bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Targeted Genotyping of Variable Number Tandem Repeats with adVNTR

Mehrdad Bakhtiari[1], Sharona Shleizer-Burko[2], Melissa Gymrek[1,2], Vikas Bansal[3], and Vineet Bafna*[1]

[1]Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA 92093, USA
[2]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[3]Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA

November 8, 2017

Abstract

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. adVNTR is available at `https://github.com/mehrdadbakhtiari/adVNTR`

Keywords. VNTR, Tandem Repeats, VNTR frameshift, Second-generation sequencing, Third-generation sequencing

*Correspondence: vbafna@eng.ucsd.edu

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 1-6bp are classified as Short Tandem Repeats (STRs), while longer RUs spanning potentially hundreds of nucleotides are denoted as *Variable Number Tandem Repeats* (VNTRs).[46,54]

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[48]. Compared to STRs, which have been extensively studied[21,49,36,52,13], VNTRs have not received as much attention. Nevertheless, multiple studies have linked variation in VNTRs with Mendelian diseases (*e.g.*, Medullary cystic kidney disease[26], Myoclonus epilepsy[29], and FSHD[33]) and complex disorders such as bipolar disorder (Table 1). In some cases, the disease associated variants correspond to point mutations in the VNTR sequence[26,42] 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[41]. Variation in this VNTR has been associated with expression of the INS gene and risk for type 1 diabetes (OR = 2.2)[14]. Notwithstanding these examples, the advent of genome-wide SNP genotyping arrays led to VNTRs being largely ignored. They have been called 'the forgotten polymorphisms'[6].

Gene	Chr	Unit len	Number of units Normal	Number of units Pathogenic	Annotation	Inheritance	Disease
PER3	1	54	4	5	coding	A	Bipolar disorder[3]
MUC1	1	60	11-12	single insertion	coding	M	MCKD1[26]
IL1RN	2	86	3-6	2	intron	A	Stroke, CAD[53]
DUX4	4	3.3kb	11-100	1-10		M	FSHD[33]
DAT1	5	44	7-11	10 (ADHD)	UTR	A	ADHD, Parkinson's, Bipolar[17,27]
MUC21	6	45	26-27	4 bp deletion	coding	A	Diffuse panbronchiolitis (DPB)[24]
CEL	9	33	11-21	single deletion	coding	M	Monogenic diabetes[42]
INS	11	14-15	26-200	26-44 (T1D)	promoter	A	T1D;T2D;Obesity[6,41,14]
DRD4	11	48	2-11	7	coding	A	OCD, ADHD[28,50]
ACAN	15	57	27-33	13-25	coding	A	Osteochondritis dissecans[16]
ZFHX3	16	12	4-5		coding	A	Kawasaki
GP1BA	17	39	1-4	2/3 genotype	coding	A	ATF in Stroke[8]
SERT	17	16-17	9/10/12		intron	A	BPSD, Alzheimer's[23,40]
SERT	17	22	14	16 (OCD)	promoter	A	OCD,Anxiety, Schizophrenia[23]
HIC1	17	70	1-4	5+/5+	promoter	A	Metastatic Colorectal Cancer[39]
MMP9	20	12	5-6		coding	A	Kawasaki
CSTB	21	12	2-3	12+	5'UTR	M	Progressive myoclonic epilepsy 1A[29]
MAOA	X	30	2-5	4	promoter	A	Bipolar disorder[7]

Table 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.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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[20]. Traditionally, VNTR genotyping required labor intensive gel-based screens which limited the size of large population based studies of VNTRs. 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; and (c) *identification of mutations*, specifically indels in coding regions. Mapping tools such as BWA[35] and Bowtie2[31] 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[12], as reads contained within the VNTR sequence have multiple equally likely mappings and therefore will be mapped randomly to different locations with low mapping quality. 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.

Other tools have addressed the problem of RU count estimation, focusing on the related problem of STR genotyping. Some of these tools do not work with large repeating patterns[52,36]. Others require all repeat units to be near-identical[13,49]. In particular, ExpansionHunter[13] 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[20] 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[15], Nanopore[10]) 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, the length-based RU count estimate from a VNTR in the SERT gene included five different values (13, 14, 15, 16, and 18) for a 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

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, or exact estimates for short VNTRs. adVNTR models can estimate RU counts for thousands of VNTRs on PacBio data. 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 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. It has three components: (i) HMM training module for model parameter estimation; (ii) read recruitment; and, (iii) estimating RU counts and variant detection. We describe each of these below.

HMM Training. The goal of training is to estimate model parameters for each VNTR and each sequencing technology. We use an HMM architecture with three parts (Fig. 1). The first part

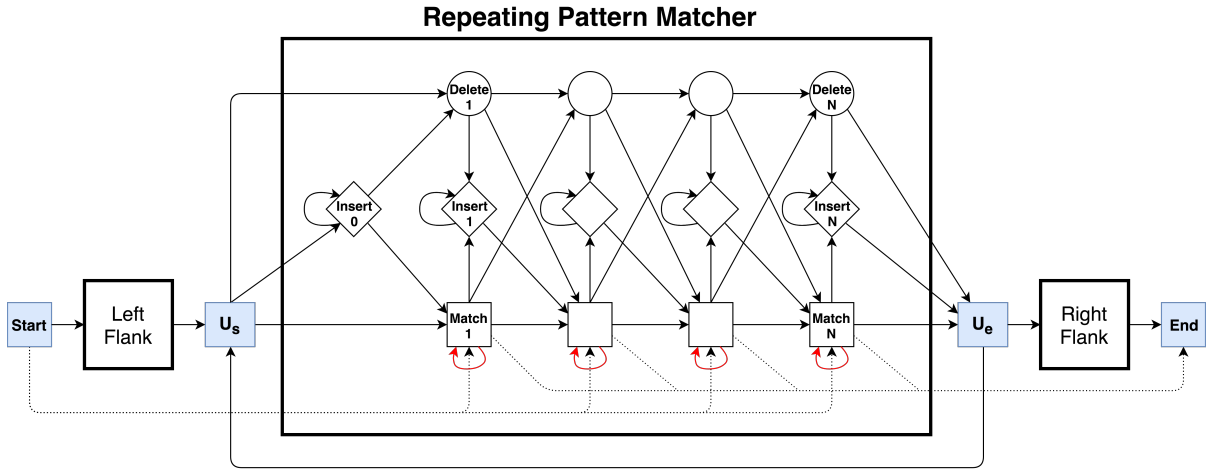

Figure 1: **The VNTR HMM.** The HMM is composed of 3 profile HMMs, one each for the left and right flanking unique regions, and one in the middle to match multiple and partial numbers of RUs. The special states U_s ('Unit-Start'), and U_e ('Unit-End') are used for RU counting. Dotted lines refer to special transitions for partial reads that do not span the entire region.

matches the 5' (left) flanking region of the VNTR. 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. 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[30] and performed

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a multiple sequence alignment. 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 match state i. We define permissible transition (arrows in Fig. 1) and match-state emission probabilities as follows:

$$T(i,j) = \frac{h(i,j) + b_0}{\sum_{i \to l}(h(i,l) + b_0)}, \quad E_i(\alpha) = \frac{h_i(\alpha) + b_1}{\sum_{\alpha'} h_i(\alpha') + Nb_1} \quad \text{for } \alpha, \alpha' \in \{A, C, G, T\}.$$

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,[22,44,2,38,32] and others assume a single haplotype for error correction[45,5]. 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[9]. 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 PacBio model parameters were updated using both simulated and real PacBio data.

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[1] 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 had high sensitivity and were filtered by aligning to the HMM using the Viterbi algorithm. We aligned 10^7 non-target genomic sequences to the HMM to form an empirical null distribution, and used 10^{-5} as the p-value cut-off, correcting for multiple ($\sim 10^3$) target VNTRs.

Estimating VNTR RU Counts. Recall 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

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

HMM. Let, $\text{Prev}_{k,j}$ denote the state j' immediately prior to j in this optimum parse. Then,

$$V_{k,j} = \max_{j'}\{V_{k',j} + \log T(j',j) + \log E_j(s_k)\}, \tag{1}$$

$$\text{Prev}_{k,j} = \arg\max_{j'}\{V_{k',j} + \log T(j',j) + \log E_j(s_k)\}, \tag{2}$$

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 for an example).

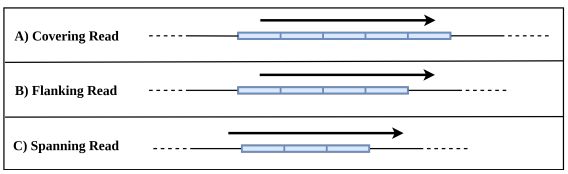

Figure 2: **Estimates of RU counts using recruited reads.** (A) $(k_1,k_2,k_3) = (1,3,1)$; RU count ≥ 5. (B) $(k_1,k_2,k_3) = (0,3,1)$; RU count ≥ 4 (C) $(k_1,k_2,k_3) = (0,3,0)$; RU count $= 3$.

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

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

The analysis of reads at a VNTR gives us a multi-set of RU counts (or lower bounds) $c_1, c_2, \ldots, c_n$. 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:

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

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:

$$\Pr((c_i,c_j)|C) = \frac{\Pr(C|(c_i,c_j))\Pr((c_i,c_j))}{\sum_{i',j'}\Pr(C|(c_{i'},c_{j'}))\Pr((c_{i'},c_{j'}))} \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We generally set equal priors. However, in the event that we only see reads with a single count c', we choose $\Pr((c', c')) = \Pr((c', X)) = \frac{1}{2}$. If we see multiple counts, we set $\Pr((c', X)) = 0$ for all $c' \in C$, and give equal priors to all other genotypes.

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 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[37]. 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

$$-2 \ln \left(\frac{\text{Binom}(d, \iota, \frac{u\ell}{v})}{\text{Binom}(d, \iota, \varepsilon_i)} \right) , \tag{4}$$

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

3 Results

HMM training. Initial HMMs were trained using a multiple alignment of RU sequences from the reference assembly hg19[30], as described in methods. Similarly, HMMs were trained for the left flanking and right flanking regions for each VNTR. To tailor HMMs for short reads, we used WGS data of a CEU trio from 1000 Genomes project and the AJ trio from Genome in a Bottle (GIAB) project. Correspondingly, to train models for PacBio reads, we re-estimated model parameters after aligning PacBio simulated reads using SimLoRD[47]. A total of 865 VNTR models were trained for VNTRs in coding and promoter regions of the genome, for both Illumina and PacBio. Subsequently, we tested performance for (a) read-recruitment, (b) counting of Repeat Units, and (c) detection of indels.

Running time. On PacBio WGS sequencing data at 30X coverage (aligned with Blasr), adVNTR took 10 hours to genotype all 865 VNTRs. For unaligned sequencing data of a CHB individual (20X) coverage, the time increased to 21:45 hours.

Read Recruitment. To evaluate read recruitment performance in *PacBio* sequencing, we simulated haplotypes with 30x coverage using SimLoRD[47] for three disease-linked VNTRs (INS, CSTB,

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

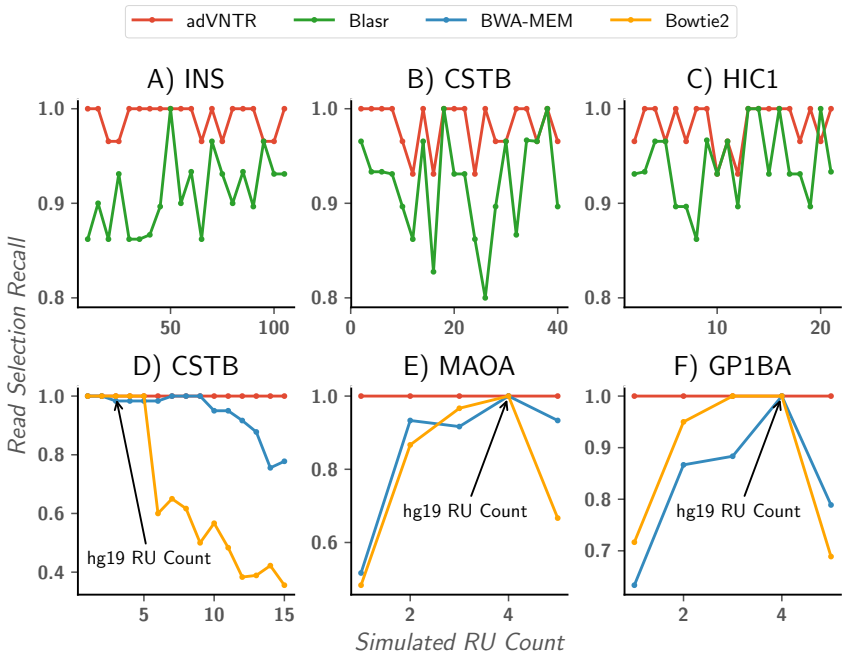

Figure 3: **Sensitivity of read recruitment at VNTR loci.** Comparison of adVNTR read selection with BWA-MEM and Bowtie2 mapping for Illumina reads (short VNTRs) and Blasr for PacBio reads (long VNTRs). Each plot shows the sensitivity of mapped/selected reads as a function of the number of repeats for different VNTRs.

and HIC1) in their known RU count range. To evaluate read recruitment and RU genotyping for *Illumina*, we used ART[25] to simulate haplotype WGS (shotgun 150bp) reads at 30x coverage for three disease-linked short VNTRs using known RU counts. Pairs of haplotypes were merged to get diploid samples. We compared adVNTR with BWA-MEM and Bowtie2 algorithms for assessing Illumina read recruitment (Fig. 3: A-C) and with Blasr[9] for PacBio reads (Fig. 3: D-F). The plots show that while adVNTR works well for a range of RU counts, other mapping tools work well only when the simulated RU count matches the reference RU count.

VNTR RU count estimation with PacBio reads. Recall that sequencing (particularly homopolymer) errors can cause lengths to change, particularly for short RU lengths and larger RU counts. To test the performance of adVNTR for RU counting on haplotype data, we compared against a naive method that estimates RU counts based on read length between the flanking regions. A total of 210 VNTRs with lengths ranging from 10bp to 90bp were selected and for each VNTR, PacBio sequence was simulated for 20 different RU counts with coverage varying from 1-40X. Fig. 4A shows RU count performance on VNTRs in INS, CSTB, and HIC1 genes for varying RU counts and sequence coverage, while Fig. 4B shows RU counting performance on all 210 VNTRs as a function of RU lengths. adVNTR estimates are uniformly good except at low sequence coverage. We tested RU counting on diploid samples by simulating different RU counts on individuals at 3 VNTRS (Table S2). 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 is not known, we checked for Mendelian inheritance consistency at four disease-linked VNTRs in the AJ trio from Genome

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

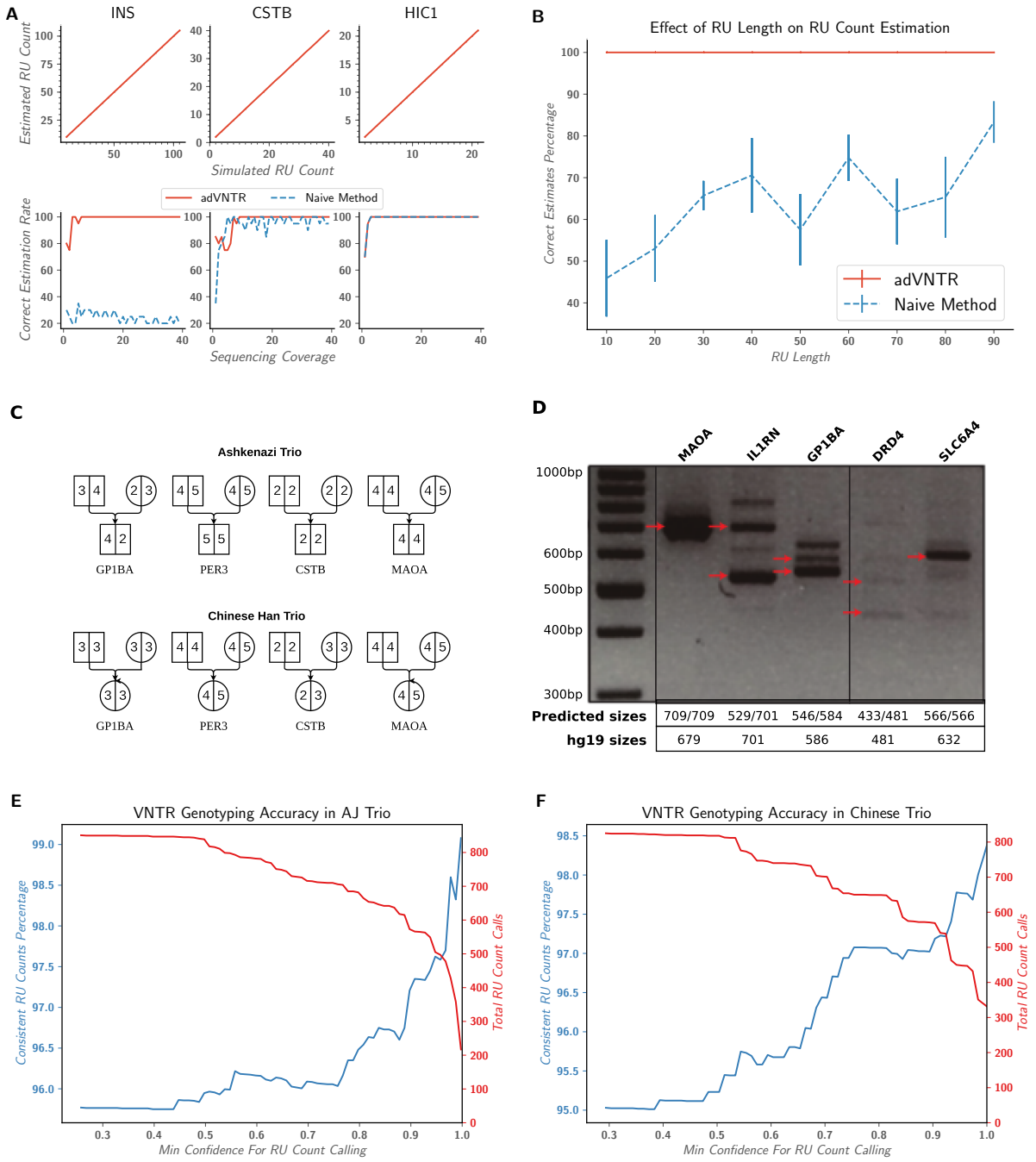

Figure 4: **VNTR genotyping using PacBio data.** (A) RU count estimation on simulated PacBio reads and effect of sequencing coverage on estimation. (B) Effect of repeating unit length on naive method. For shorter repeating units, insertion and deletion errors have more impact on RU count estimation. (C) Mendelian consistency of genotypes at 4 VNTR loci in the Chinese Han and Ashkenazi trios. (D) LR-PCR based validation of genotypes at disease-linked VNTRs in NA12878. (E) RU count calls and consistent calls ratio in AJ trio. (F) RU count calls and consistent calls ratio in Chinese trio.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in a Bottle (GIAB)[55] and a Chinese Han trio from NCBI SRA (accession PRJEB12236). Our predictions were consistent in each case (Fig. 4C). We extended this analysis to 865 VNTRs in the coding region of human genome with RU count $\geq$10. At a posterior probability threshold of 0.95, 99.1% of the calls in the AJ trio, and 98.5% of the calls in the Chinese trio are consistent with Mendelian inheritance. The few discrepancies can be attributed mainly to low coverage and missing data (Fig. 4E,F). Increasing sequence coverage to have $\geq$ 10 reads spanning the VNTR will increase the high confidence RU counts (Suppl. Fig. S1).

We also performed a long range (LR)PCR experiment on the individual NA12878 to assess the accuracy of the adVNTR genotypes. In Fig. 4D, black bands correspond to the observed PCR product lengths, and they match up with the computational product lengths based on estimated RU count (red arrows), while being different from the hg19 reference RU count. For each VNTR, there are two arrows for the predicted heterozygous RU counts, and a single arrow for the SLC6A4 VNTR that was predicted to be homozygous.

While we could not get the VNTR discovery tool VNTRseek[20] to run on our machine (personal communication), we observed that the authors had predicted 125 VNTRs in the Watson sequenced genome[51], and 75 VNTRs in two trios as being polymorphic. In contrast, analysis of the PacBio sequencing data identified >500 examples of polymorphic VNTRs with RU counts $\geq$ 10 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.

RU counting with Illumina. We tested adVNTR RU counting performance on Illumina reads simulated using ART on three VNTRs and compared it to ExpansionHunter[13] which is designed mainly for STRs (Supp. Table S2). Of the 52 samples tested, adVNTR predicted the correct genotype in all but 6 cases, with erroneous calls in the case of high RU counts where the read length does not span the VNTR perfectly. In contrast, ExpansionHunter could not predict a majority of cases as it makes the assumption that the different RUs are mostly identical in sequence (valid for STRs but not for many VNTRs).

For short VNTRs, adVNTR can be an effective tool for larger population-scale studies of VNTR genotypes using WGS data replacing labor intensive gel electrophoresis[7,8]. Fig: 5 shows the RU count frequencies for two disease-linked VNTRs (in the coding region of GP1BA and promoter of MAOA), using 150 PCR-free WGS data obtained from 1000 genomes project[11]. The 2R/3R genotypes in BP1BA are associated with Aspirin Treatment failure for stroke prevention[8]. Notably, our results suggest that the 2R genotype is absent in African populations suggesting that this shorter allele arose after the out of Africa transition.

VNTR mutation/indel detection. To test indel detection, 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[34], and GATK HaplotypeCaller[12] which uses GATK IndelRealigner, to identify frameshifts in each of the simulated datasets, and the 10 control

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

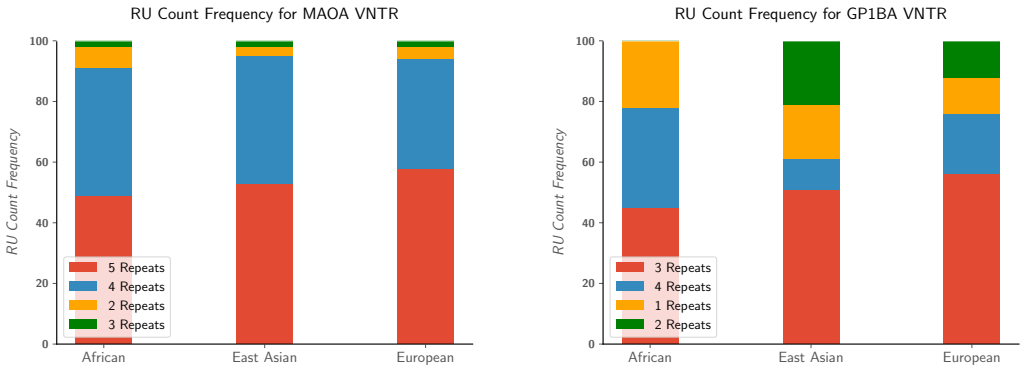

Figure 5: **Population-scale genotyping of VNTRs.** (A) RU count frequencies for MAOA VNTR. In more diverse populations, the dominant genotype has lower frequency than other populations which leads to having more diversity in RU count of the VNTR. (B) RU count frequencies for GP1BA VNTR. Lack of 2R genotype in African population suggests it may have occurred outside of Africa.

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 S3), 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. S2).[43]

As frameshifts in the VNTR region of the CEL gene have been linked to a monogenic form of diabetes[42], we tested for frameshifts in CEL using whole Exome sequencing (WES) data from 2,081 cases with Type 2 Diabetes[18] 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. S3).

4 Discussion

The main contribution of our paper is the separation of VNTR discovery from VNTR genotyping. 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. In this paper, we presented adVNTR to genotype VNTRs using different sequencing methods. Our methods use a trained model for each target, allowing us to tailor the approach for complex VNTRs, and resolves the problem of mapping reads for indel detection by collapsing all RU copies. Like other STR genotyping tools, adVNTR works best when reads can span the VNTR, but (a) indel detection is possible for long VNTRs; (b) RU counting lower bounds can still separate pathogenic cases from normal cases; and, (c) the increasing popularity of long read sequencing (esp. PacBio, and Nanopore) makes it possible to genotype over 800 coding VNTRs. Future research will focus on increasing the number of target VNTRs, and algorithmic strategies to speed up VNTR discovery.

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.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[12] 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.

[13] 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.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Daniela Galimberti, Elio Scarpini, Eliana Venturelli, Alexander Strobel, Sabine Herterich, Chiara Fenoglio, Ilaria Guidi, Diego Scalabrini, Francesca Cortini, Nereo Bresolin, et al. Association of a NOS1 promoter repeat with Alzheimer's disease. *Neurobiology of aging*, 29(9):1359–1365, 2008.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nature genetics*, 45(3):299–303, 2013.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

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

[48] 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.

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

[50] 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.

[51] 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.

[52] 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.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[53] 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.

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

[55] 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.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Material

A. Appendix

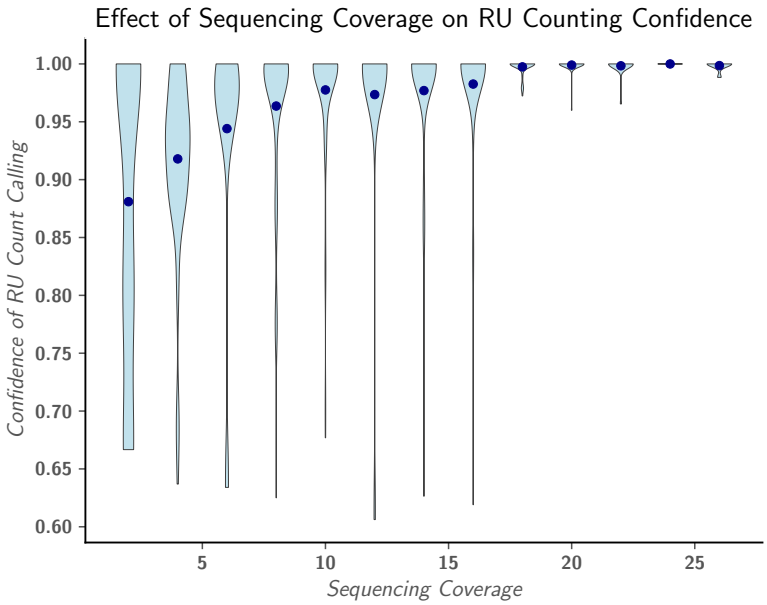

Figure S1: **Association of sequencing coverage in VNTR region and confidence in RU count calling.** The figure shows the confidence of RU count estimation in AJ trio. Most of low confidence calls are resulted from low coverage in VNTR region. With at least 10 reads that span the VNTR, we will get 0.95 confidence in RU count calling.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

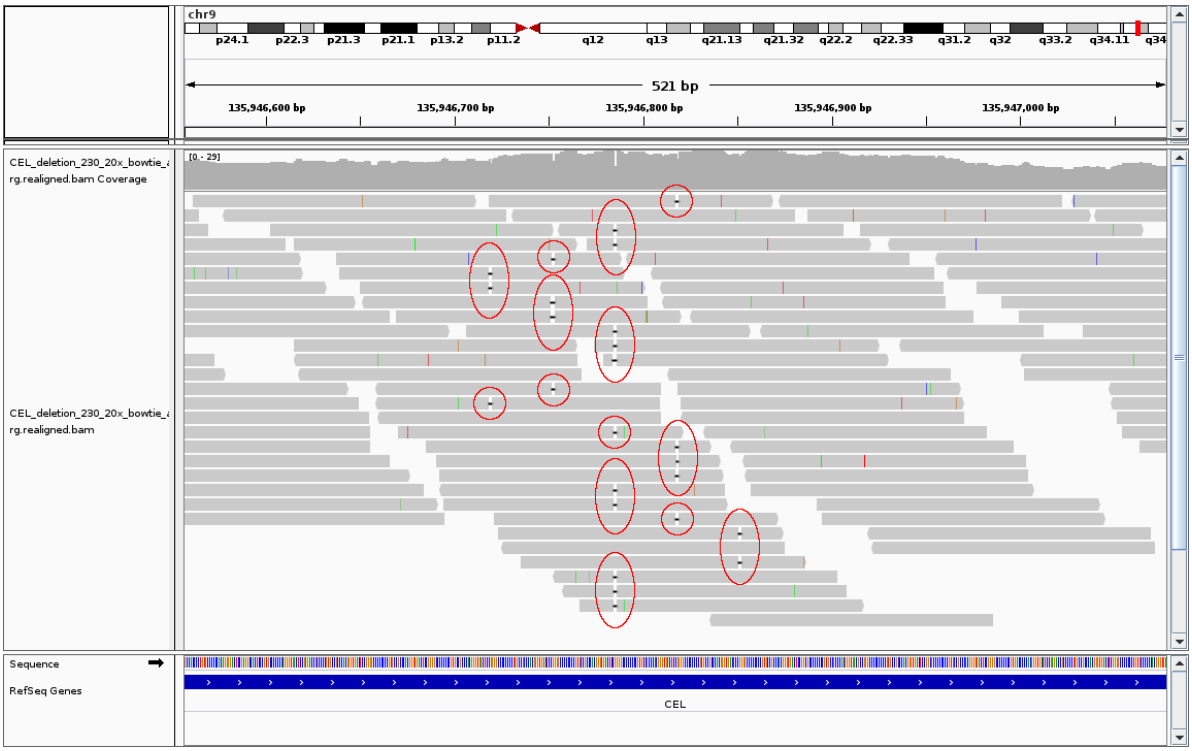

Figure S2: **Alignment stats with frameshift.** Alignment of a simulated data after running GATK IndelRealigner, when there is a deletion. With a sequencing mean of 30X, 25 reads contain the deletion but even after running realigner, deletions are mapped to five different repeating units.

```
Read1                  GGCCACCCCTGTG-CCCCCACAGGGGACTCCGA
Read2                  GGCCACCCCTGTG-CCCCCACAGGGGACTCCGA
Read3                  GGCCACCCCTGTG-CCCCCACAGGGGACTCCGA
Read4                  GGCCACCCCTGTG-CCCCCACAGGGGACTCCGA
Read5                  GGCCACCCCTGTG-CCCCCACAGGGGACTCCGA
Read6                  GGCCACCCCTGTG-CCCCCACAGGGGACTCCGA
ReferenceRepeatingUnit GGCCACCCCTGTGCCCCCCACAGGGGACTCCGA
                       ************* *******************
```

Figure S3: **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 14X and 6 reads support the deletion.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		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
CSTB	2/10	0/0	3/-5	0/0
CSTB	2/12	0/0	3/-7	0/0
CSTB	2/14	0/0	3/-9	0/-1
CSTB	3/3	0/0	-1/-1	0/0
CSTB	3/4	0/0	2/1	0/0
CSTB	3/6	0/0	2/-1	0/0
CSTB	3/8	0/0	2/-3	0/0
CSTB	3/10	0/0	2/-5	0/0

Table S2: **RU count genotyping results on simulated data.** For two cases, (MAOA 1/1 and CSTB 1/1) Expansion Hunter doesn't find any RU count.

bioRxiv preprint doi: https://doi.org/10.1101/221754; this version posted November 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		# 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

Table S3: Comparison of indel finding with Samtools