Title Page

Characterization of a Novel Hyper-Variable Variable Number Tandem Repeat in the Dopamine Transporter Gene (*SLC6A3*)

Running Title

SLC6A3 Tandem Repeats

Keywords

VNTR, heterozygosity, recombination hotspot, G-tetrad, G-quadruplex

Authors

Abner T. Apsley[1,2]; Emma R. Domico[1*]; Max A. Verbiest[3,4,5]; Carly A. Brogan [1]; Evan R. Buck[1]; Andrew J. Burich[6]; Kathleen M. Cardone [1]; Wesley J. Stone [1]; Maria Anisimova[3,5]; David J. Vandenbergh[1,2,7,8**]

[1] Department of Biobehavioral Health, The Pennsylvania State University, Pennsylvania, USA

[2] The Molecular, Cellular and Integrative Biosciences Program, The Pennsylvania State University, Pennsylvania, USA

[3] Institute of Computational Life Science, School of Life Sciences and Facility Management, Zürich University of Applied Sciences, Wädenswil, Switzerland

[4] Department of Molecular Life Sciences, Faculty of Science, University of Zurich, Zurich, Switzerland

[5] Swiss Institute of Bioinformatics, Lausanne, Switzerland

[6] Department of Information Science and Technologies - Applied Data Sciences, The Pennsylvania State University, Pennsylvania, USA

[7]Institute of the Neurosciences, The Pennsylvania State University, Pennsylvania, USA

[8] The Bioinformatics and Genomics Program, The Pennsylvania State University, Pennsylvania, USA

*Current address: Armstrong Institute for Patient Safety and Quality, Johns Hopkins University, Baltimore, MD

**Correspondence to djv4@psu.edu

ORCID #'s:

- Abner Apsley: 0000-0003-3420-7491
- Emma Domico: 0000-0003-3760-6463
- Max Verbiest: 0000-0003-3424-0136
- Maria Anisimova: 0000-0001-8145-7966
- David Vandenbergh: 0000-0002-5620-2870

Abstract

The dopamine transporter gene, *SLC6A3*, has received substantial attention in genetic association studies of various phenotypes (hypertension, ADHD, substance use disorders, etc.). Although some variable number tandem repeats (VNTRs) present in *SLC6A3* have been used as genetic markers in these association studies, results have not been consistent. We searched for unanalyzed VNTRs in *SLC6A3* that might account for the heterogeneity of existing association study results. We used the Tandem Repeat Annotation Library (TRAL) to characterize the VNTR landscape of 64 unrelated long-read haplotype-phased *SLC6A3* sequences. We further report sequence similarity of each repeat unit of the five VNTRs along with the correlations of SNP-SNP, SNP-VNTR and VNTR-VNTR alleles across the length of the gene. We present the discovery of a novel hyper-variable number tandem repeat (hyVNTR) in intron 8 of *SLC6A3*, which contains a range of 3.4-133.4 repeat copies and has a consensus sequence length of 38bp, with 84% G+C content. The 38-base repeat was predicted to form G-quadruplexes *in silico* and was confirmed by circular dichroism spectroscopy. Additionally, this hyVNTR contains multiple putative binding sites for PRDM9, which, in combination with low levels of linkage disequilibrium around the hyVNTR, suggests it is a recombination hotspot.

Introduction

Variable number of tandem repeat loci (VNTRs) are important sites of genomic variation (Gall-Duncan et al. 2022; Xiao et al. 2022). VNTRs are defined as regions of DNA where a particular nucleotide sequence is repeated in tandem and the number of copies of the repeated sequence varies between individuals (Hannan 2018). Previous studies have shown that genomic VNTRs play a role in various biological processes such as the formation of G-quadruplexes (G4s) (Li et al. 2016; Guiblet et al. 2021), DNA I-motifs (Kondo et al. 2004), recombination hotspots (Zavodna et al. 2018), gene expression control (Johansson et al. 2022; Örd et al. 2020; Lalioti et al. 1997; Borel et al. 2012; Li et al. 2016), alteration of DNA methylation (Garg et al. 2021) and histone modifications (Vasiliou et al. 2012). Interpersonal variation in VNTRs can have functional consequences by altering gene expression of nearby genes as is seen for three VNTRs in and around the Arginine Vasopressin Receptor 1A [*AVPR1A*, or proposed nomenclature *VTR1A* (Theofanopoulou et al. 2021)], which have been associated with altered gene expression and the complex trait of externalizing behavior (Landefeld et al. 2018). In addition to playing a role in biological processes, VNTRs have been implicated in psychological disorders (Hannan 2018, 2021; Gall-Duncan et al. 2022), such as Alzheimer's Disease(De Roeck et al. 2018) schizophrenia (Song et al. 2018), amyotrophic lateral sclerosis (Course et al. 2020)and myopathy with rimmed ubiquitin-positive autophagic vacuolation (MRUPAV) (Ruggieri et al. 2020). These results indicate that it is crucial to genotype polymorphic tandem repeat loci accurately to get a full picture of the genetic component of phenotypic variation in the human population.

Many different characteristics, such as location, total length, consensus sequence motif, intrapersonal repeat copy number variation, and conservation or purity of consensus sequence

motif may contribute to how VNTRs influence phenotypes (Hannan 2018). Due to the repetitive nature of VNTRs, genomic functional elements that require specific DNA motifs in tandem are likely candidates for mechanisms by which VNTRs play a role in disease/phenotype determination. For example, G4s are four-stranded secondary DNA structures that form in the presence of a repetitive guanine-rich DNA motif (Huppert and Balasubramanian 2005) and have been shown to influence biological processes such as transcription rates (Brázda et al. 2019; Huppert and Balasubramanian 2005); therefore, specific VNTR consensus sequences may aid in the formation of G4s. Additionally, the presence of multiple VNTR consensus sequences that contain protein binding site motifs may provide a source of variation in their binding affinity (Vasiliou et al. 2012).

SLC6A3, which encodes the dopamine transporter protein (DAT1), has been studied extensively in relation to its VNTRs. *SLC6A3* has 15 exons and spans approximately 52.5 kb in length on human chromosome 5 (GRCh38/hg38). It contains a VNTR in the 3'-UTR of the gene (Vandenbergh et al. 1992) with repeat copy numbers ranging from 3-11 and a consensus sequence of 40 bp in length. Genetic analyses suggested association of this polymorphic site with various phenotypes such as attention deficit hyperactivity disorder (Cook et al. 1995; Bidwell et al. 2011), Parkinson's disease (du Plessis et al. 2020), hypertension (Kim et al. 2017), depression (Kirchheiner et al. 2007), substance use disorders (van der Zwaluw et al. 2009), and other physiological and psychological ailments (Salatino-Oliveira et al. 2018). A second VNTR in intron 8 of *SLC6A3*, with repeat copy numbers of either 5 or 6, and a consensus repeat sequence length of 30 bp, was associated with cocaine dependence in Brazilian individuals (Guindalini et al. 2006). This intron 8 VNTR was also tested in many other studies for its association with disease-related phenotypes (Salatino-Oliveira et al. 2018). A third VNTR located in intron 3 of

SLC6A3 was reported to have a consensus sequence length of 63 bp and repeat copy numbers of 7 and 8 (Franke et al. 2008). Finally, in 2017, Kim and colleagues reported on a fourth VNTR present in intron 4 of *SLC6A3*, which had a consensus sequence of 75 bp and repeat copy numbers ranging from 11-32 (Kim et al. 2017). Recent work by Course and colleagues has also documented the presence of this intron 4 VNTR using long-read haplotype-phased assemblies; however, they report it as having a 38 bp consensus sequence length, essentially dividing the intron 4 VNTR reported by Kim and colleagues in half (Course et al. 2021).

Although *SLC6A3* has received a large amount of attention in relation to its VNTR landscape and the associations of these VNTR alleles with phenotypes of interest, there have been both inconsistent and contradicting results reported. One possible explanation for these inconsistent and contradictory reports may be that there is a lack of complete characterization of genetic variation in this gene. In fact, in 1993, Byerley and colleagues, using Southern blot techniques with TaqI digested DNA, reported a site in the 3'-half of *SLC6A3* that varied by as much as 7.1 kb (Byerley et al. 1993), but was not studied further. We hypothesized the presence of one or more VNTRs in this region of *SLC6A3* that may account for the heterogeneity in sequence length reported by Byerley et al., but that this VNTR has escaped detection due to the lack of technology sufficient to fully characterize its total repeat length.

In the present work, we used 64 publicly available long-read haplotype-phased genome assemblies from 32 individuals (Ebert et al. 2021 – version 1) and the Tandem Repeat Annotation Library (TRAL) Python library (Delucchi et al. 2021) to characterize the VNTR landscape of *SLC6A3*. We report the discovery of a novel hyper-variable number tandem repeat (hyVNTR) in the intron 8 region of *SLC6A3* that has a consensus sequence length of 38 bp and repeat copy numbers ranging from 3.4-133.4. Additionally, we demonstrate its ability to form G-

Quadruplexes in vitro. We also report linkage disequilibrium (LD) values of each *SLC6A3* VNTR with all other SNPs and VNTRs present in the gene, with emphasis on the hyVNTR. Potential PRDM9 binding sites – a known site for initiation of recombination – are present in this hyVNTR. These data suggest the need to assess potential functional relevance of the hyVNTR and its role in the genetic underpinnings of dopamine-related traits.

Results

SLC6A3 has a Highly Variable Sequence Length

The tandem repeat architecture of *SLC6A3* was analyzed using long-read haplotype-phased genome assemblies from 32 unrelated individuals (Ebert et al. 2021 – version 1). The assemblies were created by either CLR (continuous long reads) or CCS (circular consensus sequencing) technologies (Pacific Biosciences). Some assemblies were available for both technologies and comparison of the two versions revealed an average similarity score of 99.96% (SD = 0.06%). Using the length of *SLC6A3* in the human reference genome (GRCh38/hg38), a similarity score at this value corresponds to an average of 22.67 (SD = 29.06) mismatching bases per *SLC6A3* sequence. Additionally, an average of 20.94 (SD = 19.69) INDELS were observed across all *SLC6A3* sequences. Individual haplotype comparison values are shown in **Table S1**.

No CNVs or large deletions were found where *SLC6A3* is located, and all haplotype-phased long-read genome assemblies contained one copy of the gene. *SLC6A3* sequences ranged from 52,673 bp to 58,846 bp (mean = 54,749 bp, SD = 1,265 bp), confirming the presence of heterogeneity in sequence length among our samples.

Tandem Repeat Annotations Reveal a Novel Hyper-Variable VNTR Located in Intron 8 of *SLC6A3*

Many algorithms have been used to detect tandem repeats (TRs) in biological sequences, leading to inconsistent TR annotations of identical sequences (Schaper et al. 2012). To address this issue, we used the TRAL Python library (Delucchi et al. 2021) which compares, harmonizes, and makes non-redundant, the output of different repeat detection algorithms. We used TRAL to annotate all *SLC6A3* sequences for repeats. TRs were distributed across the length of the gene with no obvious pattern. **Figure 1** shows the location of each invariant TR (red) and VNTR (blue) displayed above the annotations from the standard Tandem Repeat Finder (TRF) results. **Table 1** shows exon/intron locations, repeat parameters and allele information for each TR that was annotated as a VNTR, and the same information for TRs that had the same copy number across all *SLC6A3* sequences is included in **Table S2**. A majority of the TRs had a consensus sequence length less than or equal to 10 bp, with 8 TRs having more than 10 bp in their consensus sequence. Only 5 TRs contained more than one copy number allele (TRs 09, 17, 21, 22, and 30) and were therefore designated as VNTRs (see **Figure 1**). Other than TR26, which is located in exon 10, all TRs are located either in introns or the 3'UTR region of the gene. **Table 2** shows the consensus sequence for each annotated VNTR.

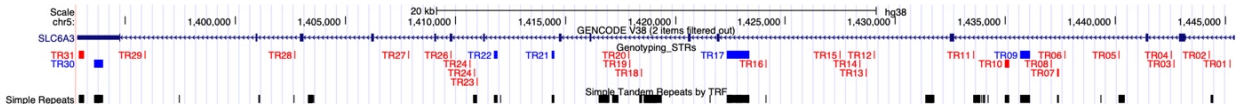

Figure - UCSC Genome Browser view of SLC6A3 with UCSC's Simple Repeats track (TRF annotation of sequence) and a custom track showing the TRs detected by TRAL with non-variable TRs in red and VNTRs in blue.

Table 1 – VNTR Characteristics

Tandem Repeat ID	Consensus Length (bp)	Number of Different Alleles	Copy Number of Major Allele	Degree of Heterozygosity	Location
TR09	66	2	7	0.469	Intron3
TR17	38	13	28	0.781	Intron4
TR21	38	43	3	0.938	Intron8
TR22	30	4	6	0.531	Intron8
TR30	40	3	10	0.406	3'UTR

Degree of heterozygosity calculated as the fraction of individuals who were heterozygous for the given locus.

Given that four of the VNTRs (TR09, TR17, TR22, and TR30) have been characterized through polymerase chain reaction (PCR) amplification (Vandenbergh et al. 1992; Guindalini et al. 2006; Kim et al. 2017), we focused on a novel hyVNTR located in intron 8 (TR21), upstream of the published VNTR in this intron (Guindalini et al. 2006). This site was not amplified by PCR despite numerous attempts to generate DNA for further analysis (data not shown). The hyVNTR in intron 8 had alleles ranging from 3.4-133.4 copies of the consensus sequence in the long-read sequence data. The human reference genome (GRCh38/hg38) shows this repeat as having 3.4 copies, which was the smallest and most common (6.1%) version of the repeat detected.

Table 2 – VNTR Consensus Sequences

Tandem Repeat ID	Consensus Sequence
TR09	TGGCCACCACCGTTCAAGGGAGCCATTTCCTCACCCAGGTGCCCAGGGAAGCATCCAGGAGGGGAC
TR17	TGTGGGCAGCGGTGGGTACCCAGCACCGTGGGCAGCAC
TR21	CCCCCACCCAGCGCCTTCCCCGCCCTGCCCCTCCAGGC
TR22	TGTGTCTGAGTGTGTATGTTGCATGGTATG
TR30	AGGAGCGTGTACTACCCCAGGACGCATGCAGGGCCCCCAC

Consensus sequences of the human reference genome (GRCh38/hg38) VNTRs are shown in the genomic orientation (antisense to *SLC6A3*).

Repeat Copy Number and Repeat Sequence Purity are Highly Variable Across *SLC6A3* VNTRs

To compare both the individual repeat sequences within each VNTR, and these sequences across all *SLC6A3* assemblies, we generated what we have termed "Mola" charts, which graphically demonstrate alignment similarity between each individual repeat sequence unit of the VNTRs and their associated consensus sequence (see Methods for details). We generated two general types of Mola charts, a multi-color chart that displays a random color for each unique repeat sequence across our sample, and a 3-color chart where, in a color gradient, blue shows a high sequence similarity and red shows a week sequence similarity to the repeat's consensus sequence. A complete documentation of each VNTR's Mola charts can be found in **Table S3**.

Multi-colored Mola charts for each VNTR are shown in **Figure 2.** Two of the VNTRs displayed both a large number [206 for TR21 (**Figure 2A**) and 91 for TR17 (**Figure 2B**)] and a large proportion of unique repeat sequences (14% for TR17 and 71% for TR21). Only one individual is homozygous for TR21 based on length, and this person's alleles differ based on sequence of the units within the alleles. In contrast, the majority of VNTRs had relatively few differences in repeat copy number (length differences) and fewer unique sequences (color differences) (**Figure 2C-E**).

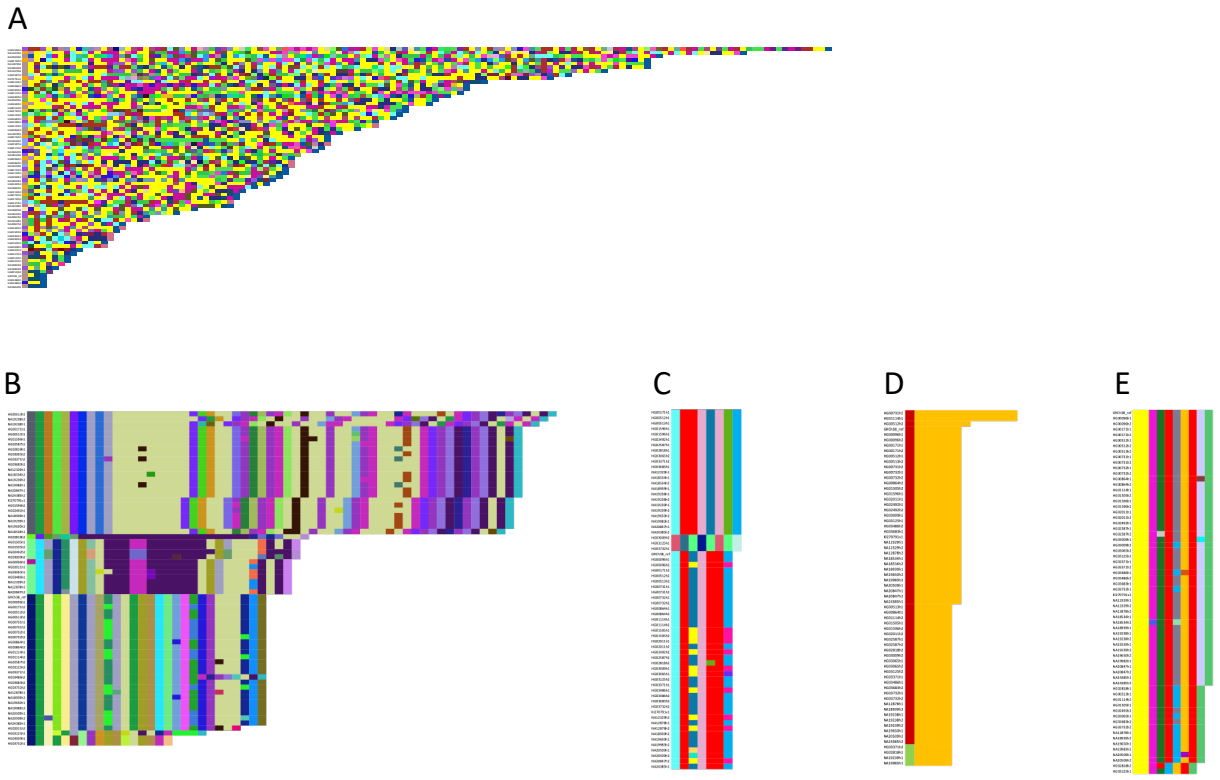

Figure 1 - Multi-colored Mola charts illustrating the degree of variability of each repeat unit in the 5 VNTRs within SLC6A3. (A) TR21, (B) TR17, (C) TR09, (D) TR22, (E) TR30. On the left of each plot are the genomes from which the haplotype was obtained (full names are available in Supplemental Table S3). The two haplotypes for each person are denoted by h1 and h2. The haplotypes are sorted based on the number of repeat units. The unit length is from TRAL, except for TR17 which is from Tandem Repeat Finder. The length of each unit varies due to insertions or deletions.

Three-colored Mola charts for each VNTR are shown in **Figure 3**. The colors are based on sequence similarity to the consensus, which emphasizes the degree of variation from the consensus present in each VNTR. In the case of TR21, no clustering of haplotypes is apparent (**Figure 3A**); however, even for the highly variable TR17 (**Figure 3B**) there are blocks of similarity both in length of the repeat and in similarity of the units. Similar similarities in length and units can also be seen for TR09, TR22, and TR30 (**Figure 3C-E**).

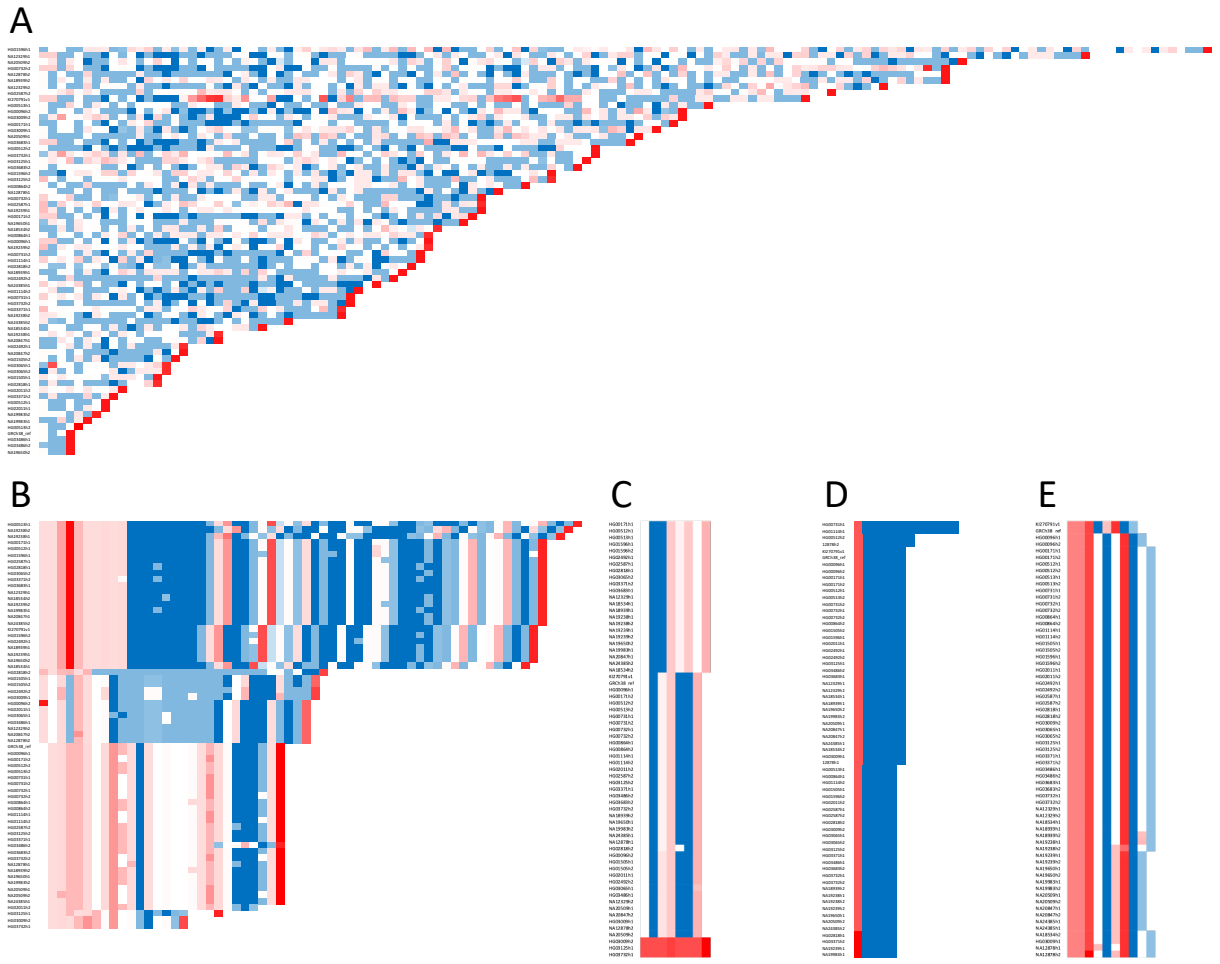

Figure 2 - 3-colored Mola charts illustrating the modes of variability in the 5 VNTRs within SLC6A3. (A) TR21, (B) TR17, (C) TR09, (D) TR22, (E) TR30. Details are as in Fig. 2, but the haplotypes are sorted based on similarity of the first repeat units in a haplotype, except for TR21, which is sorted by number of repeat units due to its complexity. (Full names are available in Supplemental Table S3)

SLC6A3 VNTRs Show Stable Copy Number Inheritance Patterns

In addition to the 64 unrelated *SLC6A3* sequences, we examined six long-read haplotype-phased genomes corresponding to three children and their parents in the original sample. All VNTR copy number alleles were inherited stably from parent to offspring. Global sequence alignments of each child's *SLC6A3* sequence with the corresponding sequence inherited from each parent show an average of 99.95% sequence similarity. In all samples, TR09, TR17, TR22 and TR30 had no intergenerational mismatches, insertions, or deletions; however, TR21

displayed between 1-5 insertions or deletions between generations, suggesting genomic instability at this site (**Table S4).**

SLC6A3 Hyper-Variable VNTR (TR21) is in a Region of Very Low Linkage Disequilibrium

Across the length of the gene, a total of 108 SNPs (excluding any SNPs found within annotated VNTRs) with a minor-allele frequency of greater than 5% in our sample were observed. Ninety-three of the observed SNPs are present in dbSNP (Build 153, https://www.ncbi.nlm.nih.gov/snp/). Correlations among SNP-SNP, SNP-VNTR, and VNTR-VNTR pairs were calculated and are shown in **Figure 4** (see **Table S5** for more detailed information). TR09 (intron 3) and TR17 (intron 4) fell within the same haplotype block and had an r^2 value of 0.96. All other VNTR-VNTR pairs had r^2 values of less than 0.08. Two SNPs (rs458609 and rs393795) were highly correlated with both TR09 and TR17 (all $r^2 > 0.96$). In contrast, the highest SNP-VNTR correlations for TR21, TR22, and TR30 were rs2937640-TR21: $r^2 = 0.11$, rs10074171-TR22: $r^2 = 0.24$, and rs11564775-TR30: $r^2 = 0.49$. All the remaining SNP-VNTR correlations for TR21, TR22, and TR30 were below 0.39. The two closest SNPs flanking TR21 (rs11564759 and rs59133686) showed correlation values with TR21 of 0.016, and 0.001, respectively; these two SNPs showed a correlation value of 0.003 with one another, despite being less than 800 nucleotides on either side of TR21. Due to the low correlations between SNPs and VNTRs in this region, we also report LD values for all SNPs within introns 7 and 8 separately (see **Table S6**). The variation in LD in this region led to a search for binding sites for PRDM9 (PR/SET Domain 9 protein), which plays a critical role in initiation of crossing over (Cheung et al. 2010; Myers et al. 2010; Altemose et al. 2017). The Genome Browser's JASPAR Transcription Factors track for hg38 reveals 139 sites across the entire gene, 6 of which are

present in TR21, which only contains 3.4 copies of the repeat. The haplotype with the largest number of repeats (133.4) is predicted to contain 421 PRDM9 binding sites (**Table S7**) using MEME (Bailey and Elkan 1994). This enrichment of PRDM9 sites within TR21, coupled with the low levels of LD within the region suggest crossing over occurs frequently in and around this site.

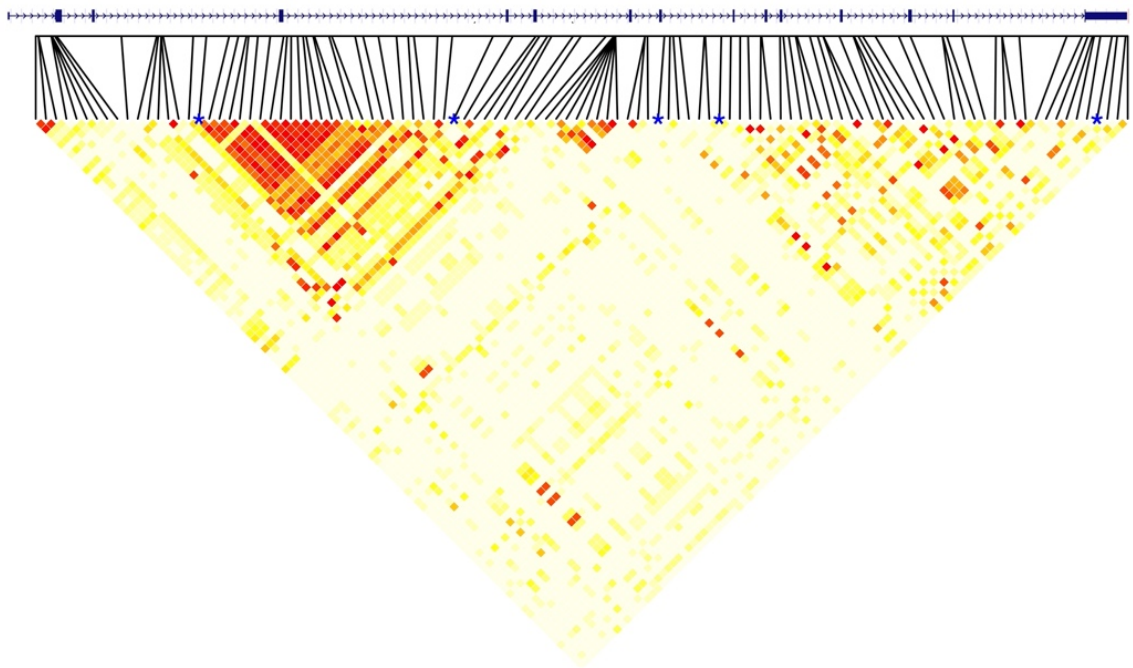

Figure 3 - Heatmap of correlations between all SNPs and VNTRs in SLC6A3. The gene is shown in the 5' to 3' orientation above the heatmap, with black lines indicating the location of each SNP. The five VNTRs are denoted with a blue asterisk above the heatmap, with TR09 on the far left and TR30 on the far right. On the heatmap, red indicates a stronger correlation and white indicates a weaker correlation.

G-Quadruplexes are Present in *SLC6A3* Novel Hyper-Variable VNTR

To explore the functional relevance of the VNTRs we detected, we used the G4-Hunter web application (Brázda et al. 2019) to predict the formation of G4s in the human reference genome (GRCh38/hg38 with 3.4 copies of TR21) and alternate human reference genome (KI270791v1_alt: 64,333-121,212 with 75.4 copies of TR21). The human reference genome results showed no regions of *SLC6A3* having over 50% coverage by predicted G4s (**Figure 5A**,

yellow). In contrast, one distinct region in the alternate human reference genome *SLC6A3* sequence displayed greater than 50% coverage by G4s (**Figure 5A**, purple). The coverage peak (which spans approximately 3,000 bp) begins at gene nucleotide coordinate 33,417 (genome coordinates chr5:1,412,023-1,415,023, GRCh38/hg38) and corresponds with the hyVNTR in intron 8 (TR21) with an average G4 coverage of 99%. A similar peak can be observed in **Figure 5B**, where the human reference genome is predicted to form around 6 G4s in TR21 at its highest, while the alternate human reference genome is predicted to form approximately 25.

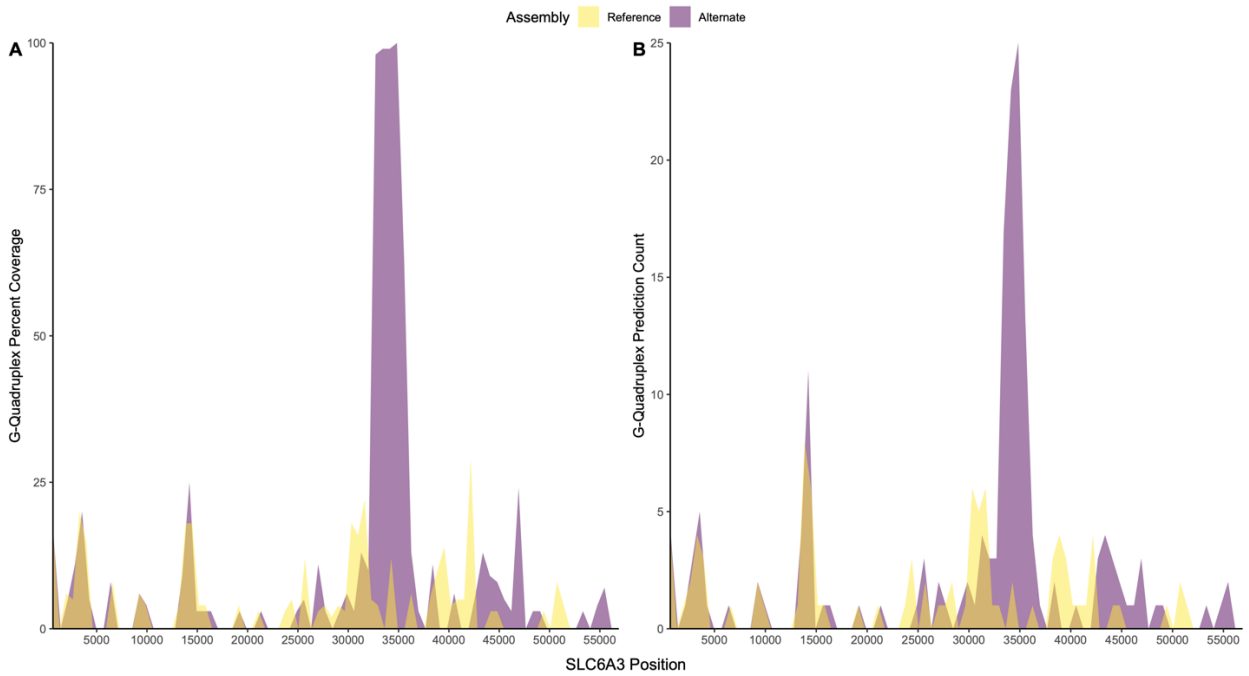

Figure 4 - G4-Hunter results for SLC6A3 sequences from the human reference genome (GRCh38/hg38; yellow) and the alternate human reference genome (KI270791v1; purple). A) shows the G4 coverage percentage in each section of SLC6A3. B) shows the G4 count in each section of SLC6A3.

In addition to using *in silico* methods to predict the presence of G4s, circular dichroism (CD) spectroscopy was used to confirm the presence of the G4s predicted in TR21. An oligonucleotide of 38 bases matching the G-rich strand of the TR21 consensus sequence generated characteristic spectra of G4 structures (i.e., 210 nm and 260 nm peaks and a 240 nm trough) (Kypr et al. 2009) **(Figure 6).** The signal strength at the two peaks and the trough was

increased with addition of potassium-containing buffer and further with additional potassium chloride, consistent with the presence of one or more G4s.

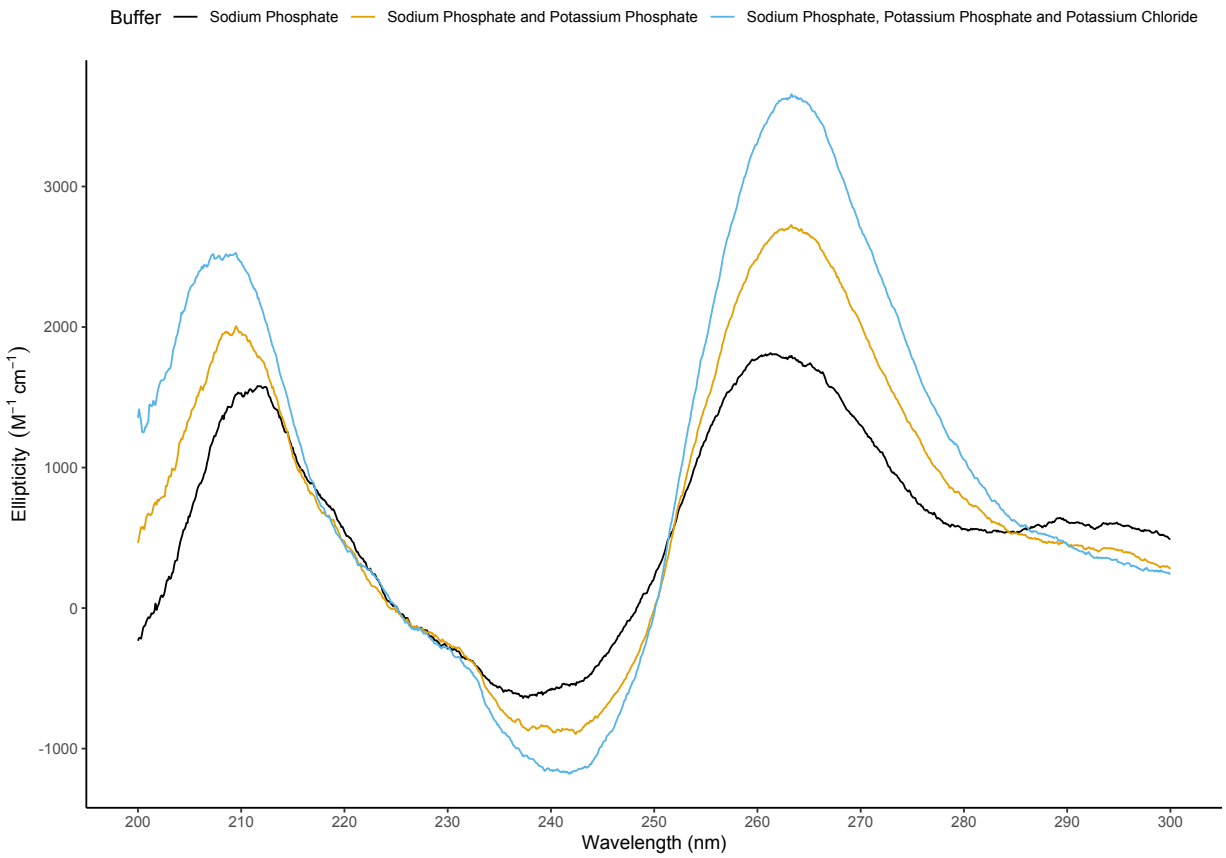

Figure 5 - Circular dichroism results of the TR21 consensus sequence oligonucleotide. Wavelength is shown on the x-axis with ellipticity on the y-axis. Three buffer solutions were used: sodium phosphate (black), sodium phosphate and potassium phosphate (orange), and sodium phosphate, potassium phosphate and potassium chloride (blue). G4 signature peaks and troughs (210 nm and 260 nm peaks and a 240 nm trough) are displayed in all three buffers.

Discussion

We used TRAL to detail the VNTR landscape of *SLC6A3* – a frequent candidate for genetic association studies – in 64 publicly available long-read haplotype-phased genome assemblies. We found a total of 31 TRs present in more than 95% of our sample, with one TR (TR11) being weakly conserved and identified in only ~50% of the sequences. Five of these TRs are variable. Additionally, we detected a novel hyVNTR (TR21) with a consensus sequence of

38bp, which presents new variation to be tested. This hyVNTR is highly variable in repeat copy number, displaying from 3.4-133.4 repeat copies in our sample, and had the highest degree of heterozygosity (**Table 1**) when compared with the four other VNTRs. To the best of our knowledge, TR21 represents a variant previously described using RFLP alleles from Southern blots (Byerley et al. 1993). The longest and shortest alleles in our data differ by approximately 5 kb, suggesting that longer alleles exist, given that the published alleles vary by 7.1 kb (Byerley et al. 1993).

Most studies of VNTRs to date have focused on their variation in repeat copy number; however, more recent data demonstrate that repeat sequence "purity" often varies within VNTRs and this variation can be functionally relevant (Gall-Duncan et al. 2022; Course et al. 2020; Song et al. 2018). In addition to the variability in VNTR copy numbers observed, individual VNTR repeats in *SLC6A3* showed a large range in the amount of sequence heterogeneity present, with TR21 having the largest amount of heterogeneity **(Figure 2**). Disruptions of PRDM9 binding site motifs in TR21 may take place as a result of the impurity of this hyVNTR's repeat sequences.

We found two general LD blocks in *SLC6A3*, the stronger of the two containing both TR09 and TR17. Two SNPs (rs458609 and rs393795) were also highly correlated with both TR09 and TR17; however, the three other VNTRs (TR21, TR22 and TR30) were not in high LD with any SNPs. TR21 and TR22 were both found in a stretch of low LD located in introns 7 and 8, and TR30 was found in the weaker LD block. TR21, the novel hyVNTR, had the smallest maximum correlation with any SNP in the gene body, similar to the VNTR in *TCHH* (Mukamel et al. 2021). This finding suggests the necessity of targeted long-read sequencing of this region in future association studies because the variability in this location is not captured by nearby SNPs. Additionally, previous work has documented the presence of recombination hotspots throughout

the body of *SLC6A3*, specifically near the site of TR21 (Zhao et al. 2019). Our report on the low LD values of TR21 and its neighboring SNPs, along with the density of potential PRDM9 binding sites within the hyVNTR, provide additional evidence that TR21 might be the site of frequent recombination which might contribute to the heterogeneity of copy numbers found in this hyVNTR.

Although each human genome has interpersonal variation, not all variants are biologically or functionally relevant. In addition to searching for recombination-related factors, we also tested for the potential of TR21 to form G4s, which have been identified to play many distinct biological roles in genome function and regulation of gene expression (Li et al. 2016; Kwok and Merrick 2017). We predicted and experimentally confirmed the presence of G4s – which could potentially alter transcription rates, mRNA splicing (Verma and Das 2018), or both – in TR21. The high GC content of this hyVNTR may also explain why the region has not yet been studied using PCR techniques. Additional studies are needed to assess the relation between *SLC6A3* gene expression in dopamine neurons as a function of repeat copy number at this site.

There are some limitations to this study. Although having 64 samples is considered small for association studies, the sample was sufficient to display the heterogeneous nature of each VNTR and to detail the VNTR landscape of *SLC6A3*. The costly nature of long-read sequencing entire genomes precludes analysis of larger numbers of chromosomes. Although both CLR and CCS sequencing technologies were used to generate the long-read haplotype-phased genome assemblies, our comparison of sequencing technologies showed an average similarity of 99.96% and there were no mismatches detected in any of the VNTR regions reported. Therefore, we felt justified in the use of both technologies for our analysis because we were primarily focused on VNTR regions. Finally, manual annotation of TR identities across genome assemblies was

required for detailed analysis. Software that can perform this task automatically would minimize the potential for inadvertent errors.

SLC6A3 and its many polymorphisms have received considerable attention (Salatino-Oliveira et al. 2018). Association studies have linked many SNP and VNTR alleles to specific phenotypes of interest, but often with inconsistent results. Analysis of VNTRs within the gene using targeted long-read sequencing, in particular for TR21, a novel hyVNTR, might lead to strengthened and consistent associations with previously studied phenotypes.

Methods

Obtaining Haplotype-Phased Long-Read Sequences of *SLC6A3*

We used 64 previously generated unrelated long-read haplotype-phased genome assemblies as our study sample (Ebert et al. 2021 – version 1). Each assembly was available as a collection of contigs mapped to their respective chromosomes from which the contig that contained *SLC6A3* on chromosome 5 was extracted. To ensure that the entirety of *SLC6A3* was contained within each extracted contig, we searched for and confirmed the presence of two 30bp sequences, located on either end of the gene (3'UTR: CAGCGGAAACGAGACAAGGAGGCTGAGGCAG (chr5:1,392,763-1,392,793) or its reverse complement, and 5'UTR: AGCCTCGGCCTCGGGCTCTTATCCAGTAGA (chr5:1445441-1445470) or its reverse complement). After extracting the *SLC6A3* sequence from each contig, we oriented each gene in the 5' to 3' direction.

Long-Read Sequencing Technology Comparison

Each CLR sequence was aligned to its corresponding CCS sequence using EMBOSS Kalign v3.3.1 (default parameters). Alignment identity matrices were used to determine the maximum number of mismatching base pairs and pairwise alignment results were used to detect INDELS present between the two sequences.

Tandem Repeat Annotation of *SLC6A3* Sequences

TRs with unit length > 6 bp were detected using PHOBOS (Mayer et al. 2010), TRF (Benson 1999), T-REKS (Jorda and Kajava 2009), and XSTREAM (Newman and Cooper 2007). Using TRALv2.0, a score was assigned to each TR using a phylogenetic model and divergence between repeat units was calculated. Repeats that were found to have arisen due to chance rather than duplication events (likelihood ratio test p-value > 0.05) and those with divergence > 0.05 were discarded. Using common-ancestry clustering, the remaining set of TRs was made nonredundant. In case of overlap between repeats, only the TR with the lowest p-value and divergence was retained. Finally, TRs were further refined by constructing cpHMMs for each repeat and re-annotating them in the original gene sequence for more sensitive detection (Schaper et al. 2014).

We assigned each annotated TR an ID if it was present in more than half of the *SLC6A3* sequences. After each TR was assigned an ID, we determined which repeats had more than one repeat copy number present across individuals in the sample. TRs with more than one allele present in our sample were considered VNTRs. In the case of TR17, TRAL identified a few different length repeats including a 78 bp repeat that in analysis with TRF was found to be two copies of a 38 bp repeat, which was used in subsequent analyses.

Comparing Sequence Identity of VNTRs with "Mola" Charts

To compare both the individual repeat sequences within each VNTR, and these sequences across all *SLC6A3* assemblies, we generated what we have termed "Mola" charts, named after the famous Panamanian textile art. "Mola" charts graphically demonstrate similarities and differences between each individual repeat sequence of the VNTRs. Two different types of Mola charts were generated which we termed "multi-colored" and "3-colored".

For the multi-colored Mola charts, every VNTR's unique sequence patterns were identified and assigned a different, random color. The unique repeats were given a color using Excel developer tools. Each individual's complete VNTR length was then plotted, using the assigned colors to represent the corresponding repeats. The resulting chart demonstrates both the diversity of sequence composition and unique repeat frequency for each VNTR. Plots with a high degree of variability in color would be considered to have a low repeat sequence "purity", and plots with a low degree of variability in color would be considered to have a high repeat sequence "purity".

Microsoft Excel was used to generate 3-color Mola charts. First, a file with separate cells for each copy of a VNTR's repeat unit from each sample was created. This file was then used to generate a global alignment score for each repeat sequence with the VNTR's consensus sequence using EMBOSS needle (Rice et al. 2000) in the Galaxy bioinformatics website (Jalili et al. 2020). A 3-color conditional formatting rule was then applied to the data so that each cell received a color, blue being the most closely aligned to the consensus, red as the least aligned, and white the median value.

VNTR Copy Number Inheritance Patterns

Haplotype-phased long-read genome assemblies of three mother-father-daughter trios (HG00512, HG00513 and HG00514; HG00731, HG00732 and HG00733; NA19238, NA19239 and NA19240) were also downloaded (Ebert et al. 2021 – version 1). Each assembly was previously constructed by Ebert and colleagues using PacBio CCS or CLR reads. *SLC6A3* was extracted from each assembly and oriented as described above. Additionally, all assemblies were annotated for TRs using TRALv2.0 as described above. The TR sites from the original sample of 64 sequences that were determined to be polymorphic in copy number were analyzed in each trio to determine patterns of inheritance. All VNTRs were present in both haplotypes of each child. After parent and child VNTR copy numbers were compared, each child's *SLC6A3* sequence was aligned to its corresponding parentally inherited sequence using EMBOSS Kalign v3.3.1 (default parameters). To further test the fidelity of each VNTR during the inheritance process, each child's VNTR was aligned to the corresponding parental VNTR using EMBOSS Needle v6.6.0 (default settings).

Linkage Disequilibrium and SNP-VNTR Copy Number Correlations

All *SLC6A3* sequences were aligned to the human reference genome (GRCh38/hg38) using the MAFFTv7.504 web interface (added flags included "reorder", "keeplength", and "addfragments"). SNP-sites software (Page et al. 2016) was used to generate a variant calling file (VCF) containing SNPs for each *SLC6A3* sequence. Linkage disequilibrium values for each SNP-SNP, SNP-VNTR, and VNTR-VNTR pair were calculated by squaring their Pearson correlation coefficient using R statistical software (R v 4.1.2). A linkage disequilibrium heatmap was produced using the *LDheatmap* function (Shin et al. 2006) in R (R v 4.1.2). All SNP-VNTR correlations were examined to determine if any previously reported SNPs (dbSNPv153) were

highly correlated with VNTR copy number. All VNTR-VNTR correlations were also examined to determine if any correlations between VNTR alleles were present. Correlations between all SNPs and VNTRs can be found in **Table S5**.

G-Quadruplex Predictions and Experimental Data

G4-Hunter software was used to predict the presence of G4s *in silico* (http://bioinformatics.ibp.cz); "window" and "threshold" parameters were set to 25 and 1.6, respectively, as has been previously recommended (Bedrat et al. 2016). Potential areas of G4 formation were marked and compared with VNTR regions previously annotated. VNTRs that contained high G4 predictions values (greater than 50% of the region covered by G4s) were tested experimentally for the presence of G4s using CD spectroscopy. An oligonucleotide matching the consensus sequence of TR21 (GCCGGGAGGGGCAGGGCGGGGAAGGCGCTGGGTGGGGG) was purchased from Integrated DNA Technologies (Coralville, Iowa; https://www.idtdna.com/pages). The oligonucleotide was tested for G4 formation using a published protocol (Kejnovská et al. 2019).

Data Access

All primary data used to generate haplotype assemblies are publicly available at INSDC under the following accessions and project IDs: Illumina high-coverage genomic sequence (PRJEB37677), HiC (ERP123231), Bionano Genomics (ERP124807), PacBio (PRJEB36100, ERP125611 and PRJNA698480), and Strand-seq (PRJEB39750). All haplotype assemblies used herein are publicly available at the following address: https://www.internationalgenome.org/data-portal/data-collection/hgsvc2

Competing Interest Statement

We have no competing interests to disclose.

Acknowledgements

This research was supported internal funds from the Pennsylvania State University and by the National Institute on Aging Grant T32 AG049676 to The Pennsylvania State University. We would also like to thank Larry John Stone for assistance with Excel data calculations.

References

Altemose N, Noor N, Bitoun E, Tumian A, Imbeault M, Chapman JR, Aricescu AR, Myers SR. 2017. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *eLife* **6**: e28383.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28–36.

Bedrat A, Lacroix L, Mergny J-L. 2016. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* **44**: 1746–1759.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Bidwell LC, Willcutt EG, Mcqueen MB, Defries JC, Olson RK, Smith SD, Pennington BF. 2011. A Family Based Association Study of DRD4, DAT1, and 5HTT and Continuous Traits of Attention-Deficit Hyperactivity Disorder. *Behav Genet* **41**: 165–74.

Borel C, Migliavacca E, Letourneau A, Gagnebin M, Béna F, Sailani MR, Dermitzakis ET, Sharp AJ, Antonarakis SE. 2012. Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB. *Hum Mutat* **33**: 1302–1309.

Brázda V, Kolomazník J, Lýsek J, Bartas M, Fojta M, Šťastný J, Mergny J-L. 2019. G4Hunter web application: a web server for G-quadruplex prediction ed. J. Hancock. *Bioinformatics* **35**: 3493–3495.

Byerley W, Hoff M, Holik J, Caron MG, Giros B. 1993. VNTR polymorphism for the human dopamine transporter gene (DAT1). *Hum Mol Genet* **2**: 335.

Cheung VG, Sherman SL, Feingold E. 2010. Genetics. Genetic control of hotspots. *Science* **327**: 791–792.

Cook EH, Stein MA, Krasowski MD, Cox NJ, Olkon DM, Kieffer JE, Leventhal BL. 1995. Association of attention-deficit disorder and the dopamine transporter gene. *Am J Hum Genet* **56**: 993–998.

Course MM, Gudsnuk K, Smukowski SN, Winston K, Desai N, Ross JP, Sulovari A, Bourassa CV, Spiegelman D, Couthouis J, et al. 2020. Evolution of a Human-Specific Tandem Repeat Associated with ALS. *Am J Hum Genet* **107**: 445–460.

Course MM, Sulovari A, Gudsnuk K, Eichler EE, Valdmanis PN. 2021. Characterizing nucleotide variation and expansion dynamics in human-specific variable number tandem repeats. *Genome Res* **31**: 1313–1324.

De Roeck A, Duchateau L, Van Dongen J, Cacace R, Bjerke M, Van den Bossche T, Cras P, Vandenberghe R, De Deyn PP, Engelborghs S, et al. 2018. An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta Neuropathol (Berl)* **135**: 827–837.

Delucchi M, Näf P, Bliven S, Anisimova M. 2021. TRAL 2.0: Tandem Repeat Detection With Circular Profile Hidden Markov Models and Evolutionary Aligner. *Front Bioinforma* **1**. https://www.frontiersin.org/articles/10.3389/fbinf.2021.691865/full (Accessed July 8, 2021).

du Plessis S, Bekker M, Buckle C, Vink M, Seedat S, Bardien S, Carr J, Abrahams S. 2020. Association Between a Variable Number Tandem Repeat Polymorphism Within the DAT1 Gene and the Mesolimbic Pathway in Parkinson's Disease. *Front Neurol* **11**: 982.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. https://science-sciencemag-org.ezaccess.libraries.psu.edu/content/early/2021/02/24/science.abf7117 (Accessed March 9, 2021).

Franke B, Hoogman M, Vasquez AA, Heister JG a. M, Savelkoul PJ, Naber M, Scheffer H, Kiemeney LA, Kan CC, Kooij JJS, et al. 2008. Association of the dopamine transporter (SLC6A3/DAT1) gene 9–6 haplotype with adult ADHD. *Am J Med Genet B Neuropsychiatr Genet* **147B**: 1576–1579.

Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. 2022. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. *Genome Res* **32**: 1–27.

Garg P, Martin-Trujillo A, Rodriguez OL, Gies SJ, Hadelia E, Jadhav B, Jain M, Paten B, Sharp AJ. 2021. Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *Am J Hum Genet* **108**: 809–824.

Guiblet WM, DeGiorgio M, Cheng X, Chiaromonte F, Eckert KA, Huang Y-F, Makova KD. 2021. Selection and thermostability suggest G-quadruplexes are novel functional elements of the human genome. *Genome Res*.

Guindalini C, Howard M, Haddley K, Laranjeira R, Collier D, Ammar N, Craig I, O'Gara C, Bubb VJ, Greenwood T, et al. 2006. A dopamine transporter gene functional variant associated with cocaine abuse in a Brazilian sample. *Proc Natl Acad Sci* **103**: 4552–4557.

Hannan AJ. 2021. Repeat DNA expands our understanding of autism spectrum disorder. *Nature* **589**: 200–202.

Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19**: 286–298.

Huppert JL, Balasubramanian S. 2005. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* **33**: 2908–2916.

Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, Nekrutenko A. 2020. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res* **48**: W395–W402.

Johansson PA, Brattås PL, Douse CH, Hsieh P, Adami A, Pontis J, Grassi D, Garza R, Sozzi E, Cataldo R, et al. 2022. A cis-acting structural variation at the ZNF558 locus controls a gene regulatory network in human brain development. *Cell Stem Cell* **29**: 52-69.e8.

Jorda J, Kajava AV. 2009. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* **25**: 2632–2638.

Kejnovská I, Renčiuk D, Palacký J, Vorlíčková M. 2019. CD Study of the G-Quadruplex Conformation. *Methods Mol Biol Clifton NJ* **2035**: 25–44.

Kim W-T, Lee S-R, Roh Y-G, Kim SI, Choi YH, Mun M-H, Jeong M-S, Koh SS, Leem S-H. 2017. Characterization of VNTRs Within the Entire Region of SLC6A3 and Its Association with Hypertension. *DNA Cell Biol* **36**: 227–236.

Kirchheiner J, Nickchen K, Sasse J, Bauer M, Roots I, Brockmöller J. 2007. A 40-basepair VNTR polymorphism in the dopamine transporter (DAT1) gene and the rapid response to antidepressant treatment. *Pharmacogenomics J* **7**: 48–55.

Kondo J, Adachi W, Umeda S, Sunami T, Takénaka A. 2004. Crystal structures of a DNA octaplex with I-motif of G-quartets and its splitting into two quadruplexes suggest a folding mechanism of eight tandem repeats. *Nucleic Acids Res* **32**: 2541–2549.

Kwok CK, Merrick CJ. 2017. G-Quadruplexes: Prediction, Characterization, and Biological Application. *Trends Biotechnol* **35**: 997–1013.

Kypr J, Kejnovská I, Renčiuk D, Vorlíčková M. 2009. Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res* **37**: 1713–1725.

Lalioti MD, Scott HS, Buresi C, Rossier C, Bottani A, Morris MA, Malafosse A, Antonarakis SE. 1997. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* **386**: 847–851.

Landefeld CC, Hodgkinson CA, Spagnolo PA, Marietta CA, Shen P-H, Sun H, Zhou Z, Lipska BK, Goldman D. 2018. Effects on gene expression and behavior of untagged short tandem repeats: the case of arginine vasopressin receptor 1a (AVPR1a) and externalizing behaviors. *Transl Psychiatry* **8**: 1–10.

Li Y, Syed J, Suzuki Y, Asamitsu S, Shioda N, Wada T, Sugiyama H. 2016. Effect of ATRX and G-Quadruplex Formation by the VNTR Sequence on α-Globin Gene Expression. *ChemBioChem* **17**: 928–935.

Mayer C, Leese F, Tollrian R. 2010. Genome-wide analysis of tandem repeats in Daphnia pulex - a comparative approach. *BMC Genomics* **11**: 277.

Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Zheng Y, McCarroll SA, Loh P-R. 2021. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. 8.

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**: 876–879.

Newman AM, Cooper JB. 2007. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* **8**: 382.

Örd T, Puurand T, Örd D, Annilo T, Möls M, Remm M, Örd T. 2020. A human-specific VNTR in the TRIB3 promoter causes gene expression variation between individuals. *PLOS Genet* **16**: e1008981.

Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genomics* **2**: e000056.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet TIG* **16**: 276–277.

Ruggieri A, Naumenko S, Smith MA, Iannibelli E, Blasevich F, Bragato C, Gibertini S, Barton K, Vorgerd M, Marcus K, et al. 2020. Multiomic elucidation of a coding 99-mer repeat-expansion skeletal muscle disease. *Acta Neuropathol (Berl)* **140**: 231–235.

Salatino-Oliveira A, Rohde LA, Hutz MH. 2018. The dopamine transporter role in psychiatric phenotypes. *Am J Med Genet B Neuropsychiatr Genet* **177**: 211–231.

Schaper E, Gascuel O, Anisimova M. 2014. Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes. *Mol Biol Evol* **31**: 1132–1148.

Schaper E, Kajava AV, Hauser A, Anisimova M. 2012. Repeat or not repeat?—Statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Res* **40**: 10005–10017.

Shin J-H, Blay S, Graham J, McNeney B. 2006. **LDheatmap** : An *R* Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. *J Stat Softw* **16**. http://www.jstatsoft.org/v16/c03/ (Accessed June 14, 2022).

Song JHT, Lowe CB, Kingsley DM. 2018. Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. *Am J Hum Genet* **103**: 421–430.

Theofanopoulou C, Gedman G, Cahill JA, Boeckx C, Jarvis ED. 2021. Universal nomenclature for oxytocin–vasotocin ligand and receptor families. *Nature* **592**: 747–755.

van der Zwaluw CS, Engels RCME, Buitelaar J, Verkes RJ, Franke B, Scholte RHJ. 2009. Polymorphisms in the dopamine transporter gene (SLC6A3/DAT1) and alcohol dependence in humans: a systematic review. *Pharmacogenomics* **10**: 853–866.

Vandenbergh DJ, Persico AM, Hawkins AL, Griffin CA, Li X, Jabs EW, Uhl GR. 1992. Human dopamine transporter gene (DAT1) maps to chromosome 5p15.3 and displays a VNTR. *Genomics* **14**: 1104–1106.

Vasiliou SA, Ali FR, Haddley K, Cardoso MC, Bubb VJ, Quinn JP. 2012. The SLC6A4 VNTR genotype determines transcription factor binding and epigenetic variation of this gene in response to cocaine in vitro. *Addict Biol* **17**: 156–170.

Verma SP, Das P. 2018. Novel splicing in IGFN1 intron 15 and role of stable G-quadruplex in the regulation of splicing in renal cell carcinoma. *PLOS ONE* **13**: e0205660.

Xiao X, Zhang C-Y, Zhang Z, Hu Z, Li M, Li T. 2022. Revisiting tandem repeats in psychiatric disorders from perspectives of genetics, physiology, and brain evolution. *Mol Psychiatry* **27**: 466–475.

Zavodna M, Bagshaw A, Brauning R, Gemmell NJ. 2018. The effects of transcription and recombination on mutational dynamics of short tandem repeats. *Nucleic Acids Res* **46**: 1321–1330.

Zhao J, Zhou Y, Xiong N, Qing H, Wang T, Lin Z. 2019. Presence of recombination hotspots throughout SLC6A3 ed. J. Subramaniam. *PLOS ONE* **14**: e0218129.