Published online DD Month YYYY

Nucleic Acids Research, YYYY, Vol. XX, No. XX **1–13**
doi:10.1093/nar/gkn000

Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences

Marzieh Eslami Rasekh [1], Yozen Hernandez [1], Samantha D. Drinan [2], Juan Fuxman Bass [1,2], Gary Benson [1,2,3] *

[1] Graduate Program in Bioinformatics, Boston University, Boston, MA 02215, USA , [2] Department of Biology, Boston University, Boston, MA 02215, USA , [3] Department of Computer Science, Boston University, Boston, MA 02215, USA

ABSTRACT

Variable Number Tandem Repeats (VNTRs) are tandem repeat (TR) loci that vary in copy number across a population. Using our program, VNTRseek, we analyzed human whole genome sequencing datasets from 2,770 individuals in order to detect minisatellite VNTRs, i.e., those with pattern sizes ranging from 7 bp to 126 bp, and with array lengths up to 230 bp. We detected 35,638 VNTR loci and classified 5,676 as common (occurring in >5% of the population). Common VNTR loci were found to be enriched in genomic regions with regulatory function, i.e., transcription start sites and enhancers. Investigation of the common VNTRs in the context of population ancestry revealed that 1,096 loci contained population-specific alleles and that those could be used to classify individuals into super-populations with near perfect accuracy. Comparison of genotyping results with proximal genes indicated that in 120 cases (118 genes), expression differences correlated with VNTR genotype. We validated our predictions in several ways, including experimentally, through identification of predicted alleles in long reads, and by comparisons showing consistency between sequencing platforms. This study is the most comprehensive analysis of minisatellites VNTRs in the human population to date.

INTRODUCTION

Over 50% of the human genome consists of repetitive DNA sequence (9, 10) and tandem repeats (TRs) comprise one such class. A TR consists of a pattern of nucleotides repeated two or more times in succession. TR loci are defined by their position on the genome, the sequence and length of their repeat unit, and copy number.

TRs are commonly divided into three classes based on pattern size: short tandem repeats (STRs), or microsatellites, with pattern lengths of six or fewer base pairs (bp), minisatellites with patterns ranging from seven to several hundred bp, and macrosatellites with patterns from hundreds to thousands of bp (11, 12).

This study focuses on the TR minisatellite class, which comprises more than a million loci in the human genome. While many minisatellite loci appear to be monoallelic with regard to copy number, a significant fraction exhibit copy number variability and are called Variable Number Tandem Repeats (VNTRs). Changes in VNTR copy number have been proposed to arise by slipped strand mispairing (13, 14, 15), unequal crossover (16, 17), and gene conversion (16, 18).

VNTRs are highly mutable, with germline mutation rates estimated between 10^{-3} and 10^{-7} per cell division (19, 20, 21, 22, 23). This mutation rate, which far exceeds that of SNPs, makes VNTRs useful for DNA fingerprinting (24, 25, 26). VNTRs have also been predicted to have high heterozygosity, ranging from 43% to 59% (27), and the copy numbers of several VNTR loci have been shown to be population-specific in humans (28, 29), suggesting that these VNTRs may be useful for population wide studies.

The Eichler group (30) has examined TR loci on human and ape genome assemblies from PacBio sequencing data and identified 1,584 human-specific VNTR loci with 52 as candidate regions associated with disease. Additionally, by comparing VNTR loci, situated in known gene enhancers, with RNA sequencing data, the authors found that expansion of VNTRs correlated with up-regulation of the corresponding genes, suggesting that TR copy number might modulate enhancer activity.

More than half of previously identified human VNTR loci (31) are located near or within genes (32) and some occur within coding exons (33, 34), so their potential effects on gene expression or the protein products are substantial. VNTRs have been shown to contain binding sites for transcription factors such as NF-κB and myc/HLH (35, 36), have been associated with changes in levels of gene expression (37), including tissue specific expression (38), and may cause splicing differences (39, 40).

VNTRs have been proposed as drivers of phenotypic variation in evolution (30, 41, 42). For example, VNTR polymorphisms may play a role in neurogenesis and account for "human-specific cognitive traits" (37). Furthermore, minisatellite VNTRs have been associated with a variety of diseases (43, 44, 45), including cancer (46, 47, 48, 49, 50), neurodegenerative disorders (51) such as Alzheimer's

*To whom correspondence should be addressed. Email: gbenson@bu.edu

disease (39, 52, 53) and Huntington's disease (36, 54), and other psychiatric conditions, such as PTSD (55), ADHD (56, 57), depression (58), and addiction (59).

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

Nonetheless, genotyping variability in repeat sites remains challenging (60, 61). Although a number of tools have been designed to detect microsatellite copy number variability such as lobSTR (62), popSTR (63), hipSTR (64), and ExpansionHunter (65), very few high-throughput tools are available for minisatellite genotyping. The adVNTR tool (66) trains a Hidden Markov Model (HMM) for each VNTR locus of interest and has been used to predict variability in 2,944 VNTRs intersecting coding regions.

VNTRseek (67), developed in our lab, uses the Tandem Repeats Finder (TRF) (68) to detect and characterize TRs inside reads and then maps read TRs to TRs in a reference set. Because it builds pattern profiles before mapping, VNTRseek is robust in the presence of SNPs and small indels.

In this paper, we present the most comprehensive catalog of minisatellite VNTRs in the human genome to date, pooling results for WGS datasets from 2,770 individuals, processed with VNTRseek on the GRCh38 human reference genome. We report a large collection of previously unknown VNTR loci, find that many VNTR loci and alleles are common in the population, show that VNTR loci are enriched in gene and regulatory sequences, provide evidence of gene expression differences correlated with VNTR genotype and evidence of population-specific VNTR alleles.

MATERIALS AND METHODS

Datasets. Datasets comprising 2,801 PCR-free, WGS samples from 2,770 individuals were used in this study (Table 1): 30 individuals from the 1000 Genomes Project Phase 3 (1), including the Utah (CEU) and Yoruban (YRI) trios (mother-father-child); 2,504 unrelated individuals mostly overlapping with the 1000 Genomes Project, recently sequenced at $>30\times$ coverage by the New York Genome Center (NYGC); 253 individuals from the Simons Genome Diversity Project (SGDP) (2), seven individuals sequenced by the Genome in a Bottle (GIAB) Consortium (3), including the Chinese (HAN) and Ashkenazi Jewish (AJ) trios and NA12878 (with ID HG001); two "haploid" hydatidiform mole cell line genomes, CHM1 (4) and CHM13 (5); tumor/normal tissues from two unrelated individuals with breast cancer (breast invasive ductal carcinoma cell line/lymphoblastoid cell line) from the Illumina Basespace public WGS datasets (6); and the AJ child sequenced with PacBio Circular Consensus Sequencing (CCS) reads (7). Duplicates of 27 genomes were present in two datasets, 1000 Genomes and NYGC. One of these, NA12878, was also included in the GIAB dataset.

Overall, read coverage ranged from approximately 27x, in the PacBio sample to 333x, in the GIAB Chinese child. Besides the PacBio data, reads consisted of three lengths, 100/101 bp (257 samples), 148/150 bp (2,508 samples), and 250 bp (35 samples). All data were downloaded as raw fastq files, except for the PacBio data which were obtained as a BAM file with reads aligned to GRCh37. SRA links to the data are given in Table 4.

The majority of the analyses in this paper were performed on the 2,504 genomes from NYGC. The 253 genomes from SGDP provided insight into under-represented populations. The 27 genomes duplicated in the 1000 Genomes and NYGC datesets were used to measure consistency across sequencing platforms. The trios from the 1000 Genomes (CEU and YRI) and GIAB (AJ and Chinese HAN) datasets were used for

Data source		Read Length (bp)	Read Coverage	Samples in Set	Ref.
1000 Genomes Phase 3 HC	Yoruban (YRI) trio	250	71–73×	3	(1)
	Utah (CEU) trio	250	55–63×	3	
	Others	250	33–66×	24	
New York Genome Center (NYGC)		150	29–101×	2,504	*
Simon's Genome Diversity Project (SGPD)		100	33–133×	253	(2)
Genome In A Bottle (GIAB)	Ashkenazim Jewish (AJ) trio	250	61–69×	3	(3)
	Chinese (HAN) trio	148 / 250	111–333×	3	
	NA12878 (HG001)	148	291×	1	
Haploid genomes	CHM1	148	40×	1	(4)
	CHM13	250	128×	1	(5)
Illumina basespace	Tumor/Normal	101	38–88×	4	(6)
PacBio	Ashkenazi Jewish (AJ) child	∼13,500	27×	1	(7)

Table 1. Data. 2,801 publicly available WGS samples, for 2,770 individuals, were used in this study. Read coverage was calculated as the product of the number of reads and the average read length, divided by the haploid genome size, as in the Lander/Waterman equation (8). All coverage values are approximate. The 1000 Genomes Phase 3 samples were released in 2015. The NYGC samples were released in 2020 by the New York Genome Center (NYGC). For the Simons Genome Diversity Project (SGDP), released in 2016, only datasets which were not present in the the 1000 Genomes datasets were used. The PacBio data were used only for comparison and validation purposes but not for our VNTR results. *These data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

analyzing Mendelian inheritance. The cancer datasets were used to find possible changes in VNTRs in tumor tissues. The PacBio data were used for validation purposes only.

Curating the reference TRs. The 22 autosomes and sex chromosomes from the human reference genome GRCh38 (69) were used to produce a reference set of TRs in TRDB (70) with the TRF software and four quality filtering steps as described in (67). In addition, centromere regions were excluded from the reference set. These filtering tools are available online in TRDB. Starting with 1,199,362 TRs found by TRF, we curated a filtered reference set with 228,486 TRs. Using VNTRseek, we classified the TRs into two subcategories, singletons and indistinguishables (Supplementary Section S2.1). A *singleton* TR appears to be unique in the genome based on a combination of its repeat pattern and flanking sequence. An *indistinguishable* TR belongs to a family of genomically dispersed TRs which share highly similar patterns and flanking sequence and may therefore produce misleading genotype calls. Indistinguishable TRs (total 37,200 or about 16% of the reference set) were flagged. Simulation testing revealed that some singletons produced false positive VNTRs. To minimize this issue, an additional filtering step was added to eliminate problematic singleton loci from the reference set (see Supplementary Section S2.2 and Supplementary Material `Reference_TRs.txt` for the reference sets).

We assumed that genotyping was possible for a reference TR locus, given a particular read length, if the TR array length plus a minimum 10 bp flank on each side, would fit within the read. The number of reference TR alleles that could be genotyped using each of the read lengths in our data is summarized in Table S1.

Genotyping TRs and VNTRs. Each dataset was processed separately with VNTRseek using default parameters: minimum and maximum flanking sequence lengths of 10 bp and 50 bp, respectively, on each side of the array, and requiring at least two reads mapped with the same array copy number to make an allele call. Output from VNTRseek included two VCF files containing genotype calls, one reporting all detected TR and VNTR loci, and the other limited to VNTR loci only. The VCF files contained two specialized FORMAT fields: SP, for number of reads *supporting* each allele, and CGL, to indicate genotype by the number of *copies gained or lost* with respect to the reference. For example, a genotype of 0 indicated detection of only the TR reference allele (zero copies gained or lost), while $0,+2$ indicated a heterozygous locus with a reference allele and an allele with a gain of two copies.

To remove clear inconsistencies, for this study we filtered the VCF files to remove *per sample* VNTR loci with more alleles than the expected number of chromosomes. The filtering criteria for these loci, termed *multis* is detailed in Supplementary Section S2.3. After multi filtering, a TR locus was labeled as a VNTR if any remaining allele, different from the reference, was observed in any sample.

Experimental validation. Accuracy of VNTRseek genotyping was experimentally tested for 13 predicted VNTR loci in the Ashkenazi Jewish (AJ) trio. The following DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: NA24385, NA24149, and NA24143 (also identified as GIAB IDs HG002, HG003, and HG004). Selection criteria required that the PCR product was not contained in a repeat region, unique primers could be designed, the primer-defined allele length difference was between 10% and 20% of the longest allele, and the primer-defined GC content was between 40% and 60%. Given these criteria, we prioritized VNTRs in genes and regulatory regions which might be of interest to researchers. Primers were designed with Primer-BLAST (71) and used to amplify the VNTR loci from the genomic DNA of each individual using the following reagents: 0.2 μL 5 U/μL DreamTaq DNA Polymerase (ThermoFisher Scientific), 4.0 μL 10X DreamTaq Buffer (ThermoFisher Scientific), 0.8 μL 10mM dNTP mix (ThermoFisher Scientific), 3.2 μL primer mix at a final concentration of 0.5 μM, 1.6 μL genomic DNA (40 ng), and 30.2 μL nuclease-free water. PCR cycling conditions were as follows: 30 seconds at 95°C, 30 seconds at 56-60°C, 20 seconds at 72°C, for 30 cycles, with an initial denaturation of 3 min at 95°C and a final extension of 7 min at 72°C. The resulting amplicons were run on a 2% agarose gel at 100 V for 2 hours and visualized with UV light using Ethidium Bromide. A complete list of loci and primers is given in the Supplementary Material as `Experiment_primers.txt`.

In addition, VNTRseek predictions in the NA12878 genome were compared to experimental validations in the paper describing the adVNTR software (66). We had three datasets for NA12878; HG001 (148 bp) from GIAB, NA12878 (150 bp) from NYGC, and NA12878 (250 bp) from 1000 Genomes. The adVNTR predictions used GRCh37 coordinates which were converted using the UCSC liftover tool (72) to coordinates to GRCh38.

Validation using long reads. Aligned PacBio reads for the AJ child (GIAB ID HG002) were processed to validate VNTRseek predictions. The read sequences were extracted from the BAM file using picard tools (73) and mapped back to the GRCh38 genome using BWA MEM default settings (74). Using bedtools (75), the reads aligning to each TR reference locus were extracted. For each read, a local wraparound dynamic programming alignment was performed using the reference pattern and the same scoring parameters used to generate the reference set (match$=+2$, mismatch$=-5$, and gap$=-7$). The number of copies of the pattern in the resulting alignment was then determined and compared with the VNTRseek predictions. If the difference between a PacBio copy number in at least one read and the VNTRseek copy number was within ± 0.25 of a copy, we considered the VNTRseek allele to be validated.

Measuring consistency of Mendelian inheritance. A locus on an autosomal chromosome is consistent with Mendelian inheritance if the genotype of a child can be explained as one allele from the mother and one from the father. Genotype consistency was evaluated for all mother-father-child trios, i.e., the AJ, CEU, HAN, and YRI trios. We evaluated loci defined by several increasingly stringent criteria: both parents heterozygous, all members of the trio heterozygous, all members of the trio heterozygous and with different genotypes. These criteria were selected to avoid false interpretations of consistency.

TR loci on the X and Y chromosome of male children were also selected for evaluation when both the son and the appropriate parent had a predicted genotype. In these cases,

inheritance consistency means a son's X chromosome allele is observed on one of the mother's X chromosomes, and a son's Y chromosome allele is observed on the father's Y chromosome.

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

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

Common and private VNTRs. To classify common and private VNTRs, we used results from the NYGC dataset (2,504 individuals) as the read length and coverage were comparable across all genomes. Additionally, these genomes contain no related individuals and represent a wide set of populations (26 populations from five continents). VNTR loci were classified as common if they were identified as VNTR in at least 5% (126) of the individuals and classified as private if they were identified as VNTR in less than 1% (25).

Annotation and enrichment. Annotation based on overlap with functional genomic regions was performed for the reference TR loci. Genomic annotations for GRCh38 were obtained from the UCSC Table Browser (76) in BED format. Known gene transcripts from GENCODE V32 (77) were used along with tracks for introns, coding exons, and $5'$ and $3'$ exons. Regulatory annotations included transcription factor binding site (TFBS) clusters (78, 79) and DNAse clusters (80) from ENCODE 3 (81), and CpG island tracks (82), comprising 25%, 15%, and 1% of the genome, respectively. Bedtools (75) was used to find overlaps between TR loci and the annotation features. Any size overlap was allowed.

LOLAweb (83) was used to determine VNTR enrichment for genomic regions in comparison to the background TR annotations, and common and private VNTR enrichment in comparison to all VNTR annotations. TRs on the sex chromosomes were excluded in the background set. To identify gene and pathway functions that could be affected by common VNTR copy number change, genes with exons or introns overlapping with common VNTRs were collected and their enrichment computed using GSEA (84) for Gene Ontology (GO) terms (85) for biological process and KEGG pathways (86) with FDR p-value ≤ 0.05.

Association of VNTR alleles with gene expression. To detect expression differences among individuals with different genotypes, mRNA expression from lymphoblastoid cell lines of 660 individuals by the Geuvadis consortium (Accession: E-GEUV-1) were downloaded (87). A total of 445 individuals overlapped with the 2,504 genomes set. We paired common VNTR loci with overlapping genes, including 1 Kb upstream or downstream (see annotation), and extracted the genotypes for each individual at those VNTRs. When no genotype was observed for an individual, we classified the genotype as *other* (assuming that the actual alleles were outside the detection range of VNTRseek because genotypes were observed in other individuals with similar coverage). VNTR–gene pairs were selected for analysis if more than one genotype was detected for that VNTR across all individuals (at least three if *other* was one of the genotypes) and if each genotype was observed in at least 20 individuals. Loci on the sex chromosomes were excluded to avoid confounding with sex-related gene expression differences.

For the genotypes in each selected pair, a linear regression on the $\log_2$(RPKM) normalized expression values was used to calculate effect size and p-values for the genotype classes, with the smallest p-value retained. We adjusted the p-values to correct for over-testing using FDR and selected the VNTR-gene pairs with FDR < 0.05.

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

In order to find super-population markers among the common VNTRs, a one-sided Fisher's exact test was used to calculate the odds ratio and p-value of each allele being in one super-population versus being collectively in all the others. We only considered alleles over-represented rather than both over- and under-represented because of an interest in identifying alleles that have a phenotypic effect. Odds ratio values were $\log_2$ transformed and p-values were adjusted for false discovery rate (FDR) (88). Any allele with FDR < 0.05 and $\log_2$(odds ratio) > 1 was chosen as a significant marker for that population.

RESULTS

In this section, we start with a summary and characterization of our VNTR predictions, followed by identification of commonly occurring VNTRs and an enrichment analysis of their association with genomic functional regions and genes sets. We next report on the effect of VNTRs on expression of nearby genes, and then identify population-specific VNTR alleles and show that they are predictive of ancestry. We conclude this section with evidence confirming the accuracy of our predictions using several validation methods.

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

WGS datasets from 2,770 human genomes were analyzed with VNTRseek to detect VNTRs. Overall, 184,315 out of 191,286 singleton reference TR loci ($\sim$96%) were genotyped across all samples (Table 2) while 5% of the loci had TR arrays too long

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

Table 2. TRs and VNTRs detected, by dataset. *TRs Genotyped* is the number of distinct TR loci genotyped across all individuals within a dataset. (All other numbers are also per dataset.) *Multis* are TR loci genotyped in a single individual with more than the expected number of alleles. They could be artifacts or indicate copy number variation in a genomic segment. Multis were excluded from further analysis on a per sample basis. *VNTRs Detected* is the number of TR loci, excluding multis, with a detected allele different from the reference.

to fit within the longest reads and could only be genotyped if they lost a sufficient number of copies.

A total of 5,198,392 VNTRs were detected, corresponding to 35,638 (∼19%) distinct VNTR loci, indicating wide occurrence of these variable repeats. Their occurrence within genes was common, totaling 7,698 protein coding genes, and 3,512 exons. The resulting genotypes were output in VCF format files (see Data Availability Section) and summarized for each genome (Supplementary Material `Summary_of_results.txt`). A website is under development to view the VNTR alleles (`http://orca.bu.edu/VNTRview/`).

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

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

VNTR detection was similarly dependent on coverage and read length, as shown in Figure 1b. However, detection was also positively correlated with population, which seemed likely due to the evolutionary distance of populations from the reference genome, which is primarily European (89). For example, in the 250 bp trios with comparable coverage, the African Yoruban genomes (YRI) had the highest number of VNTRs, followed by the Ashkenazi Jewish genomes (AJ), and finally, the Utah genomes (CEU). Notably, within each trio, the VNTR counts were similar.

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

More than two alleles are common in VNTRs

Two alleles were detected in the majority of VNTR loci across all datasets (Figure 1c). However at 10,698 loci (29%), three or more alleles were detected. In a substantial number of loci (5,395), the reference allele was never seen, but only 105 of these were in the VNTRseek detectable range for the 150 bp and 100 bp reads, which made up the bulk of our data. Intersetingly, in 1,166 loci, the reference genotype, although detectable, was not the major allele (Supplementary Material `Major_genotypes.txt`)?

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

Overall, VNTRseek found approximately 1.8-fold more alleles with copy losses (3,444,128), with respect to the reference copy number, than gains (1,958,250). Loss of one copy (2,263,608) was the most common type of VNTR polymorphism (Figure 1d). The overabundance of VNTR copy loss may actually be an underestimate. Because VNTRseek required a read to span a TR array for it to be detected, only limited gain in copies could be observed. Observing gain of one copy would have been possible in approximately 68%, 82%, and 92% of loci for samples with read lengths of 100bp, 150bp, and 250bp, respectively. By contrast, the reference locus needed to have a minimum of 2.8 copies for a loss of one copy to be observed by TRF, and only 16% of the reference loci met this criterion. Higher observed copy loss could be explained by a bias in the reference genome towards including higher copy number repeats, or by an overall mutational preference for copy loss.

VNTRs have high heterozygosity

High heterozygosity in human populations suggests higher genetic variability and may have beneficial effects on a range of traits associated with human health and disease (90). Since calculating heterozygosity for VNTRs is not straightforward (because of limitations on discovering alleles, especially within shorter reads), we used the percentage of detected, per-sample heterozygous VNTRs as an estimate for heterozygosity. At read length 250 bp, per-sample heterozygous VNTR loci comprised approximately 46-55% of the total, which is comparable to previous theoretical estimates of 43–59% (27). At shorter read lengths, the bottom of the range extended lower (∼38–57% for 150 bp reads, ∼29–51% for 100 bp reads, Supplementary Figure S1), as expected, because longer alleles were undetectable if they did not fit within a single read.

Interestingly, despite the previous comment, within genomes that were comparable in read length and coverage, the fraction of heterozygous loci clustered within populations (Supplementary Figures S2, S3, and S4), with African genomes generally having more heterozygous calls and East Asians fewer. This result is consistent with previous findings

of population differences in SNP heterozygosity among Yoruban and Ashkenazi Jewish individuals with respect to European individuals (91, 92), and suggests higher genomic diversity among African genomes, as has been previously noted (93).

Loss of heterozygosity observed in tumor samples

A significant loss of heterozygosity (LOH) was observed in predicted VNTRs of one of the tumor tissues compared to its matching normal tissue (sample ID HC1187). The percentage of heterozygous VNTRs was roughly double in the normal tissue ($\sim$38% vs $\sim$19%) (Supplementary

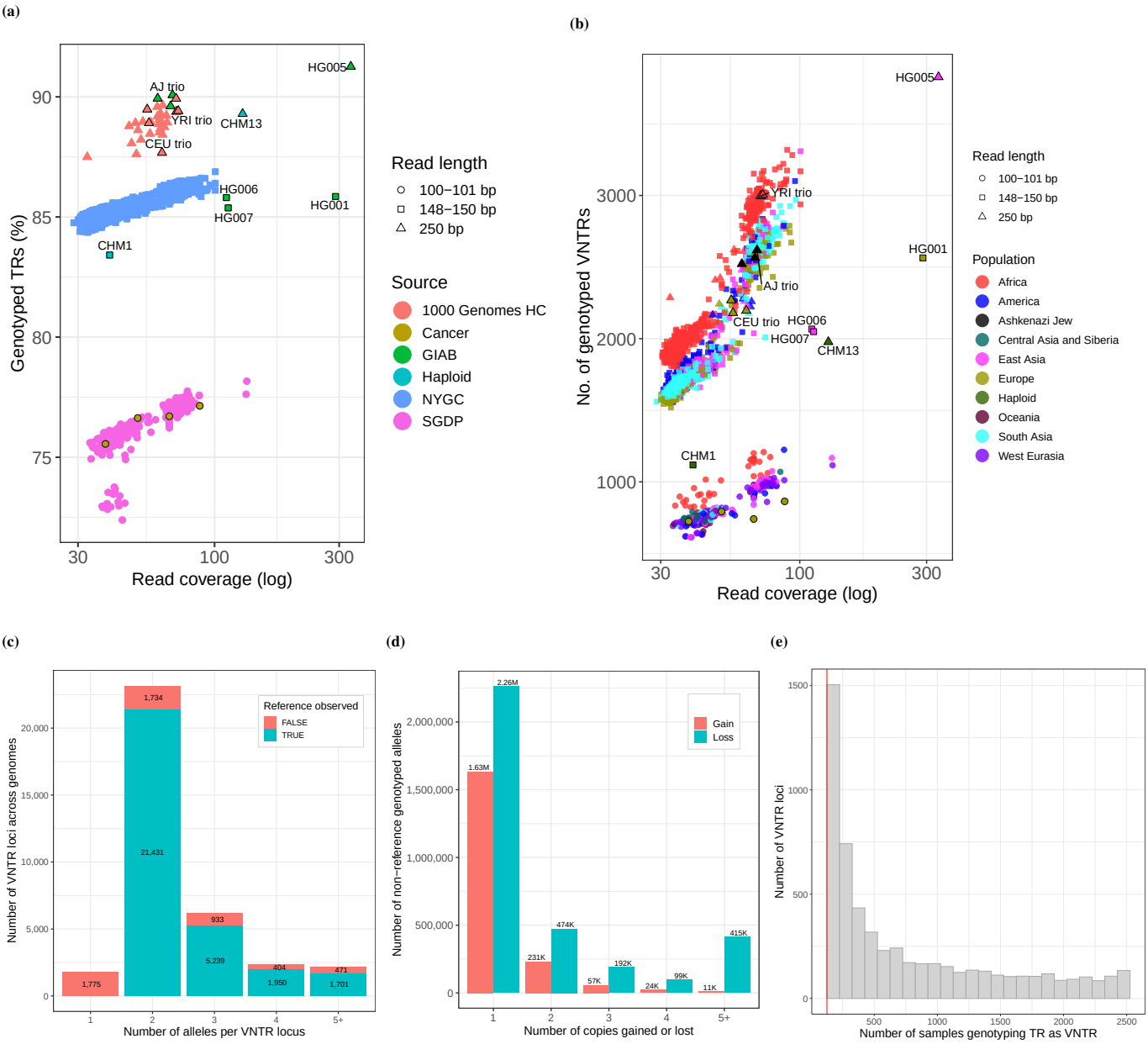

Figure 1. **(a) TR Genotyping sensitivity.** Graph shows the relationship between coverage, read length, and the percentage of TRs in the reference set that were genotyped. Each symbol represents a single sample and specific samples are labeled. Increasing read length had the largest effect on sensitivity because many reference TR alleles could not be detected at the shorter read lengths (see Supplementary Table S1). **(b) VNTRs detected.** Graph shows the relationship between coverage, read length, and the number of VNTR loci detected. Read length and coverage both had large effects. Coloring of symbols shows that population also had a strong effect, reflecting distance from the reference, which is primarily European. Note the reduced numbers for CHM1 (150 bp) and CHM13 (250 bp). Because they are "haploid" genomes, parental heterozygous loci with one reference allele would appear to be VNTRs, on average, only about half the time. **(c) Alleles detected per locus.** Each bar represents a specific number of alleles detected across all datasets. Coloring shows that proportion of loci where the reference allele was or was not observed. **(d) Copies gained or lost.** Each bar represents a specific number of copies gained or lost in non-reference VNTR alleles relative to the reference allele. Loss was always more frequently encountered. **(e) VNTR locus sample support.** Data shown are the *common* loci from the 2,504 sample NYGC dataset. Each bar represents the number of samples calling a locus as a VNTR. Bin size is 100. Bar height is number of loci with that sample support. Red line indicates the 5% cutoff for common loci (126 samples).

Tables S2, S3). Extreme loss of heterozygosity in small variants has previously been reported in these samples by Illumina Basespace (94) with the number of heterozygous small variants in HC1187 being four times lower in the tumor tissue compared to the normal. Taken together, these results suggest that VNTR LOH could be linked to tumor progression.

Additionally, in both tumors a large number of loci exhibited loss of both alleles in comparison to the normal tissue (Table S2). Given that the coverage for the tumor samples was significantly higher than for the normal tissue, it is unlikely that these observations were due to artifacts. Also, the tumor samples did not show a higher percentage of filtered multi VNTRs (too many alleles) than the normal samples (1.37% and 1.23% in normal tissue vs 1.72% and 1.71% in tumor tissue).

Common vs. private VNTRs

Following methodology used with SNPs (95), we classified VNTRs in the 2,504 healthy, unrelated individuals from the NYGC dataset (150 bp and coverage >30x) as common if they occurred in at least 5% of a population (126 individuals) and private if they occurred in less than 1% (25 individuals).

We classified 5,676 VNTRs as common (17% of the 33,403 VNTRs detected in this population) and 68% as private. Each sample averaged 1,783 common VNTRs (median 1,677) and 46 private VNTRs (median 17). A total of 3,627 common VNTRs overlapped with 2,173 protein coding genes including 254 exons. Interestingly, increasing the threshold for common VNTRs did not reduce the number dramatically (Supplementary Figure S5), suggesting that these VNTRs have not occurred randomly, but rather have undergone natural selection. Widespread occurrence of common VNTRs indicates a fitness for use in Genome Wide Association Studies (GWAS). A list of common and private VNTRs can be found in Supplementary Material `Common_VNTRs.txt` and `Private_VNTRs.txt`.

Common VNTR enrichment in functionally annotated regions

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

Our reference TR set comprised only 0.52% of the genome, however, 49% of human genes contained at least one TR and 5% of all the TFBS clusters overlapped with TRs. Moreover, high proportions of our TR reference set and common VNTRs intersected with genes (63% and 64% respectively), TFBS clusters (38% and 51%), and DNAse clusters (21% and 28%) (Table S5).

In comparison to TRs, VNTR loci were positively enriched in 1 Kbp upstream and downstream regions of genes, 5′ and 3′ UTRs, coding exons, TFBS clusters, DNAse clusters, and CpG islands (p-values<0.05) (Supplementary Tables S4 and S5). The common VNTRs, on the other hand, compared to all VNTRs, were enriched in 1 Kbp upstream regions of genes, TFBS, and CpG islands, suggesting regulatory function. Private VNTRs were less likely to occur in 1 Kbp upstream or downstream regions, inside TFBS clusters, open DNAse clusters, or CpG islands.

Focusing on the common VNTRs, we used the LOLAweb (96) online tool to perform enrichment analysis with various curated feature sets (Supplementary Section S4.2), Among the results, DNAse enrichments by tissue type (97) pointed to brain, muscle, epithelial, fibroblast, bone, hematopoietic, cervix, skin, and endothelial

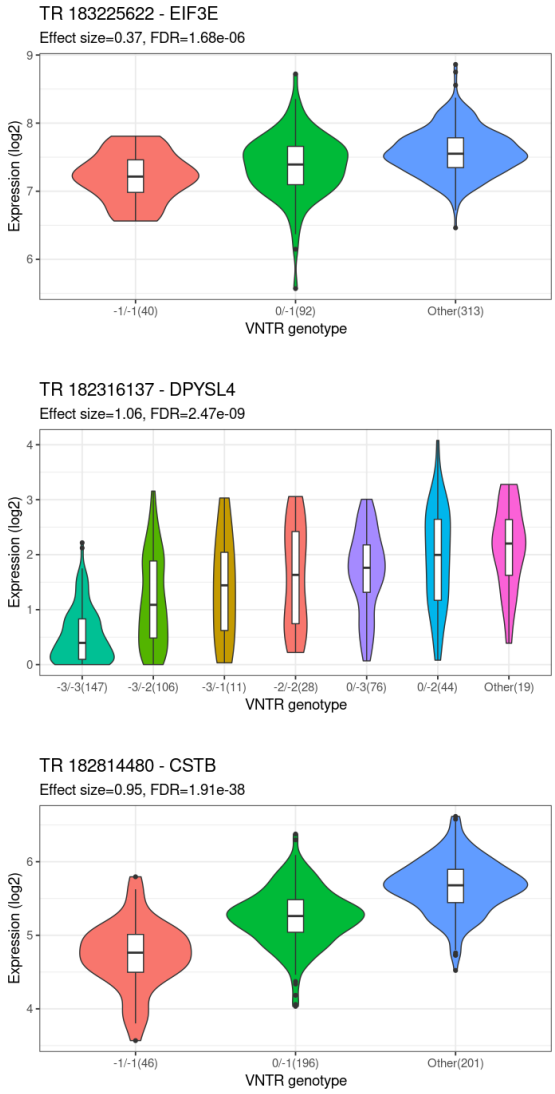

Figure 2. Gene expression differences and VNTR genotype. Shown are violin plots of gene expression values for three genes which displayed significant differential expression when samples were partitioned by VNTR allele genotype. Additional examples are shown in Supplementary Figures S22-S26. Genotype is indicated in labels on the x-axis and numbers refer to copies gained or lost relative to the reference allele. "Other" indicates a partition with undetected alleles presumed outside the range of VNTRseek detection (see text). Number of samples in each partition is shown in parenthesis. In these examples, the effect size for at least one genotype class was significant. *Top*: VNTR 183225622 is downstream of EIF3E. *Middle*: VNTR 182316137 occurs inside the first intron of DPYSL4. *Bottom*: VNTR 182814480 occurs upstream of CSTB.

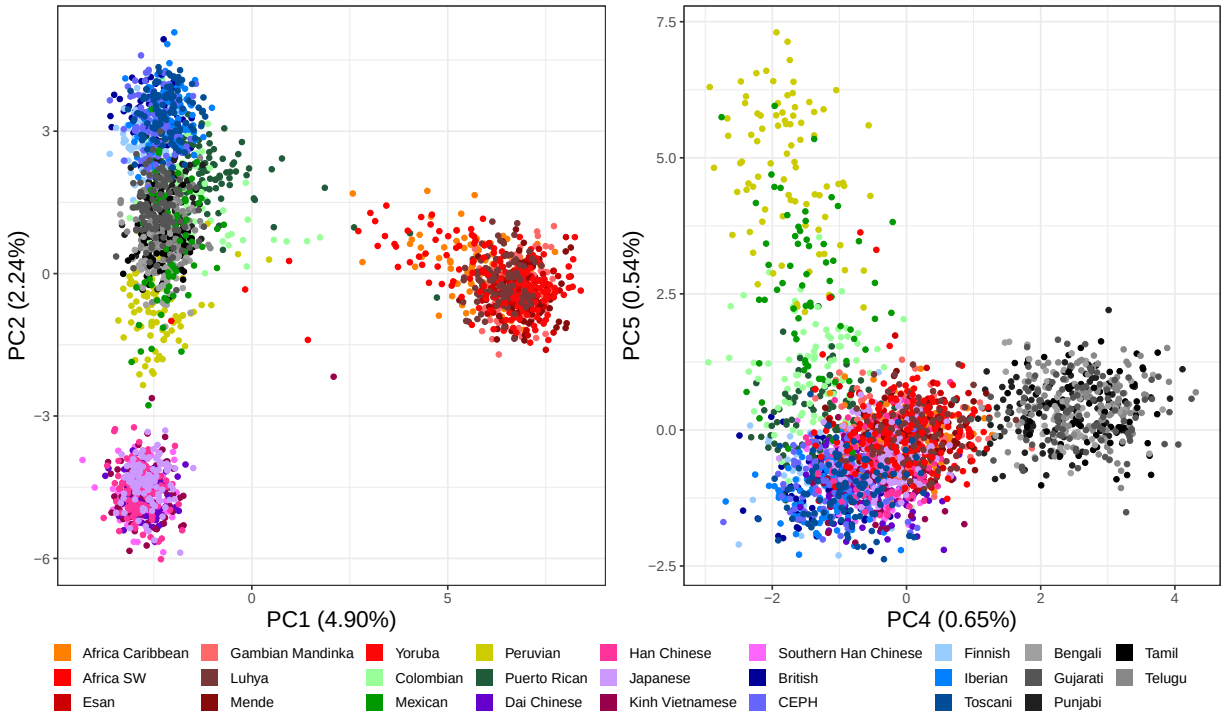

Figure 3. Principle Component Analysis (PCA) of common VNTR alleles in the NYGC population (150 bp). PCA was performed to reduce the dimensions of the data. *Left*: PC1 captured ∼5% of the variation and separated Africans from the other super-populations, suggesting that they had the greatest distance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

(Figure S4.1) with brain showing up multiple times, consistent with findings in the literature (37). These results suggest that VNTR alleles may affect gene regulation in multiple tissues.

Common VNTR genotypes are correlated with gene expression differences

Following the observation that common VNTRs were enriched in regions responsible for gene regulation and transcription, we investigated the association between VNTR alleles and gene expression. To establish a set of VNTR–gene associations, we paired common VNTRs with protein coding genes if they overlapped or occurred within 1 Kbp upstream or downstream of the gene. A total 1,419 VNTR–gene pairs across 445 individuals were examined (Supplementary Material `RNA_VNTRs.txt`).

Among those, 120 VNTR–gene pairs (118 genes) displayed significantly different expression (FDR<5%) when samples were classified by allele genotypes. The top ten differentially expressed genes were CSTB, DPYSL4, GPR63,HAGH, MRI1, PASK, PRKAR1B, PSKH1, SNX16, TRIM52-AS1 (FDR$<10^{-6}$) (Supplementary Figure S21).

Three genes are shown in Figure 2. Gene EIF3E is associated with a downstream VNTR (id 183225622), DPYSL4 is associated with VNTR (id 182316137) occurring in its first intron, and CSTB is associated with an upstream VNTR (id 182814480). Copy number expansions in the VNTR upstream of CSTB have been previously associated with progressive myoclonic epilepsy (EPM1) (98). We observed the -1 and 0 alleles, which are common in healthy individuals. However, 201 individuals had genotypes outside of our detection range which likely represented longer expansions and these individuals showed higher expression of this gene. More examples are given in Supplementary Section S7.

One in five common VNTR loci have population-specific alleles.

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

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

The first PC separated Africans, suggesting furthest evolutionary distance. The second PC separated East Asians. The third PC captured coverage bias. The fourth and fifth PCs separated South Asians and Americans, respectively. The American population had a sub-population of Puerto

Ricans that clustered with the Iberian Spanish population, suggesting mixed ancestry (99). To show the power of these alleles to predict ancestry, we next trained a decision tree model (Supplementary Figure S18) using the top 10 PCs (11% of the total variation) and achieved a recall of >98% on every population when applied to the 30% test partition (Supplementary Table S10).

A one-sided Fisher's Exact Test was applied to determine the population-specific VNTR alleles that were over-represented in one population versus all the others. A total of 3,850 VNTR alleles were identified as population-specific in one or more super-populations, corresponding to 1,096 VNTR loci (Supplementary Figures S19 and S20). The complete list of population-specific alleles can be found in Supplementary Material `Superpopulation_VNTRs.txt`. These loci overlapped with 689 genes and 51 coding exons. Africans had the highest number of population-specific alleles (266), followed by East Asians (65), while Americans had the lowest (13), suggesting more mixed ancestry. We observed 63 loci that had a population-specific allele in each population. Figure 4 illustrates seven of the top population-specific loci in a "virtual gel" representation, mimicking the appearance of bands on an agarose gel for easier interpretation. Thirty genes that displayed expression differences correlated with VNTR genotype were associated with population-specific VNTR loci (Supplementary Table S13), including the VNTR 182316137 associated with the gene DPYSL4, discussed in the previous section, which exhibited seven different alleles, five of which were population-specific.

Finally, to identify potential functional roles of the population-specific VNTR loci we performed Gene Set Enrichment Analysis (GSEA) for the associated genes against the Broad Institute MSigDB (100). Genes overlapping with the population-specific VNTRs were enriched for Endocytosis (hsa04144), Fatty acid metabolism (hsa01212), and Arrhythmogenic right ventricular cardiomyopathy (ARVC) (hsa05412) pathways (Supplementary Table S12). Among the GO biological processes affected by these genes were neurogenesis (GO:0022008; FDR=$3.62e^{-8}$), neuron differentiation (GO:0030182; FDR=$2.52e^{-7}$), and neuron development (GO:0048666; FDR=$4.31e^{-7}$). These processes are potentially related to other findings that have linked VNTRs to neurodegenerative disorders and cognitive abilities (36, 39, 51, 52, 53, 54, 55, 56, 57, 58) (Supplementary material `Population_specific_Go_BP.xlsx`). The GO term *behavior* (GO:0007610; FDR=$2.22e^{-4}$) was also found, which could be related to the association of VNTR loci with aggressive behavior (101, 102, 103). Other notable GO terms were regulation of muscle contraction (GO:0006937) and neuromuscular processes related to balancing (GO:0050885) with FDR <1%. The genes were also highly enriched in midbrain neurotype cell gene signatures (FDR=$5.49e^{-25}$), which might affect movement and emotions (104, 105, 106).

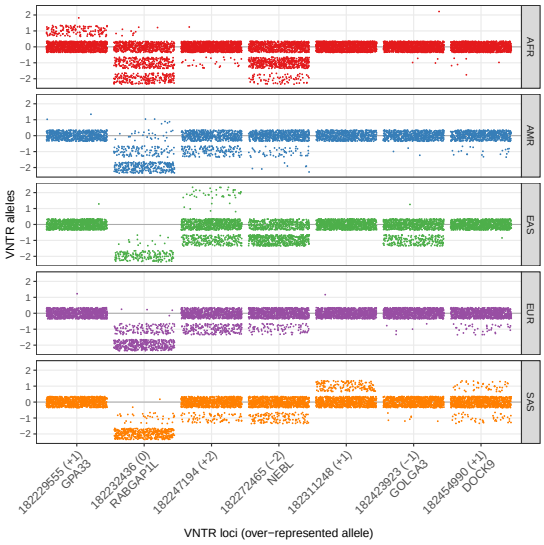

Figure 4. "Virtual gel" representation of seven population-specific VNTR alleles. Each dot represents an allele in one sample. Samples are separated vertically by super-population. Dots are jiggered in a rectangular area to reduce overlap. Population-specific alleles show up as bands over-represented in one population. Numbers and labels at bottom are VNTR locus ids with nearby genes indicated and the population-specific allele expressed as copy number change (+1, -2, *etc.*) from the reference. For example, in the leftmost column, the +1 allele was over-represented in the African population. Note that the allele bias towards pattern copy loss relative to the reference allele is apparent and that at one locus (second from left) the reference allele was the population-specific allele since almost no reference alleles were observed in the four other populations. The details of these seven loci are given in Supplementary Table S11.

Accuracy of VNTR predictions

To show the reliability of our results, we experimentally validated VNTR predictions at 13 loci in the three related AJ genomes, and also compared VNTRseek predictions to alleles experimentally validated in the literature. We additionally used accurate long reads on one genome (HG002) to find evidence of the predicted alleles. Separately, we showed the consistency of our predictions in two ways: first, we looked at inheritance consistency among four trios (mother, father, child), and second, we compared result for genomes sequenced on two different platforms.

Experimental validation. All but one of the 66 predicted VNTR alleles were confirmed at 13 loci in the three related AJ genomes (child HG002, father HG003, and mother HG004). In the remaining case, two predicted alleles were separated by only 15 nucleotides and could not be distinguished. At two loci, other bands were also observed. In one, all three family members contained an allele outside the detectable range of VNTRseek (longer than the reads). In the other, one allele that was detectable was missed in two family members. (See Table 3 for a summary of results and Supplementary Section S5.1 for gel images and more details.)

We also compared VNTRseek predictions in three datasets from the NA12878 genome with VNTRs validated in the adVNTR paper (66). Out of the original 17 VNTR loci experimentally validated in the adVNTR paper (66), four were not included in our reference set and for one, the matching TR could not be determined. In total, 11 out of 16 detectable alleles were correctly predicted, four were not found in the NA12878 sample with sufficient read size (250 bp), and one was incorrectly predicted in the HG001 sample and not found in the other two (Supplementary Table S7).

#	TR id	Pattern size	Ref. copies	Gene	VNTRseek predicted			Experimental validation		
					C	F	M	C	F	M
1*	182316181	105	4	STK32C	-2	-2	-2	-2,**-1**	-2,**-1**	-2,**+1**
2	182316985	27	6	LINC01168+	-3,0	-1,0	-3,-2	yes	yes	yes
3	182453735	30	2	DNAJC3	0,+1	0,+1	0,+1	yes	yes	yes
4	182461997	38	7	RASA3	-5,-4	-5,-4	-5,-2	yes	yes	yes
5	182493720	70	3	BEGAIN	0	0	0,-1	yes	yes	yes
6	182515357	34	8	MEGF11	-5	-5,-6	-5	yes	yes	yes
7	182608886	27	6	RPTOR	0,-2	0,-3	-2,+1	yes	yes	yes
8	182620950	48	3	RNF138	0,-1	0,-1	0,-1	yes	yes	yes
9	182982510	34	4	SLC12A7	0,-1	0	0,-1	yes	yes	yes
10♯	183046759	38	4	ARL10	0	0,-1	-1	0,**+1**	yes	-1,**+1**
11†	183081195	15	4	TENT5A	+2, -1	+2, +1	+2, -1	yes	+2	yes
12	183117043	17	9	MRM2+, LFNG+	-5, -3	-5, -3	0, -3	yes	yes	yes
13	183169331	15	4	IRF5	-2	-2, 0	-2	yes	yes	yes

Table 3. Experimental validation results: Thirteen VNTR loci were selected for experimental validation in the AJ trio. All but one of the 66 bands predicted by VNTRseek were validated. †For the remaining band, the results were questionable because the two predicted alleles for the father were only 15 nucleotides different in length, which was too close to distinguish in the image. *For all three individuals, the gel contained bands (**bold**) not predicted (or detectable) by VNTRseek. The extra band for the son corresponded to the -1 allele as found in the PacBio reads. The father's extra band appeared to match with the -1 allele. The mother's extra band appeared to be a +1 allele. ♯An extra band for the mother and son (**bold**) was not predicted by VNTRseek, although it seemed to match the +1 allele that was detectable. +These VNTRs overlapped regulatory sites that target the given genes.

Validation of predicted VNTRs using long reads. PacBio Circular Consensus Sequencing reads from the HG002 genome (7), with an average length of 13.5 Kbp and an estimated 99.8% sequence accuracy, were computationally tested to determine if they confirmed VNTRseek predicted alleles for the GIAB Illumina reads from the same genome. Overall, more than 97% of predicted alleles were confirmed, and at the predicted VNTR loci, more than 87% of alleles were confirmed (Supplementary Table S5.3).

VNTR predictions are consistent with Mendelian inheritance. We compared the predicted alleles in four trios (CEU and YRI trios from 1000 Genomes; Chinese HAN and AJ from GIAB), testing loci on autosomes and X and Y chromosomes (see Methods). In all cases, only a handful of loci were inconsistent (Table S9).

VNTR predictions are consistent across platforms. In 2015, the 1000 Genomes Phase 3 sequenced 30 genomes using Illumina HiSeq2500 at read length 250bp. In 2020, 27 of those 30 genomes were resequenced by NYGC using Illumina Novaseq 6000 at read length 150bp. Comparing VNTR loci genotyped in both platforms and non-reference alleles detectable at both read lengths, agreement ranged from 76%–91% (Supplementary Figure S17). Note, however, that read coverage was not the same for both datasets, causing variation in statistical power.

DISCUSSION

The current study represents, to our knowledge, the largest analysis of human whole genome sequencing data to detect copy number variable tandem repeats (VNTRs) and greatly expands the growing information on this class of genetic variation. The TRs genotyped consisted of some 184,000 minisatellites occupying the mid-range of pattern sizes, from seven to 126 bp. Our results reveal that nearly 20% (35,828) are variable, exhibiting at least one non-reference allele, a number much larger than has been generally understood. Moreover, we have classified a large subset of these (5,676) as *common* (occurring in >5% of the population). When considering the largest dataset in our study (2,504 individuals), we found that, on average, each genome was variable at 1,951 VNTR loci and among those, nearly 1,700 were common VNTRs.

In addition to their widespread occurrence, further evidence of minisatellite VNTR importance can be seen in the enrichment of these loci in genes and gene regulatory regions (promoters, transcription factor binding sites, DNAse hypersensitive sites, and CpG islands). Our entire set of VNTRs overlapped with 7,698 protein coding genes and 3,512 exons. The common VNTRs occurred within or were proximal to over 2,173 protein coding genes, including overlapping with 254 exons. Biological function enrichment among these genes includes neuron development and differentiation, and behavior. This is consistent with the finding that VNTR expansions in humans compared to primates are associated with gain of cognitive abilities (30), and possible involvement of VNTRS with many neurodegenerative diseases and behavioral disorders.

The overabundance of VNTR proximity to genes suggests that variability at these loci could affect gene expression and indeed, we observed that the expression levels of 118 genes were significantly correlated with the presence of specific VNTR alleles in lymphoblastoid cell lines of 445 individuals. These findings are suggestive, but more study is required, both to determine if there is more evidence of *tissue specific* gene expression variations associated with VNTR genotype (30, 38) and if such correlational differences can be definitively tied to actions associated with specific VNTR alleles such as regulator binding affinity changes in regulatory regions. For more elaborate studies such as these, it will be essential that for each sample used to measure gene expression, the raw whole genome sequencing data be available, so that specialized software programs, such as VNTRseek can be used to determine VNTR genotype.

The frequency of VNTR occurrence and possible effects on gene expression suggest that minisatellite VNTR loci could be useful in genome-wide association studies (GWAS). However, it is well known that hidden differences can lead to misinterpretation of GWAS results, and care is particularly important when those differences are tied to human ancestry. Relevant to this, we have determined that 1,096 of the common VNTR loci contain alleles showing significant population specificity and that these loci intersect with 689 genes. Understanding such hidden variability will be essential for interpreting GWAS and future studies should investigate possible haplotype linkages between specific VNTR alleles and nearby SNP alleles.

Population-specific alleles also have the potential for use in tracing early human migration. We have shown through principle component analysis with common VNTR alleles that super-populations are easily separated. Further, we have constructed a decision tree based on common VNTR alleles that obtains nearly perfect classification of individuals at the super population level. It will be interesting to see whether, with more information, classification can be refined further to encompass specific sub-populations, whether a minimal minisatellite VNTR set can be established for high accuracy population classification, and whether VNTR alleles can be used to estimate mixed ancestry as is done now with SNP haplotyping.

Despite the high sensitivity of VNTRseek, our curation of VNTR loci has certainly produced an undercount. This is true because VNTRseek requires that the tandem array fit within a read. Longer reads will help, but high-coverage long-read datasets are not yet common. Alternate methods exist (30, 66), but these have not reported an ability to handle *macrosatellite* VNTRs where the arrays and patterns are hundreds to thousands of base pairs long. For this range of the tandem repeat spectrum, new tools must be developed.

Another limitation comes from use of the Tandem Repeats Finder, which requires that the array contain at least 1.9 copies to be detected. At read length 150 bp, which included the majority of our samples, a gain of one copy compared to the reference genome could be detected in 82% of the TR loci while loss of one copy could be detected in only 16%. Despite this imbalance, one copy loss was observed nearly 40% more often than one copy gain, an important observation with regard to potential tandem repeat copy number bias in the reference genome.

Previous studies on VNTR prevalence in the human genome have been limited to a subset of minisatellites inside the transcriptome and a limited number of genomes. Here, we have shown a broader prevalence of VNTR loci and suggested their importance with regard to gene function, population studies, and GWAS. Future research can be expected to further enhance our understanding of this important class of genomic variation.

DATA AVAILABILITY

Dataset (individuals)	URL or Accession numbers
1000 Genomes phase 3 HC (30)	EBI: 1000 Genomes Project
NYGC (2,504)	Data collection 30X on GRCh38
	ENA project: PRJEB31736
SGDP (253)	IGSR: Data collection SGDP
GIAB (7)	NCBI SRA project: SRP047086
CHM1 (1)	NCBI SRA: SRX652547
CHM13 (1)	NCBI SRA: SRX1009644
Tumor/Normal Pairs (4)	Illumina basespace project
PacBio (1)	NCBI SRA: SRX5327410

Table 4. Dataset sources.

The reference TR set files, output VCF files, and the pre-processed data files along with the code to create figures and tables are published at:
https://doi.org/10.5281/zenodo.4065850
VNTRseek can be downloaded at:
https://github.com/Benson-Genomics-Lab/VNTRseek.

ACKNOWLEDGEMENT

This work was supported in part by NSF grants IIS-1423022 and DBI-1559829.

We thank Thomas Gilmore (Boston University) for helpful discussions and comments on the manuscript.

Conflict of interest statement. None declared.

REFERENCES

1. Auton, A., Brooks, L. D., Durbin, R. M., et al. (Oct, 2015) A global reference for human genetic variation. *Nature,* **526**(7571), 68–74.
2. Mallick, S., Li, H., Lipson, M., et al. (2016) The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature,* **538**(7624), 201–206.
3. Zook, J. M., Catoe, D., McDaniel, J., et al. (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data,* **3**, sdata201625.
4. Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., et al. (2014) Resolving the complexity of the human genome using single-molecule sequencing. *Nature,* **517**(7536), 608–611.
5. Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., et al. (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research,* **27**(5), 677–685.
6. Drmanac, R., Sparks, A. B., Callow, M. J., et al. (2010) Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science,* **327**(5961), 78–81.
7. Wenger, A. M., Peluso, P., Rowell, W. J., et al. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology,* **37**(10), 1155–1162.
8. Lander, E. S. and Waterman, M. S. (1988) Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics,* **2**(3), 231–239.
9. Treangen, T. J. and Salzberg, S. L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics,* **13**(1), 36.
10. de Koning, A. J., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics,* **7**(12), e1002384.
11. Lim, K. G., Kwoh, C. K., Hsu, L. Y., and Wirawan, A. (2013) Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Briefings in bioinformatics,* **14**(1), 67–81.

12. Richard, G.-F., Kerrest, A., and Dujon, B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and molecular biology reviews,* **72**(4), 686–727.
13. Taylor, J. S. and Breden, F. (2000) Slipped-strand mispairing at noncontiguous repeats in Poecilia reticulata: a model for minisatellite birth. *Genetics,* **155**(3), 1313–1320.
14. Levinson, G. and Gutman, G. A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution.. *Molecular biology and evolution,* **4**(3), 203–221.
15. Madsen, C. S., Ghivizzani, S. C., and Hauswirth, W. W. (1993) In vivo and in vitro evidence for slipped mispairing in mammalian mitochondria. *Proceedings of the National Academy of Sciences,* **90**(16), 7671–7675.
16. Jeffreys, A. J., Neil, D. L., and Neumann, R. (1998) Repeat instability at human minisatellites arising from meiotic recombination. *The EMBO Journal,* **17**(14), 4147–4157.
17. Debrauwère, H., Buard, J., Tessier, J., et al. (1999) Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. *Nature genetics,* **23**(3), 367.
18. Pâques, F., Richard, G.-F., and Haber, J. E. (2001) Expansions and contractions in 36-bp minisatellites by gene conversion in yeast. *Genetics,* **158**(1), 155–166.
19. Bustamante, A. V., Sanso, A. M., Segura, D., Parma, A. E., and Lucchesi, P. M. A. (2013) Dynamic of mutational events in variable number tandem repeats of Escherichia coli O157: H7. *BioMed research international,* **2013**.
20. Vogler, A. J., Keys, C., Nemoto, Y., et al. (2006) Effect of repeat copy number on variable-number tandem repeat mutations in Escherichia coli O157: H7. *Journal of bacteriology,* **188**(12), 4253–4263.
21. Fu, S., Octavia, S., Wang, Q., et al. (2016) Evolution of variable number tandem repeats and its relationship with genomic diversity in Salmonella Typhimurium. *Frontiers in microbiology,* **7**, 2002.
22. Verstrepen, K. J., Jansen, A., Lewitter, F., and Fink, G. R. (2005) Intragenic tandem repeats generate functional variability. *Nature genetics,* **37**(9), 986.
23. Legendre, M., Pochet, N., Pak, T., and Verstrepen, K. J. (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome research,* **17**(12), 1787–1796.
24. Panigrahi, I. (2018) Genetic Fingerprinting for Human Diseases: Applications and Implications. In *DNA Fingerprinting: Advancements and Future Endeavors* pp. 141–150 Springer.
25. Sinha, M., Rao, I. A., and Mitra, M. (2018) Molecular Basis of Identification Through DNA Fingerprinting in Humans. In *DNA Fingerprinting: Advancements and Future Endeavors* pp. 129–140 Springer.
26. Imam, J., Reyaz, R., Rana, A. K., and Yadav, V. K. (2018) DNA Fingerprinting: Discovery, Advancements, and Milestones. In *DNA Fingerprinting: Advancements and Future Endeavors* pp. 3–24 Springer.
27. Denoeud, F., Vergnaud, G., and Benson, G. (2003) Predicting human minisatellite polymorphism. *Genome research,* **13**(5), 856–867.
28. Deka, R., Chakraborty, R., and Ferrell, R. E. (1991) A population genetic study of six VNTR loci in three ethnically defined populations. *Genomics,* **11**(1), 83–92.
29. Deka, R., DeCroo, S., Yu, L. M., and Ferrell, R. E. (1992) Variable number of tandem repeat (VNTR) polymorphism at locus D17S5 (YNZ22) in four ethnically defined human populations. *Human genetics,* **90**(1-2), 86–90.
30. Sulovari, A., Li, R., Audano, P. A., et al. (2019) Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences,* **116**(46), 23243–23253.
31. Hancock, J. M. and Santibáñez-Koref, M. F. (1998) Trinucleotide Expansion Diseases in the Context of Micro-and Minisatellite Evolution Hammersmith Hospital, April 1–3, 1998. *The EMBO journal,* **17**(19), 5521–5524.
32. Duitama, J., Zablotskaya, A., Gemayel, R., et al. (2014) Large-scale analysis of tandem repeat variability in the human genome. *Nucleic acids research,* **42**(9), 5728–5741.
33. Lancaster, C. A., Peat, N., Duhig, T., et al. (1990) Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit. *Biochemical and biophysical research communications,* **173**(3), 1019–1029.
34. Van Tol, H. H., Wu, C. M., Guan, H.-C., et al. (1992) Multiple dopamine D4 receptor variants in the human population. *Nature,* **358**(6382), 149.
35. Trepicchio, W. L. and Krontiris, T. G. (1992) Members of the rel/NF-χB family of transcriptional regulatory proteins bind the HRAS1 minisatellite DNA sequence. *Nucleic acids research,* **20**(10), 2427–2434.
36. Krontiris, T. G., Devlin, B., Karp, D. D., Robert, N. J., and Risch, N. (1993) An association between the risk of cancer and mutations in the HRAS1 minisatellite locus. *New England Journal of Medicine,* **329**(8), 517–523.
37. Sonay, T. B., Carvalho, T., Robinson, M. D., et al. (2015) Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome research,* **25**(11), 1591–1599.
38. Bakhtiari, M., Park, J., Ding, Y.-C., et al. (2020) Variable Number Tandem Repeats mediate the expression of proximal genes. *bioRxiv,*.
39. De Roeck, A., Duchateau, L., Van Dongen, J., et al. (2018) An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta neuropathologica,* **135**(6), 827–837.
40. Pacheco, A., Berger, R., Freedman, R., and Law, A. J. (2019) A VNTR Regulates miR-137 expression through novel Alternative Splicing and contributes to Risk for Schizophrenia. *Scientific reports,* **9**(1), 1–12.
41. Fondon, J. W. and Garner, H. R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences,* **101**(52), 18058–18063.
42. Laidlaw, J., Gelfand, Y., Ng, K.-W., et al. (2007) Elevated Basal Slippage Mutation Rates among the *Canidae*. *Journal of Heredity,* **98**(5), 452–460 (doi:10.1093/jhered/esm017).
43. Antwi-Boasiako, C., Dzudzor, B., Kudzi, W., et al. (2018) Association between eNOS Gene Polymorphism (T786C and VNTR) and Sickle Cell Disease Patients in Ghana. *Diseases,* **6**(4), 90.
44. Ksiazek, K., Blaszczak, J., and Buraczynska, M. (2019) IL4 gene VNTR polymorphism in chronic periodontitis in end-stage renal disease patients. *Oral diseases,* **25**(1), 258–264.
45. Cong, L., Tu, G., and Liang, D. (2018) A systematic review of the relationship between the distributions of aggrecan gene VNTR polymorphism and degenerative disc disease/osteoarthritis. *Bone & joint research,* **7**(4), 308–317.
46. Ramírez-Patiño, R., Figuera, L. E., Puebla-Pérez, A. M., et al. (2013) Intron 4 VNTR (4a/b) polymorphism of the endothelial nitric oxide synthase gene is associated with breast cancer in Mexican women. *Journal of Korean medical science,* **28**(11), 1587–1594.
47. Vairaktaris, E., Serefoglou, Z. C., Yapijakis, C., et al. (2007) The Platelet Glycoprotein Ibα VNTR Polymorphism is Associated with Risk for Oral Cancer. *Anticancer research,* **27**(6B), 4121–4125.
48. Sousa, H., Santos, A. M., Catarino, R., et al. (2012) IL-1RN VNTR polymorphism and genetic susceptibility to cervical cancer in Portugal. *Molecular biology reports,* **39**(12), 10837–10842.
49. Safarinejad, M. R., Safarinejad, S., Shafiei, N., and Safarinejad, S. (2013) Effects of the T-786C, G894T, and Intron 4 VNTR (4a/b) polymorphisms of the endothelial nitric oxide synthase gene on the risk of prostate cancer. In *Urologic Oncology: Seminars and Original Investigations* Elsevier Vol. 31, pp. 1132–1140.
50. Ibrahimi, M., Moossavi, M., Mojarad, E. N., et al. (2019) Positive correlation between interleukin-1 receptor antagonist gene 86bp VNTR polymorphism and colorectal cancer susceptibility: a case-control study. *Immunologic research,* **67**(1), 151–156.
51. Marinho, F. V. C., Pinto, G. R., Oliveira, T., et al. (2019) The SLC6A3 3-UTR VNTR and intron 8 VNTR polymorphisms association in the time estimation. *Brain Structure and Function,* **224**(1), 253–262.
52. Katsumata, Y., Fardo, D. W., Bachstetter, A. D., et al. (2019) Alzheimer Disease Pathology-Associated Polymorphism in a Complex Variable Number of Tandem Repeat Region Within the MUC6 Gene, Near the AP2A2 Gene. *Journal of Neuropathology & Experimental Neurology,*.
53. Chang, H.-I., Chang, Y.-T., Tsai, S.-J., et al. (2019) MAOA-VNTR Genotype Effects on Ventral Striatum-Hippocampus Network in Alzheimer's Disease: Analysis Using Structural Covariance Network and Correlation with Neurobehavior Performance. *Molecular neurobiology,* **56**(6), 4518–4529.
54. Scott, H., Nelson, P., Hopwood, J., and Morris, C. (1991) PCR of a VNTR linked to mucopolysaccharidosis type I and Huntington disease. *Nucleic acids research,* **19**(22), 6348.
55. Hoxha, B., Goçi, A. U., Agani, F., et al. (2019) The Role of TaqI DRD2 (rs1800497) and DRD4 VNTR Polymorphisms in Posttraumatic Stress Disorder (PTSD).. *Psychiatria Danubina,* **31**(2), 263–268.
56. Šerỳ, O., Paclt, I., Drtílková, I., et al. (2015) A 40-bp VNTR polymorphism in the 3'-untranslated region of DAT1/SLC6A3 is

associated with ADHD but not with alcoholism. *Behavioral and Brain Functions,* **11**(1), 21.
57. Grünblatt, E., Werling, A. M., Roth, A., Romanos, M., and Walitza, S. (2019) Association study and a systematic meta-analysis of the VNTR polymorphism in the 3'-UTR of dopamine transporter gene and attention-deficit hyperactivity disorder. *Journal of Neural Transmission,* **126**(4), 517–529.
58. Van Assche, E., Moons, T., Van Leeuwen, K., et al. (2016) Depressive symptoms in adolescence: The role of perceived parental support, psychological control, and proactive control in interaction with 5-HTTLPR. *European Psychiatry,* **35**, 55–63.
59. Stolf, A. R., Cupertino, R. B., Müller, D., et al. (2019) Effects of DRD2 splicing-regulatory polymorphism and DRD4 48 bp VNTR on crack cocaine addiction. *Journal of Neural Transmission,* **126**(2), 193–199.
60. Gymrek, M. (2017) A genomic view of short tandem repeats. *Current Opinion in Genetics & Development,* **44**, 9–16.
61. Tørresen, O. K., Star, B., Mier, P., et al. (2019) Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic acids research,* **47**(21), 10994–11006.
62. Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012) lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research,*.
63. Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B., and Halldórsson, B. V. (2017) popSTR: population-scale detection of STR variants. *Bioinformatics,* **33**(24), 4041–4048.
64. Willems, T., Zielinski, D., Yuan, J., et al. (2017) Genome-wide profiling of heritable and extlessi extgreaterde novo extless/i extgreater STR variations. *Nature Methods,* **14**(6), 590–592.
65. Dolzhenko, E., Deshpande, V., Schlesinger, F., et al. (05, 2019) ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics,* **35**(22), 4754–4756.
66. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V., and Bafna, V. (2018) Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Research,* **28**(11), 1709–1719.
67. Gelfand, Y., Hernandez, Y., Loving, J., and Benson, G. (2014) VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Research,* **42**(14), 8884–8894.
68. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research,* **27**(2), 573–580.
69. Lander, E. S., Linton, L. M., Birren, B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**(6822), 860–921.
70. Gelfand, Y., Rodriguez, A., and Benson, G. (2007) TRDB-The Tandem Repeats Database. *Nucleic Acids Research,* **35**(suppl 1), D80–D87.
71. Ye, J., Coulouris, G., Zaretskaya, I., et al. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics,* **13**(1), 134.
72. Kent, W. J., Sugnet, C. W., Furey, T. S., et al. (2002) The human genome browser at UCSC. *Genome research,* **12**(6), 996–1006.
73. Picard toolkit. http://broadinstitute.github.io/picard/ (2019).
74. Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997,*.
75. Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* **26**(6), 841–842.
76. Karolchik, D., Hinrichs, A. S., Furey, T. S., et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic acids research,* **32**(suppl_1), D493–D496.
77. Hsu, F., Kent, W. J., Clawson, H., et al. (2006) The UCSC known genes. *Bioinformatics,* **22**(9), 1036–1046.
78. Consortium, E. P. et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**(7414), 57–74.
79. Davis, C. A., Hitz, B. C., Sloan, C. A., et al. (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research,* **46**(D1), D794–D801.
80. Thurman, R. E., Rynes, E., Humbert, R., et al. (2012) The accessible chromatin landscape of the human genome. *Nature,* **489**(7414), 75–82.
81. Kundaje, A., Meuleman, W., Ernst, J., et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature,* **518**(7539), 317–330.
82. Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *Journal of molecular biology,* **196**(2), 261–282.
83. Sheffield, N. C. and Bock, C. (2016) LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics,* **32**(4), 587–589.
84. Subramanian, A., Tamayo, P., Mootha, V. K., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences,* **102**(43), 15545–15550.
85. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000) Gene ontology: tool for the unification of biology. *Nature genetics,* **25**(1), 25.
86. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research,* **28**(1), 27–30.
87. Lappalainen, T., Sammeth, M., Friedländer, M. R., et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature,* **501**(7468), 506–511.
88. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological),* **57**(1), 289–300.
89. Günther, T. and Nettelblad, C. (2019) The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS genetics,* **15**(7), e1008302.
90. Campbell, H., Carothers, A. D., Rudan, I., et al. (2007) Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Human Molecular Genetics,* **16**(2), 233–241.
91. Herráez, D. L., Bauchet, M., Tang, K., et al. (2009) Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PloS one,* **4**(11), e7888.
92. Bray, S. M., Mulle, J. G., Dodd, A. F., et al. (2010) Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proceedings of the National Academy of Sciences,* **107**(37), 16222–16227.
93. Edea, Z., Bhuiyan, M., Dessie, T., et al. (2015) Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. *Animal: an International Journal of Animal Bioscience,* **9**(2), 218.
94. Raczy, C., Petrovski, R., Saunders, C. T., et al. (2013) Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics,* **29**(16), 2041–2043.
95. Cichon, S., Craddock, N., Daly, M., et al. (2009) Psychiatric GWAS Consortium Coordinating Committee Genomewide association studies: History, rationale, and prospects for psychiatric disorders. *Am J Psychiatry,* **166**(5), 540–556.
96. Nagraj, V., Magee, N. E., and Sheffield, N. C. (2018) LOLAweb: a containerized web server for interactive genomic locus overlap enrichment analysis. *Nucleic acids research,* **46**(W1), W194–W199.
97. Sheffield, N. C., Thurman, R. E., Song, L., et al. (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome research,* **23**(5), 777–788.
98. Lalioti, M., Antonarakis, S., and Scott, H. S. (2003) The epilepsy, the protease inhibitor and the dodecamer: progressive myoclonus epilepsy, cystatin b and a 12-mer repeat expansion. *Cytogenetic and genome research,* **100**(1-4), 213–223.
99. Sudmant, P. H., Rausch, T., Gardner, E. J., et al. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature,* **526**(7571), 75–81.
100. Liberzon, A., Subramanian, A., Pinchback, R., et al. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics,* **27**(12), 1739–1740.
101. Schlüter, T., Winz, O., Henkel, K., et al. (2016) MAOA-VNTR polymorphism modulates context-dependent dopamine release and aggressive behavior in males. *Neuroimage,* **125**, 378–385.
102. Zammit, S., Jones, G., Jones, S. J., et al. (2004) Polymorphisms in the MAOA, MAOB, and COMT genes and aggressive behavior in schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics,* **128**(1), 19–20.
103. Vernaleken, I., Schlüter, T., Eggermann, T., et al. (2015) Effect of MAOA-VNTR Polymorphism on Aggression and Dopamine Release. *Journal of Nuclear Medicine,* **56**(supplement 3), 300–300.
104. Mill, J., Asherson, P., Browes, C., D'Souza, U., and Craig, I. (2002) Expression of the dopamine transporter gene is regulated by the 3' UTR VNTR: Evidence from brain and lymphocytes using quantitative RT-PCR. *American journal of medical genetics,* **114**(8), 975–979.
105. Diatchenko, L., Nackley, A. G., Tchivileva, I. E., Shabalina, S. A., and Maixner, W. (2007) Genetic architecture of human pain perception. *TRENDS in Genetics,* **23**(12), 605–613.
106. Kang, A. M., Palmatier, M. A., and Kidd, K. K. (1999) Global variation of a 40-bp VNTR in the 3'-untranslated region of the dopamine transporter gene (SLC6A3). *Biological psychiatry,* **46**(2), 151–160.