelian disorders. While existing tools recognize VNTR carrying sequence, genotyping VNTRs (determining repeat unit count and sequence variation) from whole genome sequenced reads remains challenging. We describe a method, adVNTR, that uses Hidden Markov Models to model each VNTR, count repeat units, and detect sequence variation. adVNTR models can be developed for short-read (Illumina) and single molecule (PacBio) whole genome and exome sequencing, and show good results on multiple simulated and real data sets.

2.1 Introduction

Next Generation Sequencing (NGS) is increasingly used to identify disease causing variants in clinical and diagnostic settings, but variant detection pipelines focus primarily on single nucleotide variants (SNVs) and small indels and to a lesser extent on structural variants. The human genome contains repeated sequences such as segmental duplications, short tandem repeats, and minisatellites which pose challenges for alignment and variant calling tools. Hence, these regions are typically ignored during analysis of NGS data. In particular, *tandem repeats* correspond to locations where a short DNA sequence or *Repeat Unit* (RU) is repeated in tandem multiple times. RUs of length less than 6bp are classified as Short Tandem Repeats (STRs), while longer RUs spanning potentially hundreds of nucleotides are denoted as *Variable Number Tandem Repeats* (VNTRs)[116, 139].

VNTRs span 3% of the human genome and are often found in coding regions where the repeat unit length is a multiple of 3 resulting in tandem repeats in the amino acid sequence. More than 1,200 VNTRs with a RU length of 10 or greater exist in the coding regions of the human genome[126]. Compared to STRs, which have been extensively studied [53, 127, 84, 135, 29], VNTRs have not received as much attention. Nevertheless, multiple studies have linked variation in VNTRs with Mendelian diseases (*e.g.*, Medullary cystic kidney disease[64], Myoclonus epilepsy[72], and FSHD[78]) and complex disorders such as bipolar disorder (Table 2.1). In some cases, the disease associated variants correspond to point mutations in the VNTR sequence [64, 107] while in other cases, changes in the number of tandem repeats (RU count) show a statistical association (or causal relationship) with disease risk. For example, the insulin gene (*INS*) VNTR has an RU length of 14 bp with RU count varying from 26 to 200[104]. Variation in this VNTR has been associated with expression of the *INS* gene and risk for type 1 diabetes (OR = 2.2) [32]. Notwithstanding these examples, the advent of genome-wide SNP genotyping arrays led to VNTRs being largely ignored. They have been called 'the forgotten polymorphisms'[17].

VNTRs were originally used as markers for linkage mapping since they are highly polymorphic with respect to the number of tandem repeats at a given VNTR locus[42]. Traditionally, VNTR genotyping required labor intensive gel-based screens which limited the size of large population based studies of VNTRs [101]. Whole genome sequencing has the potential to detect and genotype all types of genetic variation, including VNTRs. However, computational identification of variation in VNTRs from sequence remains challenging. Existing variant calling methods have been developed primarily to identify short sequence variants in unique DNA sequences that fall into a reference versus alternate allele framework, which is not well suited for detecting variation in VNTR sequences.

Genotyping VNTRs in a donor genome sequenced using short (Illumina) or longer single molecule reads, requires the following: (a) *recruitment of reads* containing the VNTR sequence; (b) *counting RUs* for each of the two haplotypes; (c) *identification of indels within VNTRs*; and (d) identification of mutations within the VNTR. Mapping tools such as BWA[82] and Bowtie 2[74] can work for read recruitment for STRs, but are challenged by insertion/deletion of larger repeat units. Mapping issues also confound existing variant callers, including realignment tools such as GATK IndelRealigner[26] if the total VNTR length is larger than the read length. This is

Table 2.1: **Disease-linked VNTRs** are generally distinguished from STRs by a longer length (≥ 6) of the repeating unit. 'M' denotes Mendelian inheritance, while 'A' represents possibly complex inheritance captured via Association. As it is difficult to genotype VNTRs, most cases have been determined via association, but the inheritance mode could be high penetrance.

Gene	Chr	Unit len	Number of units		Annotation	Inheritance	Disease
			Normal	Pathogenic			
PER3	1	54	4	5	coding	A	Bipolar disorder[12]
MUC1	1	60	11-12	single insertion	coding	M	MCKD1[64]
IL1RN	2	86	3-6	2	intron	A	Stroke, CAD[137]
DUX4	4	3.3kb	11-100	1-10		M	FSHD[78]
DAT1	5	44	7-11	10 (ADHD)	UTR	A	ADHD, Parkinson's[40, 65]
MUC21	6	45	26-27	4 bp deletion	coding	A	Diffuse panbronchiolitis (DPB)[58]
CEL	9	33	11-21	single deletion	coding	M	Monogenic diabetes[107]
INS	11	14-15	26-200	26-44 (T1D)	promoter	A	T1D;T2D;Obesity[104, 32]
DRD4	11	48	2-11	7	coding	A	OCD, ADHD[71, 130]
ACAN	15	57	27-33	13-25	coding	A	Osteochondritis dissecans[37]
ZFHX3	16	12	4-5		coding	A	Kawasaki
GP1BA	17	39	1-4	2/3 genotype	coding	A	ATF in Stroke[20]
SLC6A4	17	16-17	9/10/12		intron	A	BPSD, Alzheimer's[55, 103]
SLC6A4	17	22	14	16 (OCD)	promoter	A	OCD,Anxiety, Schizophrenia[55]
HIC1	17	70	1-4	5+/5+	promoter	A	Metastatic Colorectal Cancer[100]
MMP9	20	12	5-6		coding	A	Kawasaki
CSTB	21	12	2-3	12+	5'UTR	M	Progressive myoclonic epilepsy 1A[72]
MAOA	X	30	2-5	4	promoter	A	Bipolar disorder[18]

because reads contained within the VNTR sequence have multiple equally likely mappings and therefore will be mapped randomly to different locations with low mapping quality citeKirby2013. Detection of point mutations in long VNTRs requires integrating information across the entire VNTR sequence. For VNTRs whose total sequence length (RU count times the RU length) is much longer than the read length, detection of SNVs and indels is not feasible using existing variant callers. We focus mainly on problems (a,b) relating to recruitment and RU counting. For problem (c), we focus on difficult case of large ($\geq$ 250bp) VNTRs within coding regions where the indel shifts the translation frame. We do not tackle problem (d) in this manuscript.

Other tools have addressed the problem of RU count estimation, focusing on the related problem of STR genotyping. Some of these tools do not accept large repeating patterns as input [135, 84]. Others require all repeat units to be near-identical[29, 127]. In particular, ExpansionHunter[29] looks for exact matches of short repeating sequence within flanking unique sequences, and works for STRs, but not as well with the larger VNTRs with variations in RUs (Results). VNTRseek[42] detects a VNTR-like pattern in reads and aligns it to tandem repeats, but uses a complex alignment process making it difficult to run the tool. Alignment based tools need to align reads at both unique ends, which may not be possible for short (Illumina) reads. Single molecule reads (*e.g.*, PacBio[36], Nanopore[23]) can span entire VNTR regions, but it is difficult to estimate the RU count directly since the distance between the flanking regions varies dramatically from read to read due to an excess of indel errors. For example, 14 reads spanning the *SLC6A4* VNTR in the in the PacBio sequencing data of NA12878 individual from Genome in a Bottle[143] included fifteen distinct lengths between 292bp and 385bp, leading to length-based RU count estimates 13, 14, 15, 16, and 18 for the diploid genome.

In contrast to methods like VNTRseek which seek to *discover/identify* VNTRs, we describe a method, adVNTR, for *genotyping VNTRs* at targeted loci in a donor genome. For any target VNTR in a donor, adVNTR reports an estimate of RU counts and point mutations within

the RUs. It trains Hidden Markov Models (HMMs) for each target VNTR locus, which provide the following advantages: (i) it is sufficient to match any portions of the unique flanking regions for read alignment; (ii) it is easier to separate homopolymer runs from other indels helping with frameshift detection, and to estimate RU counts even in the presence of indels; (iii) each VNTR can be modeled individually, and complex models can be constructed for VNTRs with complex structure, along with VNTR specific confidence scores. For longer VNTRs not spanned by short reads, adVNTR can still be used to detect indels, while providing lower bounds on RU counts. Also, exact estimates for RU counts could be made for shorter VNTRs. Using simulated data as well as whole-genome sequence data for a number of human individuals, we demonstrate the power of adVNTR to genotype VNTR loci in the human genome.

2.2 Method

A VNTR sequence can be represented as $SR_1R_2 \ldots R_uP$, where S and P are the unique flanking regions, and $R_i (1 \leq i \leq u)$ correspond to the tandem repeats. For each i, j, R_i is similar in sequence to R_j, and the number of occurrences, u, is denoted as the *RU count*. We do not impose a length restriction on S and P, but assume that they are long enough to be unique in the genome. For genotyping a VNTR in a donor genome, we focus primarily on estimating the diploid RU counts (u_1, u_2). However, many ($\sim 10^3$) VNTRs occur in coding regions, and mutations, particularly frameshift causing indels, are also relevant. Our method, adVNTR, models the problems of RU counting and mutation detection using HMMs trained for each target VNTR. adVNTR requires a one-time training of models for each combination of a VNTR and sequencing technology, although the user has the option to retrain models. Once models are trained, it has three stages for genotyping: (i) Read recruitment; (ii) RU count estimation; and, (iii) variant (indel) detection. We describe the training procedure and the three modules below.

2.2.1 HMM Training.

The goal of training is to estimate model parameters for each VNTR and each sequencing technology. Previous works have shown that an HMM with three groups of states could be used to find similarities between biological sequences [34]. In this model, a profile-HMMs can model a groups of sequences. Then, a new sequence can be aligned to a profile HMM to discover sequence family[66]. We use an HMM architecture with three parts, which have their own three groups of states (Fig. 2.1). The first part matches the 5' (left) flanking region of the VNTR.

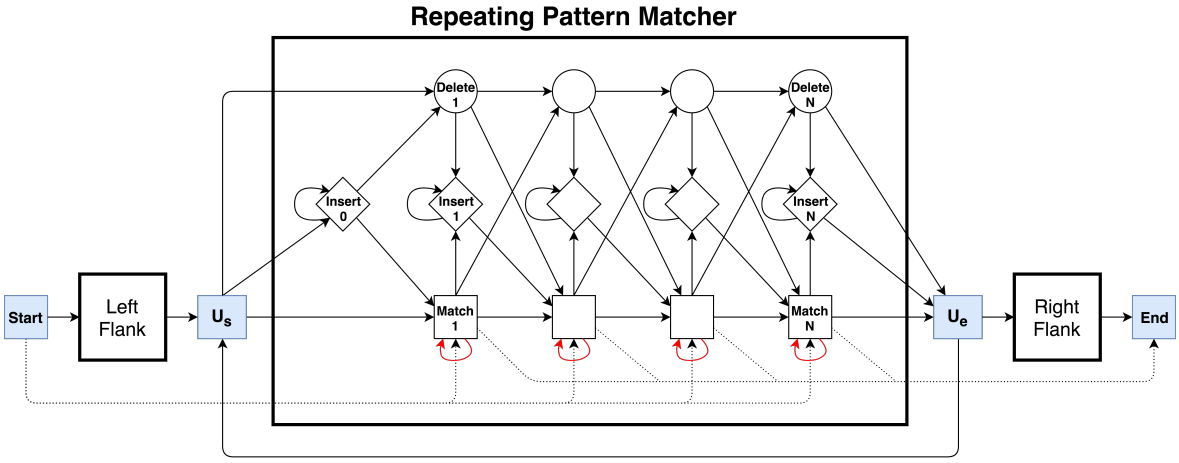


Figure 2.1: **The VNTR HMM.** The HMM is composed of 3 profile HMMs, one each for the left and right flanking unique regions, and one in the middle to match multiple and partial numbers of RUs. The special states U_s ('Unit-Start'), and U_e ('Unit-End') are used for RU counting. Dotted lines refer to special transitions for partial reads that do not span the entire region.

The second part is an HMM which matches an arbitrary number of (approximately identical) repeating units. The last part matches the 3' (right) flanking region (Fig. A.1). The RU pattern is matched with a profile HMM (*RU HMM*), with states for matches, deletions, and insertions, and its model parameters are trained first. To train RU HMM for each VNTR, we collected RU sequences from the reference assembly[73] and performed a multiple sequence alignment[35]. Let $h(i, j)$ denote the number of observed transitions from state i to state j in hidden path of each sequence in multiple alignment, and $h_i(\alpha)$ denote the number of emissions of α in state i. We define permissible transition (arrows in Fig. 2.1) and match-state emission probabilities as

follows:

$$T(i,j) = \frac{h(i,j)+b_0}{\sum_{i\to l}(h(i,l)+b_0)}, \quad E_i(\alpha) = \frac{h_i(\alpha)+b_1}{\sum_{\alpha'}(h_i(\alpha')+b_1)} \quad \text{for } \alpha,\alpha' \in \{A,C,G,T\}.$$

Non-permissible transitions have probability 0, and $h_i(\alpha) = 1/4$ for insert state i and 0 for deletions. The pseudocounts b_0 and b_1 were estimated by initially setting them to the error rate of the sequencing technology, but they (along with other model parameters) were updated after aligning Illumina or PacBio reads to the model. The RU HMM architecture was augmented by adding (a) transitions from U_e to U_s to allow matching of variable number of RU; (b) adding the HMMs for the matching of any portions of left and right flanking sequences; and (c) by adding transitions to match reads that match either the left flanking or the right flanking region. In addition, reads anchored to one of the unique regions can jump past the other HMM using dotted arrows.

While error correction tools for PacBio have been developed, most do not work for repetitive regions,[54, 111, 7, 91, 77, 92] and others assume a single haplotype for error correction[112, 14]. In contrast, the HMM allows us to model many of the common (homopolymer) errors directly. Insertion deletion errors are common in single molecule sequencing particularly in homopolymer runs of length ≥ 6, and occur mostly as insertions in the homopolymer run[21]. Consider a match state i with highest emission probability for nucleotide α. The transition probability $T(i,i)$ from a match state i to itself was set based on the match probabilities of α in previous $k = 6$ states. The model parameters were further updated using genome sequencing data of NA12878 (Supplementary Material A.1).

2.2.2 Read Recruitment.

The first step in adVNTR is to *recruit* all reads that match a portion of the VNTR sequence. Alignment-based methods do not work well due to changes in RU counts (See Results), but the

adVNTR HMM allows for variable RU count. To speed up recruitment, we used an Aho-Corasick keyword matching algorithm available as part of the BLAST package[4] to identify all reads that match a keyword from the VNTR patterns or the flanking regions. Note that the dictionary construction is a one-time process, and all reads must be scanned once for filtering. The keyword size and number of keywords were empirically chosen for each VNTR. Filtered reads were aligned to the HMM using the Viterbi algorithm. Only reads with matching probability higher than a specified threshold were retained. To compute the selection threshold for each VNTR, we aligned non-target genomic sequences that passed the keyword matching step to the HMM to form an empirical false distribution. Subsequently, we aligned VNTR encoding sequences to the HMM to form the score distribution of true reads. Then, we used a Naïve Bayes classifier to select a threshold.

2.2.3 Estimating VNTR RU Counts.

All reads covering an RU element are aligned, or 'matched' to the HMM using the Viterbi algorithm to create, in effect, a new multiple alignment. Recalling the Viterbi algorithm, let $V_{k,j}$ denote the highest (log) probability of emitting the first k letters of the sequence $s_1, s_2, \ldots s_n$ and ending in state j of an HMM. Let, $\text{Prev}_{k,j}$ denote the state j' immediately prior to j in this optimum parse. Then,

$$V_{k,j} = \max_{j'}\{V_{k',j} + \log T(j', j) + \log E_j(s_k)\}, \tag{2.1}$$

$$\text{Prev}_{k,j} = \arg\max_{j'}\{V_{k',j} + \log T(j', j) + \log E_j(s_k)\}, \tag{2.2}$$

where, $k' = k-1$ for match or insert states; $k' = k$ otherwise. For each read, the Viterbi algorithm allows for the enumeration of the maximum likelihood (ML) path by going backwards from $\text{Prev}(\text{End}, n)$. Ignoring all but the U_s and U_e states in the Viterbi path, we get a pattern of the form $U_e^{k_1}(U_sU_e)^{k_2}U_s^{k_3}$ with $k_1, k_3 \in \{0,1\}$, and $k_2 \geq 0$. We estimate the RU count of the read as

$k_1 + k_2 + k_3$, and mark it as a lower bound if $k_1 + k_3 > 0$ (see Fig. 2.2 for an example).

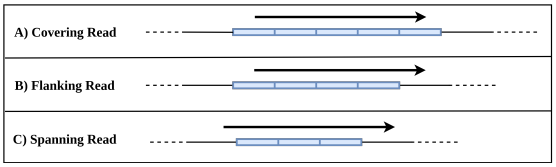


Figure 2.2: **Estimates of RU counts using recruited reads.** (A) $(k_1, k_2, k_3) = (1, 3, 1)$; RU count ≥ 5. (B) $(k_1, k_2, k_3) = (0, 3, 1)$; RU count ≥ 4 (C) $(k_1, k_2, k_3) = (0, 3, 0)$; RU count $= 3$.

One of the main reasons for erroneous RU counts is stutter during PCR amplification. The PCR amplification process is similar to replication errors that result on genetic RU count variation during cell-division, except that there are multiple rounds of amplification. In each PCR round, the number of copies might change by 1 with some probability. Once a single event has occurred and an erroneous template is generated, the event of having another change is likely to be independent of the previous event[50]. To model errors in read counts, we define parameter r_ε s.t. r_ε^Δ is the probability of RU counting error by $\pm\Delta$ in the estimation of the true count. Thus the probability of getting the correct count is $1 - r$, where

$$r = 2(r_\varepsilon + r_\varepsilon^2 + r_\varepsilon^3 + \ldots) = \frac{2r_\varepsilon}{1 - r_\varepsilon}$$

The analysis of reads at a VNTR gives us a multi-set of RU counts (or lower bounds) $c_1, c_2, \ldots, c_n$. We assume that the donor genome is diploid but do not require any phasing information in the computation of the multi-set. Additionally, we allow the possibility that all reads are sampled from one haplotype with the RU count of the missing haplotype being X. We define $C = \{c_1, c_2, \ldots, c_n\} \cup \{X\}$ and use C to get a list of possible genotypes (c_i, c_j) with $c_i \leq c_j$. Then, the

conditional likelihood of a read with RU count c is given by:

$$\Pr(\text{RU} = c|(c_i,c_j)) = \begin{cases} 1-r & c = c_i = c_j \\ \frac{1}{2}((1-r)+r_\varepsilon^{|c-c_j|}) & c = c_i \\ \frac{1}{2}((1-r)+r_\varepsilon^{|c-c_i|}) & c = c_j \\ \frac{1}{2}(r_\varepsilon^{|c-c_j|}+r_\varepsilon^{|c-c_j|}) & c \neq c_i, c \neq c_j \\ (\frac{1}{2})(1-r) & c = c_i, c_j = X \end{cases}$$

Similarly, the likelihood of a read with a lower bound c on the RU count is given by:

$$\Pr(RU \geq c|(c_i,c_j)) = \begin{cases} (1-r) & c \leq c_i \\ \frac{1}{2}(1-r) & c_i < c \leq c_j \\ r & c > c_j \end{cases}$$

The likelihood of the data C is given by $\prod_{c_k \in C} \Pr(c_k|(c_i,c_j))$. The posterior genotype probabilities can be computed using Bayes' theorem:

$$\Pr((c_i,c_j)|C) = \frac{\Pr(C|(c_i,c_j))\Pr((c_i,c_j))}{\sum_{c_{i'},c_{j'} \in C} \Pr(C|(c_{i'},c_{j'}))\Pr((c_{i'},c_{j'}))} \tag{2.3}$$

We generally set equal priors. However, in the event that we only see reads with a single count c', we choose $\Pr((c',c')) = \Pr((c',X)) = \frac{1}{2}$. The probability of "missing haplotype" event is modeled as a Bernoulli process since in genome sequencing, sampling from either chromosome is done at random and so, the probability of not observing a halplotype in each read (failure) is $1/2$. If we see multiple counts, we set $\Pr((c',X)) = 0$ for all $c' \in C$, and give equal priors to all other genotypes.

2.2.4 VNTR Mutation Detection.

It is not difficult to see that alignment based methods do not work well in VNTRs. Changes in RU counts make it difficult to align reads even for mappers that allow split-reads, as the gaps in different reads can be placed in different locations. A similar problem appears with small indels, as there are multiple ways to align reads with an indel in a Repeat Unit. The adVNTR HMM aligns all repeat units to the same HMM, and this has the effect of aligning all mutations/indels in the same column. Consider the case where reads contain a total of v nucleotides matching a VNTR RU of length ℓ, and RU count u. Moreover at a specific position covered by d Repeats, suppose we observe ι indel transitions.

For a true indel mutation, we expect $\frac{u\ell}{v}$ fraction of transitions at a location to be an indel, giving a likelihood of the observed data as $\text{Binom}(d, \iota, \frac{u\ell}{v})$. Alternatively, for a homopolymer run of $i > 0$ nucleotides, let ε_i denote the per-nucleotide indel error rate. We modeled ε_1 empirically in non-VNTR, non-polymorphic regions and confirmed prior results that ε_i increases with increasing i[87]. Thus, the likelihood of seeing ι indel transitions due to sequencing error in a homopolymer run of length i is $\text{Binom}(d, \iota, \varepsilon_i)$. We scored an indel in the VNTR using the log-likelihood ratio

$$-2\ln\left(\frac{\text{Binomial}(d, \iota, \frac{u\ell}{v})}{\text{Binomial}(d, \iota, \varepsilon_i)}\right), \tag{2.4}$$

which follows a χ^2 distribution. We select the indel if the nominal p-value is lower than 0.01.

Command line usage of adVNTR for RU count genotyping and frameshift identification is available in Supplementary Material A.4.

2.3 Results

Our method, adVNTR, requires training of separate HMM models for each combination of target VNTR and sequencing technologies. The detailed training procedure is described in

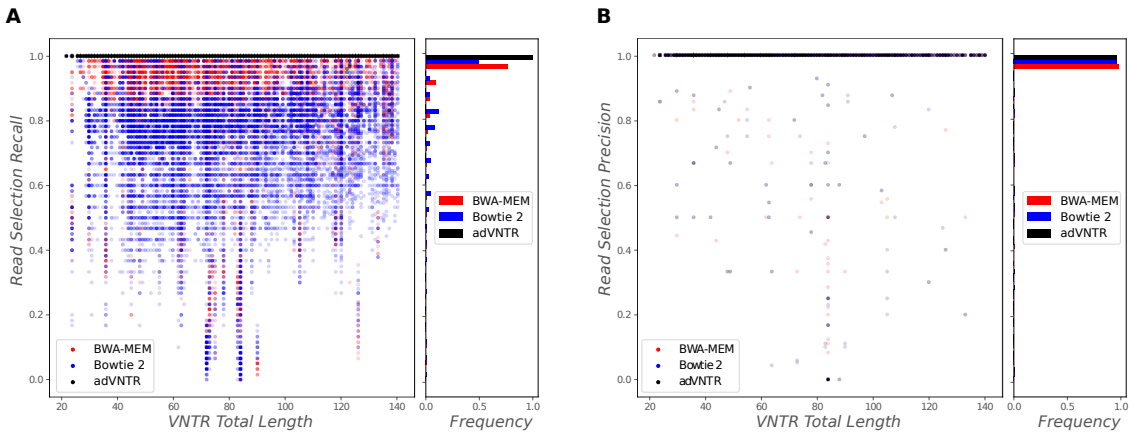


Figure 2.3: **Read recruitment quality on Illumina reads.** (A) Comparison of the recall (# true recruited reads/ # true reads) of adVNTR read recruitment against BWA-MEM and Bowtie 2, as a function of VNTR length for 1775 VNTRs with different counts (31,788 tests). Each dot corresponds to a separate test. (B) Precision (# true recruited reads/ # recruited reads) of read recruitment.

Methods. Given trained models, adVNTR genotypes the VNTRs in three stages: (i) Selection of reads that contain VNTR locus (read recruitment); (ii) RU count estimation; and, (iii) variant detection. We report results on performance of adVNTR in each of these stages using simulated and read datasets based on short-read (Illumina) and single molecule (PacBio) technologies.

2.3.1 HMM training.

Initial HMMs were trained using multiple alignments of RU sequences from the reference assembly hg19[73], as described in methods. Similarly, HMMs were trained for the left flanking and right flanking regions for each VNTR. The HMM models were augmented using data from Genome in a Bottle (GIAB) project (NA12878 WGS). VNTR models were trained for VNTRs in coding and promoter regions of the genome, for both Illumina (1755 models) and PacBio (2944 models; Supplementary Material A.2). Subsequently, we tested performance for (a) read-recruitment, (b) counting of Repeat Units, and (c) detection of indels.

2.3.2 Test Data.

To evaluate performance for *PacBio*, we simulated haplotypes for each of the 2944 VNTRs, revising the RU count to be ± 3 of the RU count in hg19, and setting 1 as the minimum RU count. We simulated haplotype reads ($15\times$ coverage) using SimLoRD[120] and aligned those reads to hg19 using BLASR[21]. For Illumina sequencing, we used ART[60] to simulate haplotype WGS (shotgun 150bp) reads at $15\times$ coverage for each VNTR and simulated VNTR haplotype with changes in RU counts similar to PacBio. Pairs of haplotypes were merged to get ($30\times$ coverage) diploid samples. The resulting data-sets were called *PacBioSim* and *IlluminaSim*, respectively (Supplementary Material A.3, Table A.1). To evaluate performance of frameshift identification, we collected a set of 115 VNTRs (Supplementary Material A.2). For each VNTR, we simulated haplotypes that contain a deletion or an insertion in the VNTR (Supplementary Material A.3). We simulated reads from each of these haplotypes and merged pairs of halpotypes to obtain diploid samples. We denote this data-set as *IlluminaFrameshift*.

2.3.3 Read recruitment.

adVNTR takes a collection of VNTR models as input, and as a first step, recruits reads that map to any of the VNTRs in the list. In testing recruitment for PacBio, we found that alignment tools such as BLASR perform well in recruiting VNTR reads even in the presence of deletions and insertions and used BLASR for all read recruitment. For *Illumina* reads, we tested adVNTR read-recruitment for all 1775 VNTRs using IlluminaSim, and compared against mapping tools BWA-MEM, Bowtie 2, and BLAST. adVNTR achieves much greater recall while maintaining or exceeding the precision of other tools (Fig. 2.3 and Fig. A.3). Specifically, adVNTR recall was 100% for 99.9% of the VNTRs, whereas the next best tool (BWA-MEM) achieved this only for 68.2% of the VNTRs. The other mapping tools lose mapping sensitivity when RU counts are increased or decreased (large indels), and perform best when the RU counts are the same as

reference (Fig. A.2A-C), partially explaining their lower recall.

2.3.4 VNTR genotyping using PacBio reads.

Recall that sequencing (particularly homopolymer) errors can cause lengths to change, particularly for short RU lengths and larger RU counts. To test adVNTR performance on PacBioSim, we compared against a naïve method that estimates RU counts based on read length between the flanking regions from the consensus of reads that cover VNTR. Detailed performance on three exemplars (*INS*, *CSTB*, and *HIC1*) gene showed high genotype accuracy for adVNTR over a wide range of RU counts, and coverage (Fig. 2.4A. Similar results were obtained for all 2944 VNTRs (Fig. 2.4B). Overall, 98.45% of adVNTR estimates were correct while 26.45% of estimates made by naïve method were correct. As it is difficult for the naïve mthod to call heterozygotes, we also compared on the subset of test data with homozygous RU counts. 97.95% of adVNTR estimates were correct, while the consensus method was correct in 66.16% of samples (Fig. A.4). adVNTR estimates were uniformly good except at low sequence coverage. To test for accuracy with changing RU counts, we simulated different RU counts for individuals at 3 VNTRs (Table A.4). adVNTR RU counts showed 100% accuracy in each of the 52 different samples tested.

To test performance on real data where the true VNTR genotype was not known, we checked for Mendelian inheritance consistency in the AJ trio from Genome in a Bottle (GIAB)[143] and a Chinese Han trio from NCBI SRA (accession PRJEB12236). On four disease related VNTRs, adVNTR predictions were consistent in each case (Fig. 2.4C). On the 2944 genic VNTRs, the trio consistency of adVNTR calls was correlated with coverage. At a posterior probability threshold of 0.99, 86.98% of the calls in the AJ trio, and 97.08% of the calls in the Chinese trio, were consistent with Mendelian inheritance (Fig.2.4E). Many of the discrepancies could be attributed to low coverage and missing data. Increasing sequence coverage threshold from 5$\times$ to 10$\times$ increased the average posterior probability from 0.91 to 0.98 and resulted in improved RU

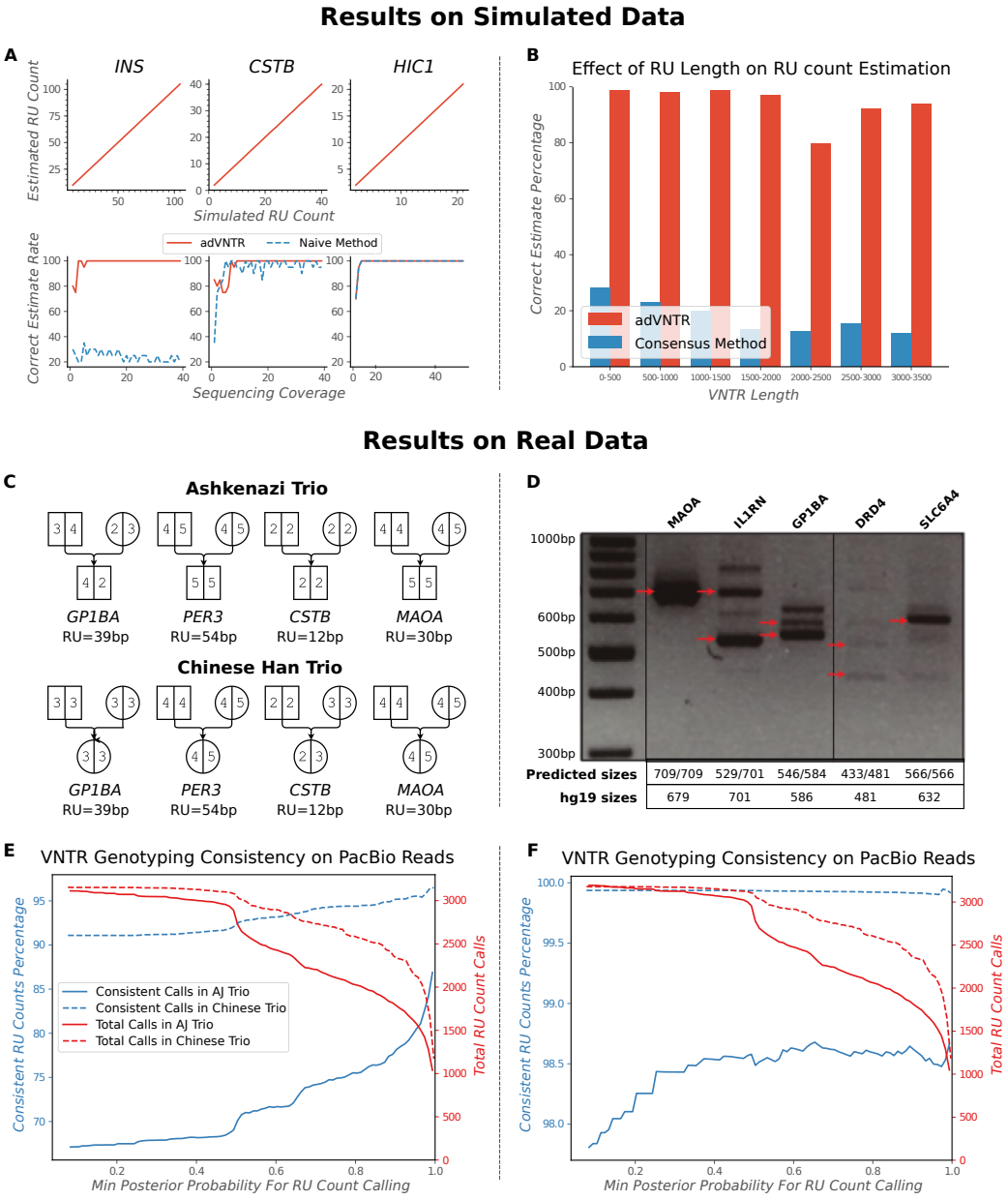


Figure 2.4: **VNTR genotyping using sequencing PacBio data.** (A) RU count estimation on simulated PacBio reads as a function of RU count and coverage for 3 medically relevant VNTRs: *INS* (RU length 14bp), *CSTB* (12 bp), and *HIC1* (70bp). adVNTR performance is compared to a naïve method. (B) The effect of RU length on count accuracy over 2944 VNTRs (30418 tests). (C) Mendelian consistency of genotypes at 4 VNTR loci in the Chinese Han and Ashkenazi trios. Note that *MAOA* results are consistent with its location on Chr X. (D) LR-PCR based validation of genotypes at 5 disease-linked VNTRs in NA12878. Red arrow correspond to VNTR lengths estimated by multiplying predicted RU counts with RU lengths. (E) Fraction of consistent calls and number of calls across 2944 VNTRs in AJ and Chinese trios from GIAB and NCBI-SRA. (F) Fraction of consistent calls allowing for off-by-one errors.

count accuracy (Fig. A.5). Also, many of these discrepancies in RU counts were off-by-one errors (Fig. A.6). These off-by-one discrepancies could be acceptable for Mendelian disease testing as the pathogenic cases often have large changes in RU counts. Treating the off by one counts as correct, we found that 98.66% and 99.91% of the high confidence calls in AJ and Chinese trios, respectively, were consistent (Fig.2.4F). Finally, some of the off-by-one counts could be natural genetic variation.

We also performed a long range (LR)PCR experiment on the individual NA12878 to assess the accuracy of the adVNTR genotypes using PacBio data (Table A.2 and Table A.3). The observed PCR product lengths (black bands in Fig. 2.4D) were consistent with the adVNTR predictions (red arrows), while being different from the hg19 reference RU count. adVNTR correctly predicted all VNTRs to be heterozygous with the exception of *SLC6A4*, that was predicted to be homozygous.

While we could not get the VNTR discovery tool VNTRseek[42] to run on our machine (personal communication), we observed that the authors had predicted 125 VNTRs in the Watson sequenced genome[133], and 75 VNTRs in two trios as being polymorphic. In contrast, analysis of the PacBio sequencing data identified $>$500 examples of polymorphic VNTRs that overlap with coding regions. The results suggest that variation in RU counts of VNTRs and their role in influencing phenotypes might be greater than previously estimated.

2.3.5 VNTR genotyping using Illumina.

The adVNTR estimate correctly matched both RU counts in 91.6% of the cases in the IlluminaSim dataset (1775 VNTRs with up to 21 diploid RU counts each) and matched at least one RU count in 97% of the cases (Fig. 2.5A,B). Most of the discrepancies occurred in VNTRs with longer lengths not covered by Illumina reads (Fig. 2.5C,D). While there was a drop in accuracy for increasing lengths, 84% of the genic VNTRs are shorter than 150bp, and could be genotyped with 94.6% accuracy. Tools such as VNTRseek require at least 20bp flanking each

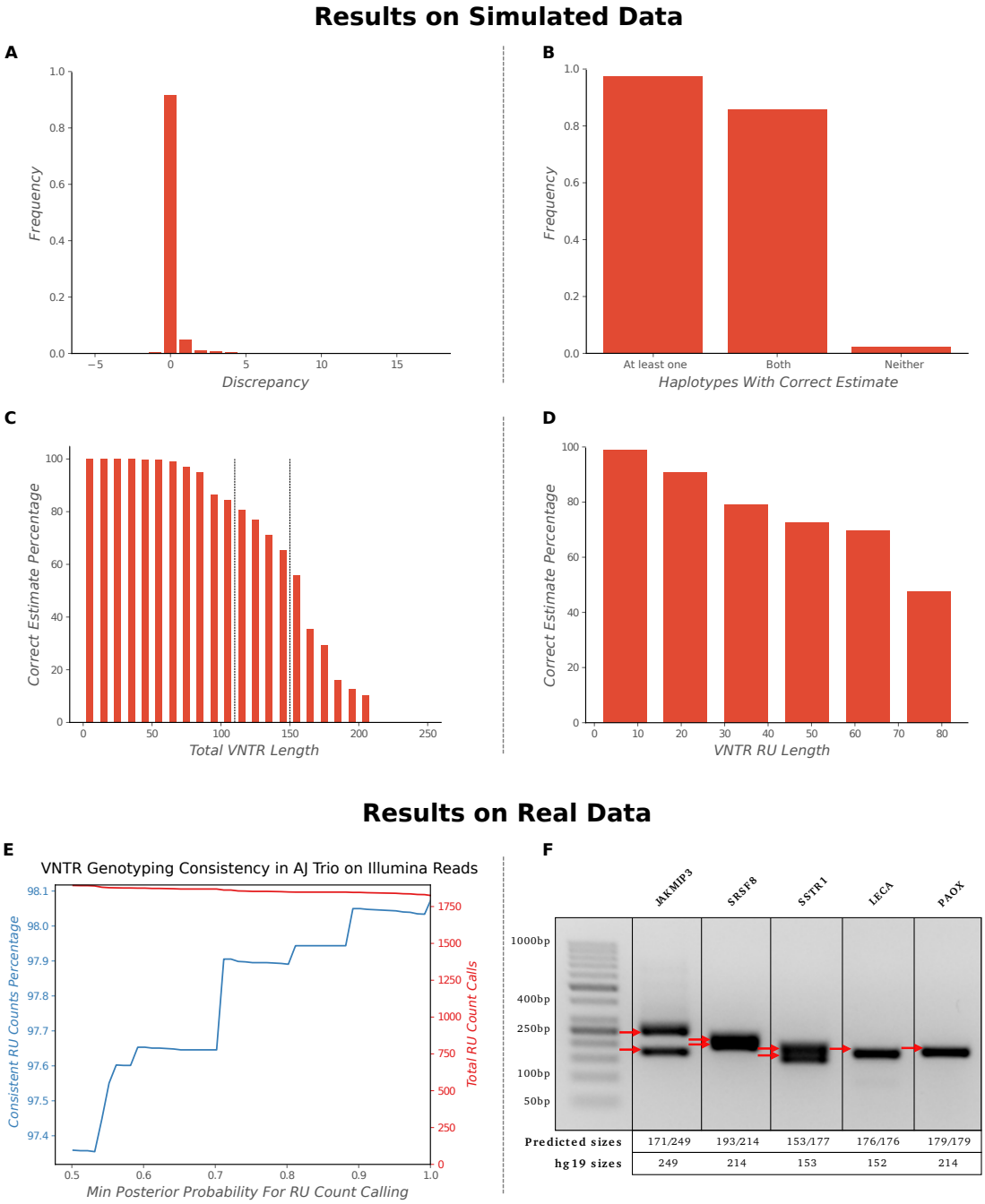


Figure 2.5: **VNTR genotyping using Illumina sequencing data.** (A-D) Correctness of RU count prediction for 1775 coding VNTRs in the IlluminaSim dataset, described by (A) RU count dscrepancy, (B) haplotypes with correct estimates, (C) correctness as a function of VNTR length, and (D) RU length. (E) Consistency of adVNTR calls on the AJ trio WGS data from GIAB. Red line describes the cumulative number of calls made at specific posterior probability cut-offs. (F) Gel electrophoresis based validation of adVNTR calls on 5 short VNTRs using WGS of individual NA12878 from GIAB. Red arrows correspond to VNTR lengths estimated by multiplying the RU lengths with the estimated RU counts.

side of the VNTR and do not return a result for VNTRs with total length greater than 110bp, while adVNTR could predict the genotype correctly in a majority of those cases (Supplementary Material A.5). ExpansionHunter, a tool designed primarily for STR genotyping [29] provided incorrect estimates in over 90% cases from this data-set (Fig. A.7). ExpansionHunter makes the assumption that the different RUs are mostly identical in sequence which is valid for STRs but not for most VNTRs, and we tested this through 52 samples on three VNTRs. adVNTR predicted the correct genotype in all but 6 cases, with erroneous calls only in the case of high RU counts where the read length did not span the VNTR perfectly, while ExpansionHunter did not return the correct estimate in most cases (Table A.4).

On the AJ trio from GIAB, 98.08% of the high confidence adVNTR calls were consistent with Mendelian inheritance (Fig. 2.5E). Note that 95.93% of all calls were high confidence (posterior probability ≥ 0.99). We validated adVNTR calls on 12 VNTRs using Gel electrophoresis (Table A.3). adVNTR predicted the correct RU counts in all cases, except in two cases where the PCR primers failed to produce a band (Fig. 2.5F, A.8). We also compared adVNTR against ExpansionHunter on 7 disease related short VNTRs in the AJ trio and obtained similar results (Table A.5).

To test adVNTR for population-scale studies of VNTR genotypes using WGS data replacing labor intensive gel electrophoresis[18, 20], we scanned the PCR-free WGS data for 150 individuals (50 in each population) obtained from 1000 genomes project[124]. We observed population specific RU counts (frequency difference $> 10\%$) in 97 of 202 VNTRs tested (Table S7). Fig. 2.6 shows the RU count frequencies for a disease-linked VNTR in the coding region of *CSTB* and a coding VNTR in *CCDC66*. The results suggest an increase in VNTRs with higher RU counts with an increase in divergence time from Africa. Thus RU3 is more prevalent in both VNTRs. We also observed RU4 in *CSTB* VNTR in the Asian and European populations, where RU counts 4 and above have been associated with progressive myocolonal epilepsy [72].

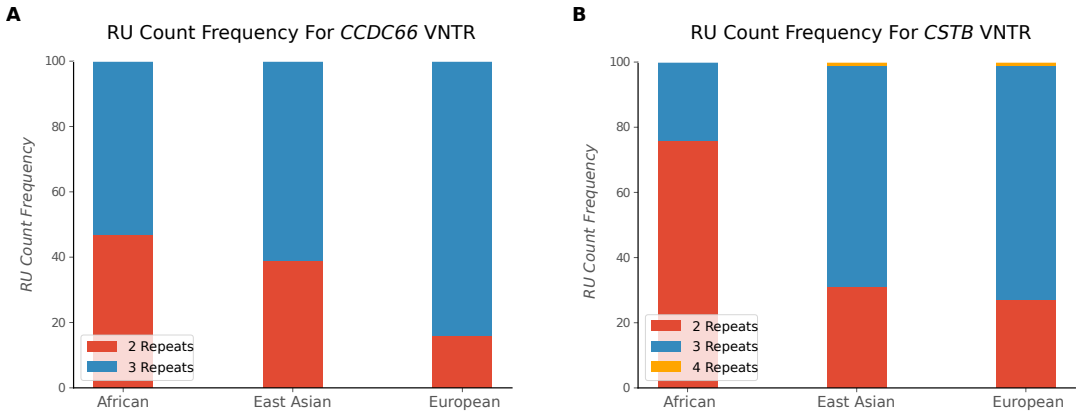


Figure 2.6: **Population-scale genotyping of VNTRs.** (A) RU count frequencies for the VNTR in *CCDC66* gene, and (B) *CSTB* in African, Asian, and European population samples from 1000 genomes project. RU counts of 4 and higher in *CSTB* are associated with myoclonal epilepsy.

2.3.6 VNTR mutation/indel detection.

As a proof of concept for other applications, we tested indel detection, focusing in particular on frameshifts in coding VNTRs. The *CEL* gene is known to contain a VNTR where a deletion changes the coding frame. We simulated Illumina reads from 20 whole genomes after introducing a single insertion or deletion in the middle of the VNTR region in the *CEL* gene. As a negative control, we simulated 10 WGS experiments with a range of sequence coverage values. We ran adVNTR, SAMtools mpileup[81], and GATK HaplotypeCaller[26] which uses GATK IndelRealigner, to identify frameshifts in each of the simulated datasets, and the 10 control datasets. On the control data, none of the tools found any variant. On the simulated indels, adVNTR made the correct prediction in each case (Suppl. Table A.6), while SAMtools and GATK were unable to predict a single insertion or deletion. This result is not surprising as the reads have poor alignment scores, and the indel can be mapped to multiple locations (Suppl. Fig. A.9)[109]. We note that mapping ambiguity in aligning each read made it difficult to pinpoint the location of single indel. However, by integrating the information across all reads, we could predict the occrrence of a frameshift in the VNTR. We next tested adVNTR frameshift prediction on the 115 VNTRs in the IlluminaFrameshift dataset, simulating 4090 total cases. Overall, the frameshifts in the VNTR regions were predicted with 51.7% sensitivity and 86.8% specificity, in contrast with

the 49.7, 43.5% sensitivity, specificity achieved by GATK. Detailed performance of methods for each VNTR is available in Table S7. Note that the performance is model specific and depends upon the similarity of different Repeat Units in a VNTR. For 29 of the 115 VNTRs, adVNTR showed high sensitivity ($\geq$90%) and specificity (100%).

As frameshifts in the VNTR region of the *CEL* gene have been linked to a monogenic form of diabetes[107], we tested for frameshifts in *CEL* using whole Exome sequencing (WES) data from 2,081 cases with Type 2 Diabetes [41] and compared the numbers to 2,090 control individuals. WES data analysis is challenging as high GC-content makes it difficult to PCR-amplify this VNTR. adVNTR found that while none of the controls had any evidence of a frameshift, 8 of the 2,081 diabetes cases showed a frameshift in this VNTR region (Suppl. Fig. A.10).

2.3.7 Compute requirements for genotyping.

adVNTR is multi-threaded. In genotyping mapped PacBio reads at 30$\times$ coverage, adVNTR took 6 hours using Intel Xeon(R) 4-core CPUs ($\leq$ 24 CPU-hours) to genotype all 2944 VNTRs, and 14:15 hours ($\leq$ 57 CPU-hours) for 70$\times$ coverage. For Illumina reads at 40$\times$ coverage, adVNTR took 87:30 cpu-hours on a single core to complete read recruitment as well as genotyping of 1775 VNTRs.

2.4 Discussion

The problem of genotyping VNTRs (determining diploid RU counts and mutations) is increasingly important for clinical pipelines seeking to find the genetic mechanisms of Mendelian disorders. As VNTRs have not been extensively studied, existing research is often focused on their discovery. One of the contributions of this paper is the separation of initial VNTR discovery from VNTR genotyping, and a focus on the genotyping problem. adVNTR genotypes VNTRs

using a hidden markov model for each target VNTR, providing a uniform training framework, but still allowing us to tailor the models for complex VNTRs on a case by case basis. The problem of mismapping due to indels introduced by changing RU counts confounds most mapping based tools, but is solved here by collapsing all RU copies and building HMMs that allow for variation in the RUs. adVNTR was tested extensively on data from different sequencing technologies, including Illumina and PacBio. As some of the data sets used were mapped only to hg19, especially the 150 whole genome sequencing data set from the Polaris project, we decided to use hg19 as the reference throughout, including simulations. Validation of the data used either orthogonal information (e.g. trios or experiments), or simulations and would not be affected by the use of GRCh38.

Like other STR genotyping tools, adVNTR works best when reads span the VNTR. However, even with this limitation, there are (a) close to $100,000$ VNTRs in the genic regions of human genome that can be spanned by Illumina reads; (b) indel detection is possible even when RU counting is not, for long VNTRs; (c) lower bounds on RU counts can separate some pathogenic cases from normal cases particularly when the normal VNTR length is shorter than the read length, while the pathogenic case is much longer (e.g. *CSTB*). Finally, dropping costs for long read sequencing (esp. PacBio, and Nanopore) will allow us to span and genotype over $158,000$ genic VNTRs.

The choice between short and long read technologies offers some trade-offs. Specifically, long reads allow for the targeted genotyping of a larger set of VNTRs ($559,804$), and are becoming increasingly cost-effective. However, the large numbers of indels in these technologies reduce the accuracy somewhat, and they are best used when there is a big difference between normal and pathogenic cases in terms of RU counts, or when the VNTRs are too long to be spanned by Illumina.

In contrast, short-read Illumina sequencing is increasingly used for Mendelian pipelines, and can be easily extended to include VNTR genotyping, with higher accuracy than PacBio.

Also, the large number of VNTRs (458, 158) that can be spanned by Illumina reads makes it the technology of choice for association testing and population based studies.

In this research, we also provided initial results on genotyping frameshift errors in coding VNTRs, focusing on the easier case when all RUs have the same length. Future work will focus on extending the target VNTRs for RU counting and frameshift detection for VNTRs that are of medical interest, population genetics of VNTRs, and algorithmic strategies for speeding up VNTR discovery and genotyping.

Software availability

adVNTR source code can be found in the Supplementary Material and it is also available at https://github.com/mehrdadbakhtiari/adVNTR.

Acknowledgements

The analyses presented in this paper are based on the use of study data downloaded from the dbGaP web site, under phs001095.v1.p1, phs001096.v1.p1 and phs001097.v1.p1.

Chapter 2, in full, contains material from Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, Vineet Bafna. ”Targeted Genotyping of Variable Number Tandem Repeats with adVNTR.”, *Genome Research*, 2018. The dissertation author was the primary author of this paper[10].

Chapter 3

Contribution of VNTR variations to gene expression mediation

Variable Number Tandem Repeats (VNTRs) account for significant genetic variation in many organisms. In humans, VNTRs have been implicated in both Mendelian and complex disorders, but are largely ignored by genomic pipelines due to the complexity of genotyping and the computational expense. We describe adVNTR-NN, a method that uses shallow neural networks to genotype a VNTR in 18 seconds on 55X whole genome data, while maintaining high accuracy.

We use adVNTR-NN to genotype 10,264 VNTRs in 652 GTEx individuals. Associating VNTR length with gene expression in 46 tissues, we identify 163 'eVNTRs'. Of the 22 eVNTRs in blood where independent data is available, 21 (95%) are replicated in terms of significance and direction of association. 49% of the eVNTR loci show a strong and likely causal impact on the expression of genes and 80% have maximum effect size at least 0.3. The impacted genes are involved in diseases including Alzheimer's, obesity and familial cancers, highlighting the importance of VNTRs for understanding the genetic basis of complex diseases.

3.1 Introduction

The human genome consists of millions of tandem repeats (TRs) of short nucleotide sequences. These are often termed as Short Tandem Repeats (STRs) if the repeating unit is $<$ 6bp, and Variable Number Tandem Repeats (VNTRs) otherwise. Together, they represent one of the largest sources of polymorphisms in humans[134, 51]. While multiple resources have been developed for genome-wide analysis of STRs, here we focus specifically on VNTRs, which have been largely missing from genome-wide studies due to technical challenges of genotyping and the computational expense.

We define VNTR genotyping in the narrower sense of determining VNTR length (number of repeating units). As VNTRs can be located in coding regions[107], untranslated regions[83], and regulatory regions proximal to a gene[43, 129], the variation in length can have a significant functional impact. Not surprisingly, VNTRs have been implicated in a large number of Mendelian diseases that affect millions of people world-wide[17, 19, 72]. They also are known to modulate quantitative phenotypes in several other organisms [38], and have shown pathogenic effects in other vertebrates including dogs [30]. VNTRs are also an important source of variations in bacteria and have commonly been used to study epidemiology and genetic diversity of Mycobacterium tuberculosis and Yersinia pestis [131, 123]. They have influenced primate and human evolution through gene regulation and differentiation of great ape populations[118]. Recent studies have identified VNTRs that have expanded in the human lineage or are differentially spliced or expressed between human and chimpanzee brains[122].

Single nucleotide polymorphisms (SNPs) that associate with gene expression, often referred to as expression Quantitative Trait Loci (eQTLs), are molecular intermediates that drive disease and variation in complex traits[99, 98, 44]. Studies have shown that causal variants for diseases often overlap with cis-eQTL variants in the affected tissue [11]. Therefore, we focus on the specific application of identifying expression mediating VNTRs ('eVNTRs'), or VNTRs

located in regulatory regions whose length is correlated with the expression of a proximal gene. Examples of 'eVNTRs' include a VNTR in the 5' UTR of *AS3MT* which is strongly associated with *AS3MT* gene expression and lies in a schizophrenia associated locus[83] and a 12-mer expansion upstream of the cystatin B (*CSTB*) gene is associated with gene expression and with progressive myoclonus epilepsy[72, 16].

Despite their importance, the full extent of VNTRs in mediating Mendelian and complex phenotypes is not known due to genotyping challenges. Traditionally, VNTR genotyping used capillary electrophoresis which did not scale to large cohorts. Despite the advent of sequence based genotyping, repetitive sequences continue to be challenging for genomic analysis. For example, 'stutter errors' due to polymerase slippage during PCR amplification change VNTR length and reduce genotyping accuracy [134]. While tools for genotyping STRs have been developed[134, 29, 52], they generally do not detect or genotype VNTRs, which have non-identical and larger repeat units. Recently, a few specialized computational methods (including our own method, adVNTR) have been published to tackle the problem of genotyping VNTRs from sequence data [10, 42]. However, these methods are too computationally intensive to scale to functional studies with hundreds of individuals and 10^4 VNTR loci (Results). There have also been recent, successful efforts to genotype VNTRs using long-read sequencing technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)[25, 94, 10]. While these methods (which include adVNTR) are quite accurate, the technologies are currently too expensive for population scale sequencing.

For these reasons, large-scale studies of VNTRs and their association with gene expression have been limited when compared to other sources of human variation such as SNPs and CNVs[11, 75, 22]. While the standard whole genome sequencing (WGS) frameworks often ignore repetitive regions, there is some progress towards 'harder' variant classes such as eSTRs[106, 53, 39] and 'eSVs'[22]. Therefore, 'missing heritability'–the gap between estimates of heritability, measured for example by twin studies[47, 138], and phenotypic variation explained by genomic variation–

remains a limitation for eQTL studies[85]. It has been speculated that the inclusion of tandem repeats in association analyses may reduce this heritability gap[56, 85, 17].

Here, we describe adVNTR-NN, a method that uses shallow neural networks for fast read recruitment followed by sensitive Hidden Markov Models for genotyping. We test the speed and accuracy of adVNTR-NN on extensive simulations to demonstrate accuracy. We use adVNTR-NN to genotype over 10,000 VNTRs in 652 individuals from the GTEx project and associate VNTR length with gene expression in 46 tissues. We additionally validate eVNTRs in blood tissues in 903 samples from an Icelandic cohort and 462 samples from the 1000 genome project with Gene expression data (Geuvadis cohort). We compare the strength of genic eVNTR association against proximal SNPs and identified many of the eVNTRs as causal. Our results suggest that it is computationally feasible to genotype VNTRs accurately in thousands of individuals, and multiple eVNTRs are likely to causally impact the expression of key genes involved in common and complex diseases.

3.2 Materials and Methods

3.2.1 Genotyping in adVNTR-NN

Filtering trade-off calculations

Let $A(r)$ denote the HMM genotyping time using r reads. The goal of filtering is to reduce the number of reads supplied to each VNTR HMM. Any filter is characterized by three parameters:

run-time: Let $P(r)$ denote the running time of the filter for r reads for each VNTR locus;

efficiency: Let f_k denote the fraction of reads that were retained for any VNTR. The efficiency is defined as $1 - f_k$ so that high efficiency implies only a small fraction being retained by the filter.

sensitivity/recall: The fraction of true VNTR overlapping reads that were accepted for each VNTR.

Consider a data-set with r unmapped reads and among the mapped reads, an average of r' reads are assigned to each VNTR locus. Assuming that the filtered reads are distributed equally among the VNTRs, each HMM will receive $f_k r + r'$ reads on the average. The total genotyping time for n VNTRs is given by:

$$T_{\text{adVNTR}}(n, r, r') = \text{indexing-time} + n\big(P(r) + A\left(f_k r + r'\right)\big), \tag{3.1}$$

Empirically, $A(r) = 0.32r$ seconds per VNTR. The keyword match filter for adVNTR achieved $f_k = 7.7 \times 10^{-5}$. For a 55X coverage WGS with $r = 4.2 \times 10^6$ reads, $P(r) = 111.22(s)$, $r' = 18$, we run the HMM on an average of $f_k r + r' = 341$ reads per VNTR on the average. The running

time is:

$$T_{\text{adVNTR}}(n,r) = 60.23 + n\left(1.853 + \frac{0.32}{60} \times 7.7 \times 10^{-5} \times 4.2 \times 10^{6} + \frac{0.32}{60} \times 18\right) \tag{3.2}$$

$$= 60.23 + 3.68n \text{ mins.}, \tag{3.3}$$

The genotyping time for n=10K VNTRs is about 631 hours per individual.

Read Filtering

For each VNTR locus V, and each read R, consider a binary classification function $f : V \times R \to \{0,1\}$, where $f(R,V) = 1$ if and only if read R maps to locus V. For each read and each of N loci $V_1, \ldots, V_N$, the neural recruitment method computes independent classification functions $f_i(V_i, R)$. Note that a read can be assigned to multiple VNTR loci, or to none. As an initial step toward this task, we perform a fast string matching based on prefix tree (trie) to assign each read to the VNTR loci that share an exact match with the read. For an efficient matching, we generate a separate aho-corasick trie[2] using every k-mer in VNTR loci as dictionary X. A trie is a rooted tree where each edge is labeled with a symbol and the string concatenation of the edge symbols on the path from the root to a leaf gives a unique word (k-mer) X. We label each leaf with a set of T VNTRs that contain corresponding k-mer. On the other hand, the string concatenation of the edge symbols from the root to a middle node gives a unique substring of X, called the string represented by the node. We add extra internal edges called failure edges to other branches of the trie that share a common prefix which allow fast transitions between failed string matches without the need for backtracking[2]. Testing whether a query q has an exact match in the trie can be done in $O(|q|)$ and we require additional $O(|T|)$ time to assign read q to all T VNTR loci that share the keyword. The overall complexity of this algorithm is linear based in the length of original dictionary (VNTRs in the database) to build the Trie and recover matches plus the length of queries (sequencing reads). Hence, after construction of the trie, the running time is

proportional to just reading in the sequences.

Neural Recruitment

To further reduce the set of reads assigned to each VNTR, we use a 2-layer feedforward Neural Network to compute f_i, using a k-mer based embedding to encode DNA strings. Specifically, we use a DNA string w of length k, consider an bijection ϕ that maps w to a unique number in $[0, 4^k - 1]$. Each read R can be defined by a collection of overlapping k-mers. We map read R to a unique vector $v_R \in \{0,1\}^{4^k}$, such that $v_R[i] = 1$ if and only if $\phi^{-1}(i) \in R$. Details of the neural network architecture and hyper-parameters are presented below.

Network Architecture

Let v denote the mapping of a read. We use a shallow architecture with an input layer used to present v to the network. We add two layers of fully connected nodes as the hidden layers, with each node being a ReLU function. In the output layer, there are two nodes *zero* and *one* which specify that whether read should be classified as true (containing VNTR) or false (Fig. 3.1). We used the training set to train the network with Adam optimization algorithm [63].

The number of hidden layers N_1 and N_2 were chosen empirically. Too many nodes would increase both training time and test time and possibly cause over-fitting. We performed the training with the number hidden nodes of each layer varying from 10 to 100 with 10 increase in each step and selected $N_1 = 100$ and $N_2 = 50$ as the best parameters according to validation performance.

Choosing the optimal k-mer length

The choice of k-mer length is important. Increasing the k-mer size could decrease sensitivity in our case as small variation will significantly change the k-mer composition, whereas lowering k-mer size reduces the features that are discriminative for a pattern[142]. In addition,

our embedding size exponentially grows with respect to the k so there is also a practical upper bound on the k. Following Zhang [142] and Dubinkina[31], we trained and tested in the range $4 \leq k < 9$. The accuracy remains comparable in this range (Fig. S23), and we chose $k = 6$ as its mean validation accuracy is the highest compared to four other values of k.

Effect of different loss functions

To choose the best loss function, we examined three regression loss functions: Mean Squared Error (MSE), Mean Squared Logarithmic Error (MSLE), and Mean Absolute Error (MAE), as well as three binary classification loss functions Hinge, Squared Hinge, and Binary Cross-Entropy. We compared the validation performance of our models for these 6 different loss functions. Each distribution in Supplementary Fig. S24 shows the accuracy on validation set across 1905 genomic loci. We analyzed these distributions using one-way analysis of variance (ANOVA) and none of them were significantly better than others. We chose binary cross-entropy as it obtained the highest mean accuracy (99.95%) among loss functions and its binary classification nature fits our requirement.

Speed and efficiency of neural network filtering

The neural-network filtering achieved a speed of $N(r) \simeq 0.03r$ seconds for r reads, greatly increasing filtering efficiency ($f_n f'_k < 10^{-6}$) to input only 14 reads per VNTR on the average when $r = 4.2 \times 10^6$. The running time using the two filters could be modeled as

$$\begin{aligned} T_{\text{adVNTR-NN}}(n,r) &= n\left(P'(r) + N\left(f'_k r\right)\right) + nA\left(f_n f'_k r\right) + nA(r') \\ &= 25.48 + 0.13n + 0.07n + 0.09n = 25.48 + 0.29n \text{ min.}, \end{aligned} \tag{3.4}$$

Simulated data for training and testing

We used ART[60] to generate $r = 6 \times 10^8$ reads from human reference genome (30X coverage) with Illumina HiSeq 2500 error profile. For each target locus, we modified the number of the repeats to be ± 3 of the original count in the reference with setting 1 as minimum number of repeats, and simulated reads from those regions. For each locus, we assigned labels to reads as being true reads or not, based on exact location. We divided the original set of reads into three parts: 70% for training, 10% for validation and 20% for testing. We trained all neural network models using the training and validation sets, and reported performance on the test dataset.

To augment the data, we added random single nucleotide variations in the genome sequences of the dataset before simulating the sequencing reads [90]. For each sequence in the dataset, we replaced its nucleotides with a random one with probability r_m. We set $r_m = 10^{-5}$, the novel base substitution mutation rate within VNTRs[61]. This method of dataset augmentation helps include 'mutated' k-mers in the embedding of reads, making the method more robust.

adVNTR-NN accuracy versus other methods

To test and compare genotyping accuracy against VNTRseek (v1.10.0), we started with a random selection of 10,000 target VNTR loci (< 140 bp) and filtered them out if a VNTR locus was marked as indistinguishable in VNTRseek. As a result, 9,638 target VNTRs remained. We used ART[60] to generate heterozygous samples by simulating 15X coverage reads from each modified haplotype which contained a non-reference allele and combined those with 15X reads that were simulated from reference. The non-reference allele for each VNTR was chosen to be in the range $[c-3, c+3]$, where c is the reference count. Together, this provided six diploid simulated data-sets for each locus, at 30X coverage.

Similarly, to test and compare genotyping accuracy against GangSTR[97] (v2.4.5) for, we selected VNTR loci with repeat unit length ≤ 20 bp. A total of 6,508 target VNTRs remained. Following the method for VNTRseek comparisons, we used ART[60] to generate a homozygous

sample and six heterozygous samples by simulating 30X paired-end reads with Illumina HiSeq 2500 error profile.

Performance test

We measured running time of adVNTR-NN and VNTRseek by running them with default parameters on a single core of Intel Xeon CPU E5-2643 v2 3.50GHz CPU. To measure the accuracy of genotyping, we ran adVNTR-NN and VNTRseek on diploid simulated data of heterozygous VNTRs and measured the number of correct calls divided by total number of VNTR loci.

3.2.2 Data and preprocessing

We accessed 30X Illumina WGS data from the GTEx cohort (652 individuals) through dbGaP (accession id phs000424.v8.p2). Specifically, we accessed CRAM files containing read alignments to the GRCh38 reference genome through cloud-hosted SRA data using fusera v1.0 and downloaded VCF files containing SNP genotype calls from dbGaP.

As genotyping VNTRs remains computationally expensive, we focused on the smaller set of VNTRs located within coding, untranslated, or promoter regions of genes, which are most likely to be involved in regulation. We identified VNTRs in coding exons and UTRs by intersecting VNTR coordinates with refseq gene coordinates downloaded from UCSC Table Browser. To identify VNTRs that appear within promoter regions, we considered 500bp upstream of the transcription start site of genes as the promoter regions. Overall, this procedure identified $13,081$ VNTRs, of which $10,262$ were within the size range for short-read genotyping (Fig. 3.1A). We subsequently added two VNTRs previously linked to a human disease to obtain $10,264$ target loci[132, 132]. We genotyped these VNTR loci in 652 individuals from GTEx cohort using adVNTR-NN on Amazon Web Services (AWS) cloud, which allowed us to do the computation in parallel for different samples.

We compared the most common allele of each VNTR with the reference allele (GRCh38) to observe representation of each VNTR in the reference. We also searched for VNTRs with multiple observed alleles to estimate a rate of polymorphism for VNTRs and find how common each allele was. To call a VNTR polymorphic, we set the minor allele frequency at 5% and any variation below that frequency was discarded. In addition, we identified the amount of base-pair difference that they make in genome of each individual by comparing the copy number difference of VNTRs between reference and the sample and multiplied that by the pattern length of each locus. We computed how many loci on average differed between an individual and reference by combining all non-reference calls in at least one haplotype from all individuals and dividing it by all called variants. VNTRs whose allele frequencies did not meet the expected percentage of homozygous versus heterozygous calls under Hardy–Weinberg equilibrium ($P < 0.05$ for two-sided binomial test) were eliminated. We further removed VNTRs that were monomorphic (only one allele) in the entire GTEx cohort or had minor allele frequency lower than 1% among the individuals with expression data in every tissue. We used the resulting 2,672 VNTRs for subsequent analysis ([9]).

We obtained processed RNA-expression data (RPKM values) from 54 tissues from dbGaP (phs000424.v7.p2) and limited analysis to 46 tissues which had data for at least 100 individuals. 'Non-expressed genes'– genes with median RPKM level zero– in each tissue were removed from analysis. For the remaining genes, we quantile-normalized RPKM values of each tissue to a normal distribution. We analyzed VNTR-Gene pairs for each VNTR and its closest gene based on refseq annotations in each of the 46 tissues.

3.2.3 Identification of eVNTRs

Before the analysis of the association of VNTR genotypes and gene expression levels, we adjusted gene expression levels for each tissue in order to control for covariates of sex, population structure, and technical variations in measuring expression. For population structure, we used the

top ten principal components (PCs) from a principal components analysis (PCA) on the matrix of SNP genotypes to provide a correction for population structure. To generate the SNP genotype matrix, we used the VCF files for GTEx cohort (accession phg001219) and filtered biallelic SNP sites MAF > 0.05 using plink [105]. To correct for non-genetic factors such as technical variations in measuring RNA expression levels (e.g batch effects, environmental variables), we applied PEER factor correction and used the top 15 factors[119]. We removed the effect of covariates by regressing them out from the RNA expression matrix of each tissue and subtracting their factor contributions and used the residuals for all eQTL association analyses.

We normalized the individual raw gene expression values to $N(0,1)$ by subtracting the mean and dividing by the standard deviation of the expression values for that cohort. For a gene-VNTR pair v, let y_{iv} denote the normalized expression value of gene in v for individual i and x_{iv} denote the genotype of the VNTR in v for individual i. Then,

$$y_{iv} = \beta_v x_{iv} + \sum_k \gamma_k \mathrm{PC}_{ik} + \sum_k \delta_k R_{ik} + \varepsilon_{iv} \tag{3.5}$$

where, PC_{ik} denotes the strength of the k-th principal component, and R_{ik} the value of the k-th PEER factor. We performed the association test for each VNTR-gene pair separately for each tissue type using Python statsmodels linear regression, Ordinary Least Squares (OLS)[115], and computed a nominal p-value of the strength of association for each VNTR-gene pair using two-sided Fisher's exact test.

Multiple Testing Correction

We used permutation tests and the Benjamini–Hochberg procedure to estimate a 5% False Discovery Rate (FDR) significance cut-off for each tissue. The significance thresholds for each of the 46 tissues ranged from 10^{-3} to 3.8×10^{-5} (Fig. S13). Overall, 759 significant tests were observed from total of 73,609 tests in all tissues and 163 unique VNTRs passed the significance

test in at least one tissue.

We performed a similar correction for the Geuvadis cohort. Specifically, we performed 100 permutations and used a Benjamini-Hochberg procedure to control the False Discovery rate at 5%. For the Icelandic cohort, only the VNTRs that showed significant associations in GTEx were tested using unmapped reads plus reads mapped to those specific loci. Hence, we used the conservative p-value cutoff from whole-blood tissue of the smaller GTEx cohort.

Fine-mapping of Causal Variants

To compare the strength of the VNTR association relative to proximal SNPs, we extracted all SNPs from 50kb 5' to the transcription start, from the gene body, and up to 50kb 3' to the end of the transcript using the GTEx variant calls. To perform a fair comparison, we used the same test and covariates for VNTRs and repeated it for each SNP by replacing the genotype to obtain the strength of association for each SNP. Then, we ranked all variants based on their association P value.

We further used a fine-mapping method, CAVIAR, as an orthogonal method to identify the causal variant for the change in gene expression level. CAVIAR is a statistical method that quantifies the probability that a variant is causal by combining association signals (i.e., summary level Z-scores) and linkage disequilibrium (LD) structure between every pair of variants[59]. We ran CAVIAR with parameter -c 1 to identify the most likely causal variant, along with the causality probability distribution for each variant site. We ranked variants based on their causality probability given by CAVIAR and called it the causality rank.

3.3 Results

3.3.1 Target VNTR loci.

Using Tandem Repeat Finder[13], 502,491 VNTRs were identified that contained at least two repeating units in the GRCh38 human assembly and had repeat unit lengths between 6bp and 100bp. Over 80% of these had total length $<$ 140bp (Fig. 3.1a) and could be genotyped using Illumina sequencing. As genotyping VNTRs remains computationally expensive, we focused on the $13,081$ VNTRs located within coding, untranslated, or promoter regions of genes (Methods) as they are most likely to be involved in gene regulation. Of those, we identified $10,262$ VNTRs that were within the size range for short-read genotyping (Fig. 3.1a). We added two additional VNTRs that were previously linked to a human disease ([9]) to obtain $10,264$ target loci [33, 132].

3.3.2 adVNTR-NN improves genotyping speed.

Our previously published tool, adVNTR, used customized Hidden Markov Models (HMMs) for each VNTR and showed excellent genotyping accuracy, based on trio-analysis, simulations and PCR[10]. However, HMMs are compute-intensive, and despite some filtering strategies used by adVNTR(Methods), the time to genotype n=10K VNTRs was about 631 hours per individual. In developing adVNTR-NN, we first made significant improvements to pre-processing time. Next, we deployed a second filtering step with a 2-layer feed-forward network trained separately for each VNTR that accepted the k-mer composition for each read and filtered it specifically for that VNTR (Fig. 3.1b,c and Methods). The neural-network filter required 0.03s per read, and filtered reads with high efficiency in filtering reads. For 55X whole genome sequencing (WGS) with $r = 4.2 \times 10^6$ unmapped reads, the NN supplied an average of 14 previously unmapped reads to each VNTR HMM. Combining with the mapped reads, each HMM received an average of 32 reads per VNTR locus. This reduced the running time for n

VNTR loci to

$$T_{\text{adVNTR-NN}}(n) = 25.48 + 0.29n \text{ mins. (Fig. 3.1d)}, \tag{3.6}$$

allowing each individual to be genotyped at $n = 10$K VNTRs in 50 CPU hours, a $13\times$ speedup over adVNTR.

3.3.3 adVNTR-NN outperforms alternative alignment methods at VNTRs.

While adVNTR was highly accurate by itself, its final accuracy depended upon reads filtered for genotyping, and specifically on false negatives–reads that were incorrectly removed by a filter. Formally, a read sampled from a VNTR was considered to be true positive (TP) if it passed the filter for that VNTR, and false negative (FN) otherwise. False positives (FP)–reads that passed the filter despite not being from the VNTR locus–were a lesser concern because they would eventually be discarded by the HMM for not aligning well to the model. However, high false-positives increase the running time. To account for this, we measured the trade-off between efficiency $(1 - (\text{TP+FP})/r)$ and recall TP/(TP+FN).

For comparisons with alternative filters, we used Bowtie2 as a representative read-mapping tool[74]. These tools are designed for fast mapping of reads and are accurate for most of the genome, but are not specifically designed for VNTR mapping genotyping (could have high FN). As a second comparison, we used adVNTR[10], which has high recall (low FN) for VNTR mapping. We used a mix of real and simulated reads to test performance (Methods).

In terms of efficiency $(1 - (\text{TP} + \text{FP})/r)$, Bowtie2 was the most efficient retaining only 0.9 in 10^6 reads for further processing for 90% of the VNTRs. Both adVNTR and adVNTR-NN were slightly less efficient retaining about 1.2 reads per million for 90% of the VNTRs. However, they had significantly better recall. adVNTR-NN filtered reads with at least 90% recall for 99% of the target VNTR loci (Fig. 3.1e). In comparison, 80% of the loci achieved that recall for adVNTR, and only 27% of the loci had a recall of 90% for Bowtie2. Notably, adVNTR-NN had much better

recall compared to adVNTR while also being more efficient, and therefore faster.

3.3.4 adVNTR-NN speed and accuracy on simulated VNTR alleles.

We had previously measured adVNTR genotyping accuracy[10] using trio-consistency, comparison to long reads and other methods. Similarly, we used a mix of whole genome sequencing data and simulated reads (Methods) to measure adVNTR-NN accuracy.

The accuracy of VNTR genotyping using short reads depends critically on total allele length and length of repeat unit itself. adVNTR was 90% accurate on reads up to 90 bp in length, but its accuracy dropped subsequently (Supp. Fig. S1). Similarly, its accuracy remained high for repeat unit length up to 40bp, as long as the total allele length did not exceed the read-length (Supp. Fig. S2). We reiterate that a majority of the known VNTRs have small allele length (Fig. 3.1a), and therefore the overall accuracy remains high.

Next, we compared the overall running time and accuracy of adVNTR-NN genotyping with VNTRseek[42], which was not available at the time of original release of adVNTR. Notably, VNTRseek combines VNTR discovery and genotyping and does not customize genotyping for each VNTR. Therefore, its running time on 55X WGS ranged from 9640-9686 minutes, and was largely independent of the number of target VNTRs (Supp. Fig. S3). This was in contrast to the 1,696 minutes required by adVNTR-NN. The speed advantage for adVNTR-NN could largely be attributed to filtering strategies which could potentially be used to improve VNTRseek genotyping time as well. On simulated heterozygous reads with 30X coverage (Methods), adVNTR-NN was highly accurate. It achieved 100% accuracy in 7343 (76%) of 9638 VNTRs compared to VNTRseek's median accuracy of 60% (Supp. Fig. S4). In contrast with adVNTR-NN, VNTRseek's genotyping accuracy was sharply asymmetric, with much lower accuracy for decreasing VNTR length (Supp. Fig. S5).

GangSTR is a method designed for Short Tandem Repeats (STRs), but can genotype repeat units up to 20bp[97]. Therefore, we compared its accuracy against adVNTR-NN for

these short motifs. GangSTR uses total allele length which could result in an incorrect call if there are significant changes in repeat unit length. Indeed, on reference data, GangSTR was accurate in 82.4% of the VNTR loci and under-counted the allele by 1 in 12.3% of the cases. adVNTR-NN called the genotype correctly in 98.5% of the loci ([9]). On simulated heterozygous reads, GangSTR accuracy lagged that of adVNTR-NN (Supp. Fig. S6).

3.3.5 adVNTR-NN consistency on trio data.

A reference database of VNTR allele counts is not available for testing performance on real data. Instead, we tested for consistency of adVNTR-NN calls on 10,264 VNTRs WGS data of 537 trios from 1000 Genomes Project[1] (5,511,768 tests total). We observed 98.4% consistency in the calls obtained by adVNTR-NN. The inconsistent alleles had longer length (median 90bp) in contrast to the length of the consistent alleles (median 52bp, Supp. Fig. S7) a range in which VNTR genotyping is more likely to be erroneous. Moreover, in a third of the inconsistent cases (0.5% of total), the RU count of the inconsistent allele was ± 1 of a parent's RU count, suggestive of a de novo mutation. Comparing adVNTR-NN genotypes with adVNTR, the calls were identical in 99.81% of the loci showing high similarity in accuracy between two genotyping methods ([9]).

3.3.6 Data-sets for identifying eVNTRs.

To identify expression-mediating VNTR Loci (eVNTRs), we primarily used data from the GTEx project[11] (Methods). The GTEx project provided WGS for 652 individuals as well as RNA-seq for each of these individuals from 46 tissue types including whole-blood. A majority (86.0%) of the donors were of European origin; another 11.5% were African American and the remaining were Asian and American Indian. For validation, we used a second cohort of 903 Icelandic individuals[48] with associated whole blood RNA expression data and WGS. We also

chose a smaller, third cohort from the Geuvadis[75] project which provided gene-expression data in lymphoblastoid cell-lines for 462 samples, where the WGS for the samples was available from the 1000 genomes project[1]. The Geuvadis cohort was dominated by individuals of European ancestry (80.7% of cohort). Most of the remaining (19.3%) were of African ancestry. Due to the match of tissue type and ethnicity, the Icelandic and Geuvadis whole blood data were used for validation of methods for identifying eVNTRs discovered from the GTEx project.

3.3.7 eVNTR identification.

We genotyped 10,264 VNTR loci in all 652 samples from GTEx to study the role of VNTRs in mediating gene expression of proximal genes. As expected, the most frequent allele matched the reference allele in 96.8% of the cases (Supp. Fig. S8).

Despite the GTEx data being predominatly European, 51% of the target VNTRs were polymorphic. Consistent with evolutionary constraints, VNTRs in promoters were most likely to be polymorphic (57%) followed by Untranslated regions (UTRs) (51%) and coding exons (47%) (Fig. 3.2a). Each individual in the GTEx cohort had a non-reference allele in at least 839 (8.2%) of the tested VNTR loci, with an average of 1,259 (12.3%) non-reference VNTRs per individual. Altogether, the 10,264 VNTRs inserted or deleted an average of 47,197bp per individual (Fig. 3.1f). As this represents $< 10\%$ of all VNTRs, the results highlight VNTRs as an important source of genomic variation. The minimum variation in a non-reference VNTR allele involved at least 6 basepairs and the average change in each variant site was 37bp or about 3 repeat units (Supp. Fig. S9).

To perform association analysis, we excluded 1817 (17.7% of total) VNTRs that were monomorphic , 1445 (14.1%) VNTRs that violated Hardy-Weinberg equilibrium constraints and 4330 (42.2%) VNTRs that had minor allele frequency $<1\%$ after removing individuals in the GTEx cohort with no expression data for the specific gene (Methods). We investigated VNTRs that violated HWE. Similar to trio-inconsistent VNTRs but distinct from all VNTRs,

these VNTRs were longer, had long common alleles (Supp. Fig. S10, S11), or their flanking regions had a strong (>5 bp) match to the sequence of the repeating units (Supp. Fig. S12).

The filtering resulted in a set of 2,672 VNTRs (26%) available for association analysis. We used linear regression to measure the strength of association between average VNTR length of the two haplotypes, and adjusted gene expression level of the closest gene (Fig. 3.2b and Methods). To account for confounding factors, we included sex and population principal components of each individual as covariates. We also added PEER (probabilistic estimation of expression residuals) factors to account for experimental variations in measuring RNA expression levels (e.g batch effects, environmental variables)[119]. Briefly, PEER infers hidden covariates influencing gene expression levels, and we removed their effect by producing a residual gene expression matrix and using it for linear regression (See Methods).

We measured association with gene expression in each of the 46 tissues. To control False Discovery Rate (FDR), we used the Benjamini-Hochberg procedure to identify a tissue-specific 5% FDR cutoff (Supp. Fig. S13 and Methods). Combining data from all tissues, 759 tests tied to 163 unique VNTR loci passed the significance threshold (Fig. 3.2c). We refer to these (VNTR, gene) pairs as eVNTRs. Unlike VNTRs that failed HWE (median length: 92bp), eVNTR allele lengths were much smaller (median:48bp, Supp. Fig. S10), and in a range where VNTR genotyping is highly accurate (Supp. Fig. S1).

Not surprisingly, a larger fraction (6.8%; Fig. 3.2a) of the UTR and regulatory (6.0%) variants were associated, compared to coding VNTRs (4.9%). The strength of association did not depend upon the location of the VNTRs (Supp. Fig. S14). However, VNTRs within 100bp of the Transcription Start Sites (TSS) were twice as likely to be eVNTRs compared to other locations ($P = 6 \times 10^{-6}$; Fisher's exact test), consistent with their known roles in core-promoters[121].

The number of eVNTRs observed in each tissue type generally correlated with the number of individuals samples for each tissue type (Supp. Fig. S15). Consistent with previous results on eQTLs[11], and eSTRs[39], testis and fibroblasts had the largest number of eVNTRs, while

fewer eVNTRs were identified in whole blood and skeletal muscle, relative to the sample size. Only 4% of the eVNTRs were tissue specific (Fig. 3.2d). We used the method mash[128] to test for reproducibility in other tissues. Mash exploits the power gains that come from cross-sharing the effect of an eVNTR in multiple tissues. The analysis suggested that many (38%) eVNTRs were significant in at least half (23) of the tissues tested (Supp. Fig. S16).

Twenty-three of the 163 unique eVNTRs showed significant association in whole blood (Table 3.1), a tissue type in which we could validate the eVNTRs using independent data from the Icelandic cohort of 903 individuals. The VNTRs that showed significant associations in GTEx were replicated on the Icelandic cohort using the conservative p-value cutoff from the smaller GTEx cohort. Two of the 23 VNTR loci could not be used for replication in the Icelandic cohort due to missing expression data for *TRIM15* and *SNHG16* genes. 18 (86%) of the 21 VNTRs showed significance at a similar level and same direction of effect in Icelanders, highlighting the strong reproducibility of the associations. The Geuvadis data were acquired for a smaller cohort compared to the Icelandic data and measured expression in lymphoblastoid cells–transformed B cells, which are a component of whole blood tissue. Therefore, we recomputed 5% FDR cut-offs using the Benjamini-Hochberg method on 100 permuted samples. Despite the caveats, 12 of the eVNTRs were replicated. Combined, 91% (20/22) of eVNTRs could be replicated in an independent cohort where data was available. We also tested for correlation of effect-sizes between the Icelandic and GTEx data and found strong correlation (Supp. Fig. S17; Spearman correlation coefficient 0.88; p-value $= 1.15 \times 10^{-7}$). A similarly strong correlation was observed between the Geuvadis cohort and GTEx (Supp. Fig. S18; Spearman correlation coefficient 0.70; p-value: 4.57×10^{-4}).In all cases, the direction of effect was also maintained.

STR genotyping software such as HipSTR[134] can also genotype repeats up to 6bp. Therefore we compared GTEx association results on hexamer repeats from a recent eSTR study[39]. 15 loci were identified as eSTR/eVNTR in at least one of the two studies (Supp. Table B.2). Despite differences in genotyping methods, filtering, FDR controls, choice of co-

variates, and reference assemblies, all 15 loci were at least nominally significant in both tests, and 6 of 15 were identified as eSTRs/eVNTRs in both studies.

In 65% of the cases, VNTR length had a positive correlation with gene expression; the remaining cases had a negative correlation (Fig. 3.2e). This was consistent with the hypothesis that many VNTRs encode transcription factor binding sites and increasing length improved the TF binding affinity. Moreover, the overall effect size was also large and 80% of the eVNTRs had a maximum effect-size 0.3 or higher.

We computed correlation of eVNTR effect size between each pair of tissues using the Spearman rank test. Despite the multi-tissue activity of most eVNTRs, each tissue showed distinct behavior with low correlation to most other tissues (Fig. 3.2f). Similar tissue types were expectedly correlated (e.g. brain). Some correlations were seen among glandular tissues (salivary, prostate, pituitary) and also between adipose tissue and nearby tissues and organs (heart, esophagus muscularis, artery, breast). Fotsing et al.[39] used eSTRs to cluster a subset of 17 tissue types. When restricted to that subset (Supp. Fig. S19), the eVNTR clustering was highly consistent with the eSTR clustering. Both analyses showed distinct clades for (a) the two skin tissues with esophageal-mucosa possibly due to an abundance of squamous cells and (b) the two adipose tissues with esophageal-muscularis. Moreover the second clade was part of a larger one containing the arterial tissues, the tibial nerve, thyroid and lung in both analyses. Thus, even though most eVNTRs are shared across tissues, we hypothesize that the combined effect of active eVNTRs is tissue-specific and leads to unique regulatory program for each tissue type.

Similar to SNPs, and due in part to power considerations, VNTR loci generally showed a negative correlation between Minor Allele Frequency (MAF) and effect size, so that common variants generally had low effect size with larger effects mainly shown by rare variants[15] (Fig. 3.2g). However, we still observed many eVNTRs where common VNTR (MAF > 0.05) showed large effects. These eVNTRs had highly significant p-values (Supp. Fig. S20) and in many cases, the proximal genes were associated with known diseases or phenotypes (Supplementary

Table B.1). As these represent potentially the most interesting eVNTR findings, we tested them further for causality and function.

3.3.8 VNTRs mediate expression of key genes.

Only a small number of examples have been reported where VNTR repeat unit counts have a causative on gene expression[83]. Each of these cases has been discovered by gel analysis or Sanger sequencing on individual loci in specifically chosen cohort. One well known example is the *AS3MT* gene which is involved in early brain development, where the VNTR was associated with expression and was in LD with SNPs associating with schizophrenia[83].

To investigate causality, we ranked each eVNTR against all SNPs within 100kbp by (a) comparing the relative significance of association with gene expression (r_1); and (b) using the tool CAVIAR[59] to measure the causality of association (r_2)(Methods). Remarkably, the two rankings were very similar with mean discrepancy $2|r_1 - r_2|/(r_1 + r_2) = 2.3 \times 10^{-3}$ across the 163 eVNTRs. We used the harmonic mean $(2/(1/r_1 + 1/r_2))$ of the two ranks to order the eVNTRs. Of the 163 VNTRs, 81 of the eVNTRs were ranked 1 which are likely causal (Supp. Fig. S21), indicating that the 49.6% of the eVNTRs had the highest posterior probability of causality compared to all other variants tested. Separating tissue types, 170 (22%) of the 759 significant associations were possibly causal. These results suggest a large fraction of causal eVNTRs even with the caveat that we only tested 'genic' VNTRs.

Looking at individual eVNTRs, we recapitulated a previous result by identifying an eVNTR in the *AS3MT* gene. The lowest association p-value measured in any tissue using 652 samples was 3.9×10^{-54}, which was orders of magnitude higher than the significance reported with 322 samples[83](Fig. 3.3a,b). Its CAVIAR rank was 1 and it had an effect size of 0.33 in Brain Cortex in contrast to the effect-size of 0.16 for the top SNP in Brain Cortex. Finally, the VNTR is located in a regulatory region of the genome as identified by H3K27Ac and DNase marks (Fig. 3.3c).

The other eVNTRs, including the 81 with CAVIAR rank 1, represent novel findings. Many mediate the expression of genes (Supplementary Table B.1) involved in key functions. For example, Proopiomelanocortin (*POMC*) is a precursor protein for many peptide hormones with multiple roles including regulation of appetite and satiety[57]. Hypermethylation of *POMC* (and reduced expression) in peripheral blood cells and melanocyte-stimulating hormone positive neurons was strongly associated with obesity and body mass index[70]. Surprisingly, *POMC* over-expression also predisposed lean rats into diet-induced obesity[80]. Our analysis identified a VNTR in the coding region of the *POMC* gene as the causal variant governing expression levels in 15 tissues, including adipose and nerve tissues. The 6R allele had 1.8-fold higher expression in blood and nerve cells (Fig. 3.3d), and the correlation with expression was much stronger than neighboring SNPs (Fig. 3.3e). The eVNTR had an effect size of 0.48 in Nerve tissue, compared to 0.27 for the top SNP using the same model. Moreover, the VNTR was located within an H3K27Ac mark that was topologically close to the promoter of the gene based on chromatin conformation (Fig. 3.3f).

The *ZNF232* gene is differentially expressed in ovarian and breast cancers[114, 117]. Also, the chr17 locus containing the gene has been associated with Alzheimer's in a recent large meta-GWAS study on the UK Biobank data[88]. We identified an eVNTR in the promoter region where expanded alleles (at least 5 repeat units) had 2-fold higher median expression relative to RU3 (Fig. 3.3g). The VNTR was ranked 1 in 40 of 46 tissues including 7 brain sections, and specifically the Hippocampus, which is the affected region in Alzheimer's[102, 46] (Fig. 3.3h) and was also ranked 1 in ovary and breast (Supplementary Table B.1). In Hippocampus, the eVNTR effect size was 0.34 for eVNTR compared to 0.07 for top SNP using the same model.

The *RPA2* gene product is part of the Replication Protein A complex involved in DNA damage checkpointing[76]. Its over-expression is identified as a prognostic marker for colon cancer and bladder cancers[45]. A VNTR that overlapped the Transcription Start Site (TSS) of *RP2A* with lower VNTR length showed 1.9-fold higher expression of *RPA2* in multiple tissues including

colon (Supp. Fig. S22 and Supplementary Table B.1). Supplementary Table B.1 identifies other important genes including *NBPF3* (Neuroblastoma[125]), *TBC1D7* (lung cancer[113]), *ZNF490* (colorectal cancer[49]), *MSH3* (myotonic dystrophy[95]) and others. We note that the VNTR in *MSH3* is a 9bp repeat that is distinct from the trinucleotide expansion mediated by *MSH3*[136]. Taken together, our results suggest that VNTRs mediate the expression of key genes.

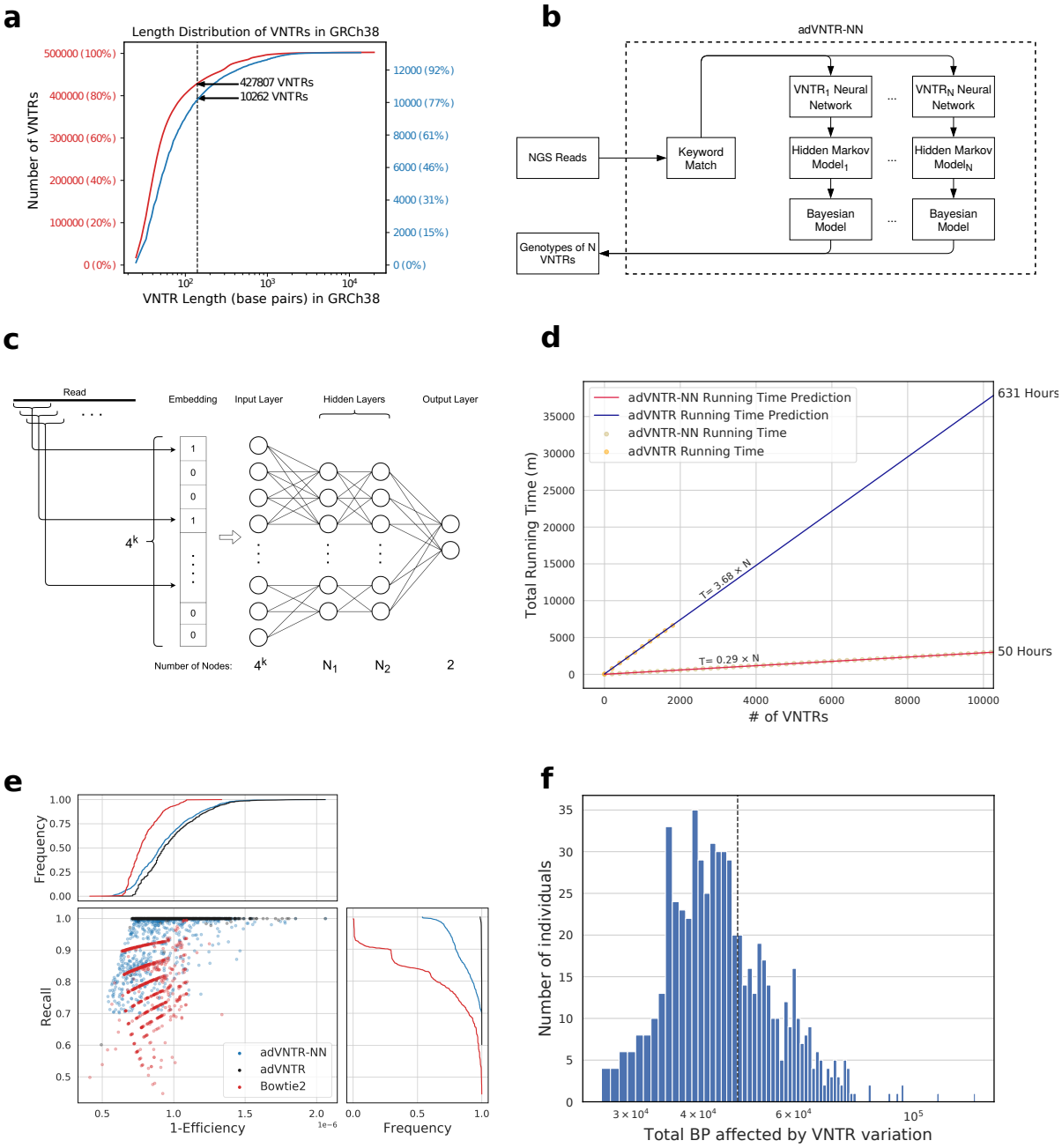
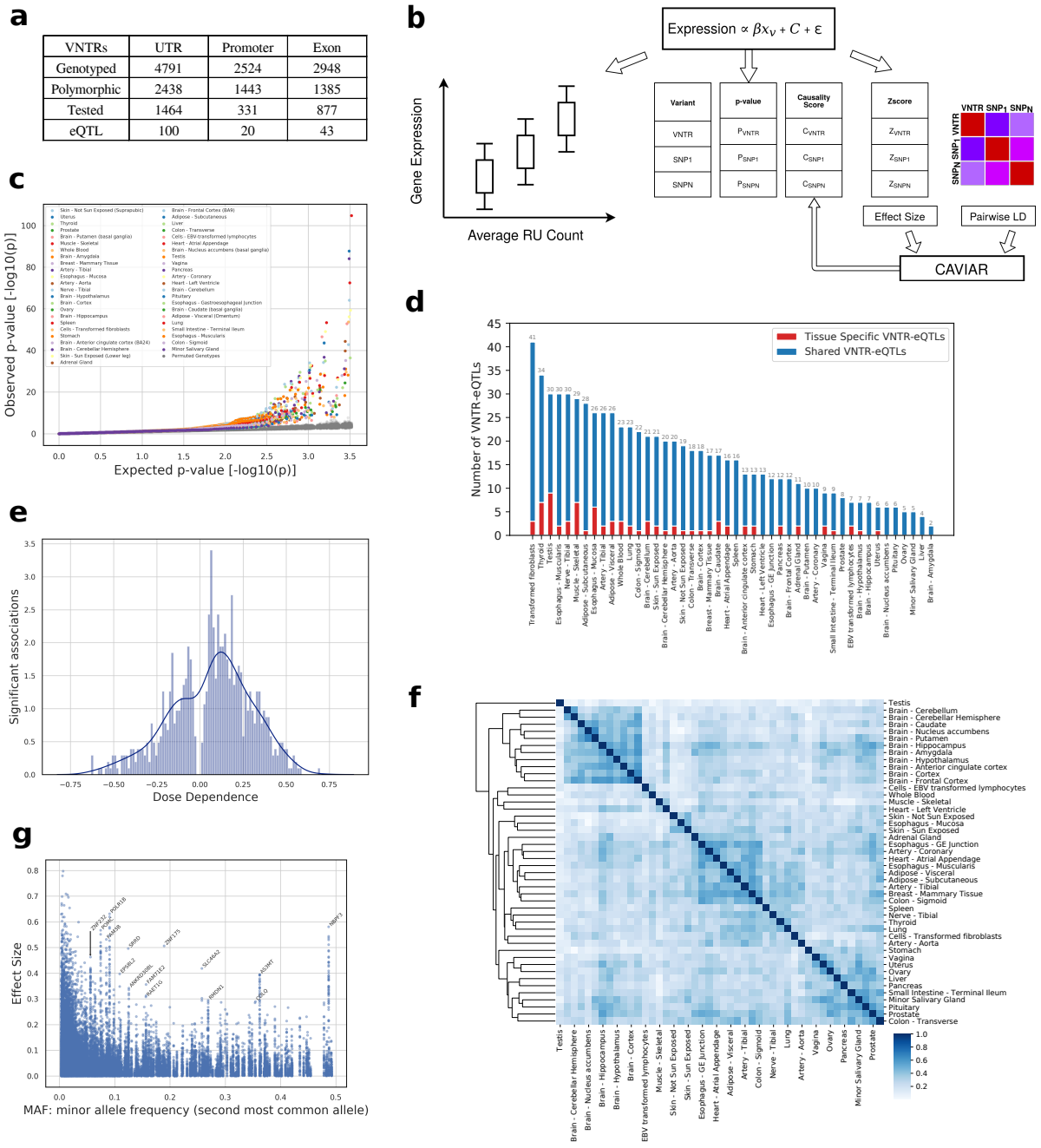


Figure 3.1: **Genome-wide VNTR genotyping performance.** (a) Length distribution of all known VNTRs (red) and selected targeted VNTRs (blue) across the GRCh38 human genome in base pairs. (b) The genotyping pipeline. (c) Neural network architecture for each VNTR which uses a mapping of reads to a k-mer composition vector. (d) Improvement in running time after using neural network and kmer matching. (e) Accuracy and efficiency of read recruitment in simulated data. The scatter plot shows 1-efficiency $((\mathrm{TP}+\mathrm{FP})/R)$ and recall (TP/(TP+FN)) of classification with different methods. High efficiency is related directly with running time. Each of 10,264 points represents a VNTR locus (method) and are shown once for each method. The side and top panels show cumulative distributions of recall and 1-efficiency. (f) Base-pairs (log-scale) affected by VNTRs per individual in the GTEx cohort.



VNTRs	UTR	Promoter	Exon
Genotyped	4791	2524	2948
Polymorphic	2438	1443	1385
Tested	1464	331	877
eQTL	100	20	43

Figure 3.2: **Effect of VNTR genotypes on mediating gene expression.** (a) Location of target VNTRs and e-VNTRs relative to the proximal genes. (b) Pipeline to identify eVNTRs and assign causality scores. Ancestry, Sex, and PEER factors are included in C as covariates. We associate VNTR genotype with expression residuals after correcting for the effect of C. (c) Quantile-quantile plot showing p-values of association signals separated by tissue. Green line represents the p-values using 100 permutations. (d) Number of unique and shared eVNTRs in each tissue. (e) Trend of RU count correlation with gene expression level. (f) Spearman correlation of eVNTRs effects sizes for each pair of tissues. (g) Scatter-plot correlating effect size versus Minor Allele Frequency (MAF).

Table 3.1: **Replication of whole blood VNTRs in independent cohorts.** Each row describes an eVNTR in whole blood from GTEx project(n=652 individuals) identified with false discovery rate (FDR) < 0.05 based on 100 permutations. Replication of the signal in whole blood tissue of the Icelandic cohort of 903 samples and in lymphoblastoid cell-lines from the Geuvadis cohort (462 samples) with the same direction of effect and FDR < 0.05. For the Icelandic cohort, only the VNTRs that showed significant associations in GTEx were tested using unmapped reads plus reads mapped to those specific loci. Hence, we used the conservative p-value cutoff from the smaller GTEx cohort. Length (respectively, RU length) refers to the total (respectively, repeat-unit length) of the VNTR.

							Replication	
	Locus	Length	RU Length	Effect Size	Gene	Annotation	Icelandic	Geuvadis
1	chr1:21440112-21440147	35	6	0.43	*NBPF3*	UTR	Y	Y
2	chr2:24084339-24084414	75	25	-0.12	*TP53I3*	UTR	Y	Y
3	chr2:25161573-25161616	43	9	0.22	*POMC*	Coding	Y	Y
4	chr2:112542424-112542500	76	25	-0.18	*POLR1B*	Coding	Y	Y
5	chr3:56557249-56557289	40	20	-0.12	*CCDC66*	Coding	Y	Y
6	chr6:13328502-13328532	30	6	0.12	*TBC1D7*	UTR	Y	Y
7	chr7:64337190-64337240	50	13	0.09	*ZNF736*	UTR	Y	Y
8	chr8:86508719-86508765	46	23	0.13	*RMDN1*	UTR	Y	Y
9	chr10:102869497-102869605	108	36	0.22	*AS3MT*	Coding	Y	Y
10	chr21:46228815-46228863	48	9	-0.03	*LSS*	UTR	Y	Y
11	chr17:75589192-75589228	36	6	-0.06	*MYO15B*	Coding	Y	-
12	chr1:46609102-46609134	32	16	0.09	*MOB3C*	UTR	Y	N
13	chr5:80654880-80654954	74	9	0.04	*MSH3*	Coding	Y	N
14	chr9:137063433-137063550	117	39	-0.15	*SAPCD2*	UTR	Y	N
15	chr14:61762420-61762454	34	17	0.03	*SNAPC1*	UTR	Y	N
16	chr19:12577507-12577551	44	22	-0.09	*ZNF490*	UTR	Y	N
17	chr21:41316673-41316756	83	13	-0.19	*FAM3B*	UTR	Y	N
18	chr22:37805258-37805313	55	6	0.11	*H1F0*	UTR	Y	N
19	chr1:202187007-202187042	35	7	0.06	*PTPRVP*	UTR	N	Y
20	chr17:18208488-18208544	56	7	-0.13	*ALKBH5*	UTR	N	Y
21	chr17:76564106-76564152	46	9	0.11	*SNHG16*	UTR	-	N
22	chr17:56978047-56978107	60	20	0.15	*SCPEP1*	UTR	N	N
23	chr6:30163542-30163579	37	12	0.14	*TRIM15*	UTR	-	-

3.4 Discussion

VNTRs are the "hidden polymorphisms." Despite high mutation rates and known examples of function modifications, VNTR genotyping is not a component of Mendelian or GWAS pipelines. This is primarily due to technical challenges. Here, we use a combination of fast filtering followed by a hidden markov model-based genotyping to accurately determine VNTR genotypes. Our method, adVNTR-NN, can genotype 10K VNTRs for an individual in 50 cpu-hours with high accuracy. We used adVNTR-NN to genotype close to $2,000$ human samples at 10K loci. The use of neural networks as a filtering strategy is novel, and we believe that further improvements could lead to another order of magnitude reduction in compute time, making it

practical to genotype $\geq 10^5$ individuals in the future.

Some VNTRs have complex multi-repeat structure making it difficult to map reads and count the repeating units. However, unlike other VNTR genotyping methods, our method customizes the genotyping for each VNTR. Future research will focus on improving the genotyping for the hard cases, possibly by building HMMs with separate profiles for each distinct repeating unit, as well as the use of long-reads to improve anchoring to the correct locations. We pursue a targeted genotyping approach which has the disadvantage of not being able to discover new VNTRs, and we rely on other methods for the initial discovery of VNTRs. However, we note that the discovery is a one-time process while genotyping must be repeated for each cohort, and therefore, it makes sense to separate the two problems. For maximum sensitivity, discovery of VNTRs could be performed on a new cohort prior to genotyping. Even if the reference contained 0 copies, knowledge of the repeat pattern and location would allow us to genotype donors with multiple repeat units.

The relatively large number of VNTRs violating HWE suggests that genotyping accuracy could be improved by filtering problematic VNTRs. We are developing strategies to filter VNTRs based on similarity to other VNTRs, matching sequence of repeat-units and flanking regions, and other tests for long alleles. As more data is collected, we will be able to assess the accuracy of these strategies.

adVNTR-NN can be used for association of a VNTR genotype chosen from a large collection of target VNTRs, against categorical or quantitative phenotypes. We used it to identify eVNTRs, where VNTR allele changes associated strongly with gene expression. It is possible that the largest allele or some other regrouping has the strongest effect for some VNTRs, and this idea may be used to strengthen the eVNTR association. However, we did not have a consistent strategy for grouping the VNTRs and therefore did not try this approach for the VNTRs in our study. Nevertheless, for individual VNTRs that are on the borderline for significance, this approach could be tried prior to functional tests.

We found that VNTRs were strongly associated with the expression of proximal genes with over 6.1% of the 2672 VNTRs tested showing genome wide significant association. Nearly half of the eVNTR loci were more significant compared to neighboring SNPs. While the high fraction of causal eVNTRs can partly be explained by the choice of 'genic' VNTRs for testing, we believe that non-genic regions will identify additional causal eVNTRs. In testing for causality, it would be best to compare against all other forms of variation including SNPs (which include small indels), structural variations, and other STRs. However, there is significant complexity in calling these variants. For example many STRs and even VNTRs are mis-annotated as structural variants. We will address these concerns in future work. In summary, ongoing technical innovations in speed and accuracy of VNTR genotyping are likely to improve our understanding of human genetic variation, and provide novel insights into the function and regulation of key genes and complex phenotypes.

Acknowledgements

The research was supported in part by grants HG010149, and R01GM114362 from the NIH.

Chapter 3, in full, contains material from Mehrdad Bakhtiari, Jonghun Park, Yuan-Chun Ding, Sharona Shleizer-Burko, Susan L. Neuhausen, Bjarni V. Halldórsson, Kári Stefánsson, Melissa Gymrek, Vineet Bafna. "Variable Number Tandem Repeats mediate the expression of proximal genes." *Nature Communications*, 2021[9]. The dissertation author was the primary author of this paper.

Code availability

adVNTR-NN is available at https://github.com/mehrdadbakhtiari/adVNTR. v1.4.0 of the software was used for this paper. Code for eVNTR analysis and generating figures is available at

https://github.com/mehrdadbakhtiari/VNTR-eQTL[8].

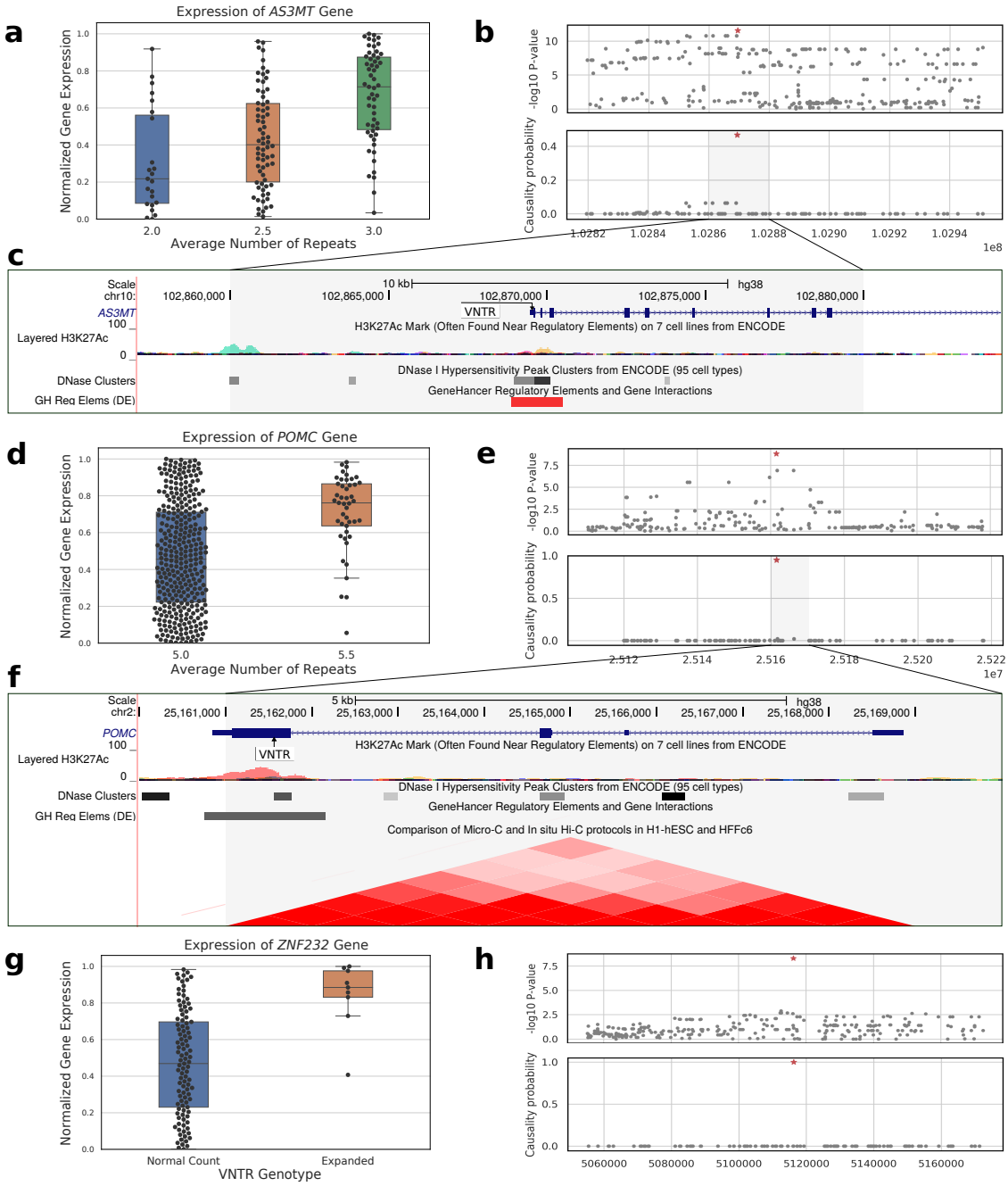


Figure 3.3: **Causal effect of VNTR genotypes on mediating expression of key genes.** (a) Association of *AS3MT* VNTR genotype with gene expression in Brain-Cortex (n=148 samples, Fisher's two-sided *P*: 2.78×10^{-12}). Box plots display the median, 25th and 75th percentiles. (b) Association with gene expression (upper panel) and CAVIAR causality probability of proximal SNPs– all SNPs in 100kbp window on either side of the *AS3MT* VNTR (red-star). (c) Location of *AS3MT* VNTR relative to known regulatory elements. (d,e): Association with gene expression of the *POMC* VNTR (n=378 samples, Fisher's two-sided *P*: 1.53×10^{-9}) and its causality probability relative to proximal SNPs. Box plots display the median, 25th and 75th percentiles. (f) Location of *POMC* VNTR relative to other regulatory regions and its spatial proximity with the promoter region revealed via Hi-C. (g,h) Association with gene expression of the *ZNF232* VNTR (n=114 samples, Fisher's two-sided *P*: 5.47×10^{-9}) and its causality score relative to proximal SNPs. Box plots display the median, 25th and 75th percentiles.

Chapter 4

Case study of VNTR variation effects on Breast Cancer

4.1 Introduction

For carriers of pathogenic BRCA1 or BRCA2 mutations (BRCA), the lifetime risk of developing breast cancer (up to an 80% lifetime risk) is a six-fold increase over that of average risk women and ovarian cancer risk (up to a 44% lifetime risk) is up to a 30-fold increase [68]. Despite higher average risk, penetrance is incomplete (not all carriers will develop cancer) and age at cancer diagnosis varies. The limited understanding of factors that modify cancer risks in BRCA carriers hampers clinical decision-making ability, including decisions about the appropriate type and timing of preventive strategies. Therefore, there is a critical clinically-relevant need for more refined estimates of risk.

The variation in risk, even in identical mutation carriers, suggests that modifier factors, both genetic and environmental, affect cancer risks [79]. Studies to identify "modifier genes" that govern the phenotypic expression of BRCA mutation carriers have been ongoing for the past decade, conducted largely through the Consortium of Investigators of Modifiers of BRCA1/2

(CIMBA) [6]. Through genome-wide association studies (GWAS), single nucleotide polymorphisms (SNPs) have been identified that, when combined into a polygenic risk score (PRS), better define those at higher and lower risk of developing breast cancer (e.g., [24, 69, 89]). However, these modifier variants explain only a portion of the variation in risk, particularly for women carrying BRCA1 mutations [93]. Identifying additional genetic modifiers will facilitate better risk estimates for clinical decision-making on timing and options for prevention.

A limited number of studies of VNTRs and breast cancer risk have been published to date that reported an association of rare alleles in a HRAS1 VNTR and development of cancers, including breast cancer [67], and a meta-analysis of 13 breast cancer studies found an association with breast cancer risk [141]. Functional analysis showed that this HRAS VNTR altered CpG DNA methylation. Another breast cancer-associated VNTR is a CAG-repeat polymorphism in the androgen receptor; a meta-analysis of 17 studies found an association of longer CAG repeats with an increased risk of breast cancer in Caucasian women [86]. For MNS16A, a VNTR in the hTERT promoter, a meta-analysis found that it was significantly associated with development of breast cancer (OR 1.46; 95% CI, 1.16-1.84) [140]. These studies were targeted at individual VNTRs located at or around candidate genes. Thus, although there is ample suggestive evidence that VNTRs may increase risk of developing breast cancer, there has been no systematic, genome-wide investigation of VNTRs such as proposed herein.

In this chapter, we will systematically assess the role of Variable Number Tandem Repeats (VNTRs) as genetic modifiers of breast cancer risk in BRCA1 and BRCA2 pathogenic mutation carriers, and as a determinant of age at diagnosis. VNTRs are known to modulate biologic processes including gene transcription and protein function [17, 53]. These eVNTRs (expression Quantitative Trait Loci) also mediate risks of developing various cancers [110] including breast cancer [67]. Unlike SNPs, where fine-mapping is often needed to identify the causal SNP, the VNTR at a locus, is more likely to be the causal alteration because of the disruption due to the multiple alleles and insertion/ deletion of multiple base pairs[17]. This will be the first study to

conduct a genome-wide investigation to identify VNTRs that may act as genetic risk factors for the development of breast and ovarian cancers. We hypothesize that VNTRs act as modifiers of risk of developing breast cancer in female BRCA-mutation carriers.

4.2 Materials and Methods

One reason for the lack of genome-wide investigation of VNTRs has been feasibility due to a lack of high-throughput genotyping and computational methodologies. Historically, VNTR genotyping required gel-based screens which are not amenable to high-throughput genotyping [101]. Microarray-based technologies, such as used for SNP GWAS studies, do not exist for VNTR genotyping. With the advent of high throughput sequencing, it is possible to identify variation ranging from SNPs to large structural variation. With a robust method that we developed to study VNTRs (adVNTR), we could genotype VNTRs from next generation sequencing (NGS) data including whole genome (WGS), whole exome (WES), and targeted sequencing data.

adVNTR uses Hidden Markov models (HMMs) to model each VNTR, count repeat units, and detect sequence variation. Using this approach, we conducted a pilot study using targeted-capture sequence of VNTRs, called genotypes with adVNTR, and explored the association of VNTRs and breast cancer in BRCA1 carriers.

To identify VNTRs that show association with breast cancer risk in women carrying pathogenic BRCA1 and BRCA2 mutations, we conducted targeted sequencing of 6271 VNTRs located in gene coding and regulatory regions on 552 DNA samples from female BRCA pathogenic mutation carriers. We used adVNTR to assign VNTR genotypes and conducted a GWAS of VNTRs using modified statistical association tests. We then used retrospective likelihood approaches within a survival analysis framework to test the associations.

4.3 Improving VNTR genotyping accuracy

With the exception of single nucleotide polymorphisms, identifying genomic variations have not reached near perfect accuracy. For example, the sensitivity of identifying insertion or deletion within the genome is estimated at 75% to 97% depending on the variant caller and sequencing coverage[3]. Similarly, the accuracy of VNTR genotyping using short reads is reported to be 98.08% in the most accurate experiments[10]. Here, we took additional steps to close the accuracy gap as much as possible to achieve the perfect genotyping with the exception of experimental shortcomings (e.g. PCR stutter error or low sequencing coverage)[50].

After genotyping 6271 VNTRs in 552 samples, we found the following categories of erroneous genotype calls reported with their frequency:

Lack of enough sequencing coverage

One assumption in the sequencing is that each amplification that leads to read generation happens independent of others and has a $1/2$ probability of occurrence in each haplotype in the sample. So, if r reads are generated, it is reasonable to assume $\frac{1}{2^r}$ probability that all reads are sequenced from one haplotype and the data from the other haplotype is missing in the data. Therefore, when the coverage $<$ 10X for hundreds of loci, we expect to miss reads from one haplotype in a few of them. If these cases are heterozygous VNTRs in the sample, then missing a read from one haplotype will lead to an erroneous homozygous call since all the reads support one specific allele.

To find the expected number of VNTRs with this error, we looked at the coverage of reads for every 6271 target VNTR in 552 samples. While the majority of them had high coverage (with a mean 57.9X and a median of 35X), 10.6% of the calls had less than 7 reads in at least one haplotype. Assuming less than 50% of them are polymorphic with a heterozygous underlying genotype, we expect that less than 1% of the genotyped VNTR suffer from this error[9]. In

addition, there are post processing methods to deal with the trade-off between the number of genotyped loci to include in the downstream analysis and the probability of missing reads from one haplotype [96], where including loci with more reads reduces the probability of this error while removing many correct calls. Using these methods, a user specific threshold could be chosen to filter less reliable calls.

Similar VNTRs appeared multiple times on genome

Some of the VNTRs are located within a large segmental duplication or other evolutionary process have copied them to another chromosome or another location within the same chromosome. Therefore, the reads sequenced from these VNTRs are indistinguishable due to the similarity of pattern and flanking regions even though they may have different genotypes.

We found out 1495 VNTRs (23.8%) out of 6271 target VNTRs are located in more than one place in the genome (e.g as part of SINE or LINE elements in the genome or segmental duplications that contain VNTRs). Although in some cases it may be possible to differentiate the reads of similar VNTRs depending on the evolutionary history of VNTRs, we removed the duplicated VNTRs from our downstream analysis.

Incorrectly recruited reads

Initially, adVNTR's read recruitment was designed based on the assumption that mutations occur with the same rate in the VNTR region. However, as it later turned out, mutations are up to 10^3X more likely within the tandem repeats of the VNTR compared to the flanking regions [10]. Assuming a high mutation rate within the flanking regions results in accepting the reads that do not have a perfect match in the flanking regions and are not sequenced from the VNTR region. Using these erroneous recruited reads will later interfere with the statistical model estimating the underlying genotype in the sample. To resolve this, we introduced additional filters in the

read recruitment stage to enforce high similarity in the flanking regions observed in the reads. To quantify its effect, we measured the sequence identity of flanking regions within $115,988$ recruited reads for every genotyped VNTR in 552 samples before enforcing the filter. Although the alignment score had a median of 98% meaning majority of the reads had near identical flanking regions, 7.4% of the reads had low quality alignments for flanking regions of VNTRs. By enforcing high similarity in flanking regions ($> 95\%$) we could eliminate the incorrectly recruited reads.

Erroneous heterozygous call with weak support

adVNTR was initially developed to leverage whole genome sequencing data which resulted in expecting $\sim 30X$ sequencing coverage. So, all the parameters of the method and the model were trained to find erroneous reads or strong support for an allele using WGS data with a reasonable variation in coverage. However, using the targeted sequencing approach in this project we generated more than 100%X sequencing data for VNTRs and we observed close to 200 reads covering some VNTRs. While in a typical whole genome sequencing experiment having 5 reads is a sign of strong support for an allele, having 5 reads with PCR stutter error is to some degree expected out of 200 reads. So in a case where the sample is homozygous at a particular locus, sequencing would result in an imbalanced support for the correct allele and a few reads supporting the erroneous genotype. To resolve this issue and account for this scenario, we did introduce additional parameters looking at the assignments of reads to each haplotype so that adVNTR models could use data generated with targeted sequencing approach.

Multiple repeating unit types

Another novel property of VNTRs that we have found is that while repeating units are not identical, they can be grouped together as some of them are more similar to each other.

Additionally, not all repeats of the same group appear in tandem and more different patterns can occur in between (Fig. 4.1). We found out that when we merge all such repeats to generate one profile HMM, the resulting profile does not represent every unit equally which in turn makes the alignments (and thus counting number of repeats) imperfect.

To tackle this problem, we developed an enhanced approach using Hidden Markov Models (HMMs) that instead of merging all repeating units together to generate a profile of repeating pattern, uses multiple profile of patterns that each represent one repeating group (Fig. 4.1). Using the enhanced model each repeating unit will match with its corresponding HMM with high quality alignment and it becomes possible to discard all the reads with low quality matches.

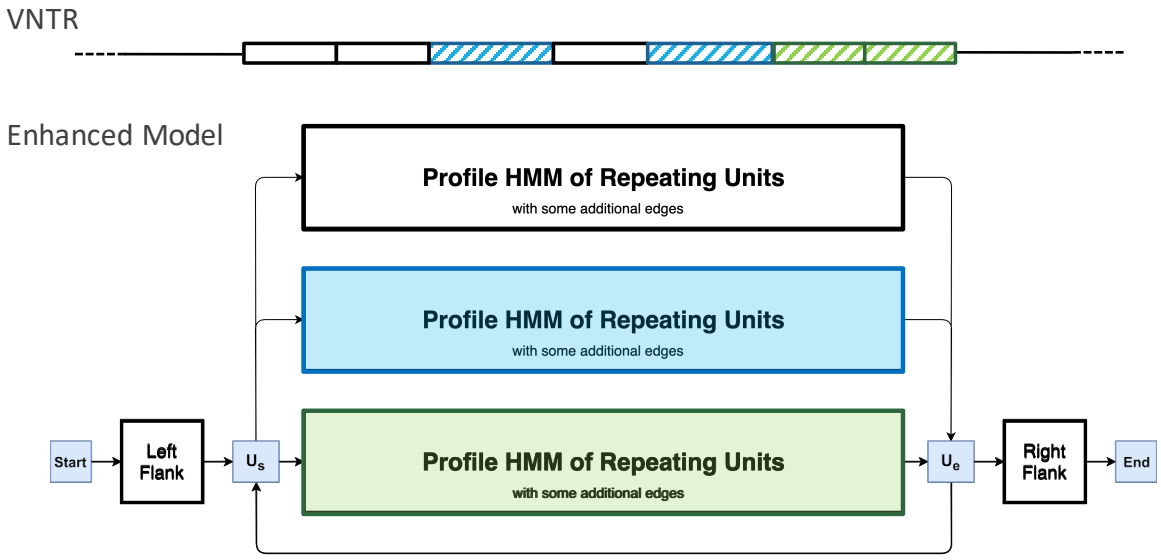


Figure 4.1: **The Enhanced VNTR HMM.** The HMM is composed of 3 profile HMMs, one each for the left and right flanking unique regions, and one in the middle to match multiple and partial numbers of RUs. Since not all repeating units are similar, we divide them into groups and introduce a new profile HMM for each group. The special states U_s ('Unit-Start'), and U_e ('Unit-End') are then connected to all possible repeating unit HMMs.

Similarity of repeating pattern and flanking regions

One reason that adVNTR's HMMs fails to accurately count the number of repeats is when a prefix of repeating pattern is similar to the prefix of the right flanking region. In this case, the

HMM cannot clearly identify the sequence as the repeating unit or flanking region and may make erroneous assignment. Likewise, the similarity between the suffix of repeating pattern and the suffix left flanking region results in the same confusion in counting the repeats. The relation between these similarities and erroneous calls is measured in Fig. B.12, which shows while most of the VNTRs do not suffer from this error, there are cases in which distinguishing repeating units and VNTR borders are not perfect.

Additionally, we found out this ambiguity is more likely when the VNTR region is made up of complex repeats in the genome and multiple VNTRs appear within one another (e.g. the repeating unit itself is a VNTR) or have intersection with each other (e.g. smaller part of one repeating units is repeated multiple times as another VNTR). We found out 1586 of our target VNTRs (25.3%) have these properties and thus more likely to suffer from this problem. To systematically solve the issue, we modified the model so that a specific number of base pairs should be aligned in order to make a call on whether a pattern belongs to the repeating units or flanks. So, part of the information in reads could be discarded if it is not possible to resolve the ambiguity. To modify this method, we made this threshold a VNTR specific parameter that is dynamically determined based on VNTR sequence. We define it as the shortest prefix (or suffix, when applicable) where repeating unit and flanking region are at least 70% different so the read sequence can be assigned to the repeating unit or flank unambiguously.

Presence of multiple alleles longer than read length

The last limitation that we observed using adVNTR was the inability to distinguish between genotypes that are both longer than the length of a single sequencing read. However, since we divide the genotypes of a VNTR into two groups of *short* and *long* alleles, the length of a read length can still be considered as the threshold dividing the risk group and rest of the samples. Since we focus on the short VNTR in the genome, the majority of alleles will be shorter than sequencing reads and we do not find cases where two frequent alleles are longer than the

Table 4.1: **Number of alleles in the 2494 VNTRs in female BRCA1 mutation carriers.**

	Number of repeat alleles observed in a VNTR										
Alleles	2	3	4	5	6	7	8	9	10	11+	Total
Number of VNTRs	305	522	430	326	233	157	104	86	91	240	2494

read length. However, we believe further improvements similar to other variant callers can make it feasible to infer accurate lengths up to 800bp using the variations of insert size of paired reads [97].

4.4 Identifying VNTRs associated with the risk of developing breast cancer

We excluded monomorphic markers and those where there were fewer than 3 heterozygous genotypes((heterozygosity < 0.01), as well as those that violated Hardy-Weinberg equilibrium (HWE) measured by p-value < 0.01. These resulted in a total of 2494 VNTRs (40% of the target loci) for the association tests (Table 4.1).

We performed statistical analysis to identify the association between the VNTR genotype and breast cancer risk using a retrospective likelihood approach which models the likelihood of observing the VNTR genotypes given the observed disease phenotypes [5, 28]. In the model, the breast cancer incidence is assumed to depend on the underlying VNTR genotype through a Cox Proportional Hazards model. Women are censored at the first of: (a) age at diagnosis of breast cancer; age at prophylactic surgery; or age at last-follow-up. We considered participants with a first breast cancer as affected. In the primary association, the VNTR genotype was defined as a continuous variable using the average length of a participant's two alleles in the genotype[9]. We considered the study group, country of residence, and race/ethnicity determined by the top three principal components (PCs) from the principal component analysis of SNPs as the covariates in the model. Analyses were done separately for each VNTR. We then adjusted probability

values for multiple comparisons using the False Discovery Rate (FDR) method of Benjamini and Hochberg. For VNTRs with associations with FDR < 0.25 in the primary association analysis, we performed a secondary association analysis to identify the specific risk groups of repeat alleles using a sliding window method of dichotomizing repeat alleles into short (shorter or equal to a repeat cut-point) and long (longer than a repeat cut-point) alleles using a sliding cut-point along the observed small-to-large repeat length distribution of the VNTR [108]. This will convert the VNTR genotype of an individual to homozygous-short-allele genotype (S/S), heterozygous-short-and-long-allele genotype (S/L), or homozygous-long-allele genotype (L/L); in the secondary analysis, the effect of the long allele will be modeled as a per-allele hazard ratio by comparing women carrying two, one, or zero copies of the L alleles and breast cancer incidence. The optimal repeat allele cut point was then determined by the smallest p-value among the multiple association tests. This secondary analysis allowed us to identify critical cut points along the continuous repeat allele distribution in a VNTR for which breast cancer risk may be modified and then to estimate the effect size of association related to the specific repeat alleles.

In the primary analysis to test the association of breast cancer risk and a continuous VNTR variable for 2494 makers using the average repeat length of the two alleles of the VNTR genotype, we found 46 VNTR markers with p value ¡ 0.01, and 9 with FDRs ¡ 0.25. After performing the secondary test for them (Fig. 4.2), we experimentally validated the genotypes we obtained for the samples using NGS data. To validate the calls, we tested 10 samples such that they represent all observed alleles for the VNTRs. We then designed primers and tested a set of primers and accuracy of the VNTR genotype call. We ran a 2% agarose gel to measure the PCR product length and observed concordance for all VNTRs except one, which due to having a small repeating unit is not clearly separable with agarose gel.

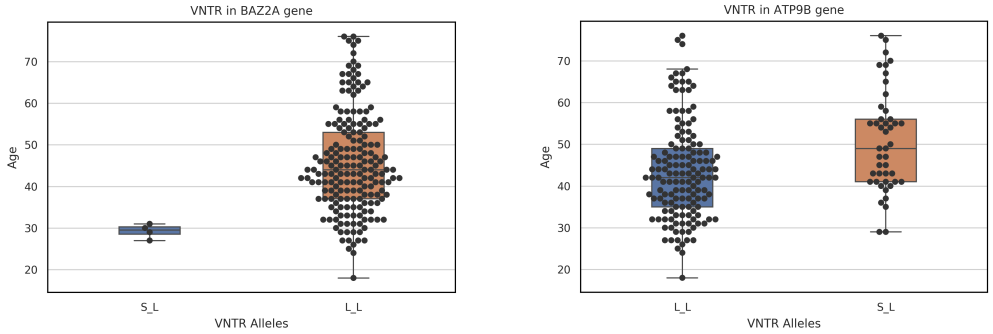


Figure 4.2: **Association of VNTR genotypes with risk of developing breast cancer.** The age onset of cancer is shown for each sample on the y-axis. For each VNTR, alleles are divided into two groups of short and long genotypes and samples are assigned to the corresponding homozygous or heterozygous genotype. In BAZ2A gene, the shorter alleles are associated with earlier age onset of cancer while longer allele of ATP9B VNTR is associated with earlier age onset of cancer.

Acknowledgements

Chapter 4, in part, is based on the ongoing research from Mehrdad Bakhtiari and Vineet Bafna, which is currently being prepared for submission for publication of the material. The dissertation author was a joint primary investigator and author of this material.

Appendix A

Supplemetary Material for Chapter 2

A.1 Model Structure and Parameter Setting

Each VNTR is represented by three Hidden Markov Models. A detailed sketch of the Repeat Match HMM is shown in Fig. 2.1. Here, we show the structure of two other parts in Fig. A.1. We repeated the blue silent states (*Start*, U_S, U_e, and *End*) to show how these three models are connected.

To set the transition and emission probabilities of repeat matcher, we used the parameter obtained by pair HMM of repeating units in reference genome. We set pseudocounts equal to error rate of sequencing technology in all three HMMs to allow for mutations and sequencing errors. After the initialization of each model, we updated them using sequencing data of NA12878 (Table A.2). To update each model, we ran read recruitment on sequencing data of NA12878 and extracted repeating units as described in Methods. Then, we aligned the repeating units to the HMM, and used the new aligned reads to update HMM parameters. We measure fitness of model by the sum of log-likelihood of the recruited reads, as follows:

$$\text{fitness} = \sum_{r \in reads} \log\left(\textbf{likelihood}(r)\right),$$

where likelihood of read r is defined as the probability of most likely path in the HMM to emit r. We continued to iterate the model alignment, and parameter update steps until convergence of fitness values.

As described in Methods, we compute the likelihood using the Viterbi algorithm. Let $V_{k,j}$ denote the highest (log) probability of emitting the first k letters of the sequence $s_1, s_2, \ldots s_n$ and ending in state j of an HMM. Let, $\text{Prev}_{k,j}$ denote the state j' immediately prior to j in this

optimum parse. Then,

$$V_{k,j} = \max_{j'}\{V_{k',j} + \log T(j',j) + \log E_j(s_k)\},$$

$$\text{Prev}_{k,j} = \arg\max_{j'}\{V_{k',j} + \log T(j',j) + \log E_j(s_k)\},$$

where, $k' = k-1$ for match or insert states; $k' = k$ otherwise. Then, for a read sequence r with length n, $max_j V_{n,j}$ over all states j in the HMM determines the maximum likelihood.

A.2 Selecting Target VNTRs

We selected sets of target models that could be analyzed based on their characteristics and the sequencing technologies as follows: We started with the human VNTR list created by Tandem Repeat Finder. To select the most important loci, we considered VNTRs that had an intersection with coding regions of human genome. Next, we excluded cases where the flanking regions of VNTR were not known (e.g. VNTR is close to telomere; the flanking region doesn't exist in reference genome; and there is a sequence of 'N' adjacent to the VNTR.). Finally, we added 17 VNTRs that are in promoter or intron of the genes but are known to be linked to a disease (Table 2.1). We removed VNTRs that appear multiple times in different loci of the genome with identical patterns and flanking regions, but with different number of copies. To find such similar VNTRs, we compared each pair of VNTRs by comparing the flanking regions and repeating unit with BLAT [62] and eliminating the VNTRs if their similarity was higher than 75%.

This procedure resulted in 2944 'coding' VNTRs out of 3147 VNTRs that intersected with coding regions of human genome. The 2944 VNTRs were used for PacBio analysis. For Illumia analysis, we used a subset of 1775 VNTRs of the 2944, whose length was shorter than 140bp. Finally to create a difficult test case for testing frame-shifts, we selected 115 of 2944 VNTRs for which the total length was $\geq$ 250bp, and all Repeat Units had the same length, and

used those to simulate indel (frameshift) data-sets.

A.3 Test Datasets

Multiple test cases were generated using the three lists containing 2944, 1775, and 115 VNTRs, respectively as described in the previous section. We started by generating a distinct human genomic sequence `VNTR_I_X_reference.fa` for each $I \in [1, 2944]$ and each value $X \in [-3, 3]$ (20,608 total sequences). Each sequence VNTR_I_X_reference.fa was identical to the human reference except that it had X' copies for I-th VNTR, where X' takes the RU count in reference genome $\pm X$. To increase the RU count of a VNTR, we added the repeating units from the first repeat to the last unit, one at a time. We additionally generated ~ 4920 reference sequences `VNTR_I_Deletion_P.fa` and `VNTR_I_Insertion_P.fa` for all $I \in [1, 115]$ VNTRs indexing the third list, and a single insertion or deletion at the Pth base pair of the Ith VNTR. We set P to every position in the VNTR that was a multiple of 10 and was at least 140bp apart from each side of the VNTR. These reference templates were used for generating simulated datasets as follows:

IlluminaSim Dataset. We used the following command to simulate the reads from haplotypes using ART:

```
art_illumina -ss HSXt -sam -i VNTR_I_X_reference.fa -l 150 -f 15 \
-o VNTR_I_X_set
```

Then, we merged every pair of haploid datasets with RU counts X and Y to get diploid sequencing data with genotype (X,Y) for VNTR I by appending VNTR_I_X_set.fq to the end of VNTR_I_Y_set.fq to get VNTR_I_XY_set.fq. Then, we aligned these diploid reads to the reference genome using Bowtie 2 as follows:

```
bowtie2 -x hg19_bowtie2_index -U VNTR_I_XY_set.fq -S VNTR_I_XY_aln.sam
```

PacBioSim Dataset. We used the following command to simulated the reads for *I*th VNTR using SimLoRD:

```
simlord -rr VNTR_I_X_reference.fa -pi 0.12 -pd 0.02 -ps 0.02 \
-c 15 VNTR_I_X_pb_set
```

Next, we merged each pair of reads (fastq files) to get the diploid set of reads at 30× coverage.

PacBioLong Dataset. The dataset is similar to PacBioSim but with higher RU counts for 3 VNTRs 120, 40, and 25 for VNTRs in *INS*, *CSTB*, and *HIC1* genes, which represent the largest expansion known for these VNTRs. Again, we used SimLord to generate reads.

```
simlord -rr VNTR_I_X_reference.fa -pi 0.12 -pd 0.02 -ps 0.02 \
-c 30 VNTR_I_X_pb_set
```

PacBio Coverage Dataset. We simulated different levels of coverage for the three VNTRs using:

```
simlord -rr VNTR_I_X_reference.fa -pi 0.12 -pd 0.02 -ps 0.02 \
-c C VNTR_I_X_C_set
```

Here, $1 \leq C \leq 40\times$.

IlluminaFrameshift Dataset. We simulated these datasets using following commands:

```
art_illumina -ss HSXt -sam -i VNTR_I_Deletion_P.fa -l 150 -f 15 \
-o VNTR_I_Deletion_p
art_illumina -ss HSXt -sam -i VNTR_I_Insertion_P.fa -l 150 -f 15 \
-o VNTR_I_Insertion_p
```

We also simulated reads from reference genome without the frameshift:

```
art_illumina -ss HSXt -sam -i hg19.fa -l 150 -f 15 -o normal_haplotype
```

Finally, we merged fastq read files of a haplotype with frameshift with that of normal haplotype to get the diploid sample at 30× coverage and aligned the reads with Bowtie 2 similar to "IlluminaSim Dataset".

Table A.1: **Simulated dataset summary.**

Dataset Name	Profile	Depth	# of VNTRs
Illumina Genotyping Dataset	HiSeqX TruSeq	30X	1775
PacBio Genotyping Dataset	PacBio	30X	2944
PacBio Long Expansion Dataset	PacBio	30X	3
PacBio Coverage Dataset	PacBio	$1 \leq C \leq 40$	3
Frameshift Dataset	HiSeqX TruSeq	$C \in \{10, 20, 30, 40\}$	123

WGS data used for testing was taken from Genome in a Bottle, NCBI sequence read archive, Polaris, while exome data was obtained from GoT2D. See Table A.2

Table A.2: **Real sequencing data used in tests.**

Samples	Study	Profile	PCR free	Depth	Access
AJ Child	GIAB	PacBio	-	70X	http://jimb.stanford.edu/giab-resources
AJ Father	GIAB	PacBio	-	30X	http://jimb.stanford.edu/giab-resources
AJ Mother	GIAB	PacBio	-	30X	http://jimb.stanford.edu/giab-resources
Chinese Child	PRJEB12236	PacBio	-	70X	ncbi.nlm.nih.gov/sra/ERX1322863
Chinese Father	PRJEB12236	PacBio	-	35X	ncbi.nlm.nih.gov/sra/ERX1322861
Chinese Mother	PRJEB12236	PacBio	-	35X	ncbi.nlm.nih.gov/sra/ERX1322862
AJ Child	GIAB	HiSeq 2500	Y	40X	http://jimb.stanford.edu/giab-resources
AJ Father	GIAB	HiSeq 2500	Y	40X	http://jimb.stanford.edu/giab-resources
AJ Mother	GIAB	HiSeq 2500	Y	40X	http://jimb.stanford.edu/giab-resources
NA12878	GIAB	PacBio	-	70X	http://jimb.stanford.edu/giab-resources
NA12878	GIAB	HiSeq 2500	Y	30X	http://jimb.stanford.edu/giab-resources
Subset of 1KGP	Polaris	HiSeq X	Y	30-40X	ebi.ac.uk/ena/data/view/PRJEB20654
Diabetes WES	GoT2D	HiSeq 2000	N	82X	phs001095, phs001096, and phs001097

A.4 Running adVNTR

adVNTR is available at https://github.com/mehrdadbakhtiari/adVNTR. As stated in the repository, the best way to install it is to use conda package manager and running `conda install`

advntr. After installation, advntr command invokes the program with four possible commands genotype, addmodel, viewmodel, and delmodel. Detail of each command as well as complete tutorial on installation and usage are available at http://advntr.readthedocs.io/. Also, passing -h argument to each command will show the correct command line usage of the command. We used following commands to run adVNTR on each simulated datasets:

IlluminaSim Dataset

```
advntr genotype --alignment_file VNTR_I_XY_aln.bam -vid I \
-wd ./working_dir
```

PacBioSim and PacBioLong Datasets

```
advntr genotype --alignment_file VNTR_I_X_pb_set.fastq.bam -vid I \
-p -wd ./working_dir
```

PacBio Coverage Datasets

```
advntr genotype --alignment_file VNTR_I_X_C_set.fastq.bam -vid I \
-p -wd ./working_dir
```

IlluminaFrameshift Dataset

```
advntr genotype --alignment_file VNTR_I_Insertion_p.bam -vid I \
-fs -wd ./working_dir
```

A.5 VNTRseek

In order to make a call on a VNTR, VNTRseek requires both ends to be anchored with a minimum of 20bp on each side of VNTR. This limits the length of VNTRs that can be identified using VNTRseek is limited to 110bp using Illumina sequencing technology. Also, it compares each VNTR in the sequencing reads to every VNTR in reference genome which makes the process computationally demanding, and inaccessible for large data-sets. For these reasons, extensive VNTRseek comparisons were not conducted.

A.6 Supplementary Figures and Tables

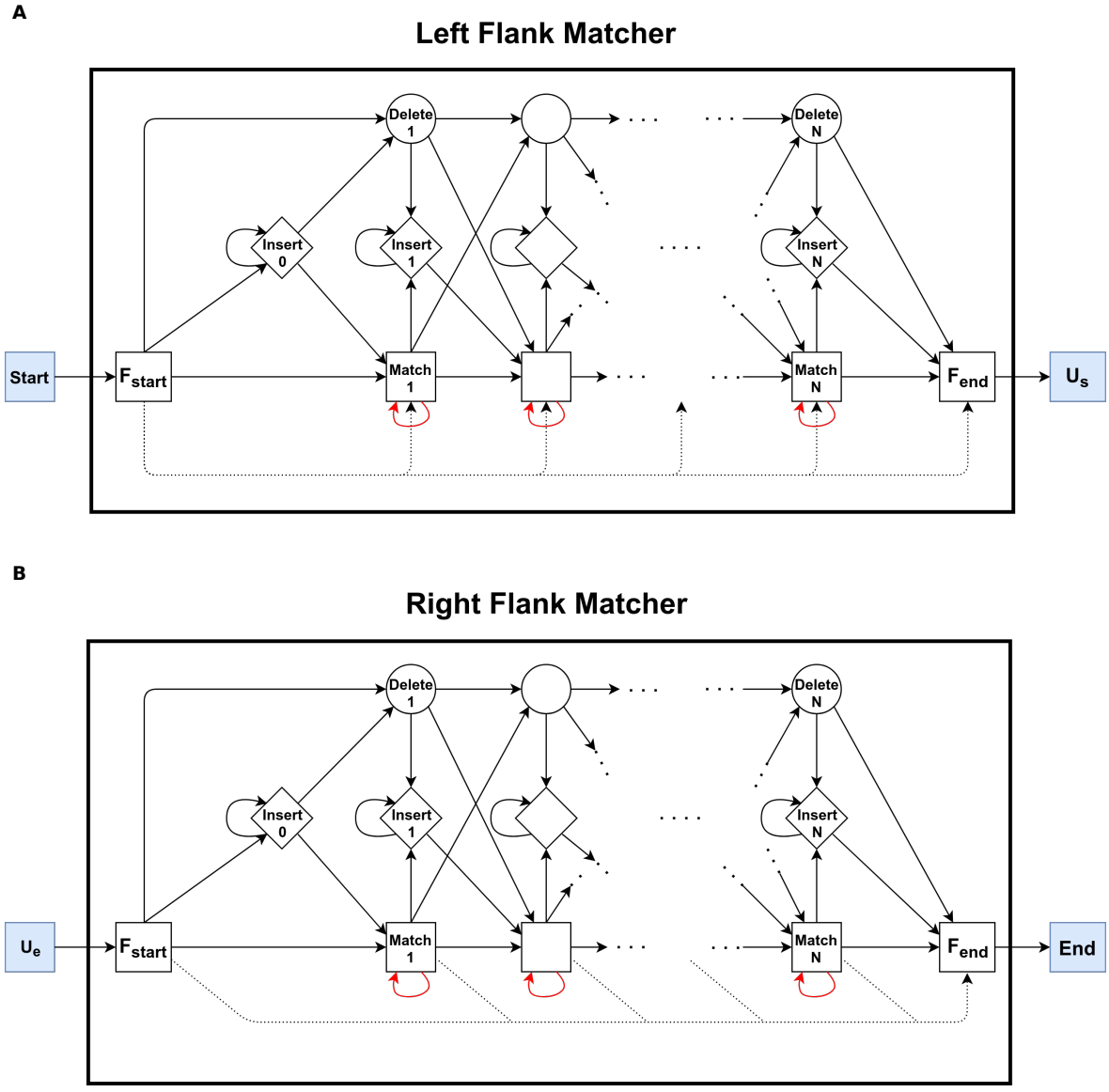


Figure A.1: **Flanking region matcher HMMs.** (A) Shows the structure of Left Flank Matcher, which matches a suffix of left flanking region of the VNTR. In this part, the dotted edges allows skipping of adjacent base pairs at the beginning of the flanking region, and the rest of region (base pairs on the right) should be matched to the states and this is how matching of a suffix is insured. (B) Shows the structure of Right Flank Matcher, the model that matches a prefix of right flanking region of the VNTR. Here, dotted edges ensure the matching of a prefix of the flanking region sequence.

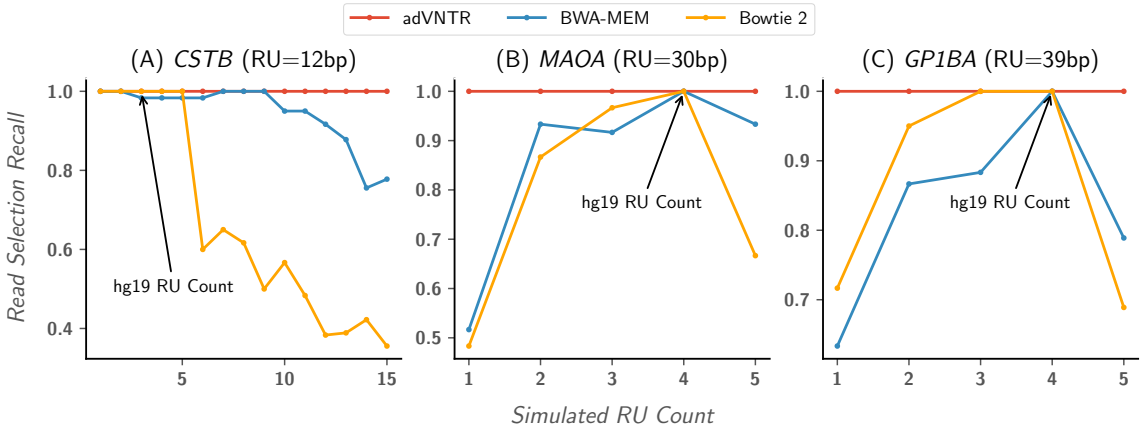


Figure A.2: **Sensitivity of Illumina read recruitment at specific VNTR loci.** Comparison of adVNTR read selection with BWA-MEM and Bowtie 2 mapping for Illumina reads (short VNTRs). Each plot shows the sensitivity of mapped/selected reads as a function of the number of repeats for different VNTRs. These plots show examples of alignment tools' behavior when RU count of VNTR deviates from the RU count in the reference genome. (A) Shows the comparison for the VNTR in *CSTB* gene, in which the pathogenic cases have more then 12 repeats and as it is shown alignment tools perform poorly in those cases. (B) Shows the comparison for the VNTR in *MAOA* gene, where the 4 repeats corresponds to both pathogenic case and number of repeats in reference genome. However, other tools perform poorly in normal cases. (C) Shows the comparison for the VNTR in *GP1BA* gene, and again, alignment tools only perform well when RU count is same as RU count in reference genome.

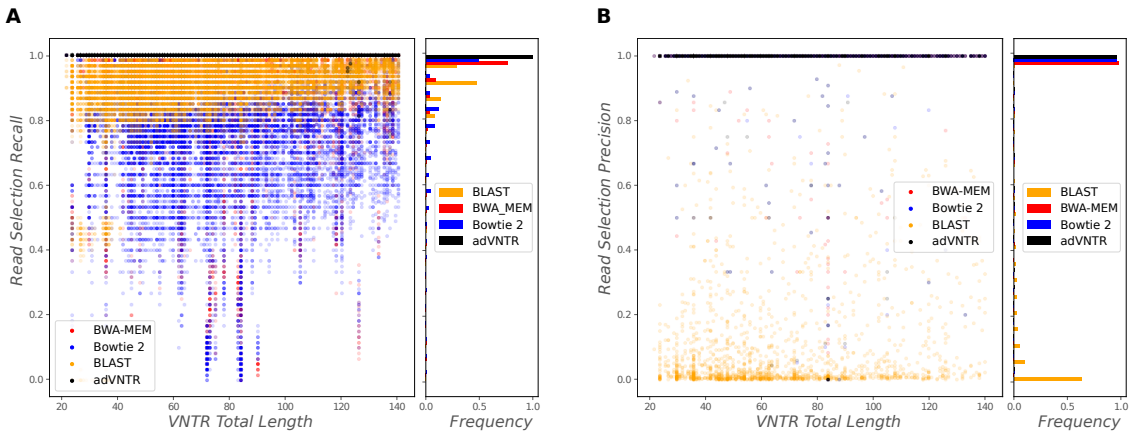


Figure A.3: **Read recruitment accuracy on Illumina reads.** (A) Shows the comparison of the recall of adVNTR read recruitment with BWA-MEM, Bowtie 2, and BLAST. (B) Shows the precision for read recruitment. These figures show that adVNTR has much higher recall compare to standard alignment tools without losing precision.

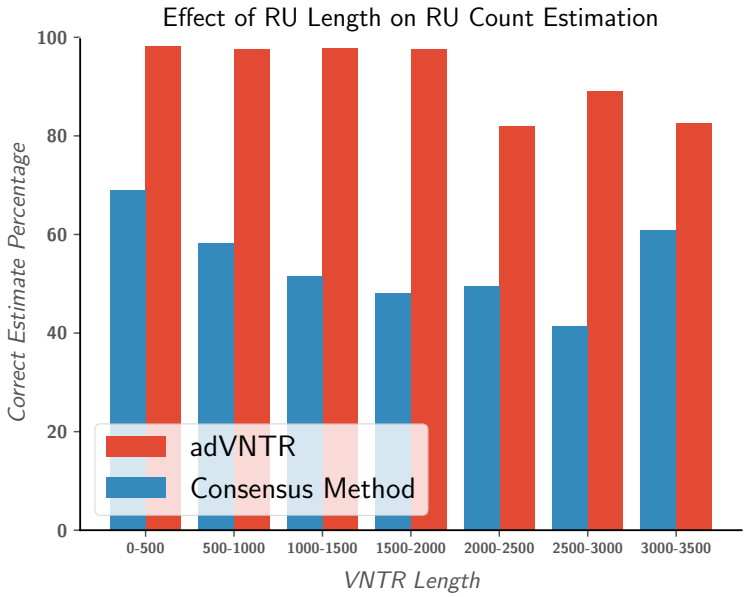


Figure A.4: **Comparison of adVNTR genotyping with consensus method on homozygous simulated data.** adVNTR and consensus method comparison on homozygous testcases in *PacBioSim*.

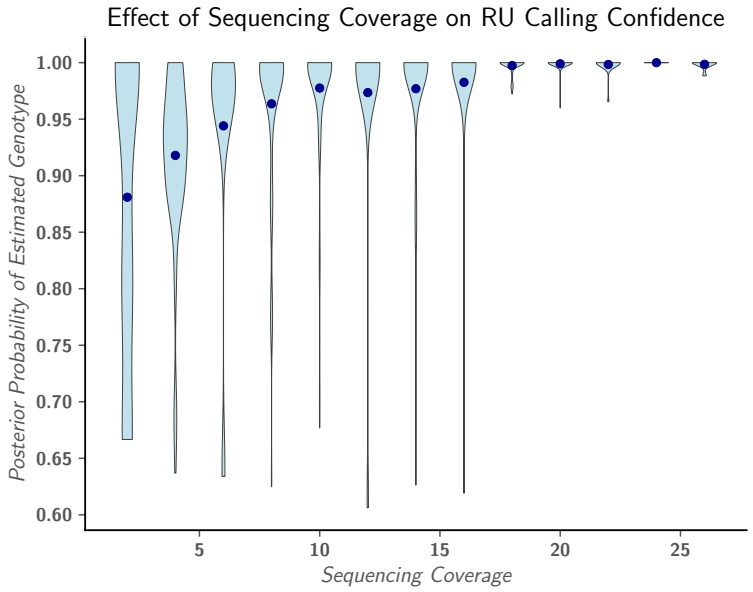


Figure A.5: **Association of PacBio sequencing coverage in VNTR region and posterior probability of RU count calling.** The figure shows posterior probability of RU count estimation in AJ trio sequencing data form GIAB. Most of calls with low posterior probability (low confidence calls) result from low coverage in VNTR region. With at least 10 reads that span the VNTR, we will get 0.98 posterior probability for estimated genotype.

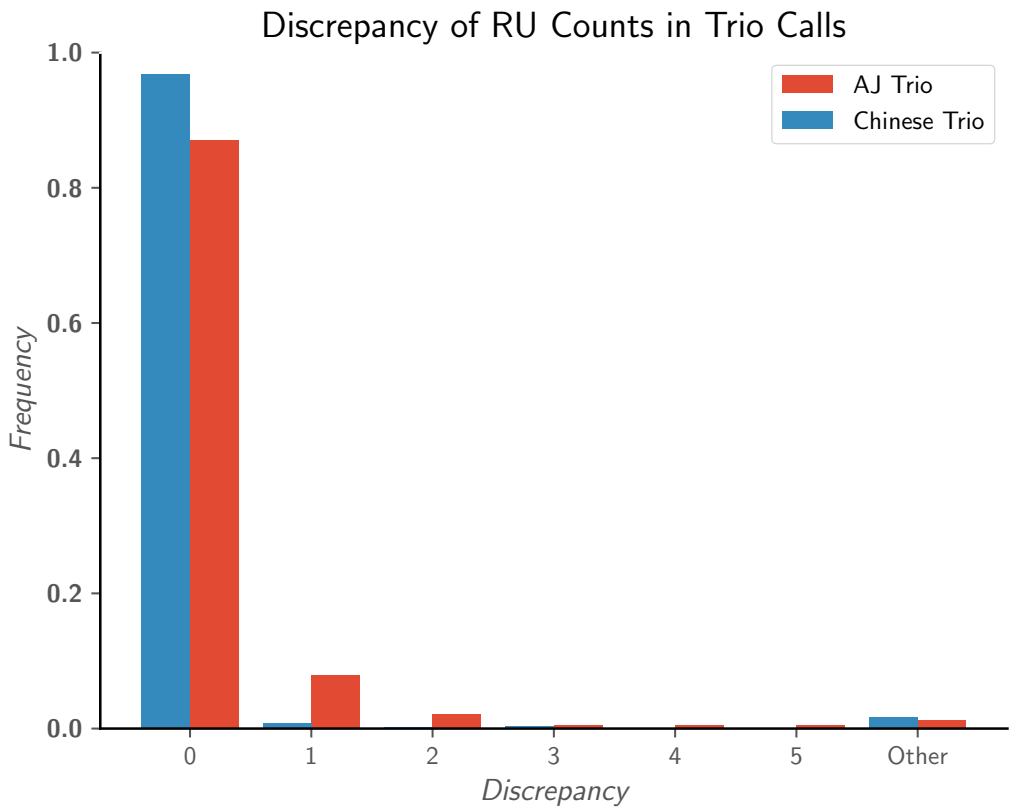


Figure A.6: **Distribution of discrepancies on trio calls using PacBio reads.** This figure shows the distribution of discrepancies in adVNTR estimates on AJ and Chinese trios. As shown in the figure, most of non consistent calls in AJ trio have one discrepancy in estimated RU counts.

Table A.3: **Primers for gel electrophoresis validation.** Last column shows whether we used the primers were used for a long range PCR. We used long range PCR to validate adVNTR calls on longer VNTRs (using PacBio reads).

Gene	Locus (hg19)	Forward primer	Reverse primer	hg19 product length	Long reads
MAOA	chrX:43514348-43514468	GGCTACACCCACG TCTACTC	CACTCTTGGAGTC GGAGTCA	679	Y
IL1RN	chr2:113888105-113888449	ATTCCTGTCCTGG TAGTTCTCC	AGAGGGGAGGGTC AGGTTAAT	701	Y
GP1BA	chr17:4837118-4837278	AGGACTGTGGTCA AGTTCCC	GCTTTGGTGGCTG ATCAAGT	586	Y
DRD4	chr11:639988-640180	CCGTGTGCTCCTT CTTCCTA	GACAGGAACCCAC CGACC	481	Y
SLC6A4	chr17:28564157-28564483	AGGGACTGAGCTG GACAAC	AGGCAGCAGACAA CTGTGTT	632	Y
JAKMIP3	chr10:133954073-133954190	CAAACAGACAGGA CGGACC	GTGCCCGAGTCAG CTATCA	249	N
SRSF8	chr11:94800727-94800790	CAGGTGGCGCGCT ATG	GAGACCGGCTATA GCGAGAA	214	N
SSTR1	chr14:38679763-38679811	CGTCTTCCGTAAT GGCACCT	CCCTGGATACCGT CCCTTT	153	N
C14orf180	chr14:105055118-105055145	CCTATACTGCGGC CGGG	CCTAGTTAGCCCT CAGGCAG	265	N
EIF3G	chr19:10229726-10229768	GGCAGAAGGGGAA AAACAGA	AGCTGACTCCTCC TTCCTAC	247	N
STK39	chr2:169103796-169103845	AACTGTTGAAGCC AGTAGGC	AGTTTCAAGTGGA AGGTCGT	408	N
BRWD1	chr21:40585353-40585415	TGCCCTATTTGTT CATTGGACT	TCCTTGCCAACAA GTCACTAC	249	N
CSTB	chr21:45196323-45196359	GAGGCACTTTGGC TTCGGA	GCGCCCGGAAAGA CGATA	193	N
UBXN11	chr1:26608801-26608909	GCCTTTCCTACGT GCCTG	AGATCTTCAGCAC ATTCCCG	321	N
CLCA4	chr1:87045895-87045932	CTCAGAAGAAAAT GCAACCCAC	CACAGACAATACC AGCGTAGA	214	N
LCE4A	chr1:152681679-152681727	ATCCCCAAGTATC CCCAAA	GACCTATGGTGTC TGTGGTG	152	N
PAOX	chr10:135202324-135202464	CAGTGGTTCCTTG CTGAGAA	GGCAATGAACCCA CAGAGAA	214	N

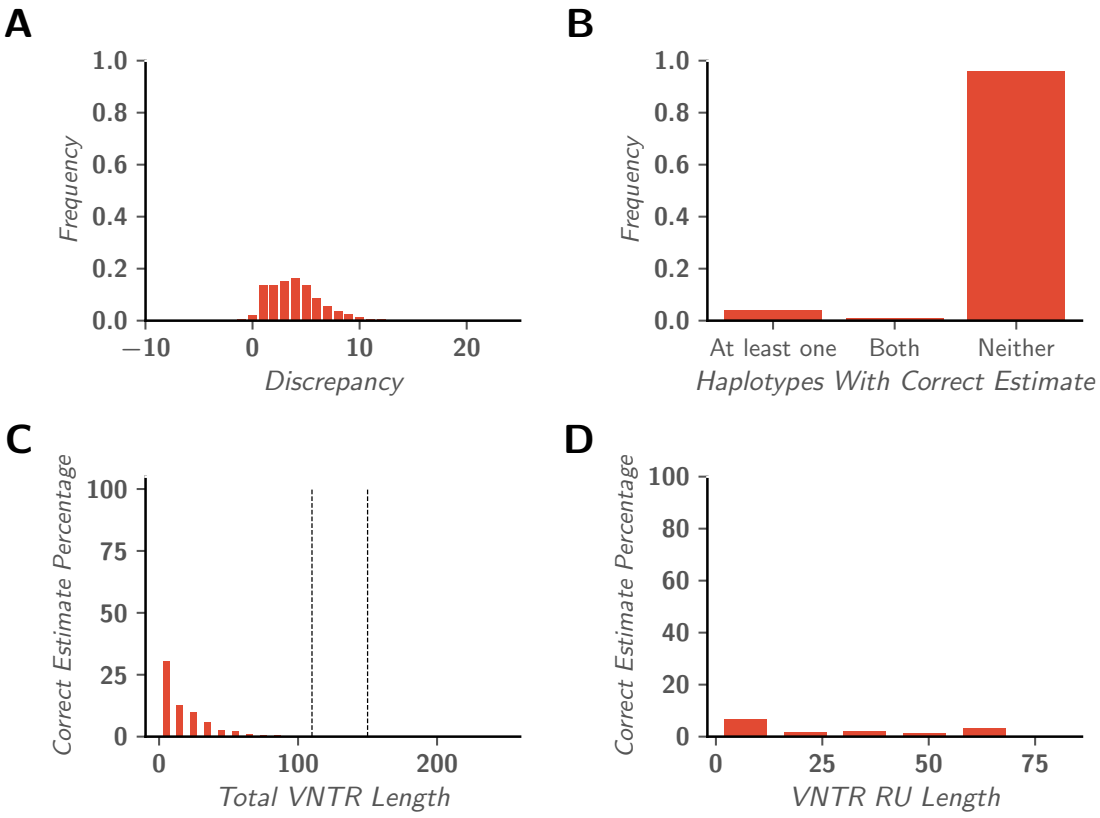


Figure A.7: **Expansion Hunter's performance on VNTR genotyping using Illumina reads.** Expansion Hunter's performance on IlluminaSim dataset.

Table A.4: **VNTR genotyping results on simulated data.** For two cases, (*MAOA* 1/1 and *CSTB* 1/1) Expansion Hunter doesn't find any RU count.

		RU Count Discrepancy		
		PacBio Dataset	Illumina Dataset	
VNTR	Simulated Genotype	adVNTR	Expansion Hunter	adVNTR
MAOA	1/1	0/0	-/-	0/0
MAOA	1/2	0/0	0/-1	0/0
MAOA	1/3	0/0	0/-2	0/0
MAOA	1/4	0/0	0/-3	0/0
MAOA	1/5	0/0	0/-4	0/0
MAOA	2/2	0/0	-1/-1	0/0
MAOA	2/3	0/0	0/-1	0/0
MAOA	2/4	0/0	-1/-3	0/0
MAOA	2/5	0/0	-1/-4	0/0
MAOA	3/3	0/0	-2/-2	0/0
MAOA	3/4	0/0	-2/-3	0/0
MAOA	3/5	0/0	-2/-4	0/0
MAOA	4/4	0/0	-3/-3	0/0
MAOA	4/5	0/0	-3/-4	0/0
MAOA	5/5	0/0	-4/-4	-1/-1
GP1BA	1/1	0/0	0/0	0/0
GP1BA	1/2	0/0	0/0	0/0
GP1BA	1/3	0/0	0/-1	0/0
GP1BA	1/4	0/0	1/-2	0/-1
GP1BA	2/2	0/0	0/0	0/0
GP1BA	2/3	0/0	0/-1	0/0
GP1BA	2/4	0/0	0/-2	0/-1
GP1BA	3/3	0/0	-1/-1	0/0
GP1BA	3/4	0/0	-1/-2	0/0
GP1BA	4/4	0/0	-2/-2	-1/0
CSTB	1/1	0/0	-/-	0/0
CSTB	1/2	0/0	1/0	0/0
CSTB	1/3	0/0	2/0	0/0
CSTB	1/4	0/0	3/0	0/0
CSTB	1/5	0/0	4/0	0/0
CSTB	1/6	0/0	4/-1	0/0
CSTB	1/7	0/0	3/-3	0/0
CSTB	1/8	0/0	4/-3	0/0
CSTB	1/9	0/0	3/-5	0/0
CSTB	1/10	0/0	4/-5	0/0
CSTB	1/11	0/0	4/-6	0/0
CSTB	1/12	0/0	4/-7	0/0
CSTB	1/13	0/0	4/-8	0/0
CSTB	1/14	0/0	3/-10	0/-1
CSTB	2/2	0/0	0/0	0/0
CSTB	2/3	0/0	1/0	0/0
CSTB	2/4	0/0	1/-1	0/0
CSTB	2/6	0/0	3/-1	0/0
CSTB	2/8	0/0	3/-3	0/0

Table A.5: **Genotyping comparison on AJ trio using Illumina reads from GIAB.** Table shows the genotype found by adVNTR and ExpansionHunter in disease causing VNTRs that are shorter than Illumina reads. -/- denotes ExpansionHunter has not found any genotype for the VNTR. It worths mentioning the genotypes found by adVNTR for *MAOA* are not inconsistent as this VNTR is located on ChrX and the son has haploid RU counts inherited from mother.

	Estimated Genotype					
		adVNTR			ExpansionHunter	
VNTR	AJ Child	AJ Mother	AJ Father	AJ Child	AJ Mother	AJ Father
DRD4	4/5	4/5	4/4	-/-	-/-	-/-
ZFHX3	4/4	4/4	4/4	3/3	-/-	3/3
GP1BA	2/5	2/3	3/4	2/2	1/1	2/2
SLC6A4	13/13	11/13	13/13	-/-	-/-	-/-
MMP9	3/3	3/3	3/3	-/-	-/-	-/-
CSTB	2/2	2/2	2/2	3/3	2/2	1/1
MAOA	5/5	4/5	4/4	-/-	-/-	-/-

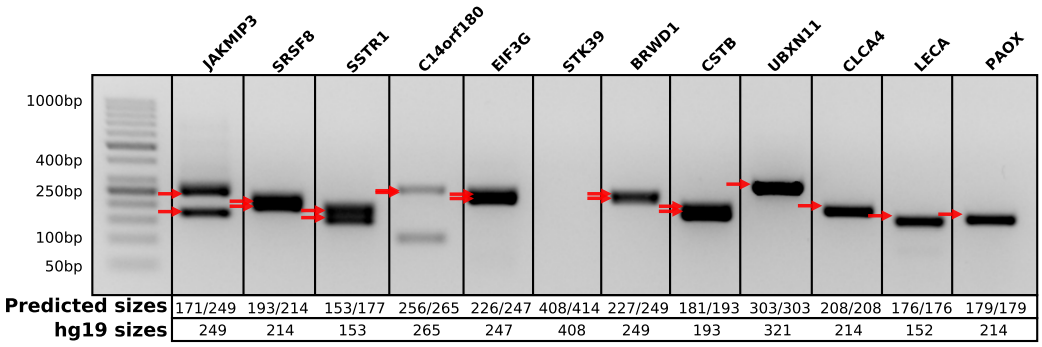


Figure A.8: **Validation of adVNTR genotyping on short VNTRs.** In experiment for *C14orf180* the primers were repeated in another region of genome which resulted in having extra band. Even with zero copy of VNTR patterns, the distance of primers around VNTR is 238bp which means the extra band ($\sim$100bp) is resulted from another region of genome. Also, PCR amplification failed for *STK39* and no band is visible. Results of all other 10 experiments are consistent with adVNTR's estimates.

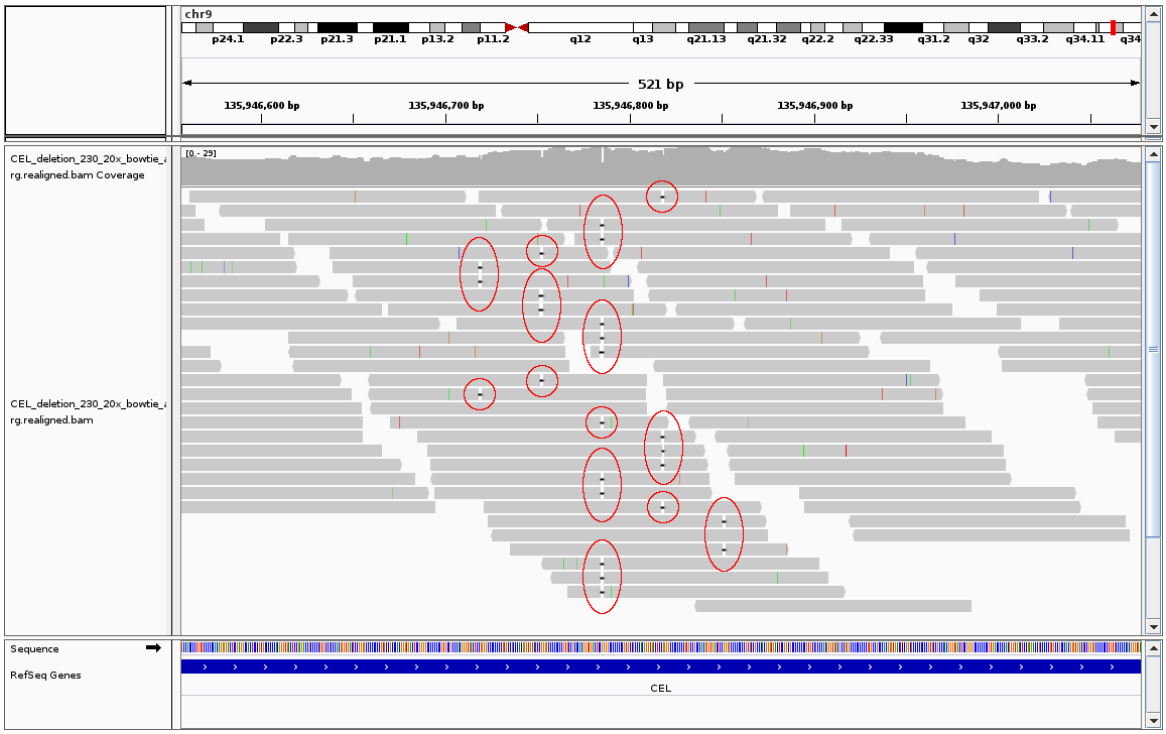


Figure A.9: **Alignment in VNTR region with the presence of a frameshift.** Alignment of a simulated data after running GATK IndelRealigner, when there is a deletion. With a sequencing mean of 30×, 25 reads contain the deletion but even after running realigner, deletions are mapped to five different repeating units.

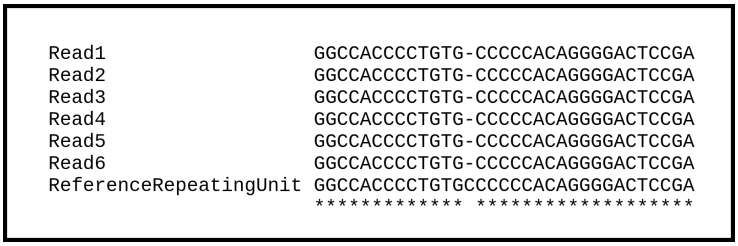


Figure A.10: **Frameshift in *CEL* gene.** Multiple alignment of sequenced reads and reference repeating unit shows a deletion in diabetes patient genome. Due to low PCR amplification in GC rich VNTR region (84.8%), the coverage of VNTR region is 14× and 6 reads support the deletion.

Table A.6: Comparison of indel detection with SAMtools and GATK

		# of Samples	# of samples that frameshift has been identified		
			Samtools	Our Method	GATK
10X	Insertions	20	0	20	0
	Deletions	20	0	20	0
20X	Insertions	20	0	20	0
	Deletions	20	0	20	0
30X	Insertions	20	0	20	0
	Deletions	20	0	20	0
40X	Insertions	20	0	20	0
	Deletions	20	0	20	0

Appendix B

Supplemetary Material for Chapter 3

B.1 Supplementary Figures

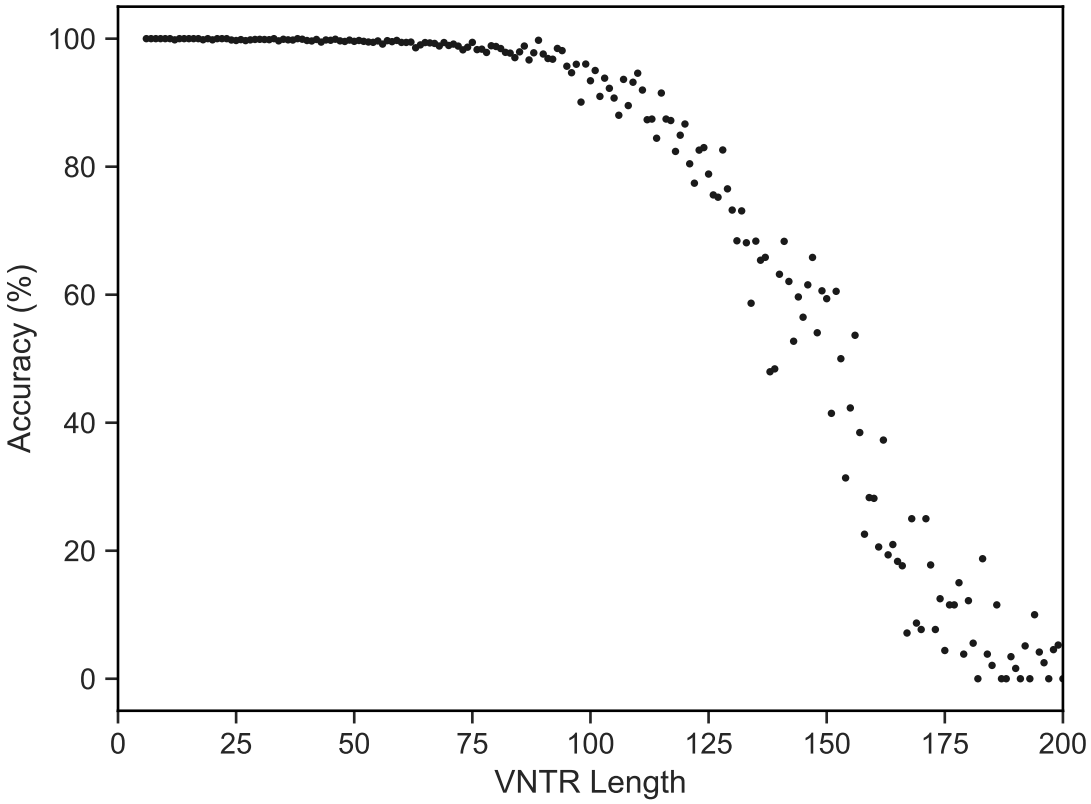


Figure B.1: **Distribution of genotyping accuracy of adVNTR-NN stratified by VNTR length on simulated VNTRs.** Heterozygous alleles were simulated by inserting or deleting repeating units in one reference allele to transform its RU count c to c+x, where c is the hg19 reference count, and $x \in [-3,3]$. Source data are provided as a Source Data file.

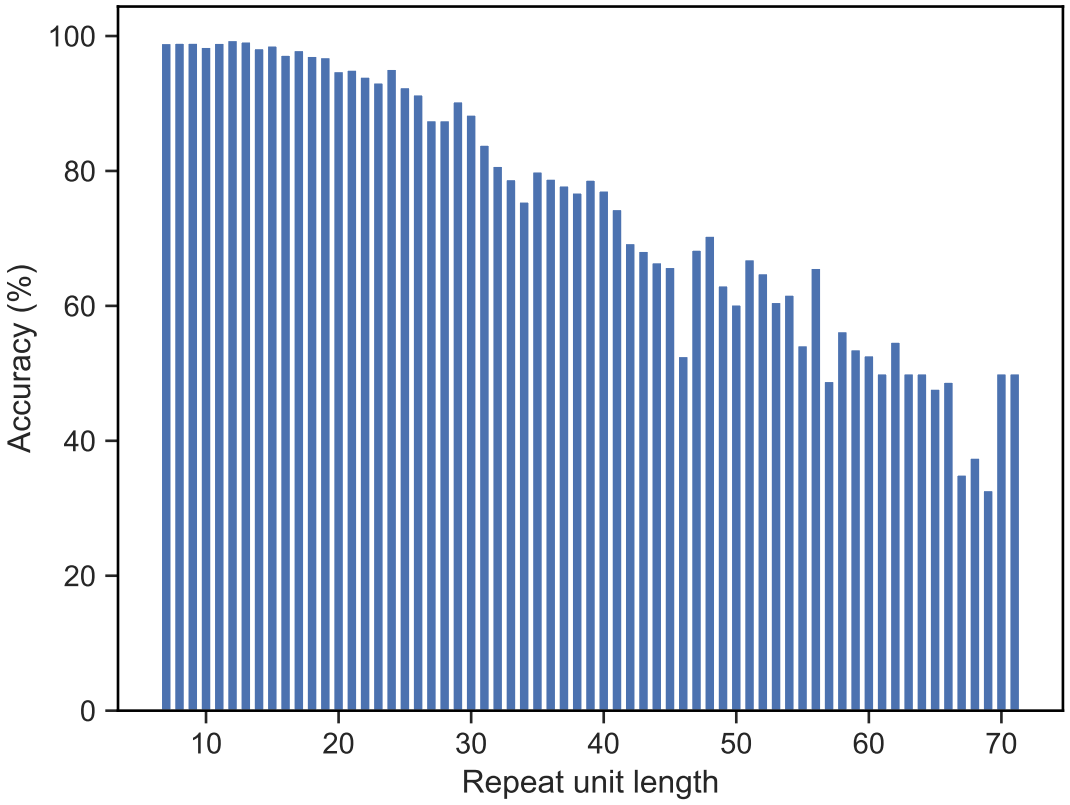


Figure B.2: **Distribution of genotyping accuracy of adVNTR-NN stratified by repeat length for simulated heterozygous reads.** Heterozygous alleles were simulated by inserting or deleting repeating units in one reference allele to transform its RU count c to c+x, where c is the hg19 reference count, and $x \in [-3, 3]$.

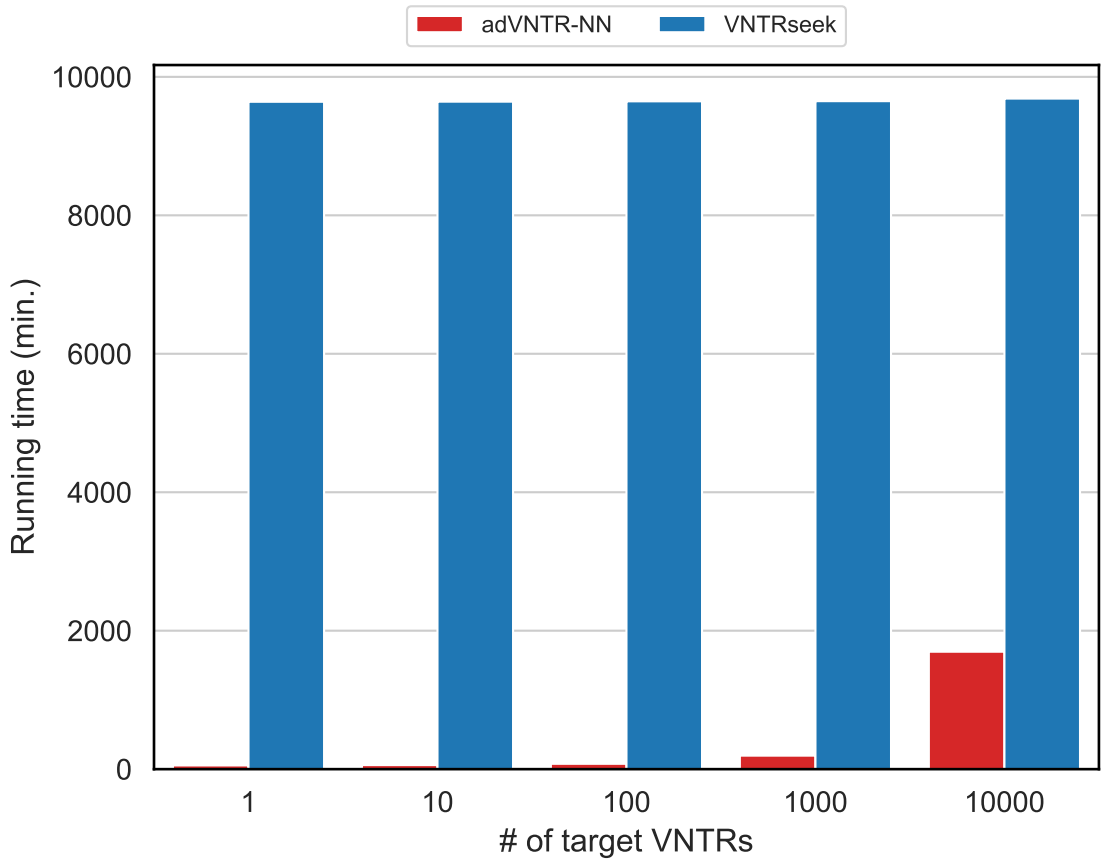


Figure B.3: **adVNTR-NN and VNTRseek running time comparison.** Running time comparison on 1, 10, 100, 1,000, and 10,000 VNTR loci of one individual (NA24149) with 1.16×10^9 reads.

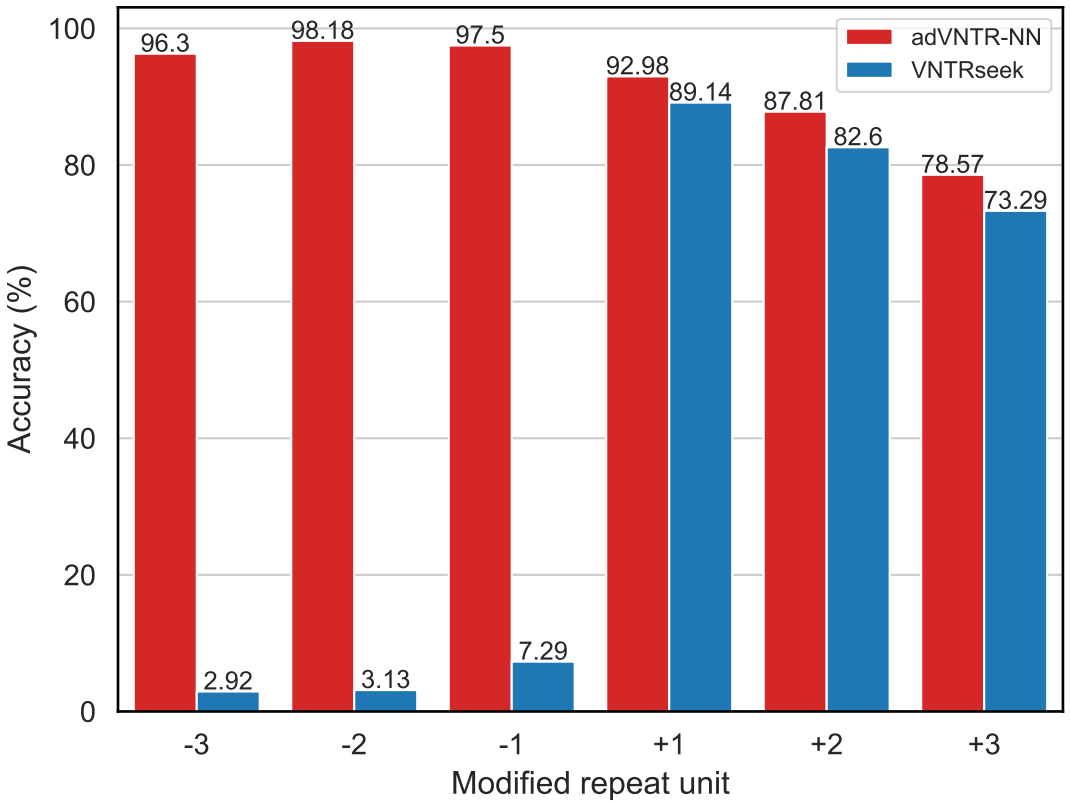


Figure B.4: **adVNTR-NN and VNTRseek genotyping accuracy on simulated heterozygous reads.** The genotyping accuracy for each scenario is defined by the the number of VNTR loci correctly genotyped correctly divided by the number of VNTR loci. Six different heterozygous VNTR scenarios were tested; specifically, c/c-3, c/c-2, c/c-1, c/c+1, c/c+2, c/c+3, where c is the hg19 reference count. The number of VNTR loci modified for contraction scenarios were 9,638 (c-1), 5,078 (c-2), and 2,084 (c-3), with the reductions happening due to a requirement of at least 1 repeating copy for each VNTR allele. All expansion scenarios had 9,638 VNTRs.

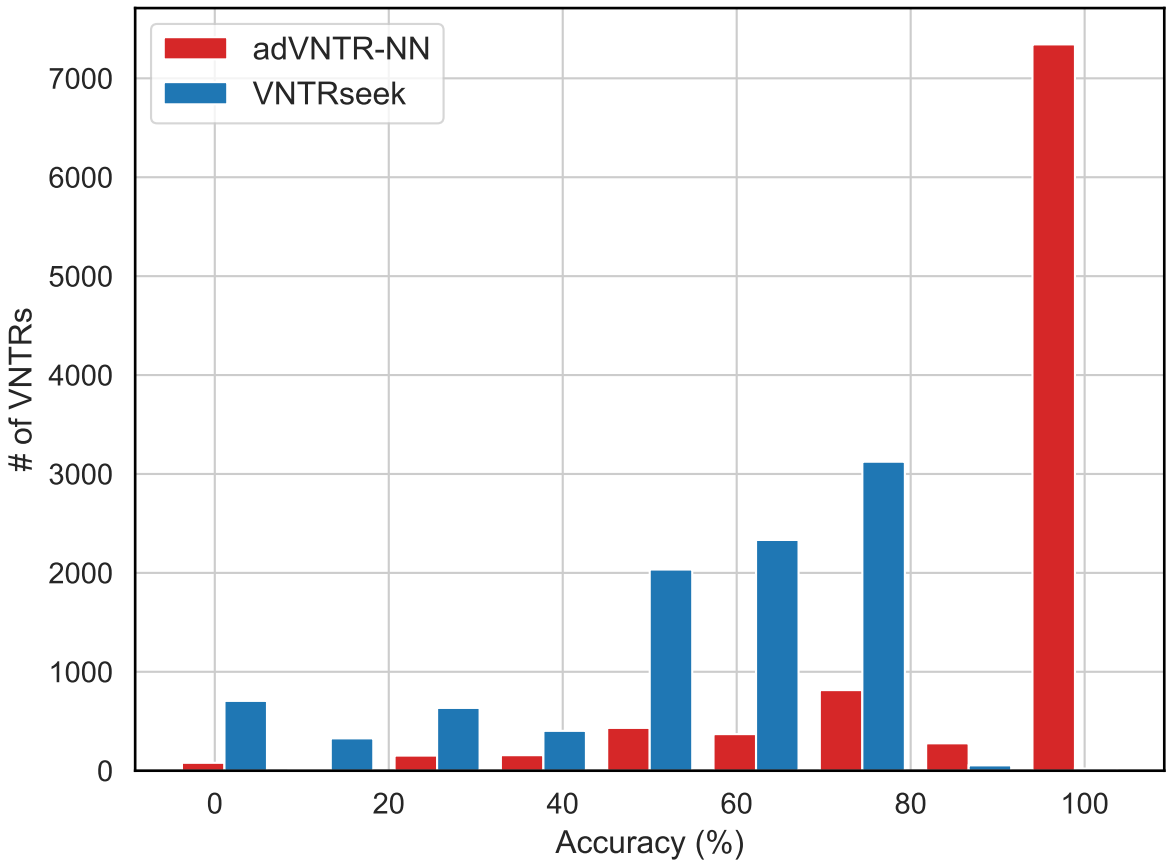


Figure B.5: **Distribution of genotyping accuracy of adVNTR-NN and VNTRseek on simulated heterozygous reads.** The genotyping accuracy for each VNTR is defined by the number of loci genotyped correctly divided by the number of loci. Six different heterozygous VNTR scenarios were tested; specifically, c/c-3, c/c-2, c/c-1, c/c+1, c/c+2, c/c+3, where c is the hg19 reference count. The number of VNTR loci modified for contraction scenarios were 9,638 (c-1), 5,078 (c-2), and 2,084 (c-3), with the reductions happening due to a requirement of at least 1 repeating copy for each VNTR allele. All expansion scenarios had 9,638 VNTRs. adVNTR-NN had 100% accuracy in 7,343 (76%) of 9,638 VNTRs.

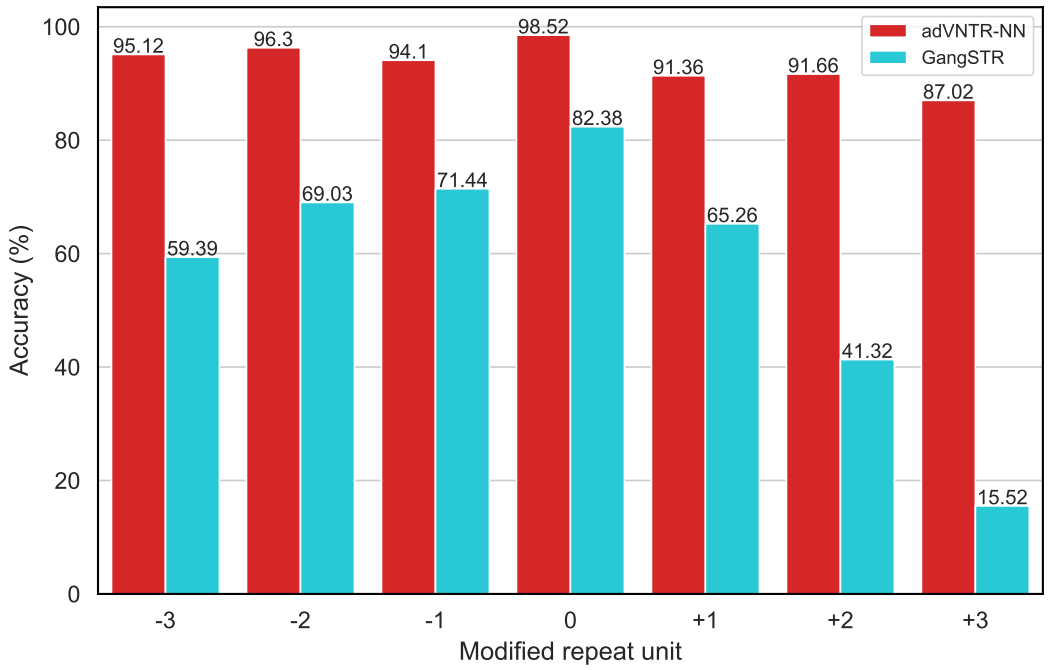


Figure B.6: **Comparison of adVNTR-NN versus GangSTR accuracy on simulated heterozygous reads for short RU lengths (≤ 20).** Seven scenarios were tested; specifically, c/c-3, c/c-2, c/c, c/c-1, c/c+1, c/c+2, c/c+3, where c is the hg38 reference count. The genotyping accuracy for each scenario is defined by the number of VNTR loci genotyped correctly divided by the number of VNTR loci. The number of VNTR loci modified for contraction scenarios were 6,508 (c-1), 4,763 (c-2), and 2,805 (c-3), with the reductions happening due to a requirement of at least 1 repeating copy for each VNTR allele. All expansion scenarios and the homozygous case had 6,508 VNTRs.

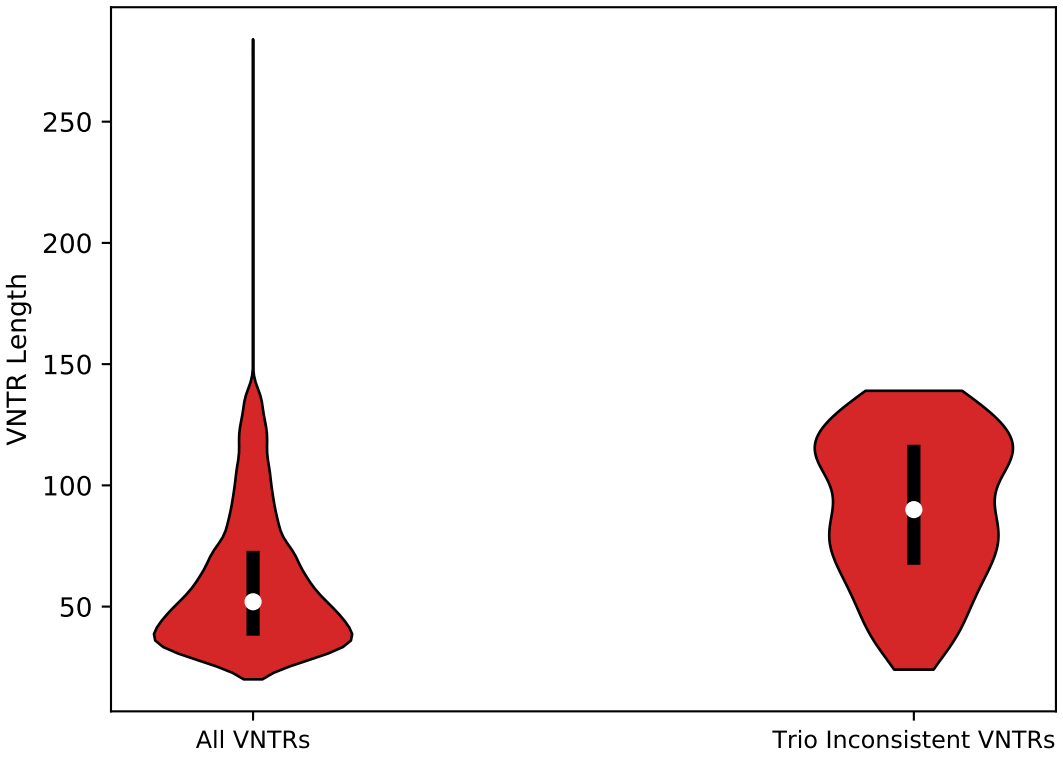


Figure B.7: **Length distribution of VNTRs.** The violin plots show the distribution of VNTR lengths in 537 Trios from the 1000 Genomes Project (n=10,264 VNTRs for each trio). White dots show median values and boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). VNTRs that showed consistency with Mendelian inheritance patterns had a median length of 52bp, while inconsistent calls have a median of 90bp.

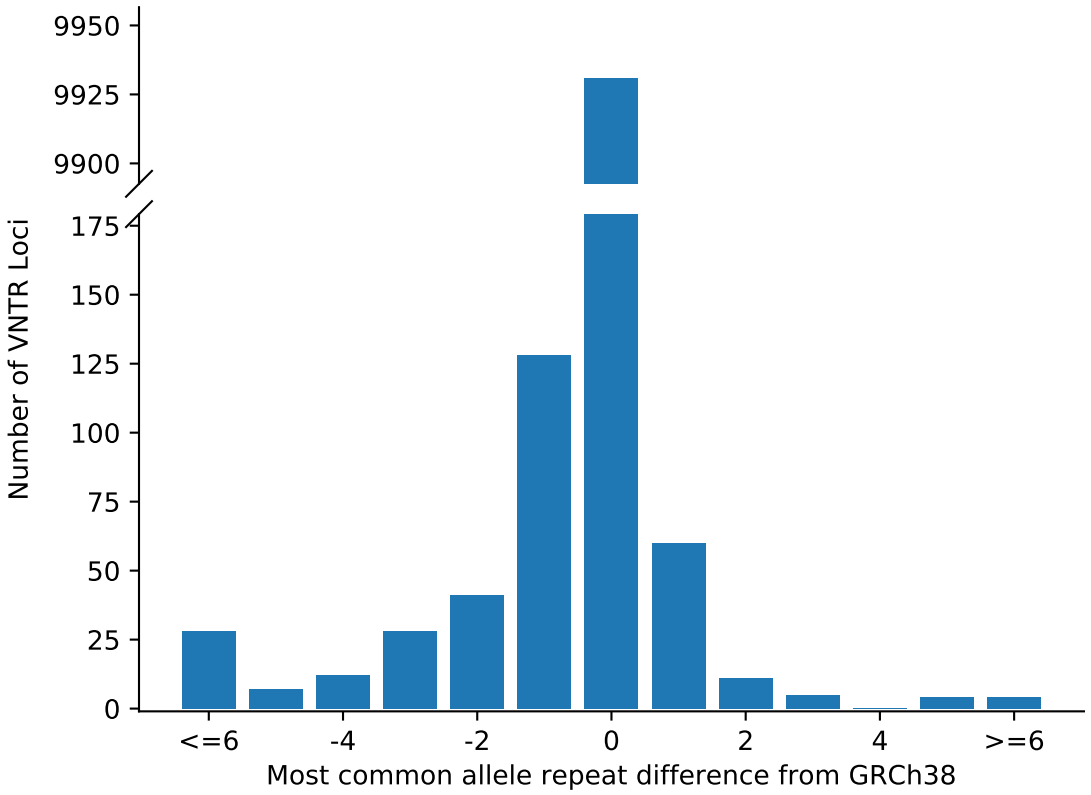


Figure B.8: **Genotype difference in VNTR loci between donors and GRCh38.** For each VNTR, the difference between the most common allele in the GTEx cohort and the GRCh38 reference repeat count was recorded. The plot shows the distribution of the differences.

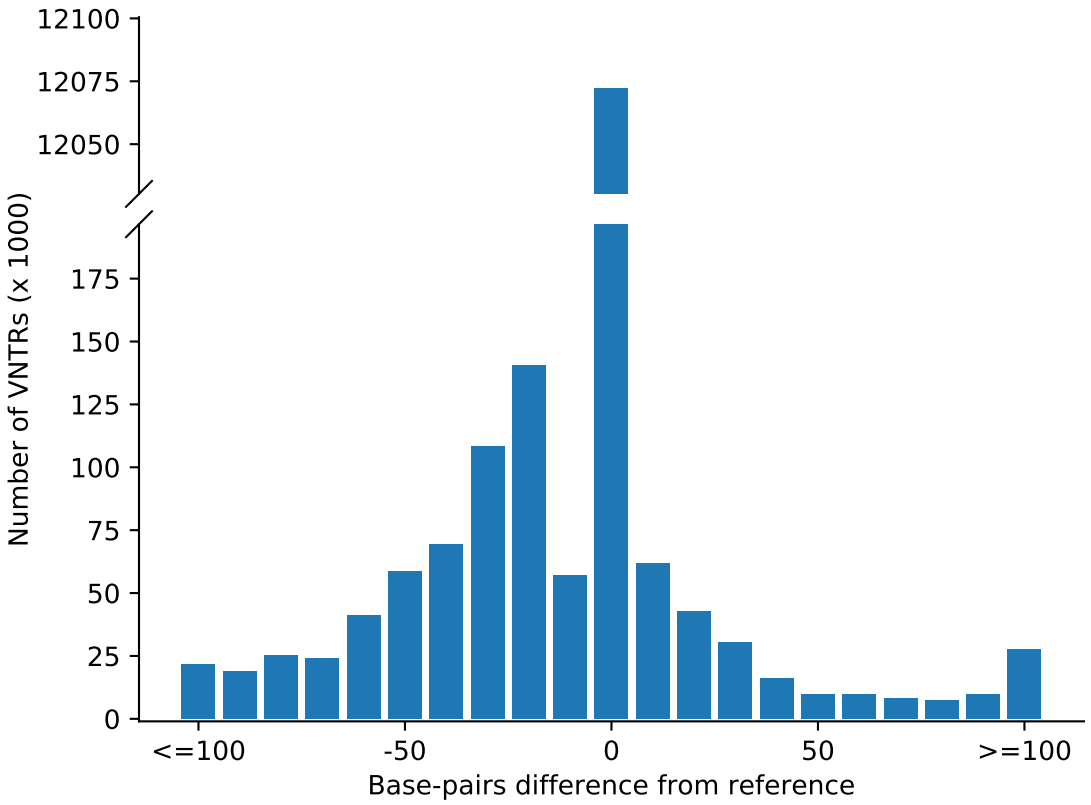


Figure B.9: **Base pair difference in VNTR loci between donors and GRCh38.** For each VNTR and each individual allele in a GTEx donor, the difference in length from the GRCh38 reference VNTR length was recorded. The plot shows a distribution of differences.

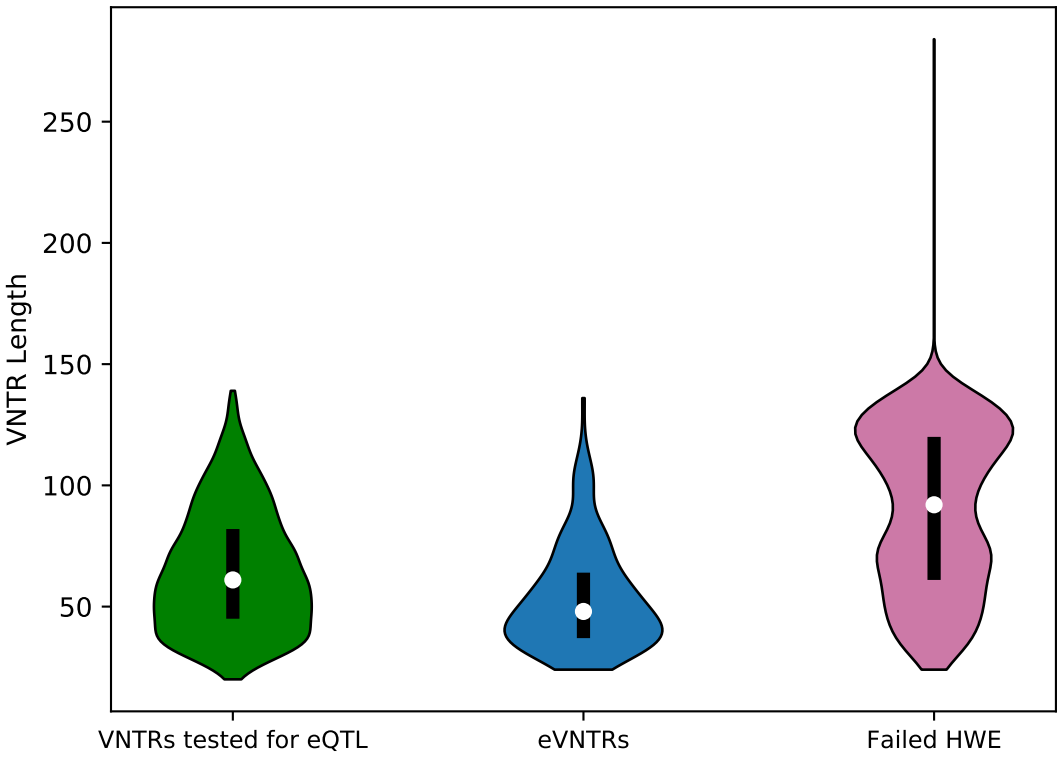


Figure B.10: **Length distribution of VNTRs in the GTEx cohort (n=4,280 VNTRs).** White dots show median values and boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). The length distribution for all VNTRs that passed filters had a median of 61, slightly larger than eVNTRs (median length: 48bp). In contrast, the VNTRs that failed the HWE test had a length distribution (median: 92bp), which was similar to VNTRs that showed inconsistent Mendelian inheritance patterns in trios from the 1000 Genome data (median: 90 bp; Fig. B.7).

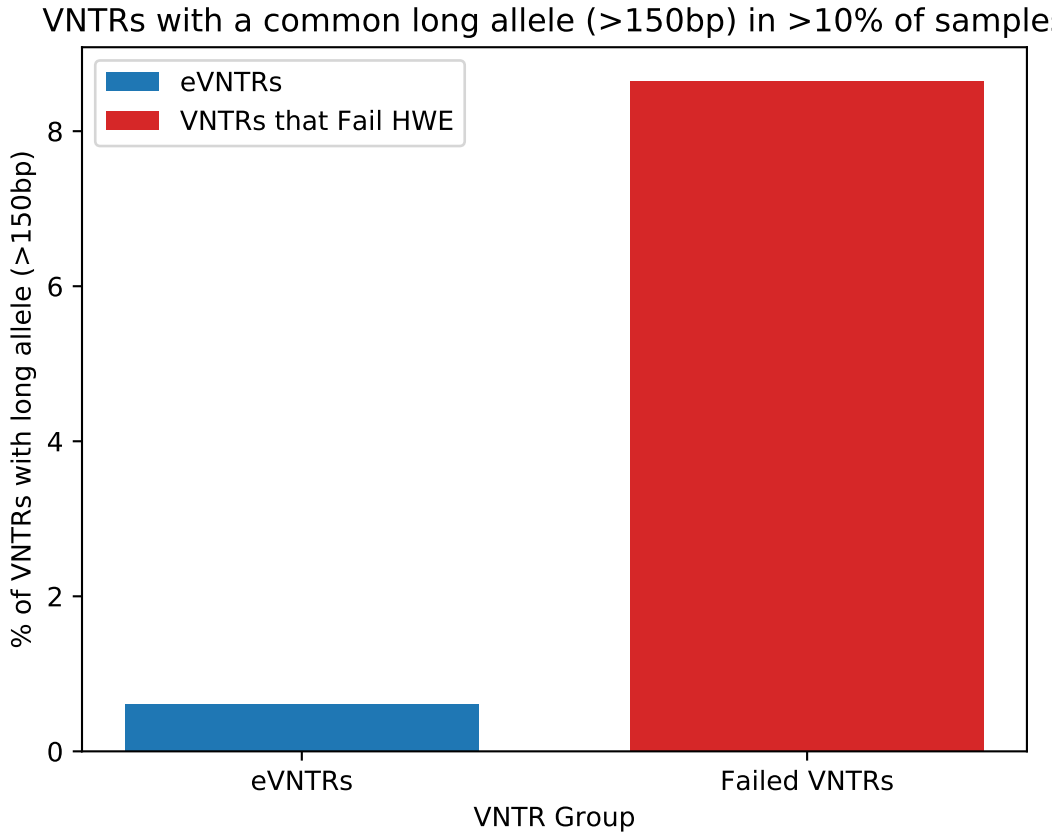


Figure B.11: **Fraction of VNTRs with a common long allele.** Only one (0.6%) of the eVNTRs had a common allele (present in $>$ 10% of samples) that was longer than a read length, while (125) 8.06% of VNTRs that fail HWE test had a common long allele. Shorter alleles are genotyped more accurately.

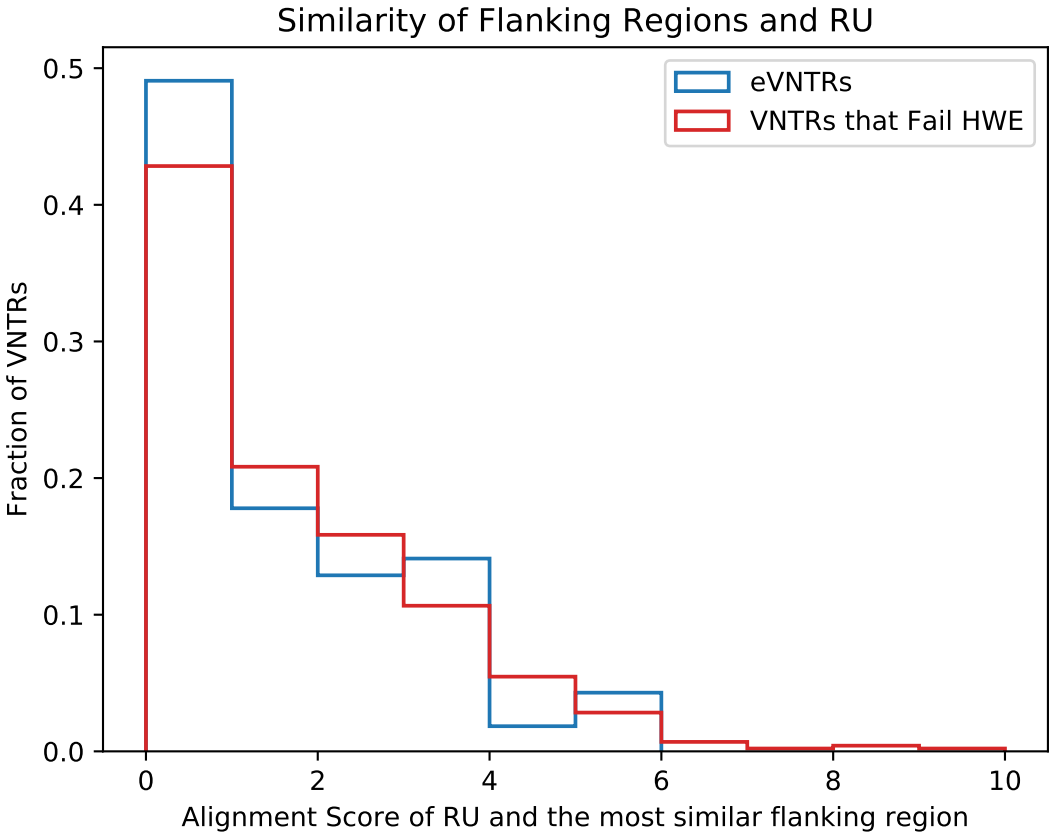


Figure B.12: **Similarity of VNTR repeating pattern with flanking region.** Distribution of the number of bases in the repeat unit of VNTRs that identically match a flanking region. Higher similarity of repeating unit and flanking region makes it more challenging to distinguish the VNTR boundary and make an accurate genotype call.

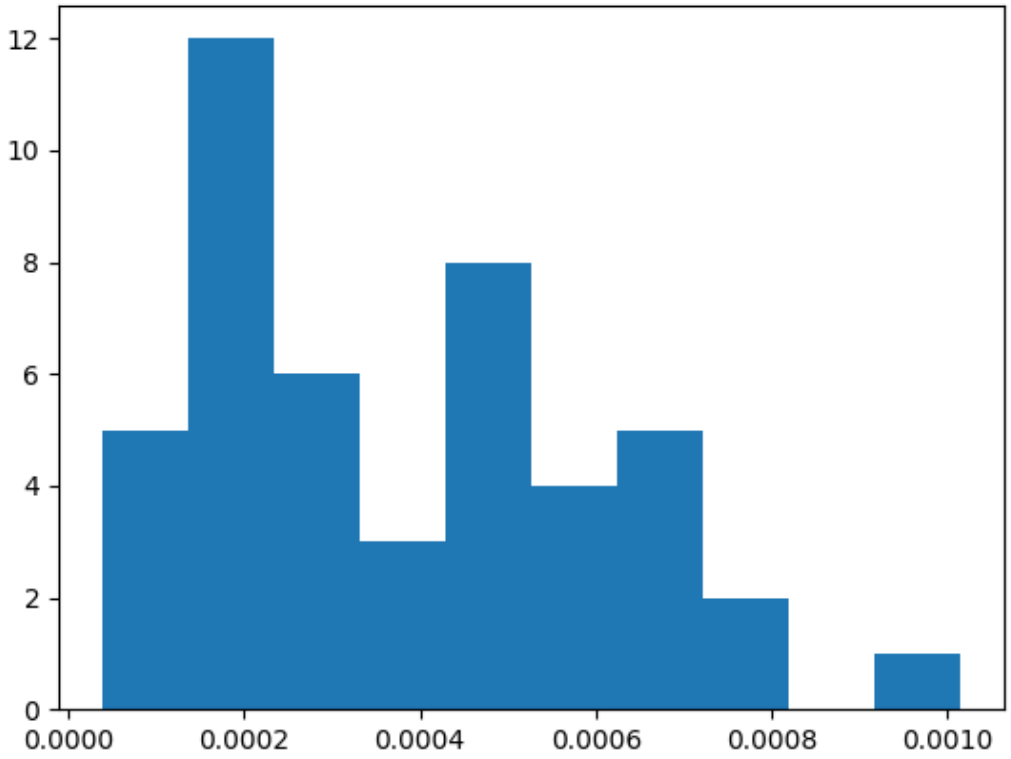


Figure B.13: **Distribution of significance thresholds for association test.** Significance thresholds for each of the 46 tissues.

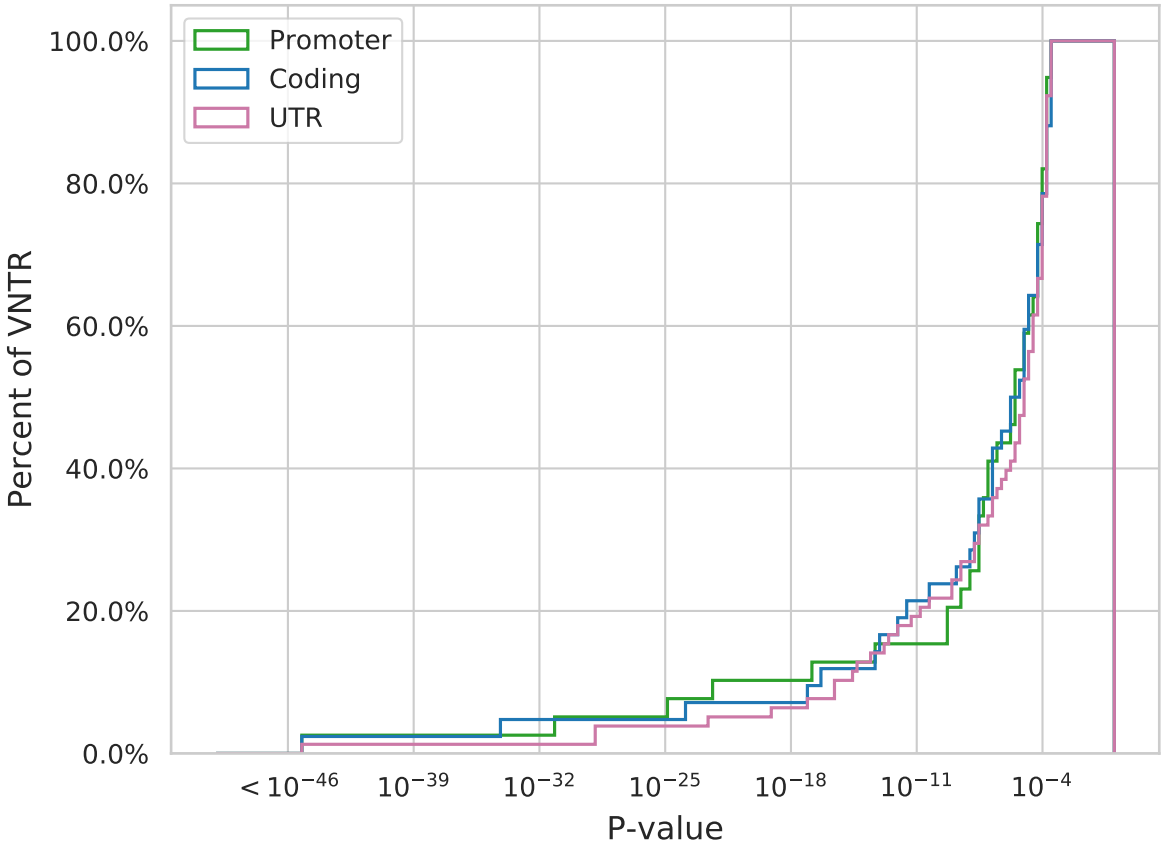


Figure B.14: **Cumulative distribution of eVNTR p-values for different classes.** The plots suggest that the relative location of a genic VNTR does not significantly change the strength of association with gene expression.

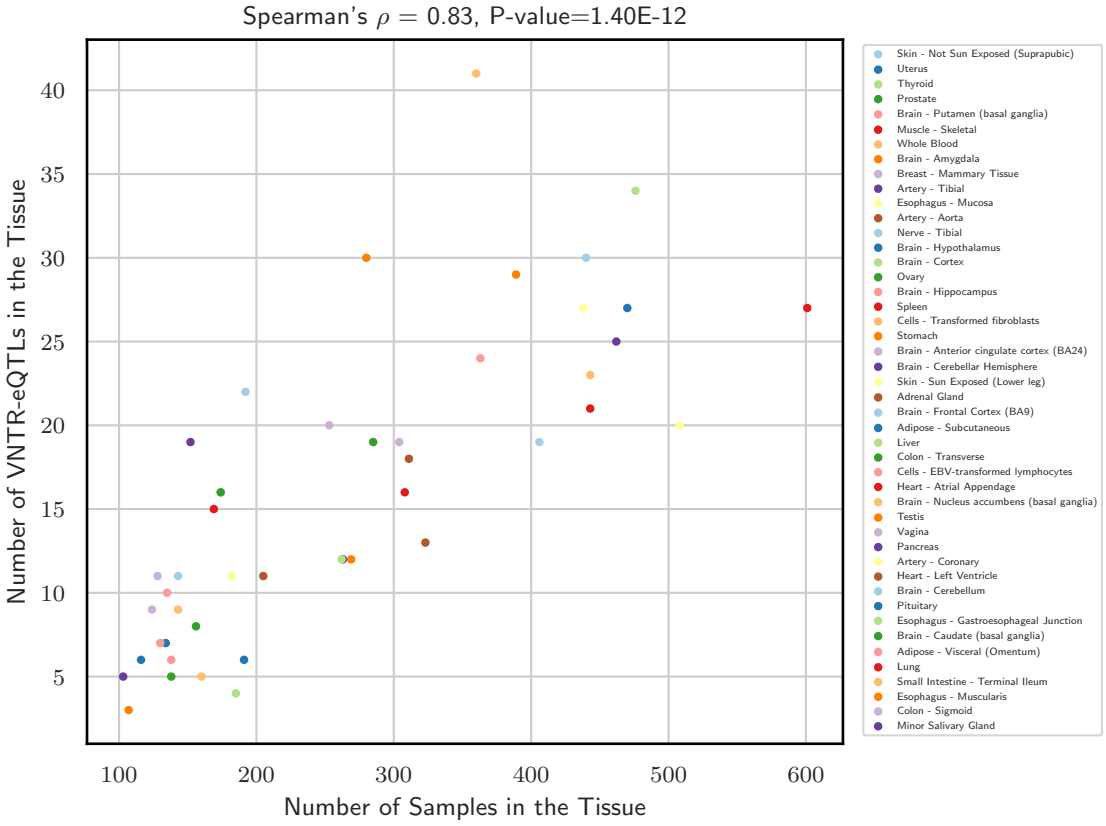


Figure B.15: **Correlation between number of eVNTRs and sample-size.** Overall, we see a strong correlation between the number of samples and eVNTRs. Testis and transformed-fibroblasts had relatively higher number of eVNTRs, while fewer eVNTRs were identified in Whole blood and Skeletal muscle, relative to the sample size.

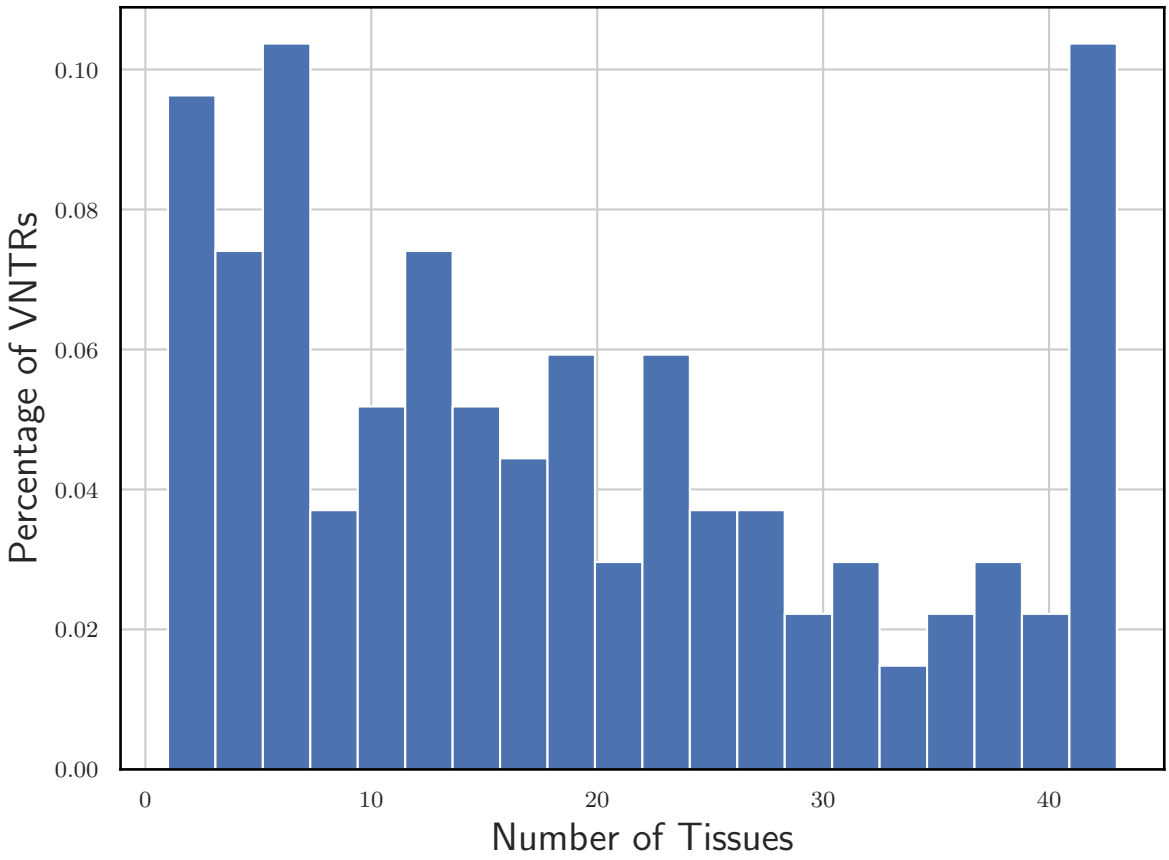


Figure B.16: **Tissue sharing of eVNTRs.** The fraction of eVNTRs that are active in a specific number of tissues as determined by mash. 38% of eVNTRs were significant in at least half (23) of all tissues.

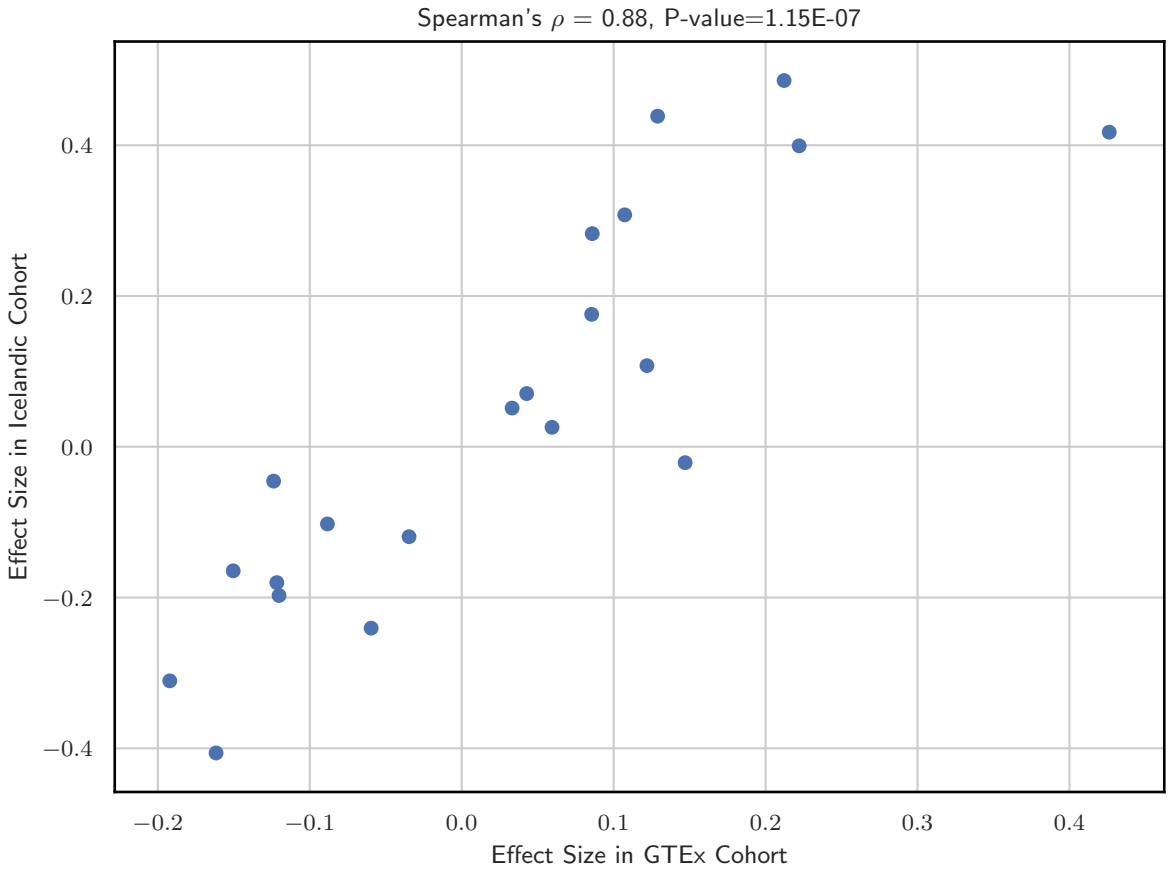


Figure B.17: **Reproducibility of effect sizes in Icelandic Cohort.** The scatter plot compares the effect sizes of each eVNTR association in the GTEx cohort (x-axis) against the Icelandic cohort (y-axis) for blood tissue. The Spearman's correlation coefficient was 0.88 (p-val: $1.15E-07$).

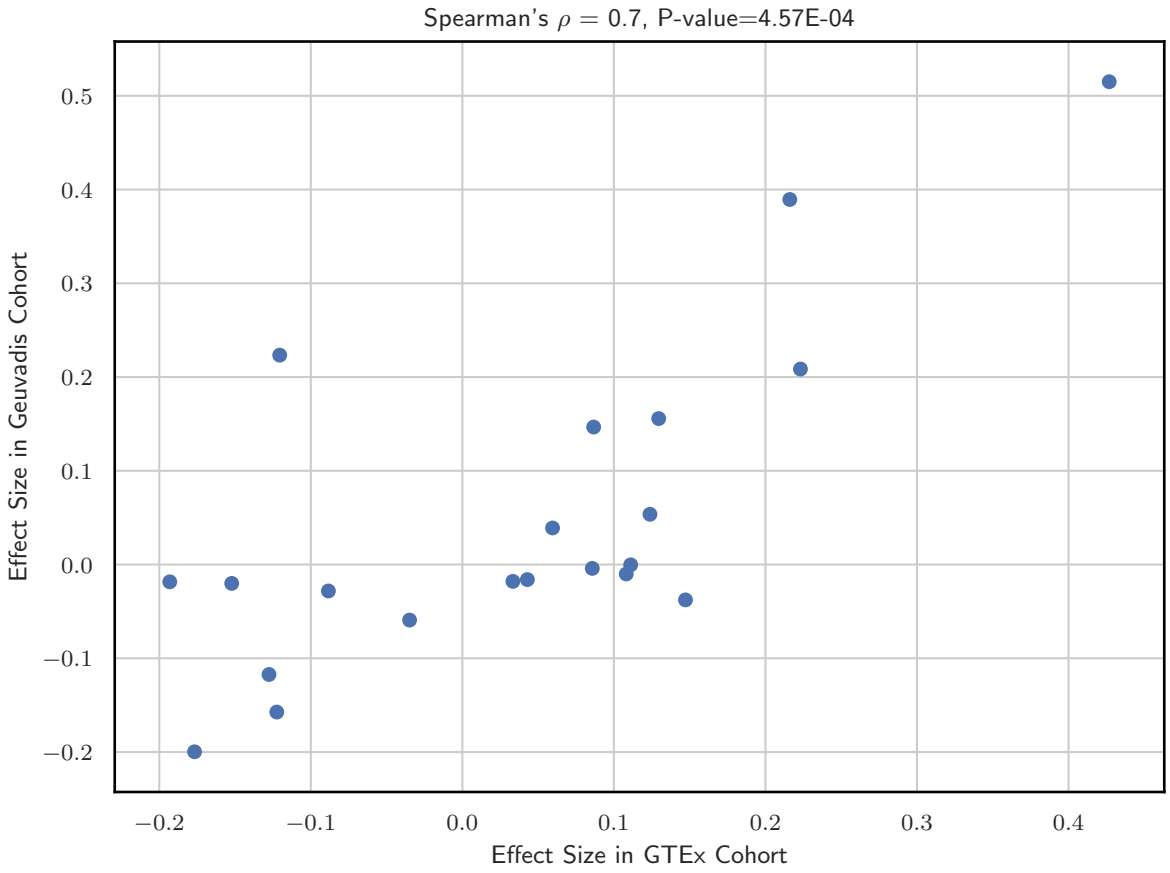


Figure B.18: **Reproducibility of effect sizes in the Geuvadis Cohort.** The scatter plot compares the effect sizes of each eVNTR associations in GTEx cohort (x-axis) against the Geuvadis cohort (y-axis) for blood tissue. The Spearman's correlation coefficient was 0.7 (p-val: $4.57E-04$).

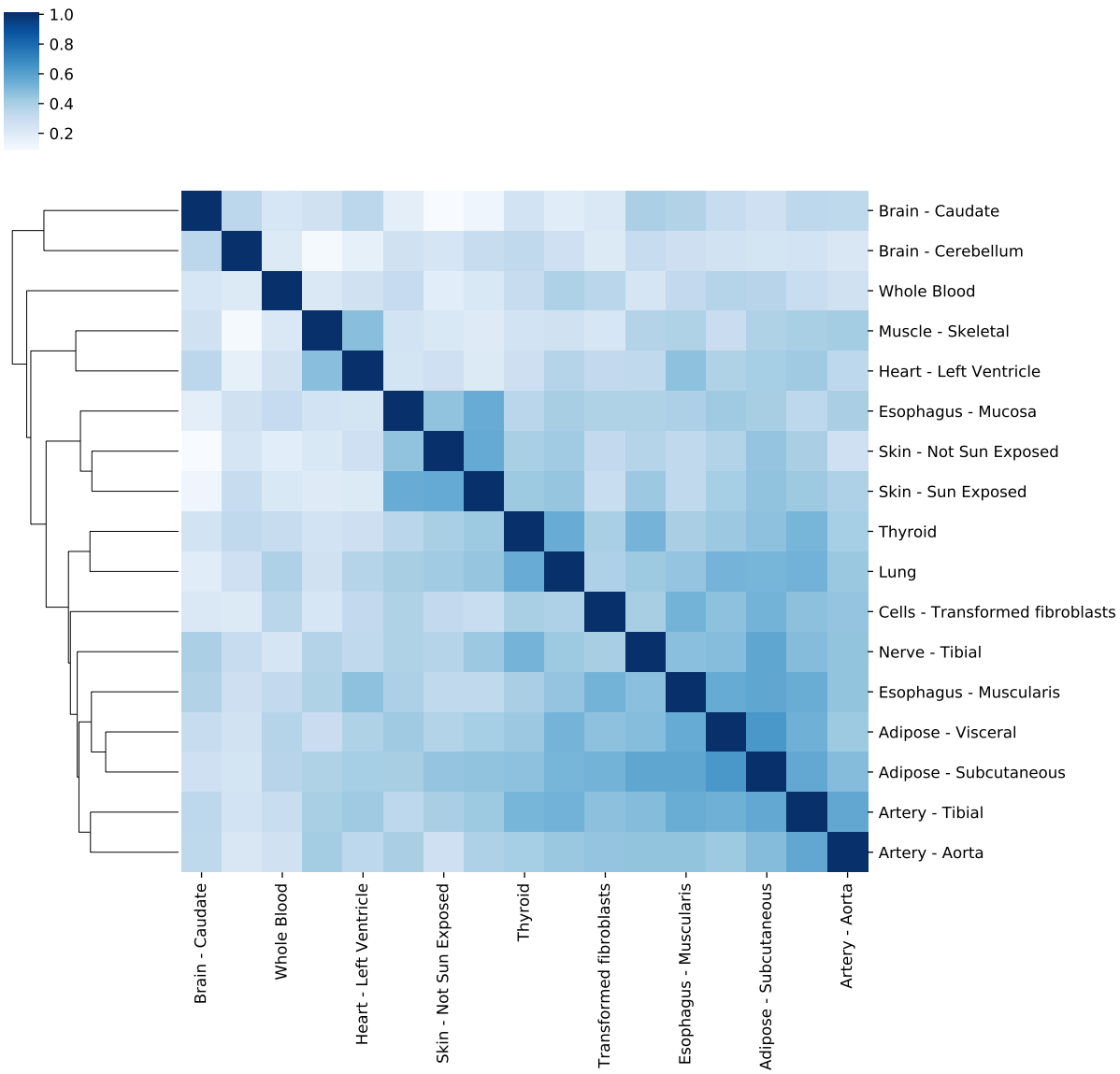


Figure B.19: **Spearman correlation of eVNTRs effect sizes for pairs of tissues.** The correlation was restricted to the subset of 17 tissue types used in Fotsing[39], Fig. 1d for comparison.

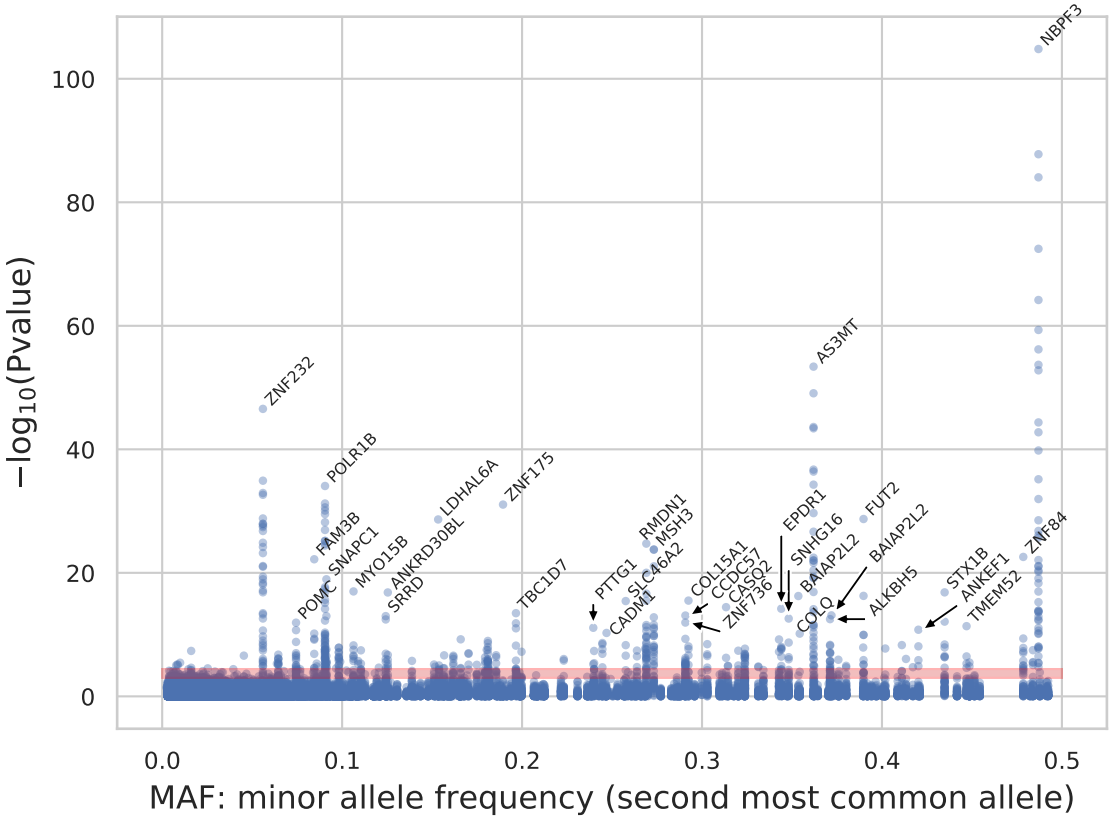


Figure B.20: **Significance of VNTR association with gene expression plotted against Minor Allele Frequency.** The shaded region represents tissue specific false discovery rate cut-offs. Note that all significant tests for a single VNTR appear in a single column.

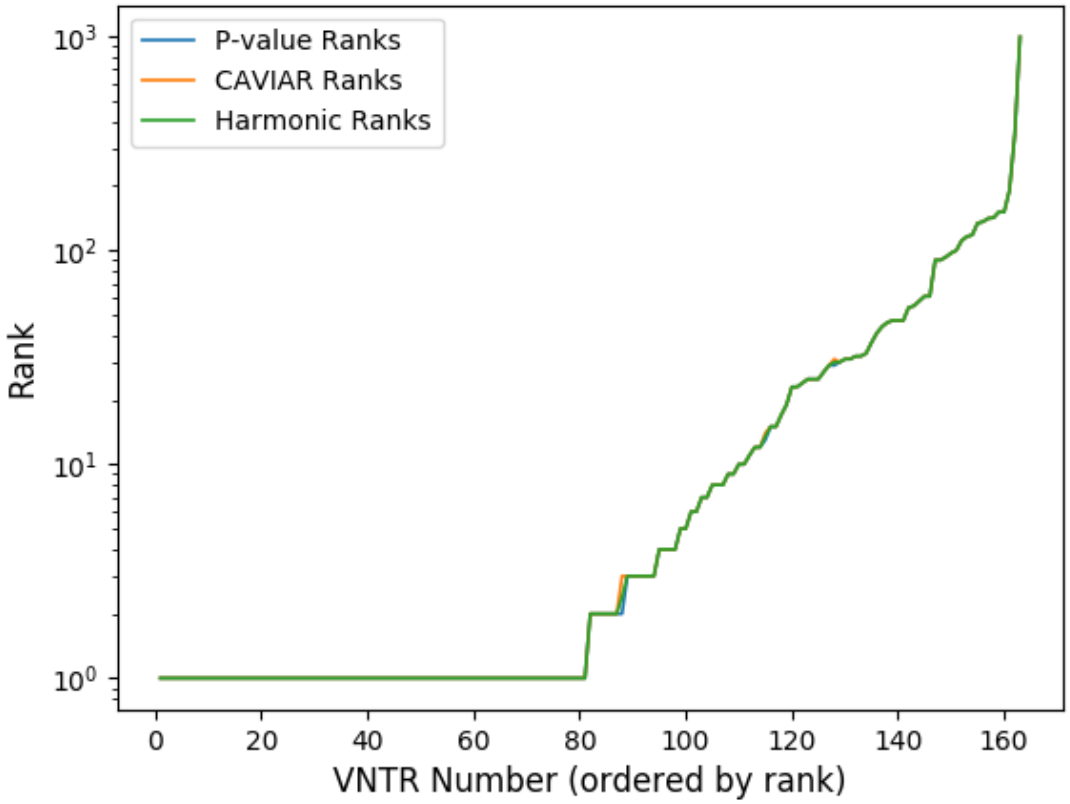


Figure B.21: Causality rank of eVNTRs measured using strength of association (blue), CAVIAR (red), and mean harmonic rank (green). The P-value and CAVIAR based ranks coincide.

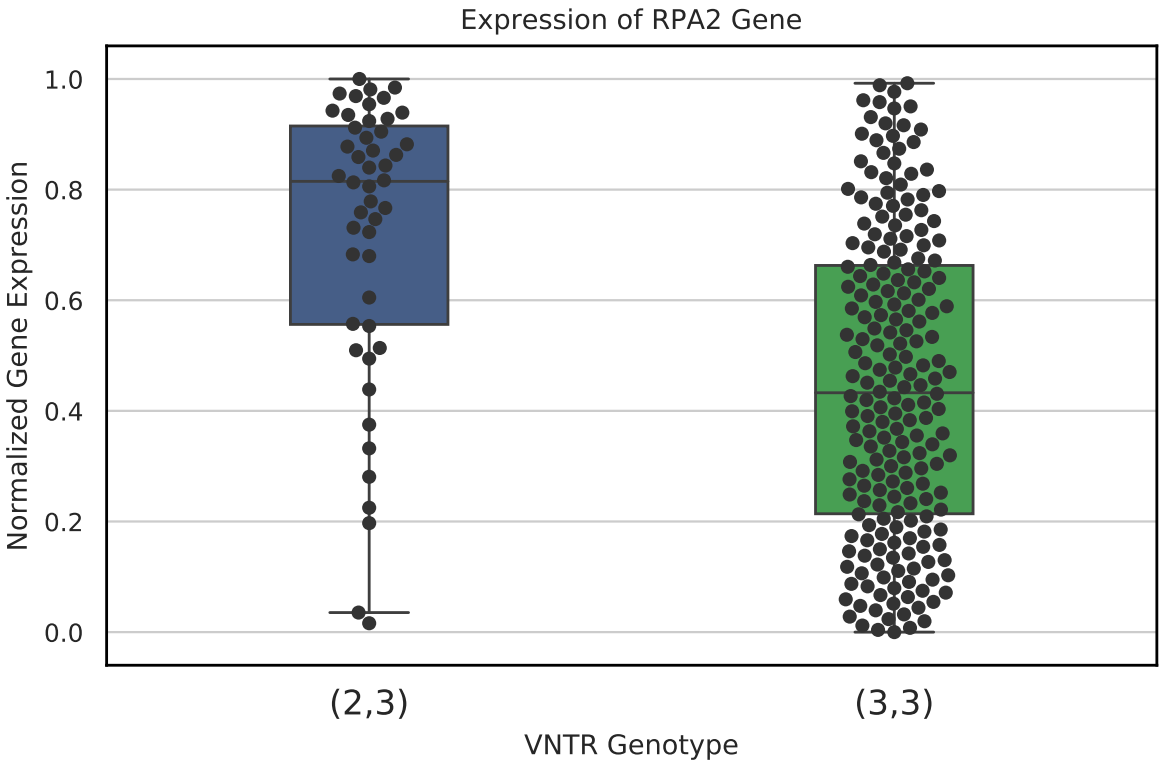


Figure B.22: **Association of RPA2 VNTR genotype with gene expression level. n=254 samples, P-value** 3.79×10^{-25}. Increase RPA2 expression has been associated with worse survival outcomes in colon cancer[45]. Only two samples had a homozygous (2, 2) genotype. Their normalized expression levels were 0.85 and 0.99, which is consistent with the trend. However, they were excluded from analysis due to the small counts. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to $Q1 - 1.5 \times IQR$ (bottom) and $Q3 + 1.5 \times IQR$ (top), where IQR is the interquartile range ($Q3 - Q1$).

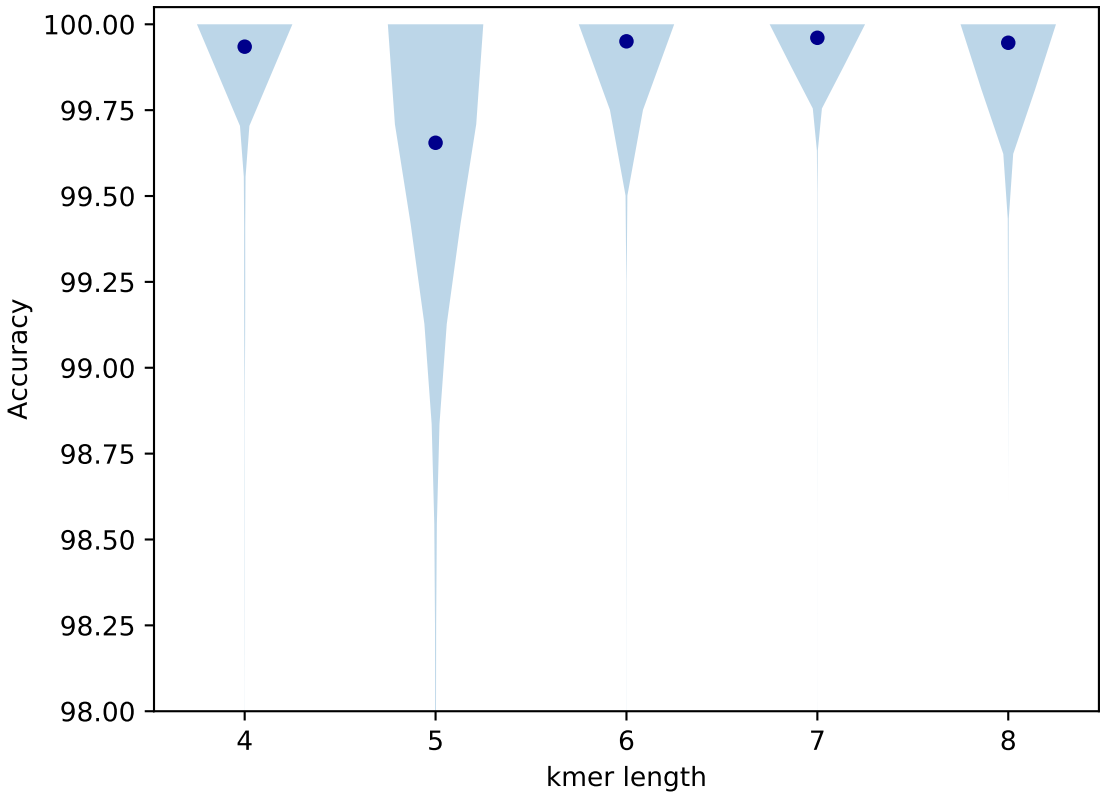


Figure B.23: **Effect of kmer length on accuracy.** Performance of the neural network model on validation set for different k-mer lengths. k=6 was used for all test runs as it had the highest mean accuracy of 99.95%.

B.2 Supplementary Tables

Table B.1: **eVNTRs with known phenotypes.** Top 20 VNTRs loci that were previously linked to a phenotype.

Locus	RU Length	Effect Size	Gene	Tissues	Phenotype	P-Value	CAVIAR Rank	Causality Probability
chr17:5116259-5116362	34	0.46	ZNF232 (UTR)	(40)	Alzheimer's Disease	2.82×10^{-47}	1	1
					Breast Cancer			
chr1:21440112-21440147	6	0.58	NBPF3 (UTR)	Cells - EBV	Neuroblastoma	1.54×10^{-105}	1	0.63
chr2:112542424-112542500	25	-0.63	RPA2 (Exon)	Brain, Colon	Colon cancer	8.67×10^{-35}	1	0.48
				Pancreas	Bladder cancer			
chr10:102869497-102869605	36	0.39	AS3MT (Exon)	4 Brain parts	Schizophrenia	4.10×10^{-54}	1	0.52
chr2:25161573-25161616	9	0.57	POMC (Exon)	(15)	Obesity	1.18×10^{-12}	1	0.99
chr19:12577507-12577551	22	-0.20	ZNF490 (UTR)	Nerve, Breast	Colorectal cancer	3.66×10^{-09}	1	0.95
				Whole Blood	germline mutation			
chr6:13328502-13328532	6	0.17	TBC1D7 (UTR)	Brain (1)		3.39×10^{-14}	1	0.48
chr2:24084339-24084414	25	-0.27	TP53I3/PIG3 (UTR)	Nerve	Lung adenocarcinoma	5.84×10^{-10}	1	0.20
				Brain (BA24)				
chr5:80654880-80654954	9	0.09	MSH3 (Exon)	(4)	Myotonic dystrophy	1.61×10^{-24}	1	0.99

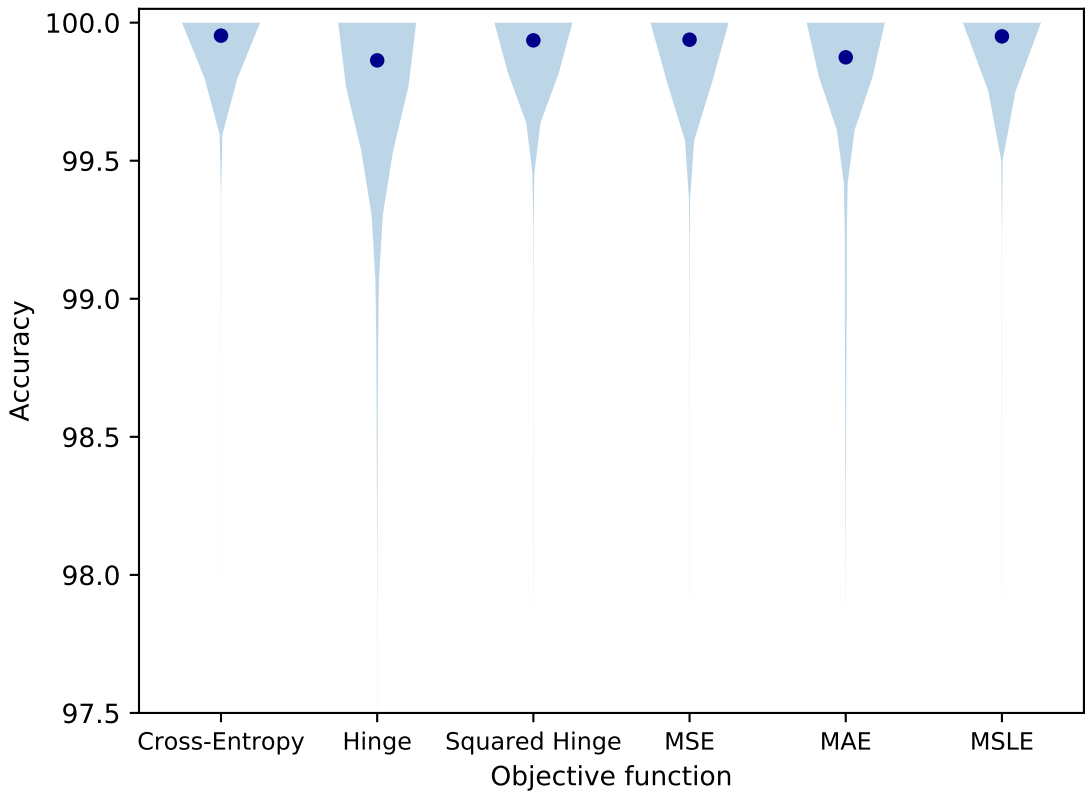


Figure B.24: **Effect of loss function on accuracy.** Performance of the neural network model on validation set for different loss functions. The mean of each distribution is shown by a blue dot. Binary cross-entropy was used as the loss function for all tests.

Table B.2: **Comparison of hexamer eVNTRs using differing methods.** Each row describes a 6-bp variant identified either as eVNTR here or as an eSTR in Fotsing[39]. Fotsing et al. identified eSTRs with false discovery rate (FDR) $< 10\%$ in contrast to our cut-off of 5% FDR. Therefore, the nominal p-value of each association is presented for easier comparison. Despite differing methodologies and versions of GTEx, the loci are at least nominally significant ($p < 0.05$) in both tests.

	Locus	adVNTR P-value	HipSTR P-value[39]	Replication	
				eVNTR	eSTR
1	chr6:13328502-13328532	3.39×10^{-14}	7.89×10^{-13}	Y	Y
2	chr5:160421950-160422000	7.86×10^{-12}	8.59×10^{-11}	Y	Y
3	chr22:37510301-37510338	2.29×10^{-9}	5.39×10^{-7}	Y	Y
4	chr17:63703959-63703989	4.14×10^{-7}	9.17×10^{-4}	Y	N
5	chr10:70132751-70132793	2.25×10^{-5}	3.74×10^{-3}	Y	N
6	chr11:6390700-6390749	5.80×10^{-5}	2.22×10^{-5}	Y	Y
7	chr20:35652812-35652848	1.12×10^{-4}	3.99×10^{-2}	Y	N
8	chr6:148343091-148343168	1.80×10^{-4}	3.78×10^{-3}	Y	N
9	chr22:37805258-37805313	2.12×10^{-4}	4.82×10^{-10}	Y	Y
10	chr3:51993818-51993872	2.79×10^{-4}	9.70×10^{-9}	Y	Y
11	chr16:71922603-71922638	7.05×10^{-4}	4.91×10^{-5}	N	Y
12	chr13:113119135-113119222	5.27×10^{-3}	9.45×10^{-8}	N	Y
13	chr16:67416367-67416422	9.93×10^{-3}	7.44×10^{-5}	N	Y
14	chr1:151511435-151511510	1.56×10^{-2}	6.37×10^{-5}	N	Y
15	chr11:8964363-8964423	1.56×10^{-2}	1.55×10^{-6}	N	Y

Bibliography

[1] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[2] A.V. Aho and M.J. Corasick. Efficient String Matching: An Aid to Bibliographic Search. *Communications of the ACM*, 1975.

[3] Cornelis A Albers, Gerton Lunter, Daniel G MacArthur, Gilean McVean, Willem H Ouwehand, and Richard Durbin. Dindel: accurate indel calls from short-read data. *Genome research*, 21(6):961–973, 2011.

[4] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[5] Antonis C Antoniou, Karoline B Kuchenbaecker, Penny Soucy, Jonathan Beesley, Xiaoqing Chen, Lesley McGuffog, Andrew Lee, Daniel Barrowdale, Sue Healey, Olga M Sinilnikova, et al. Common variants at 12p11, 12q24, 9p21, 9q31. 2 and in ZNF365 are associated with breast cancer risk for BRCA1 and/or BRCA2 mutation carriers. *Breast Cancer Research*, 14(1):1–18, 2012.

[6] Antonis C. Antoniou, Xianshu Wang, Zachary S. Fredericksen, Lesley McGuffog, Robert Tarrell, Olga M. Sinilnikova, Sue Healey, Jonathan Morrison, Christiana Kartsonaki, Timothy Lesnick, Maya Ghoussaini, Daniel Barrowdale, Susan Peock, Margaret Cook, Clare Oliver, Debra Frost, Diana Eccles, D. Gareth Evans, Ros Eeles, Louise Izatt, Carol Chu, Fiona Douglas, Joan Paterson, Dominique Stoppa-Lyonnet, Claude Houdayer, Sylvie Mazoyer, Sophie Giraud, Christine Lasset, Audrey Remenieras, Olivier Caron, Agnès Hardouin, Pascaline Berthet, Frans B.L. Hogervorst, Matti A. Rookus, Agnes Jager, Ans Van Den Ouweland, Nicoline Hoogerbrugge, Rob B. Van Der Luijt, Hanne Meijers-Heijboer, Encarna B. G'mez García, Peter Devilee, Maaike P.G. Vreeswijk, Jan Lubinski, Anna Jakubowska, Jacek Gronwald, Tomasz Huzarski, Tomasz Byrski, Bohdan G'rski, Cezary Cybulski, Amanda B. Spurdle, Helene Holland, David E. Goldgar, Esther M. John, John L. Hopper, Melissa Southey, Saundra S. Buys, Mary B. Daly, Mary Beth Terry, Rita K. Schmutzler, Barbara Wappenschmidt, Christoph Engel, Alfons Meindl, Sabine Preisler-Adams, Norbert Arnold, Dieter Niederacher, Christian Sutter, Susan M. Domchek,

Katherine L. Nathanson, Timothy Rebbeck, Joanne L. Blum, Marion Piedmonte, Gustavo C. Rodriguez, Katie Wakeley, John F. Boggess, Jack Basil, Stephanie V. Blank, Eitan Friedman, Bella Kaufman, Yael Laitman, Roni Milgrom, Irene L. Andrulis, Gord Glendon, Hilmi Ozcelik, Tomas Kirchhoff, Joseph Vijai, Mia M. Gaudet, David Altshuler, Candace Guiducci, Niklas Loman, Katja Harbst, Johanna Rantala, Hans Ehrencrona, Anne Marie Gerdes, Mads Thomassen, Lone Sunde, Paolo Peterlongo, Siranoush Manoukian, Bernardo Bonanni, Alessandra Viel, Paolo Radice, Trinidad Caldes, Miguel De La Hoya, Christian F. Singer, Anneliese Fink-Retter, Mark H. Greene, Phuong L. Mai, Jennifer T. Loud, Lucia Guidugli, Noralane M. Lindor, Thomas V.O. Hansen, Finn C. Nielsen, Ignacio Blanco, Conxi Lazaro, Judy Garber, Susan J. Ramus, Simon A. Gayther, Catherine Phelan, Stephen Narod, Csilla I. Szabo, Javier Benitez, Ana Osorio, Heli Nevanlinna, Tuomas Heikkinen, Maria A. Caligo, Mary S. Beattie, Ute Hamann, Andrew K. Godwin, Marco Montagna, Cinzia Casella, Susan L. Neuhausen, Beth Y. Karlan, Nadine Tung, Amanda E. Toland, Jeffrey Weitzel, Olofunmilayo Olopade, Jacques Simard, Penny Soucy, Wendy S. Rubinstein, Adalgeir Arason, Gad Rennert, Nicholas G. Martin, Grant W. Montgomery, Jenny Chang-Claude, Dieter Flesch-Janys, Hiltrud Brauch, Gianluca Severi, Laura Baglietto, Angela Cox, Simon S. Cross, Penelope Miron, Sue M. Gerty, William Tapper, Drakoulis Yannoukakos, George Fountzilas, Peter A. Fasching, Matthias W. Beckmann, Isabel Dos Santos Silva, Julian Peto, Diether Lambrechts, Robert Paridaens, Thomas Rüdiger, Asta Försti, Robert Winqvist, Katri Pylkäs, Robert B. Diasio, Adam M. Lee, Jeanette Eckel-Passow, Celine Vachon, Fiona Blows, Kristy Driver, Alison Dunning, Paul P.D. Pharoah, Kenneth Offit, V. Shane Pankratz, Hakon Hakonarson, Georgia Chenevix-Trench, Douglas F. Easton, and Fergus J. Couch. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor–negative breast cancer in the general population. *Nature genetics*, 42(10):885–892, 2010.

[7] Kin Fai Au, Jason G Underwood, Lawrence Lee, and Wing Hung Wong. Improving PacBio long read accuracy by short read alignment. *PloS one*, 7(10):e46679, 2012.

[8] Mehrdad Bakhtiari and Jonghun Park. Variable Number Tandem Repeats mediate the expression of proximal genes, 2020.

[9] Mehrdad Bakhtiari, Jonghun Park, Yuan-Chun Ding, Sharona Shleizer-Burko, Susan L Neuhausen, Bjarni V Halldórsson, Kári Stefánsson, Melissa Gymrek, and Vineet Bafna. Variable number tandem repeats mediate the expression of proximal genes. *Nature Communications*, 12(1):1–12, 2021.

[10] Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, and Vineet Bafna. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Research*, 28(11):1709–1719, 2018.

[11] A. Battle, C. D. Brown, B. E. Engelhardt, S. B. Montgomery, F. Aguet, K. G. Ardlie, B. B. Cummings, E. T. Gelfand, G. Getz, K. Hadley, R. E. Handsaker, K. H. Huang, S. Kashin, K. J. Karczewski, M. Lek, X. Li, D. G. MacArthur, J. L. Nedzel, D. T. Nguyen, M. S. Noble, A. V. Segr?, C. A. Trowbridge, T. Tukiainen, N. S. Abell, B. Balliu, R. Barshir,

O. Basha, A. Battle, G. K. Bogu, A. Brown, C. D. Brown, S. E. Castel, L. S. Chen, C. Chiang, D. F. Conrad, N. J. Cox, F. N. Damani, J. R. Davis, O. Delaneau, E. T. Dermitzakis, B. E. Engelhardt, E. Eskin, P. G. Ferreira, L. Fr?sard, E. R. Gamazon, D. Garrido-Mart?n, A. D. H. Gewirtz, G. Gliner, M. J. Gloudemans, R. Guigo, I. M. Hall, B. Han, Y. He, F. Hormozdiari, C. Howald, H. Kyung Im, B. Jo, E. Yong Kang, Y. Kim, S. Kim-Hellmuth, T. Lappalainen, G. Li, X. Li, B. Liu, S. Mangul, M. I. McCarthy, I. C. McDowell, P. Mohammadi, J. Monlong, S. B. Montgomery, M. Mu?oz-Aguirre, A. W. Ndungu, D. L. Nicolae, A. B. Nobel, M. Oliva, H. Ongen, J. J. Palowitch, N. Panousis, P. Papasaikas, Y. Park, P. Parsana, A. J. Payne, C. B. Peterson, J. Quan, F. Reverter, C. Sabatti, A. Saha, M. Sammeth, A. J. Scott, A. A. Shabalin, R. Sodaei, M. Stephens, B. E. Stranger, B. J. Strober, J. H. Sul, E. K. Tsang, S. Urbut, M. van de Bunt, G. Wang, X. Wen, F. A. Wright, H. S. Xi, E. Yeger-Lotem, Z. Zappala, J. B. Zaugg, Y. H. Zhou, J. M. Akey, D. Bates, J. Chan, L. S. Chen, M. Claussnitzer, K. Demanelis, M. Diegel, J. A. Doherty, A. P. Feinberg, M. S. Fernando, J. Halow, K. D. Hansen, E. Haugen, P. F. Hickey, L. Hou, F. Jasmine, R. Jian, L. Jiang, A. Johnson, R. Kaul, M. Kellis, M. G. Kibriya, K. Lee, J. Billy Li, Q. Li, X. Li, J. Lin, S. Lin, S. Linder, C. Linke, Y. Liu, M. T. Maurano, B. Molinie, S. B. Montgomery, J. Nelson, F. J. Neri, M. Oliva, Y. Park, B. L. Pierce, N. J. Rinaldi, L. F. Rizzardi, R. Sandstrom, A. Skol, K. S. Smith, M. P. Snyder, J. Stamatoyannopoulos, B. E. Stranger, H. Tang, E. K. Tsang, L. Wang, M. Wang, N. Van Wittenberghe, F. Wu, R. Zhang, C. R. Nierras, P. A. Branton, L. J. Carithers, P. Guan, H. M. Moore, A. Rao, J. B. Vaught, S. E. Gould, N. C. Lockart, C. Martin, J. P. Struewing, S. Volpi, A. M. Addington, S. E. Koester, A. R. Little, L. E. Brigham, R. Hasz, M. Hunter, C. Johns, M. Johnson, G. Kopen, W. F. Leinweber, J. T. Lonsdale, A. McDonald, B. Mestichelli, K. Myer, B. Roe, M. Salvatore, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, J. Bridge, B. A. Foster, B. M. Gillard, E. Karasik, R. Kumar, M. Miklos, M. T. Moser, S. D. Jewell, R. G. Montroy, D. C. Rohrer, D. R. Valley, D. A. Davis, D. C. Mash, A. H. Undale, A. M. Smith, D. E. Tabor, N. V. Roche, J. A. McLean, N. Vatanian, K. L. Robinson, L. Sobin, M. E. Barcus, K. M. Valentino, L. Qi, S. Hunter, P. Hariharan, S. Singh, K. S. Um, T. Matose, M. M. Tomaszewski, L. K. Barker, M. Mosavel, L. A. Siminoff, H. M. Traino, P. Flicek, T. Juettemann, M. Ruffier, D. Sheppard, K. Taylor, S. J. Trevanion, D. R. Zerbino, B. Craft, M. Goldman, M. Haeussler, W. J. Kent, C. M. Lee, B. Paten, K. R. Rosenbloom, J. Vivian, J. Zhu, B. Craft, M. Goldman, M. Haeussler, W. J. Kent, C. M. Lee, B. Paten, K. R. Rosenbloom, J. Vivian, J. Zhu, F. Aguet, A. A. Brown, S. E. Castel, J. R. Davis, Y. He, B. Jo, P. Mohammadi, Y. Park, P. Parsana, A. V. Segr?, B. J. Strober, Z. Zappala, B. B. Cummings, E. T. Gelfand, K. Hadley, K. H. Huang, M. Lek, X. Li, J. L. Nedzel, D. Y. Nguyen, M. S. Noble, T. J. Sullivan, T. Tukiainen, D. G. MacArthur, G. Getz, A. Addington, P. Guan, S. Koester, A. R. Little, N. C. Lockhart, H. M. Moore, A. Rao, J. P. Struewing, S. Volpi, L. E. Brigham, R. Hasz, M. Hunter, C. Johns, M. Johnson, G. Kopen, W. F. Leinweber, J. T. Lonsdale, A. McDonald, B. Mestichelli, K. Myer, B. Roe, M. Salvatore, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, J. Bridge, B. A. Foster, B. M. Gillard, E. Karasik, R. Kumar, M. Miklos, M. T. Moser, S. D. Jewell, R. G. Montroy, D. C. Rohrer, D. Valley, D. C. Mash, D. A. Davis, L. Sobin, M. E. Barcus, P. A. Branton, N. S. Abell, B. Balliu,

O. Delaneau, L. Fr?sard, E. R. Gamazon, D. Garrido-Mart?n, A. D. H. Gewirtz, G. Gliner, M. J. Gloudemans, B. Han, A. Z. He, F. Hormozdiari, X. Li, B. Liu, E. Y. Kang, I. C. McDowell, H. Ongen, J. J. Palowitch, C. B. Peterson, G. Quon, S. Ripke, A. Saha, A. A. Shabalin, T. C. Shimko, J. H. Sul, N. A. Teran, E. K. Tsang, H. Zhang, Y. H. Zhou, C. D. Bustamante, N. J. Cox, R. Guig?, M. Kellis, M. I. McCarthy, D. F. Conrad, E. Eskin, G. Li, A. B. Nobel, C. Sabatti, B. E. Stranger, X. Wen, F. A. Wright, K. G. Ardlie, E. T. Dermitzakis, and T. Lappalainen. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 10 2017.

[12] Francesco Benedetti, Sara Dallaspezia, Cristina Colombo, Adele Pirovano, Elena Marino, and Enrico Smeraldi. A length polymorphism in the circadian clock gene Per3 influences age at onset of bipolar disorder. *Neuroscience letters*, 445(2):184–187, 2008.

[13] Gary Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27(2):573, 1999.

[14] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, 33(6):623–630, 2015.

[15] Lorenzo Bomba, Klaudia Walter, and Nicole Soranzo. The impact of rare and low-frequency genetic variants in common disease. *Genome biology*, 18(1):77, 2017.

[16] Christelle Borel, Eugenia Migliavacca, Audrey Letourneau, Maryline Gagnebin, Frédérique Béna, M Reza Sailani, Emmanouil T Dermitzakis, Andrew J Sharp, and Stylianos E Antonarakis. Tandem repeat sequence variation as causative Cis-eQTLs for protein-coding gene expression variation: The case of CSTB. *Human mutation*, 33(8):1302–1309, 2012.

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

[18] Amy L Byrd and Stephen B Manuck. MAOA, childhood maltreatment, and antisocial behavior: meta-analysis of a gene-environment interaction. *Biological psychiatry*, 75(1):9–17, 2014.

[19] Cristiano Capurso, Vincenzo Solfrizzi, Anna Maria Colacicco, Alessia D'introno, Vincenza Frisardi, Bruno P Imbimbo, Maria Lorusso, Gianluigi Vendemiale, Marta Denitto, Andrea Santamato, et al. Interleukin 6–174 G/C promoter and variable number of tandem repeats (VNTR) gene polymorphisms in sporadic Alzheimer's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 34(1):177–182, 2010.

[20] A Cervera, D Tassies, V Obach, S Amaro, JC Reverter, and A Chamorro. The BC genotype of the VNTR polymorphism of platelet glycoprotein Ibα is overrepresented in patients with recurrent stroke regardless of aspirin therapy. *Cerebrovascular Diseases*, 24(2-3):242–246, 2007.

[21] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics*, 13(1):238, 2012.

[22] Colby Chiang, Alexandra J Scott, Joe R Davis, Emily K Tsang, Xin Li, Yungil Kim, Tarik Hadzic, Farhan N Damani, Liron Ganel, Stephen B Montgomery, et al. The impact of structural variation on human gene expression. *Nature genetics*, 49(5):692, 2017.

[23] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology*, 4(4):265–270, 2009.

[24] Fergus J. Couch, Xianshu Wang, Lesley McGuffog, Andrew Lee, Curtis Olswold, Karoline B. Kuchenbaecker, Penny Soucy, Zachary Fredericksen, Daniel Barrowdale, Joe Dennis, Mia M. Gaudet, Ed Dicks, Matthew Kosel, Sue Healey, Olga M. Sinilnikova, Adam Lee, François Bacot, Daniel Vincent, Frans B.L. Hogervorst, Susan Peock, Dominique Stoppa-Lyonnet, Anna Jakubowska, Paolo Radice, Rita Katharina Schmutzler, Susan M. Domchek, Marion Piedmonte, Christian F. Singer, Eitan Friedman, Mads Thomassen, Thomas V.O. Hansen, Susan L. Neuhausen, Csilla I. Szabo, Ignacio Blanco, Mark H. Greene, Beth Y. Karlan, Judy Garber, Catherine M. Phelan, Jeffrey N. Weitzel, Marco Montagna, Edith Olah, Irene L. Andrulis, Andrew K. Godwin, Drakoulis Yannoukakos, David E. Goldgar, Trinidad Caldes, Heli Nevanlinna, Ana Osorio, Mary Beth Terry, Mary B. Daly, Elizabeth J. van Rensburg, Ute Hamann, Susan J. Ramus, Amanda Ewart Toland, Maria A. Caligo, Olufunmilayo I. Olopade, Nadine Tung, Kathleen Claes, Mary S. Beattie, Melissa C. Southey, Evgeny N. Imyanitov, Marc Tischkowitz, Ramunas Janavicius, Esther M. John, Ava Kwong, Orland Diez, Judith Balmaña, Rosa B. Barkardottir, Banu K. Arun, Gad Rennert, Soo Hwang Teo, Patricia A. Ganz, Ian Campbell, Annemarie H. van der Hout, Carolien H.M. van Deurzen, Caroline Seynaeve, Encarna B. Gómez Garcia, Flora E. van Leeuwen, Hanne E.J. Meijers-Heijboer, Johannes J.P. Gille, Margreet G.E.M. Ausems, Marinus J. Blok, Marjolijn J.L. Ligtenberg, Matti A. Rookus, Peter Devilee, Senno Verhoef, Theo A.M. van Os, Juul T. Wijnen, Debra Frost, Steve Ellis, Elena Fineberg, Radka Platte, D. Gareth Evans, Louise Izatt, Rosalind A. Eeles, Julian Adlard, Diana M. Eccles, Jackie Cook, Carole Brewer, Fiona Douglas, Shirley Hodgson, Patrick J. Morrison, Lucy E. Side, Alan Donaldson, Catherine Houghton, Mark T. Rogers, Huw Dorkins, Jacqueline Eason, Helen Gregory, Emma McCann, Alex Murray, Alain Calender, Agnès Hardouin, Pascaline Berthet, Capucine Delnatte, Catherine Nogues, Christine Lasset, Claude Houdayer, Dominique Leroux, Etienne Rouleau, Fabienne Prieur, Francesca Damiola, Hagay Sobol, Isabelle Coupier, Laurence Venat-Bouvet, Laurent Castera, Marion Gauthier-Villars, Mélanie Léoné, Pascal Pujol, Sylvie Mazoyer, Yves Jean Bignon, Elzbieta Złowocka-Perłowska, Jacek Gronwald, Jan Lubinski, Katarzyna Durda, Katarzyna Jaworska, Tomasz Huzarski, Amanda B. Spurdle, Bernard Peissel, Bernardo Bonanni, Giulia Melloni, Laura Ottini, Laura Papi, Liliana Varesco, Maria Grazia Tibiletti, Paolo Peterlongo, Sara Volorio, Siranoush Manoukian, Valeria Pensotti, Norbert Arnold, Christoph Engel, Helmut Deissler, Dorothea Gadzicki, Andrea Gehrig, Karin Kast, Kerstin Rhiem, Alfons Meindl, Dieter

Niederacher, Nina Ditsch, Hansjoerg Plendl, Sabine Preisler-Adams, Stefanie Engert, Christian Sutter, Raymonda Varon-Mateeva, Barbara Wappenschmidt, Bernhard H.F. Weber, Brita Arver, Marie Stenmark-Askmalm, Niklas Loman, Richard Rosenquist, Zakaria Einbeigi, Katherine L. Nathanson, Timothy R. Rebbeck, Stephanie V. Blank, David E. Cohn, Gustavo C. Rodriguez, Laurie Small, Michael Friedlander, Victoria L. Bae-Jump, Anneliese Fink-Retter, Christine Rappaport, Daphne Gschwantler-Kaulich, Georg Pfeiler, Muy Kheng Tea, Noralane M. Lindor, Bella Kaufman, Shani Shimon Paluch, Yael Laitman, Anne Bine Skytte, Anne Marie Gerdes, Inge Sokilde Pedersen, Sanne Traasdahl Moeller, Torben A. Kruse, Uffe Birk Jensen, Joseph Vijai, Kara Sarrel, Mark Robson, Noah Kauff, Anna Marie Mulligan, Gord Glendon, Hilmi Ozcelik, Bent Ejlertsen, Finn C. Nielsen, Lars Jønson, Mette K. Andersen, Yuan Chun Ding, Linda Steele, Lenka Foretova, Alex Teulé, Conxi Lazaro, Joan Brunet, Miquel Angel Pujana, Phuong L. Mai, Jennifer T. Loud, Christine Walsh, Jenny Lester, Sandra Orsulic, Steven A. Narod, Josef Herzog, Sharon R. Sand, Silvia Tognazzo, Simona Agata, Tibor Vaszko, Joellen Weaver, Alexandra V. Stavropoulou, Saundra S. Buys, Atocha Romero, Miguel de la Hoya, Kristiina Aittomäki, Taru A. Muranen, Mercedes Duran, Wendy K. Chung, Adriana Lasa, Cecilia M. Dorfling, Alexander Miron, Javier Benitez, Leigha Senter, Dezheng Huo, Salina B. Chan, Anna P. Sokolenko, Jocelyne Chiquette, Laima Tihomirova, Tara M. Friebel, Bjarni A. Agnarsson, Karen H. Lu, Flavio Lejbkowicz, Paul A. James, Per Hall, Alison M. Dunning, Daniel Tessier, Julie Cunningham, Susan L. Slager, Chen Wang, Steven Hart, Kristen Stevens, Jacques Simard, Tomi Pastinen, Vernon S. Pankratz, Kenneth Offit, Douglas F. Easton, Georgia Chenevix-Trench, and Antonis C. Antoniou. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet*, 9(3):e1003212, 2013.

[25] Arne De Roeck, Wouter De Coster, Liene Bossaerts, Rita Cacace, Tim De Pooter, Jasper Van Dongen, Svenn D'Hert, Peter De Rijk, Mojca Strazisar, Christine Van Broeckhoven, et al. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome biology*, 20(1):239, 2019.

[26] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[27] Daniel Dieringer and Christian Schlötterer. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome research*, 13(10):2242–2251, 2003.

[28] Yuan C Ding, Lesley McGuffog, Sue Healey, Eitan Friedman, Yael Laitman, Bella Kaufman, Annelie Liljegren, Annika Lindblom, Håkan Olsson, Ulf Kristoffersson, et al. A nonsynonymous polymorphism in IRS1 modifies risk of developing breast and ovarian cancers in BRCA1 and ovarian cancer in BRCA2 mutation carriers. *Cancer Epidemiology and Prevention Biomarkers*, 21(8):1362–1370, 2012.

[29] Egor Dolzhenko, Joke JFA van Vugt, Richard J Shaw, Mitchell A Bekritsky, Marka van Blitterswijk, Giuseppe Narzisi, Subramanian S Ajay, Vani Rajan, Bryan Lajoie, Nathan H Johnson, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, pages gr–225672, 2017.

[30] Cord Drögemüller, Elinor K Karlsson, Marjo K Hytönen, Michele Perloski, Gaudenz Dolf, Kirsi Sainio, Hannes Lohi, Kerstin Lindblad-Toh, and Tosso Leeb. A mutation in hairless dogs implicates FOXI3 in ectodermal development. *Science*, 321(5895):1462–1462, 2008.

[31] Veronika B Dubinkina, Dmitry S Ischenko, Vladimir I Ulyantsev, Alexander V Tyakht, and Dmitry G Alexeev. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC bioinformatics*, 17(1):38, 2016.

[32] Ivana Durinovic-Belló, RP Wu, VH Gersuk, S Sanda, HG Shilling, and GT Nepom. Insulin gene VNTR genotype associates with frequency and phenotype of the autoimmune response to proinsulin. *Genes and immunity*, 11(2):188–193, 2010.

[33] Mark TW Ebbert, Stefan L Farrugia, Jonathon P Sens, Karen Jansen-West, Tania F Gendron, Mercedes Prudencio, Ian J McLaughlin, Brett Bowman, Matthew Seetin, Mariely DeJesus-Hernandez, et al. Long-read sequencing across the C9orf72 'GGGGCC'repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Molecular neurodegeneration*, 13(1):46, 2018.

[34] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[35] Sean R Eddy et al. Multiple alignment using hidden Markov models. In *Ismb*, volume 3, pages 114–120, 1995.

[36] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[37] O Eser, B Eser, M Cosar, MO Erdogan, A Aslan, H Yildiz, M Solak, and A Haktanir. Short aggrecan gene repetitive alleles associated with lumbar degenerative disc disease in Turkish patients. *Genet Mol Res*, 10(3):1923–1930, 2011.

[38] John W Fondon and Harold R Garner. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences*, 101(52):18058–18063, 2004.

[39] Stephanie Feupe Fotsing, Jonathan Margoliash, Catherine Wang, Shubham Saini, Richard Yanicky, Sharona Shleizer-Burko, Alon Goren, and Melissa Gymrek. The impact of short tandem repeat variation on gene expression. *Nature genetics*, 51(11):1652–1659, 2019.

[40] Barbara Franke, Alejandro Arias Vasquez, Stefan Johansson, Martine Hoogman, Jasmin Romanos, Andrea Boreatti-Hümmer, Monika Heine, Christian P Jacob, Klaus-Peter Lesch, Miguel Casas, et al. Multicenter analysis of the SLC6A3/DAT1 VNTR haplotype in persistent ADHD suggests differential involvement of the gene in childhood and persistent ADHD. *Neuropsychopharmacology*, 35(3):656, 2010.

[41] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, 2016.

[42] Yevgeniy Gelfand, Yozen Hernandez, Joshua Loving, and Gary Benson. VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic acids research*, 42(14):8884–8894, 2014.

[43] Rita Gemayel, Marcelo D Vinces, Matthieu Legendre, and Kevin J Verstrepen. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics*, 44:445–477, 2010.

[44] Yoav Gilad, Scott A Rifkin, and Jonathan K Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in genetics*, 24(8):408–415, 2008.

[45] Nikolaos Givalos, Hariklia Gakiopoulou, Melina Skliri, Katerina Bousboukea, Anastasia E. Konstantinidou, Penelope Korkolopoulou, Maria Lelouda, Gregory Kouraklis, Efstratios Patsouris, and Gabriel Karatzas. Replication protein A is an independent prognostic indicator with potential therapeutic implications in colon cancer. *Modern Pathology*, 2007.

[46] Teresa Gómez-Isla, Joseph L Price, Daniel W McKeel Jr, John C Morris, John H Growdon, and Bradley T Hyman. Profound loss of layer II entorhinal cortex neurons occurs in very mild Alzheimer's disease. *Journal of Neuroscience*, 16(14):4491–4500, 1996.

[47] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44(10):1084–1089, 2012.

[48] Daniel F. Gudbjartsson, Hannes Helgason, Sigurjon A. Gudjonsson, Florian Zink, Asmundur Oddson, Arnaldur Gylfason, Soren Besenbacher, Gisli Magnusson, Bjarni V. Halldorsson, Eirikur Hjartarson, Gunnar Th Sigurdsson, Simon N. Stacey, Michael L. Frigge, Hilma Holm, Jona Saemundsdottir, Hafdis Th Helgadottir, Hrefna Johannsdottir, Gunnlaugur Sigfusson, Gudmundur Thorgeirsson, Jon Th Sverrisson, Solveig Gretarsdottir, G. Bragi Walters, Thorunn Rafnar, Bjarni Thjodleifsson, Einar S. Bjornsson, Sigurdur Olafsson, Hildur Thorarinsdottir, Thora Steingrimsdottir, Thora S. Gudmundsdottir, Asgeir Theodors, Jon G. Jonasson, Asgeir Sigurdsson, Gyda Bjornsdottir, Jon J. Jonsson, Olafur Thorarensen, Petur Ludvigsson, Hakon Gudbjartsson, Gudmundur I. Eyjolfsson, Olof Sigurdardottir, Isleifur Olafsson, David O. Arnar, Olafur Th Magnusson, Augustine

Kong, Gisli Masson, Unnur Thorsteinsdottir, Agnar Helgason, Patrick Sulem, and Kari Stefansson. Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, 47(5):435–444, may 2015.

[49] Alexandra E. Gylfe, Riku Katainen, Johanna Kondelin, Tomas Tanskanen, Tatiana Cajuso, Ulrika Hänninen, Jussi Taipale, Minna Taipale, Laura Renkonen-Sinisalo, Heikki Järvinen, Jukka Pekka Mecklin, Outi Kilpivaara, Esa Pitkänen, Pia Vahteristo, Sari Tuupanen, Auli Karhu, and Lauri A. Aaltonen. Eleven Candidate Susceptibility Genes for Common Familial Colorectal Cancer. *PLoS Genetics*, 9(10), 2013.

[50] Melissa Gymrek. PCR-free library preparation greatly reduces stutter noise at short tandem repeats. *bioRxiv*, page 043448, 2016.

[51] Melissa Gymrek. A genomic view of short tandem repeats, 2017.

[52] Melissa Gymrek, David Golan, Saharon Rosset, and Yaniv Erlich. lobSTR: a short tandem repeat profiler for personal genomes. *Genome research*, 22(6):1154–1162, 2012.

[53] Melissa Gymrek, Thomas Willems, Audrey Guilmatre, Haoyang Zeng, Barak Markus, Stoyan Georgiev, Mark J Daly, Alkes L Price, Jonathan K Pritchard, Andrew J Sharp, and Yaniv Erlich. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics*, 48(1):22–29, 2016.

[54] Thomas Hackl, Rainer Hedrich, Jörg Schultz, and Frank Förster. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, 30(21):3004–3011, 2014.

[55] K Haddley, VJ Bubb, G Breen, UM Parades-Esquivel, and JP Quinn. Behavioural genetics of the serotonin transporter. In *Behavioral Neurogenetics*, pages 503–535. Springer, 2011.

[56] Anthony J Hannan. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends in genetics*, 26(2):59–65, 2010.

[57] Ruo-Han Hao, Tie-Lin Yang, Yu Rong, Shi Yao, Shan-Shan Dong, Hao Chen, and Yan Guo. Gene expression profiles indicate tissue-specific obesity regulation changes and strong obesity relevant tissues. *International Journal of Obesity*, 42(3):363–369, 2018.

[58] Minako Hijikata, Ikumi Matsushita, Goh Tanaka, Tomoko Tsuchiya, Hideyuki Ito, Katsushi Tokunaga, Jun Ohashi, Sakae Homma, Yoichiro Kobashi, Yoshio Taguchi, et al. Molecular cloning of two novel mucin-like genes in the disease-susceptibility locus for diffuse panbronchiolitis. *Human genetics*, 129(2):117–128, 2011.

[59] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[60] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2011.

[61] Alec J Jeffreys, Victoria Wilson, and Swee Lay Thein. Hypervariable 'minisatellite'regions in human DNA. *Nature*, 314(6006):67–73, 1985.

[62] W James Kent. BLAT-the BLAST-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[63] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[64] Andrew Kirby, Andreas Gnirke, David B Jaffe, Veronika Barešová, Nathalie Pochet, Brendan Blumenstiel, Chun Ye, Daniel Aird, Christine Stevens, James T Robinson, et al. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nature genetics*, 45(3):299–303, 2013.

[65] J Kirchheiner, K Nickchen, J Sasse, M Bauer, I Roots, and J Brockmöller. A 40-basepair VNTR polymorphism in the dopamine transporter (DAT1) gene and the rapid response to antidepressant treatment. *The pharmacogenomics journal*, 7(1):48, 2007.

[66] Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of molecular biology*, 235(5):1501–1531, 1994.

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

[68] Karoline B. Kuchenbaecker, John L. Hopper, Daniel R. Barnes, Kelly Anne Phillips, Thea M. Mooij, Marie José Roos-Blom, Sarah Jervis, Flora E. Van Leeuwen, Roger L. Milne, Nadine Andrieu, David E. Goldgar, Mary Beth Terry, Matti A. Rookus, Douglas F. Easton, and Antonis C. Antoniou. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *Jama*, 317(23):2402–2416, 2017.

[69] Karoline B. Kuchenbaecker, Lesley McGuffog, Daniel Barrowdale, Andrew Lee, Penny Soucy, Joe Dennis, Susan M. Domchek, Mark Robson, Amanda B. Spurdle, Susan J. Ramus, Nasim Mavaddat, Mary Beth Terry, Susan L. Neuhausen, Rita Katharina Schmutzler, Jacques Simard, Paul D.P. Pharoah, Kenneth Offit, Fergus J. Couch, Georgia Chenevix-Trench, Douglas F. Easton, Antonis C. Antoniou, Sue Healey, Michael Lush, Ute Hamann, Melissa Southey, Esther M. John, Wendy K. Chung, Mary B. Daly, Saundra S. Buys, David E. Goldgar, Cecilia M. Dorfling, Elizabeth J. van Rensburg, Yuan Chun Ding, Bent Ejlertsen, Anne Marie Gerdes, Thomas V.O. Hansen, Susan Slager, Emily Hallberg, Javier Benitez, Ana Osorio, Nancy Cohen, William Lawler, Jeffrey N. Weitzel, Paolo Peterlongo, Valeria Pensotti, Riccardo Dolcetti, Monica Barile, Bernardo Bonanni, Jacopo Azzollini, Siranoush Manoukian, Bernard Peissel, Paolo Radice, Antonella Savarese,

Laura Papi, Giuseppe Giannini, Florentia Fostira, Irene Konstantopoulou, Julian Adlard, Carole Brewer, Jackie Cook, Rosemarie Davidson, Diana Eccles, Ros Eeles, Steve Ellis, EMBRACE, Debra Frost, Shirley Hodgson, Louise Izatt, Fiona Lalloo, Kai ren Ong, Andrew K. Godwin, Norbert Arnold, Bernd Dworniczak, Christoph Engel, Andrea Gehrig, Eric Hahnen, Jan Hauke, Karin Kast, Alfons Meindl, Dieter Niederacher, Raymonda Varon-Mateeva, Shan Wang-Gohrke, Barbara Wappenschmidt, Laure Barjhoux, Marie Agnès Collonge-Rame, Camille Elan, GEMO Study Collaborators, Lisa Golmard, Emmanuelle Barouk-Simonet, Fabienne Lesueur, Sylvie Mazoyer, Joanna Sokolowska, Dominique StoppaLyonnet, Claudine Isaacs, Kathleen B.M. Claes, Bruce Poppe, Miguel de la Hoya, Vanesa Garcia-Barberan, Kristiina Aittomäki, Heli Nevanlinna, Margreet G.E.M. Ausems, J. L. de Lange, Encarna B. Gomez Garcia, HEBON, Frans B.L. Hogervorst, Carolien M. Kets, Hanne E.J. Meijers-Heijboer, Jan C. Oosterwijk, Matti A. Rookus, Christi J. van Asperen, Ans M.W. van den Ouweland, Helena C. van Doorn, Theo A.M. van Os, Ava Kwong, Edith Olah, Orland Diez, Joan Brunet, Conxi Lazaro, Alex Teulé, Jacek Gronwald, Anna Jakubowska, Katarzyna Kaczmarek, Jan Lubinski, Grzegorz Sukiennicki, Rosa B. Barkardottir, Jocelyne Chiquette, Simona Agata, Marco Montagna, Manuel R. Teixeira, KCon Fab Investigators, Sue Kyung Park, Curtis Olswold, Marc Tischkowitz, Lenka Foretova, Pragna Gaddam, Joseph Vijai, Georg Pfeiler, Christine Rappaport-Fuerhauser, Christian F. Singer, Muy Kheng M. Tea, Mark H. Greene, Jennifer T. Loud, Gad Rennert, Evgeny N. Imyanitov, Peter J. Hulick, John L. Hays, Marion Piedmonte, Gustavo C. Rodriguez, Julie Martyn, Gord Glendon, Anna Marie Mulligan, Irene L. Andrulis, Amanda Ewart Toland, Uffe Birk Jensen, Torben A. Kruse, Inge Sokilde Pedersen, Mads Thomassen, Maria A. Caligo, Soo Hwang Teo, Raanan Berger, Eitan Friedman, Yael Laitman, Brita Arver, Ake Borg, Hans Ehrencrona, Johanna Rantala, Olufunmilayo I. Olopade, Patricia A. Ganz, Robert L. Nussbaum, Angela R. Bradbury, Katherine L. Nathanson, Banu K. Arun, Paul James, Beth Y. Karlan, and Jenny Lester. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *JNCI: Journal of the National Cancer Institute*, 109(7), 2017.

[70] Peter Kühnen, Daniela Handke, Robert A Waterland, Branwen J Hennig, Matt Silver, Anthony J Fulford, Paula Dominguez-Salas, Sophie E Moore, Andrew M Prentice, Joachim Spranger, et al. Interindividual variation in DNA methylation at a putative POMC metastable epiallele is associated with obesity. *Cell metabolism*, 24(3):502–509, 2016.

[71] GJ LaHoste, JMet Swanson, SB Wigal, C Glabe, T Wigal, N King, and JL Kennedy. Dopamine D4 receptor gene polymorphism is associated with attention deficit hyperactivity disorder. *Mol Psychiatry*, 1(2):121–124, 1996.

[72] Maria D Lalioti, Hamish S Scott, Catherine Buresi, Colette Rossier, Armand Bottani, Michael A Morris, Alain Malafosse, and Stylianos E Antonarakis. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*, 386(6627):847, 1997.

[73] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer

Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[74] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[75] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[76] Dong-Hyun Lee, Yunfeng Pan, Shlomo Kanner, Patrick Sung, James A Borowiec, and Dipanjan Chowdhury. A PP4 phosphatase complex dephosphorylates RPA2 to facilitate DNA repair via homologous recombination. *Nature structural & molecular biology*, 17(3):365, 2010.

[77] Hayan Lee, James Gurtowski, Shinjae Yoo, Shoshana Marcus, W Richard McCombie, and Michael Schatz. Error correction and assembly complexity of single molecule sequencing reads. *BioRxiv*, page 006395, 2014.

[78] Richard JLF Lemmers, Peggy de Kievit, Lodewijk Sandkuijl, George W Padberg, Gert-Jan B van Ommen, Rune R Frants, and Silvère M van der Maarel. Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nature genetics*, 32(2):235, 2002.

[79] Ephrat Levy-Lahad and E Friedman. Cancer risks among BRCA1 and BRCA2 mutation carriers. *British journal of cancer*, 96(1):11–15, 2007.

[80] G. Li, Y. Zhang, K. Y. Cheng, and P. J. Scarpace. Lean rats with hypothalamic pro-opiomelanocortin overexpression exhibit greater diet-induced obesity and impaired central melanocortin responsiveness. *Diabetologia*, 50(7):1490–1499, 2007.

[81] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[82] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[83] Ming Li, Andrew E Jaffe, Richard E Straub, Ran Tao, Joo Heon Shin, Yanhong Wang, Qiang Chen, Chao Li, Yankai Jia, Kazutaka Ohi, et al. A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24. 32 schizophrenia-associated locus. *Nature medicine*, 22(6):649, 2016.

[84] Qian Liu, Peng Zhang, Depeng Wang, Weihong Gu, and Kai Wang. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome medicine*, 9(1):65, 2017.

[85] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[86] Qixing Mao, Mantang Qiu, Gaochao Dong, Wenjie Xia, Shuai Zhang, Youtao Xu, Jie Wang, Yin Rong, Lin Xu, and Feng Jiang. CAG repeat polymorphisms in the androgen receptor and breast cancer risk in women: a meta-analysis of 17 studies. *OncoTargets and therapy*, 8:2111, 2015.

[87] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[88] Riccardo E. Marioni, Sarah E. Harris, Qian Zhang, Allan F. McRae, Saskia P. Hagenaars, W. David Hill, Gail Davies, Craig W. Ritchie, Catharine R. Gale, John M. Starr, Alison M. Goate, David J. Porteous, Jian Yang, Kathryn L. Evans, Ian J. Deary, Naomi R. Wray, and Peter M. Visscher. GWAS on family history of Alzheimer's disease. *Translational Psychiatry*, 2018.

[89] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P. Tyrer, Ting Huei Chen, Qin Wang, Manjeet K. Bolla, Xin Yang, Muriel A. Adank, Thomas Ahearn, Kristiina Aittomäki, Jamie Allen, Irene L. Andrulis, Hoda Anton-Culver, Natalia N. Antonenkova, Volker Arndt, Kristan J. Aronson, Paul L. Auer, Päivi Auvinen, Myrto Barrdahl, Laura E. Beane Freeman, Matthias W. Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Leslie Bernstein, Carl Blomqvist, Natalia V. Bogdanova, Stig E. Bojesen, Bernardo Bonanni, Anne Lise Børresen-Dale, Hiltrud Brauch, Michael Bremer, Hermann Brenner, Adam Brentnall, Ian W. Brock, Angela Brooks-Wilson, Sara Y. Brucker, Thomas Brüning, Barbara Burwinkel, Daniele Campa, Brian D. Carter, Jose E. Castelao, Stephen J. Chanock, Rowan Chlebowski, Hans Christiansen, Christine L. Clarke, J. Margriet Collée, Emilie Cordina-Duverger, Sten Cornelissen, Fergus J. Couch, Angela Cox, Simon S. Cross, Kamila Czene, Mary B. Daly, Peter Devilee, Thilo Dörk, Isabel dos Santos-Silva, Martine Dumont, Lorraine Durcan, Miriam Dwek, Diana M. Eccles, Arif B. Ekici, A. Heather Eliassen, Carolina Ellberg, Christoph Engel, Mikael Eriksson, D. Gareth Evans, Peter A. Fasching, Jonine Figueroa, Olivia Fletcher, Henrik Flyger, Asta Försti, Lin Fritschi, Marike Gabrielson, Manuela Gago-Dominguez, Susan M. Gapstur, José A. García-Sáenz, Mia M. Gaudet, Vassilios Georgoulias, Graham G. Giles, Irina R. Gilyazova, Gord Glendon, Mark S. Goldberg, David E. Goldgar, Anna González-Neira, Grethe I. Grenaker Alnæs, Mervi Grip, Jacek Gronwald, Anne Grundy, Pascal Guénel, Lothar Haeberle, Eric Hahnen, Christopher A. Haiman, Niclas Håkansson, Ute Hamann, Susan E. Hankinson, Elaine F. Harkness, Steven N. Hart, Wei He, Alexander Hein, Jane Heyworth, Peter Hillemanns, Antoinette Hollestelle, Maartje J. Hooning, Robert N. Hoover, John L. Hopper, Anthony Howell, Guanmengqian Huang, Keith Humphreys,

David J. Hunter, Milena Jakimovska, Anna Jakubowska, Wolfgang Janni, Esther M. John, Nichola Johnson, Michael E. Jones, Arja Jukkola-Vuorinen, Audrey Jung, Rudolf Kaaks, Katarzyna Kaczmarek, Vesa Kataja, Renske Keeman, Michael J. Kerin, Elza Khusnutdinova, Johanna I. Kiiski, Julia A. Knight, Yon Dschun Ko, Veli Matti Kosma, Stella Koutros, Vessela N. Kristensen, Ute Krüger, Tabea Kühl, Diether Lambrechts, Loic Le Marchand, Eunjung Lee, Flavio Lejbkowicz, Jenna Lilyquist, Annika Lindblom, Sara Lindström, Jolanta Lissowska, Wing Yee Lo, Sibylle Loibl, Jirong Long, Jan Lubiński, Michael P. Lux, Robert J. MacInnis, Tom Maishman, Enes Makalic, Ivana Maleva Kostovska, Arto Mannermaa, Siranoush Manoukian, Sara Margolin, John W.M. Martens, Maria Elena Martinez, Dimitrios Mavroudis, Catriona McLean, Alfons Meindl, Usha Menon, Pooja Middha, Nicola Miller, Fernando Moreno, Anna Marie Mulligan, Claire Mulot, Victor M. Muñoz-Garzon, Susan L. Neuhausen, Heli Nevanlinna, Patrick Neven, William G. Newman, Sune F. Nielsen, Børge G. Nordestgaard, Aaron Norman, Kenneth Offit, Janet E. Olson, Håkan Olsson, Nick Orr, V. Shane Pankratz, Tjoung Won Park-Simon, Jose I.A. Perez, Clara Pérez-Barrios, Paolo Peterlongo, Julian Peto, Mila Pinchev, Dijana Plaseska-Karanfilska, Eric C. Polley, Ross Prentice, Nadege Presneau, Darya Prokofyeva, Kristen Purrington, Katri Pylkäs, Brigitte Rack, Paolo Radice, Rohini Rau-Murthy, Gad Rennert, Hedy S. Rennert, Valerie Rhenius, Mark Robson, Atocha Romero, Kathryn J. Ruddy, Matthias Ruebner, Emmanouil Saloustros, Dale P. Sandler, Elinor J. Sawyer, Daniel F. Schmidt, Rita K. Schmutzler, Andreas Schneeweiss, Minouk J. Schoemaker, Fredrick Schumacher, Peter Schürmann, Lukas Schwentner, Christopher Scott, Rodney J. Scott, Caroline Seynaeve, Mitul Shah, Mark E. Sherman, Martha J. Shrubsole, Xiao Ou Shu, Susan Slager, Ann Smeets, Christof Sohn, Penny Soucy, Melissa C. Southey, John J. Spinelli, Christa Stegmaier, Jennifer Stone, Anthony J. Swerdlow, Rulla M. Tamimi, William J. Tapper, Jack A. Taylor, Mary Beth Terry, Kathrin Thöne, Rob A.E.M. Tollenaar, Ian Tomlinson, Thérèse Truong, Maria Tzardi, Hans Ulrich Ulmer, Michael Untch, Celine M. Vachon, Elke M. van Veen, Joseph Vijai, Clarice R. Weinberg, Camilla Wendt, Alice S. Whittemore, Hans Wildiers, Walter Willett, Robert Winqvist, Alicja Wolk, Xiaohong R. Yang, Drakoulis Yannoukakos, Yan Zhang, Wei Zheng, Argyrios Ziogas, Alison M. Dunning, Deborah J. Thompson, Georgia Chenevix-Trench, Jenny Chang-Claude, Marjanka K. Schmidt, Per Hall, Roger L. Milne, Paul D.P. Pharoah, Antonis C. Antoniou, Nilanjan Chatterjee, Peter Kraft, Montserrat García-Closas, Jacques Simard, and Douglas F. Easton. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*, 104(1):21–34, 2019.

[90] Romain Menegaux and Jean-Philippe Vert. Continuous Embeddings of DNA Sequencing Reads and Application to Metagenomics. *Journal of Computational Biology*, 2018.

[91] Giles Miclotte, Mahdi Heydari, Piet Demeester, Stephane Rombauts, Yves Van de Peer, Pieter Audenaert, and Jan Fostier. Jabba: hybrid error correction for long sequencing reads. *Algorithms for Molecular Biology*, 11(1):10, 2016.

[92] Jason R Miller, Peng Zhou, Joann Mudge, James Gurtowski, Hayan Lee, Thiruvarangan Ramaraj, Brian P Walenz, Junqi Liu, Robert M Stupar, Roxanne Denny, et al. Hybrid

assembly with long and short reads improves discovery of gene family expansions. *BMC genomics*, 18(1):541, 2017.

[93] Roger L Milne and Antonis C Antoniou. Modifiers of breast and ovarian cancer risks for BRCA1 and BRCA2 mutation carriers. *Endocrine-related cancer*, 23(10):T69–T84, 2016.

[94] Satomi Mitsuhashi, Martin C Frith, Takeshi Mizuguchi, Satoko Miyatake, Tomoko Toyota, Hiroaki Adachi, Yoko Oma, Yoshihiro Kino, Hiroaki Mitsuhashi, and Naomichi Matsumoto. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome biology*, 20(1):58, 2019.

[95] Fernando Morales, Melissa Vásquez, Carolina Santamaría, Patricia Cuenca, Eyleen Corrales, and Darren G Monckton. A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA repair*, 40:57–66, 2016.

[96] Nima Mousavi, Jonathan Margoliash, Neha Pusarla, Shubham Saini, Richard Yanicky, and Melissa Gymrek. TRTools: a toolkit for genome-wide analysis of tandem repeats. *bioRxiv*, 2020.

[97] Nima Mousavi, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic acids research*, 47(15):e90–e90, 2019.

[98] Alexandra C Nica, Stephen B Montgomery, Antigone S Dimas, Barbara E Stranger, Claude Beazley, Inês Barroso, and Emmanouil T Dermitzakis. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet*, 6(4):e1000895, 2010.

[99] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4), 2010.

[100] Satoshi Okazaki, Marta Schirripa, Fotios Loupakis, Shu Cao, Wu Zhang, Dongyun Yang, Yan Ning, Martin D Berger, Yuji Miyamoto, Mitsukuni Suenaga, et al. Tandem repeat variation near the HIC1 (hypermethylated in cancer 1) promoter predicts outcome of oxaliplatin-based chemotherapy in patients with metastatic colorectal cancer. *Cancer*, 2017.

[101] Masato Orita, Hiroyuki Iwahana, Hiroshi Kanazawa, Kenshi Hayashi, and Takao Sekiya. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences*, 86(8):2766–2770, 1989.

[102] Anna A Pimenova, Towfique Raj, and Alison M Goate. Untangling genetic risk for Alzheimer's disease. *Biological psychiatry*, 83(4):300–310, 2018.

[103] Antonia L Pritchard, Colin W Pritchard, Peter Bentham, and Corinne L Lendon. Role of serotonin transporter polymorphisms in the behavioural and psychological symptoms in probable Alzheimer disease patients. *Dementia and geriatric cognitive disorders*, 24(3):201–206, 2007.

[104] Alberto Pugliese, Markus Zeller, Alarico Fernandez, Laura J Zalcberg, Richard J Bartlett, Camillo Ricordi, Massimo Pietropaolo, George S Eisenbarth, Simon T Bennett, and Dhavalkumar D Patel. The insulin gene is transcribed in the human thymus and transcription levels correlate with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. *Nature genetics*, 15(3):293–297, 1997.

[105] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[106] Javier Quilez, Audrey Guilmatre, Paras Garg, Gareth Highnam, Melissa Gymrek, Yaniv Erlich, Ricky S Joshi, David Mittelman, and Andrew J Sharp. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic acids research*, 44(8):3750–3762, 2016.

[107] Helge Ræder, Stefan Johansson, Pål I Holm, Ingfrid S Haldorsen, Eric Mas, Véronique Sbarra, Ingrid Nermoen, Stig Å Eide, Louise Grevle, Lise Bjørkhaug, et al. Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nature genetics*, 38(1):54, 2006.

[108] Timothy R Rebbeck, Philip W Kantoff, Krishna Krithivas, Susan Neuhausen, M Anne Blackwood, Andrew K Godwin, Mary B Daly, Steven A Narod, Judy E Garber, Henry T Lynch, et al. Modification of BRCA1-associated breast cancer risk by the polymorphic androgen-receptor CAG repeat. *The American Journal of Human Genetics*, 64(5):1371–1377, 1999.

[109] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

[110] Anna M Rose, Abhay Krishan, CF Chakarova, Leire Moya, SK Chambers, M Hollands, JC Illingworth, SMG Williams, HE McCabe, AZ Shah, et al. MSR1 repeats modulate gene expression and affect risk of breast and prostate cancer. *Annals of Oncology*, 29(5):1292–1303, 2018.

[111] Leena Salmela and Eric Rivals. LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, 30(24):3506–3514, 2014.

[112] Leena Salmela, Riku Walve, Eric Rivals, and Esko Ukkonen. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, 33(6):799–806, 2016.

[113] Nagato Sato, Junkichi Koinuma, Tomoo Ito, Eiju Tsuchiya, Satoshi Kondo, Yusuke Nakamura, and Yataro Daigo. Activation of an oncogenic TBC1D7 (TBC1 domain family, member 7) protein in pulmonary carcinogenesis. *Genes Chromosomes and Cancer*, 2010.

[114] Aurora Savino, Lidia Avalle, Emanuele Monteleone, Irene Miglio, Alberto Griffa, Giulia Accetta, Paolo Provero, and Valeria Poli. Network analysis allows to unravel breast cancer molecular features and to identify novel targets. *bioRxiv*, page 570051, 2019.

[115] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Scipy, 2010.

[116] Mark D Shriver, Li Jin, Ranajit Chakraborty, and Eric Boerwinkle. VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics*, 134(3):983–993, 1993.

[117] Amy PN Skubitz, Stefan E Pambuccian, Peter A Argenta, and Keith M Skubitz. Differential gene expression identifies subgroups of ovarian carcinoma. *Translational research*, 148(5):223–248, 2006.

[118] Tugce Bilgin Sonay, Tiago Carvalho, Mark D Robinson, Maja P Greminger, Michael Krützen, David Comas, Gareth Highnam, David Mittelman, Andrew Sharp, Tomàs Marques-Bonet, et al. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome research*, 25(11):1591–1599, 2015.

[119] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500, 2012.

[120] Bianca K Stöcker, Johannes Köster, and Sven Rahmann. SimLoRD: Simulation of Long Read Data. *Bioinformatics*, 32(17):2704–2706, 2016.

[121] Barbara E Stranger, Stephen B Montgomery, Antigone S Dimas, Leopold Parts, Oliver Stegle, Catherine E Ingle, Magda Sekowska, George Davey Smith, David Evans, Maria Gutierrez-Arcelus, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS genetics*, 8(4), 2012.

[122] Arvis Sulovari, Ruiyang Li, Peter A Audano, David Porubsky, Mitchell R Vollger, Glennis A Logsdon, Wesley C Warren, Alex A Pollen, Mark JP Chaisson, Evan E Eichler, et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences*, 116(46):23243–23253, 2019.

[123] Philip Supply, Sarah Lesjean, Evgueni Savine, Kristin Kremer, Dick Van Soolingen, and Camille Locht. Automated high-throughput genotyping for study of global epidemiology of mycobacterium tuberculosis based on mycobacterial interspersed repetitive units. *Journal of clinical microbiology*, 39(10):3563–3571, 2001.

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

[125] Yukiko Tomioka, Shusuke Numata, Makoto Kinoshita, Hidehiro Umehara, Shin ya Watanabe, Masahito Nakataki, Yoshimi Iwayama, Tomoko Toyota, Masashi Ikeda, Hidenaga Yamamori, Shinji Shimodera, Atsushi Tajima, Ryota Hashimoto, Nakao Iwata, Takeo Yoshikawa, and Tetsuro Ohmori. Decreased serum pyridoxal levels in schizophrenia: Meta-analysis and Mendelian randomization analysis. *Journal of Psychiatry and Neuroscience*, 43(3):194–200, may 2018.

[126] Cath Tyner, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Christopher Eisenhart, Clayton M Fischer, David Gibson, Jairo Navarro Gonzalez, Luvina Guruvadoo, et al. The UCSC Genome Browser database: 2017 update. *Nucleic acids research*, 45(D1):D626–D634, 2016.

[127] Ajay Ummat and Ali Bashir. Resolving complex tandem repeats with long reads. *Bioinformatics*, 30(24):3491–3498, 2014.

[128] Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics*, 51(1):187–195, 2019.

[129] Petros Vafiadis, Simon T. Bennett, John A. Todd, Joseph Nadeau, Rosemarie Grabs, Cynthia G. Goodyer, Saman Wickramasinghe, Eleanor Colle, and Constantin Polychronakos. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. *Nature Genetics*, 1997.

[130] Biju Viswanath, Meera Purushottam, Thennarasu Kandavel, YC Janardhan Reddy, Sanjeev Jain, et al. DRD4 gene and obsessive compulsive disorder: do symptom dimensions have specific genetic correlates? *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 41:18–23, 2013.

[131] Amy J Vogler, Christine E Keys, Christopher Allender, Ira Bailey, Jessica Girard, Talima Pearson, Kimothy L Smith, David M Wagner, and Paul Keim. Mutations, mutation rates, and evolution at the hypervariable vntr loci of yersinia pestis. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 616(1-2):145–158, 2007.

[132] Ying Wang, Shuichi Kikuchi, Hiromichi Suzuki, Sohji Nagase, and Akio Koyama. Endothelial nitric oxide synthase gene polymorphism in intron 4 affects the progression of renal failure in non-diabetic renal diseases. *Nephrology Dialysis Transplantation*, 14(12):2898–2902, 1999.

[133] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel DNA sequencing. *nature*, 452(7189):872–876, 2008.

[134] Thomas Willems, Melissa Gymrek, Gareth Highnam, David Mittelman, Yaniv Erlich, 1000 Genomes Project Consortium, et al. The landscape of human STR variation. *Genome research*, 24(11):1894–1904, 2014.

[135] Thomas Willems, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. Genome-wide profiling of heritable and de novo STR variations. *Nature Methods*, 2017.

[136] G. M. Williams and J. A. Surtees. MSH3 Promotes Dynamic Behavior of Trinucleotide Repeat Tracts In Vivo. *Genetics*, 200(3):737–754, Jul 2015.

[137] Bradford B Worrall, Thomas G Brott, Robert D Brown, W Mark Brown, Stephen S Rich, Sampath Arepalli, Fabienne Wavrant-De Vrièze, Jaime Duckworth, Andrew B Singleton, John Hardy, et al. IL1RN VNTR polymorphism in ischemic stroke. *Stroke*, 38(4):1189–1196, 2007.

[138] Fred A Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, et al. Heritability and genomics of gene expression in peripheral blood. *Nature genetics*, 46(5):430–437, 2014.

[139] Jonathan M Wright. Mutation at VNTRs: Are minisatellites the evolutionary progeny of microsatellites? *Genome*, 37(2):345–347, 1994.

[140] Xiaoping Xia, Rui Rui, Sheng Quan, Rong Zhong, Li Zou, Jiao Lou, Xuzai Lu, Juntao Ke, Ti Zhang, Yu Zhang, et al. MNS16A tandem repeats minisatellite of human telomerase gene and cancer risk: a meta-analysis. *PloS one*, 8(8):e73367, 2013.

[141] Chun Zhang, Guo-Qiang Lv, Xian-Ming Yu, Yuan-Long Gu, Jian-Ping Li, Liang-Feng Du, and Ping Zhou. Current evidence on the relationship between HRAS1 polymorphism and breast cancer risk: a meta-analysis. *Breast cancer research and treatment*, 128(2):467–472, 2011.

[142] Qian Zhang, Se-Ran Jun, Michael Leuze, David Ussery, and Intawat Nookaew. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. *Scientific reports*, 7:40712, 2017.

[143] Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, 3, 2016.