Structural characterization of a complex repeat at the *CACNA1C* pan-psychiatric locus

Raquel Moya[a], Xiaohan Wang[b,c], Richard W. Tsien[b,c], Matthew T. Maurano[a,d,1]

[a] Institute for Systems Genetics, NYU School of Medicine, New York, NY 10016, USA;

[b] Neuroscience Institute, NYU School of Medicine, New York, NY 10016, USA;

[c] Department of Neuroscience and Physiology, New York University, New York, NY 10016;

[d] Department of Pathology, NYU School of Medicine, New York, NY 10016, USA

[1] To whom correspondence may be addressed. Email: maurano@nyu.edu

Author contributions: R.M., R.W.T. and M.T.M. conceived and designed the study. R.M. and M.T.M. designed and performed computational analyses. R.M., X.W., R.W.T and M.T.M. interpreted results. R.M. and M.T.M. wrote the manuscript.

ABSTRACT

Genetic variation within intron 3 of *CACNA1C* surrounding a variable-number tandem repeat (VNTR) is associated with schizophrenia and bipolar disorder, but the causal variant(s) and their effects remain unclear. We fine-mapped the association at *CACNA1C* including this VNTR using sequences from 155 long-read genome assemblies. Across global populations, we found 7 alleles (called Types) of the *CACNA1C* VNTR where sequence differences within repeat units revealed distinct VNTR structure. Within the most common VNTR Type, a previously identified polymorphism of repeat unit composition (termed Variable Region 2) was in complete linkage disequilibrium with fine-mapped schizophrenia SNPs. Applying a genotyping strategy to GTEx data that capitalizes on sequence differences between repeat units, we found that the eQTL in brain tissues at Variable Region 2 had a similar effect size and significance as SNP eQTLs. We show that the risk allele of Variable Region 2 decreases *CACNA1C* gene expression. Our long-read-informed approach to genotype structurally complex VNTR alleles in large cohorts permits investigation of other variants missed by short-read sequencing. Our work suggests an effect on gene expression arising from sequence variation within a VNTR and provides a detailed characterization of new alleles at a flagship psychiatric GWAS locus.

INTRODUCTION

Variable number tandem repeats (VNTRs) are one class of repetitive DNA with a repeat unit longer than 6 bp. Studies of VNTRs reveal coding and regulatory domains critical for organism health and evolutionary adaptation. In humans, length polymorphisms of five coding VNTRs, rather than nearby SNPs, drive associations with human phenotypes (Mukamel et al. 2021). Methods to infer tandem repeat length using high-coverage whole-genome DNA sequencing data identify widespread associations between VNTRs and nearby gene expression (Lu et al. 2021; Garg et al. 2022; Bakhtiari et al. 2021; Lu et al. 2023; Mukamel et al. 2023; Ren et al. 2023). Disease studies found non-coding VNTRs in Amyotrophic Lateral Sclerosis (ALS), Alzheimer's disease, and progressive myoclonic epilepsy that are proposed to act in a variety of regulatory roles (Course et al. 2020; De Roeck et al. 2018; Lalioti et al. 1997). In other eukaryotes, tandem repeats generate functional diversity that impacts organism health (Drögemüller et al. 2008; Lohi et al. 2005; Fondon and Garner 2004; Babb et al. 2017; Vinces et al. 2009; Verstrepen et al. 2005; Fidalgo et al. 2006). Thus several lines of evidence suggest the involvement of VNTRs in phenotypic diversity and disease.

However, the complexity of VNTRs introduces genotyping challenges. Standard variant-calling pipelines ignore repetitive regions. Specialized tools estimate VNTR length from short-read data, but accurate inference of repeat structure and variation is difficult when VNTRs exceed sequencing read lengths (Lu et al. 2021; Bakhtiari et al. 2021, 2018). High-quality long-read assemblies enable new genome-scale assessment of polymorphic VNTR sequences by resolving repeat unit order, sequence, and length (Liao et al. 2023; Nurk et al. 2022; Ebert et al. 2021; Chaisson et al. 2019; Bakhtiari et al. 2018), but this leaves the bulk of variation in VNTRs unexplored.

Many genomic loci are associated with psychiatric disorders, but the causal variants driving these associations are complex to decipher, particularly in the appropriate cellular and genomic context. A top genome-wide association study (GWAS) signal for schizophrenia, bipolar disorder, and major depression (Stahl et al. 2019; Mullins et al. 2021; Pardiñas et al. 2018; Ruderfer et al. 2014; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Cross-Disorder Group of the Psychiatric Genomics Consortium 2013; Liu et al. 2011; Ripke et al. 2011; Psychiatric GWAS Consortium Bipolar Disorder Working Group 2011; Ferreira et al. 2008) lies deep within the large ~330-kb third intron of the ~600-kb *CACNA1C* gene (**Fig. 1A**). *CACNA1C* encodes the pore-forming subunit of CaV1.2, the predominant L-type voltage-gated calcium channel expressed in human central nervous system neurons (Hodge et al. 2019; Miller et al. 2014). CaV1.2 channels on the postsynaptic membrane play a dominant role in triggering a signaling cascade that culminates in the phosphorylation of the nuclear transcription factor CREB (Deisseroth et al. 1996; Wheeler et al. 2012; Ma et al. 2014; Li et al. 2016) and nuclear activation (Vecsey et al. 2007), which is important for learning and memory across the evolutionary tree (Yin et al. 1995; Impey et al. 1996; Bartsch et al. 1998). A single missense mutation in *CACNA1C* is responsible for the monogenic disorder Timothy Syndrome (Splawski et al. 2004, 2005; Barrett and Tsien 2008). The GWAS signal overlaps an expression quantitative trait locus (eQTL) (Lu et al. 2023), suggesting that the SNPs themselves may affect *CACNA1C* expression. The same association locus also harbors a human-specific VNTR (repeat unit length = 30 bp, average length = 6 kb) that is collapsed to 300 bp in the GRCh38/hg38 reference genome (Song et al. 2018). Specific repeat unit sequences within this VNTR could increase schizophrenia risk by decreasing *CACNA1C* expression (Song et al. 2018). Thus *CACNA1C* has well-established relevance for psychiatric disorders, and other genes at the locus are at great distance, but it remains a mystery exactly how these circumstantial functional signals might act individually or together to increase risk for psychiatric disorders.

Here we analyzed the *CACNA1C* VNTR in 155 long-read haplotype assemblies to characterize its structure, variation, and relationship to the schizophrenia disease association. Our analysis uncovers unexpected patterns of structural diversity within the repeat, and delineates a series of common alleles at two variable regions within the predominant repeat sequence type. We show that one of these variable regions is tightly correlated with both the schizophrenia association and gene expression in brain. Our work thus serves as an example for investigation of other difficult association loci.

RESULTS

Analysis of the *CACNA1C* VNTR in long-read assemblies.

To comprehensively map the repeat structure and variation of the *CACNA1C* VNTR, we analyzed phased, long-read genome assemblies from the Human Genome Structural Variation Consortium Phase 2 (HGSVC2) (Ebert et al. 2021), the Human Pangenome Reference Consortium (HPRC) (Liao et al. 2023), and the Telomere-to-Telomere project (Nurk et al. 2022). This collection contained 155 genomes assembled from 78 distinct individuals, including 3 trios and the haploid hydatidiform mole CHM13 (**Fig. 1B**, **Table S1**). For each assembly, we identified the contig containing the previously identified consensus repeat unit of this VNTR (Song et al. 2018) and extracted the full repeat sequence for analysis (**Data S1**, **Table S2**). We confirmed that each identified VNTR sequence was present on a single contig and flanked by unique sequence.

These sequences recapitulated the known repeat unit sequence motif, with strict conservation across at 22 out of 30 positions (**Fig. 1C**). We identified 34,172 repeat units in total and 158 distinct repeat units, most of which were 30 bp in length (**Fig. 1D**, **Table S3**). We identified 35 distinct repeat units with frequency > 0.0005 that comprise > 98% of *CACNA1C* VNTR sequences. The most common repeat unit, defined as the consensus, was present at 32% of positions across all assemblies and differed from other units by an average of 2.3 substitutions. Seven out of eight positions that were most frequently mutated relative to the consensus repeat unit showed a consistent change to one nucleotide in all respective repeat units. Infrequent repeat units fell into two categories: repeat units of size other than 30 bp (n = 11) and infrequent 30-bp units (frequency ≤ 0.0005, n = 111).

The genomes of chimpanzees and other non-human primates at the orthologous site do not have a repetitive sequence. Chimpanzees and gorillas have the same single 30-bp sequence that differs from the human *CACNA1C* VNTR consensus repeat unit sequence at three nucleotides

(**Fig. 1D**) (Song et al. 2018). The T-to-C and T-to-G mutations at the 3rd and 12th positions of the 30-bp chimpanzee sequence were observed in 42% and 21% of human repeat units, respectively. In contrast, the A at 17th position of the 30-bp chimpanzee sequence is a mutation never observed in 34,172 human repeat units from this dataset. The G at the 17th position in the human consensus repeat unit is rarely mutated; it is a C in only 0.44% of repeat units. Thus, while the *CACNA1C* repeat has a strong consensus sequence, there is significant variability.

Structural characterization of the *CACNA1C* VNTR.

We iteratively performed multiple sequence alignment of all 155 repeat sequences, and converged on 7 distinct Types of the *CACNA1C* VNTR (**Fig. 1E**, **Fig. S1A**). Alignments within each Type showed a high proportion of a single repeat unit at each non-gap position (**Fig. 1E**). Each VNTR Type was supported by at least one PacBio HiFi assembly (**Fig. S1B**). For each VNTR Type, we generated a consensus sequence from the most frequent repeat unit at each position that best represented its distinct length and repeat unit order (**Fig. 1E**, **Data S2**).

Type 1, representing about 90% of sequences, defined the most common length (5.7 kb) (**Fig. S1C**). Type 2 and Type 3 sequences were 30-40% longer than Type 1 sequences with less consistency in length within each Type (**Fig. S1A**). Type 4 sequences (frequency = 0.02) had identical lengths (7.2 kb) and very few sequence differences among them. Type 5 sequences (frequency = 0.03) were the longest (12.87-61.198 kb) and most variable (s.d. = 21 kb) (**Fig. S1A**).

All VNTR Types had similar amounts of the most common repeat unit, while the amount of other units differed (**Fig. 2A**). A minority of repeat units (37%, n = 13/35) were found in all VNTR Types, while most of the remaining repeat units were segregated by VNTR Type (**Fig. 2B**) and occurred at low frequencies (**Fig. 2C**). In particular, Types 4-7 contained a greater proportion of repeat units found less frequently in Types 1-3. By contrast, the relative quantity of each repeat unit in Types 1, 2, and 3 was nearly identical (**Fig. 2A**).

These Types represented dissimilar structural *CACNA1C* VNTR alleles. Only the first repeat unit was shared among all Types, though the first 12 repeat units were shared among most Types (**Fig. 2D**). The 3' end was also variable and no position was shared across all Types (**Fig. 2E**). Types 6 and 7 were defined by single sequences that were distinct from other Types in length (**Fig. 1E**), repeat unit order (**Fig. S1A**), and repeat unit quantities (**Fig. 2A**). Overall, these sequences attest to an exceptional diversity in repeat structure.

Kilobase-scale duplications within the *CACNA1C* VNTR.

Given the high degree of similarity between relative quantities of distinct repeat units (**Fig. 2A**) in Types 1, 2, and 3, we reasoned that they might be structurally related. We computed dosage of a sliding window (width = 6 repeat units) across these three consensus sequences to systematically assess their shared and unique sequences. This identified two separate kilobase-sized tandem duplications that distinguished Type 2 and Type 3 from Type 1, suggesting that Type 2 and Type 3 sequences are derived from Type 1 (**Fig. 3A**). Within Type 2, individual repeat sequences showed significant variation (**Fig. S1A**). We identified a single Type 3 sequence (NA21309_paternal) that had a third copy of Duplication 2 with the same breakpoints (**Fig. 3B**).

Two smaller tandem duplications were identified in Types 1, 2, and 3 using this approach. A 420-bp duplication (Tandem Duplication 1) was not exactly in tandem; it was separated by one repeat unit. A 300-bp duplication (Tandem Duplication 2) overlapped part of VR1. Their partial overlap with Duplication 1 and Duplication 2 in Type 2 and Type 3 suggested that the smaller tandem duplications preceded larger ones in time. This pattern reveals that one mechanism of expansion of this VNTR is successive tandem duplication of incrementally larger VNTR segments.

The alignment of Type 5 sequences also showed significant heterogeneity (**Fig. S1A**). Scanning for duplications within Type 5 revealed multiple kilobase-sized tandem duplicated segments (**Fig. 3C**). To facilitate analysis of Type 5, we established a Type 5 exemplar sequence from HG00735_paternal omitting several small insertions not found in other Type 5 sequences (**Fig. 3D**). HG00735_paternal was selected because it had a dosage of the sliding window equal to 1 across most of the sequence. The other Type 5 sequences also showed structural variety: HG02717_paternal contained large deletions, HG03579_maternal had a kilobase-sized tandem Duplication 1, and HG02818_maternal contained a complex tiling path starting with Duplication 2 and proceeding with multiple smaller interspersed duplications, followed by two more copies (a quadruplication) of an 8.79-kb segment, suggesting a complex mechanism of expansion resulting in this sequence (**Fig. 3D**). This degree of structural variation might indicate that Type 5 is particularly prone to rearrangement.

Type 5 was also distinguished from Types 1-3 by the high density of repeat unit mismatches throughout the sequence relative to the exemplar sequence. Relative to HG00735_paternal, the other three Type 5 sequences had repeat unit mismatches occur across the entire sequence. In

contrast, Types 1-3 had repeat unit mismatches concentrated in a 690-bp segment, leading us to reason that this segment had a different repeat unit composition in Types 1, 2, and 3.

We also scanned for duplications within the Type 4, 6, and 7 consensus sequences, and did not find large tandem duplications (**Fig. S2**). Thus the *CACNA1C* VNTR is characterized by a predominant structure (Types 1-3) that account for 95% of sequences, but the remaining 5% harbor a significant amount of structural diversity.

***CACNA1C* repeat polymorphism.**

We investigated the extent to which the Type 1 VNTR sequences that comprise 88% of our dataset varied across individuals. We computed a positional variability score along their multiple sequence alignment (**Fig. 4A**). Two previously identified regions (Song et al. 2018) stood out because of their high variability: Variable Region 1 (VR1) and Variable Region 2 (VR2) (**Fig. 4B**, **Fig. S3**). We then characterized multiple alleles at each of these regions.

We tabulated 18 distinct VR1 sequences spanning 17 aligned repeat units. We grouped them into two alleles and identified consensus sequences for VR1A and VR1B (**Fig. S4A-C**, **Table S4**). VR1A and VR1B differed by 11 repeat units and 330 positions at the nucleotide level. VR1B was a truncation of VR1A (**Fig. 4C**), suggesting that VR1B resulted from one or multiple ancestral deletion events. VR1A had a frequency of 64%, which was slightly higher than VR1B (**Fig. 4D**). Most VR1 sequences matched the consensus sequence and any VR1 sequence that differed from it had an average of 2 repeat units changed. One VR1 sequence found in two individuals was left unclassified due to the high number of mismatches to both VR1A and VR1B (**Fig. S4A,C**).

For VR2, we tabulated 71 distinct VR2 sequences spanning 45 aligned repeat units. We grouped VR2 sequences into four alleles: VR2A, VR2B, VR2C, VR2D (**Fig. 4E**, **Fig. S5A-C**, **Table S5**). The large number of repeat unit differences between VR2 alleles corresponded to a large number of differences at the nucleotide level (including alignment gaps) (**Fig. S5B**, **Table S6**). Only 25% of sequences matched a VR2 allele sequence exactly, but all sequences had fewer than 25% of repeat units deviating from their respective VR2 allele sequence that corresponded to a low number of intra-allele nucleotide differences (**Table S7**). One VR2 sequence was unclassified due to the high number of mismatches to any VR2 allele, and was identified in the same sequences as the unclassified VR1 sequence (**Fig. S3**). The four VR2 alleles formed two common alleles and

two rarer alleles (**Fig. 4F**) with similar lengths and defined by different repeat units and their orders (**Fig. 4E**), suggesting that VR2 derived from a more complex mutational process than VR1.

Comparison of two full *CACNA1C* VNTR sequences with different VR1 and VR2 alleles confirmed that VR1 and VR2 alleles corresponded to distinct sequences at the nucleotide level (**Fig. S4B** and **Fig. S5B**). Ordering Type 1 sequences by VR2 allele revealed high LD between VR1 and VR2, with VR1A found in the same sequences as VR2A, VR2C and VR2D, and VR1B with VR2B (**Fig. S3**). VR1 and VR2 alleles were significantly expanded compared to the previously published description (Song et al. 2018) (**Fig. S6**).

We then scanned Type 2 and Type 3 sequences for the VR alleles identified in Type 1. We classified both VR1 and VR2 all Type 2 and Type 3 sequences except one Type 2 sequence with a VR2 sequence that had a high number of mismatches to every existing VR2 allele (**Fig. S7A-D**). We identified 6 distinct VR1 sequences and 10 distinct VR2 sequences (**Fig. S4D** and **Fig. S5D**). Only two VR1 sequences and one VR2 sequence had been previously identified in Type 1 sequences. Type 2 and Type 3 sequences had duplications of VR1 (**Fig. S7A-B**) and different frequencies of VR2 compared to Type 1 (**Fig. S7E**). At least two copies of VR1A were found in half of Type 2 sequences (n = 4/8 sequences) and all Type 3 sequences (n = 4/4 sequences) (**Fig. S7A-B**). The remaining Type 2 sequences contained two copies of VR1B (**Fig. S7A**). Consistent with the triplication we observed in one Type 3 sequence (NA21309_paternal), there were 3 copies of VR1A in this sequence. Therefore, Type 2 and Type 3 sequences are defined by the kilobase-sized tandem duplications that include VR1. The VR1A sequences within a single VNTR sequence were not exact copies: most often, one VR1 sequence matched the consensus VR1A, while the other had more repeat unit mismatches to the consensus VR1A (**Fig. S7A-B**) and was not observed in Type 1 but was specific to either Type 2 or Type 3 (**Fig. S4D**). VR1B, which was duplicated in Type 2 sequences, did not show this pattern (**Fig. S7B**). Instead, VR1B sequences in Type 2 were exact copies, perhaps because VR1B is shorter and therefore has less scope to accumulate mutations. Interestingly, VR1 shared a right breakpoint with Duplication 2 and Triplication in Type 3, suggesting this site might be prone to instability. One full VR2 sequence belonging to an existing VR2 allele was found within most Type 2 and all Type 3 sequences (**Fig. S7C-D**). Type 3 sequences contained exclusively VR2C (n = 4/4 VR2 sequences) and Type 2 sequences had four with VR2B, three with VR2C, and one VR2 sequence that was dissimilar to all four VR2 alleles (**Fig. S5D**, **Fig. S7C**). The same correlation between VR1 and VR2 alleles was observed in Type 2 and Type 3 sequences. These results indicate that despite shared repeat structure, Types 1, 2, and 3 sequences are also distinguished from each other by their VR2 allele

frequencies (**Fig. S7E**). Additionally accounting for VR allele frequencies in Type 2 and Type 3 did not change the global VR1 allele frequencies reported in Type 1 alone; VR2 allele frequencies changed slightly by an increased frequency of VR2C and a comparable decrease in frequency of VR2A (**Fig. S7F-G**).

Variable Region 2 is associated with schizophrenia and *CACNA1C* expression.

We examined the relationship between VR1 and VR2 and the nearby schizophrenia-associated SNPs (**Fig. 5A**). We focused on a subset of long-read assemblies with available phased SNP genotypes (n = 68 HGSVC2 haplotypes) (Ebert et al. 2021). This subset of *CACNA1C* VNTR sequences had comparable VR allele frequencies to the whole dataset (**Fig. 5B,D**). We calculated linkage disequilibrium (LD) between each VR and the genome-wide significant SNPs ($P < 5x10^{-8}$) (Pardiñas et al. 2018) in a 1.4 Mb region around *CACNA1C* (**Fig. 5C,E**, **Fig. S8**). Fine-mapped schizophrenia SNPs showed high LD with VR2 and moderate LD with VR1, while SNPs not associated with disease were less correlated (**Fig. 5C,E**).

The *CACNA1C* VNTR lay among *CACNA1C* eQTLs in three brain tissues profiled by the GTEx project (GTEx Consortium 2020): cerebellum (n = 125), cerebellar hemisphere (n = 125 with 117 eQTLs shared between these replicate tissues), and putamen (n = 1) (**Fig. 5A**). Intersecting these data with GWAS results showed that schizophrenia risk alleles were associated with reduced *CACNA1C* expression. Comparison of eQTL *P* values to LD with VR2 showed that the eQTL was composed of two distinct signals (**Fig. 5F, Fig. S9**), only one of which was correlated with schizophrenia GWAS *P* values. The degree of LD with VR2 was strongly correlated with the statistical significance of the eQTL for *CACNA1C* gene expression ($r^2 = 0.526$, $P < 2.2x10^{-16}$). Thus, our analysis identifies VR2 as a potentially functional variant at this locus.

Ancestry and history of the *CACNA1C* VNTR.

Our dataset comprised predominantly of individuals with African ancestry and a minority with Asian and European (**Fig. 6A**). While Type 1 sequences reflected this overall composition, Type 2 and Type 3 showed a dramatically different ancestry breakdown (**Fig. 6B**): Type 2 was found mostly in African individuals, while Type 3 was found mostly in Asian individuals, suggesting two separate divergence events that are now segregated geographically. Types 4, 5, 6, and Type 7,

were also found in African individuals, which supports the finding that African individuals will contribute more novel alleles of a structural variant compared to non-African individuals (Audano et al. 2019) and that the highest amount of genetic diversity exists within humans of African descent (Sherman et al. 2019).

The ancestry of Type 1 can be further broken down by VR2 allele (**Fig. 6C**). VR2B showed more representation of African and South Asian ancestries, while VR2A was less prevalent in Admixed American and East Asian haplotypes. VR2C showed a large representation of East Asian ancestry. VR2D was found exclusively in African and Admixed American ancestries and had an ancestry breakdown nearly identical to that of Type 5. The two unclassified VR alleles in Type 1 were exclusively African, consistent with high genetic diversity rather than sequencing errors

Their occurrence in different modern populations as well as their distinct structural differences suggests that Types 4-7 emerged in separate populations from Type 3 and possibly from each other. The diploid genotypes of *CACNA1C* VNTR sequences (n = 77 individuals excluding CHM13) also suggested this. In this limited set of individuals, Types 2-7 are observed heterozygous with Type 1 sequences except in 3 individuals with Type 2/4, 2/5, and 2/6 genotypes. The ubiquity of Type 1 indicated that it may have emerged earliest.

To directly investigate the timing of the repeat expansion, we investigated four ancient genomes: three Neandertals (Mafessoni et al. 2020; Prüfer et al. 2017, 2014) and one Denisovan (Meyer et al. 2012), dating to 50-120 kya. DNA sequencing coverage depth over the *CACNA1C* VNTR region indicated the presence of repeat at lengths ranging from 3,360 kb to 10,560 kb, similar to that found in modern humans (**Fig. 6D-E, Fig. S10**). In summary, our results show that the *CACNA1C* VNTR had already expanded prior to the divergence of early hominin populations.

DISCUSSION

Annotation of the *CACNA1C* VNTR sequence from long-read genomes shows it to be a complex repeat with high structural diversity between alleles. Six previously unknown alleles of this VNTR reveal structural patterns of shared and unique VNTR segments, which suggest that Type 2 and Type 3 derive from Type 1, while the other Types emerged independently. Our annotation of variable regions VR1 and VR2 within the most common Type discovered two additional VR2 alleles and a few unclassified VR sequences that are found in 2% (n = 3/155) of *CACNA1C* VNTR

sequences. The catalog of their alleles shows some diversity at VR1 and extensive diversity of alleles and their sequences at VR2. Further investigation will be needed to determine whether the allelic diversity at VR2 is related to a functional property or if it is a byproduct of the emergence of this sequence in humans. The two alleles at VR1 and four alleles at VR2 tag 4 versions of Type 1 sequences, two of which are common. We explicate a more complete compendium of 158 repeat unit variants and their frequencies in individuals from global populations, many of which are infrequent yet consistently observed across *CACNA1C* VNTR sequences. The impossibility of resolving this complex variation of the *CACNA1C* VNTR without long-read, haplotype-resolved genomes underscores the value of this record of its Types, internal duplications, variable regions, and repeat unit variants for genotyping other *CACNA1C* VNTR sequences.

Among 1,584 tandem repeats that are expanded in humans, the one in *CACNA1C* intron 3 is among the 13% (n = 207/1,584) that are collapsed in GRCh38/hg38 and has a maximum length that is ~6X longer than any human-specific VNTR sequence from six long-read haplotypes of Yoruban, Chinese, and Puerto Rican origin (Sulovari et al. 2019). This VNTR has an average repeat unit size relative to other human-specific VNTRs; 30 bp is larger than 46% of human-specific tandem repeats (n = 735/1,584) (Sulovari et al. 2019). The number of Types of the *CACNA1C* VNTR is approximately the expected number of distinct alleles per VNTR (7.5-16.7 alleles) estimated using an efficient set of repeat unit variants defined across all VNTRs in a reference panel of long-read genomes (Ren et al. 2023). In contrast, the amount of repeat unit diversity (n = 158 distinct repeat units) of the *CACNA1C* VNTR greatly exceeded the average number of distinct repeat units per VNTR (8.97 ± 26.57 repeat units) in a genome-wide analysis (Ren et al. 2023). Taken together, these analyses contextualize the *CACNA1C* VNTR as one of the human-specific VNTRs with the highest length variability and repeat unit diversity in the genome. One caveat is that both of these studies analyzed a subset of the 155 long-read haplotype assemblies described in this manuscript; thus, the metrics were limited to the allelic diversity of the dataset.

Relative to the initial 27 *CACNA1C* VNTR sequences analyzed (Song et al. 2018), which in retrospect were all Type 1 sequences, this dataset of 155 *CACNA1C* VNTR sequences has the same consensus repeat unit and a similar length distribution as 362 *CACNA1C* VNTRs analyzed using Southern blot (Song et al. 2018). It is unclear whether the initial study examined any individuals that overlap with our dataset. Unlike many known diseases caused by a repetitive sequence, the length of the *CACNA1C* VNTR has not been associated with alleles at schizophrenia and bipolar disorder GWAS SNPs. This was first tested for the modest length range (3.2-6 kb) of

21 *CACNA1C* VNTR sequences in the initial study (Song et al. 2018). We recapitulated no association between a wider length range (3.21-61.198 kb) and *CACNA1C* gene expression (which is correlated with schizophrenia GWAS *P* values). Instead, given its striking sequence diversity we turned to our map of the *CACNA1C* VNTR sequence for further investigation of the *CACNA1C* GWAS signal. Previously, the *CACNA1C* VNTR sequence was associated with schizophrenia and bipolar disorder through individual repeat units with a propensity to occur in DNA sequencing reads from homozygous risk or homozygous protective individuals at four nearby GWAS SNPs (Song et al. 2018). Our data expands upon this finding by fine-mapping the entire association at this locus, including VR1 and VR2 of the *CACNA1C* VNTR that tag different versions of the sequence. By quantifying LD between VR2 and SNPs at this locus, we show that VR2 is in perfect LD with 3 FINEMAP SNPs that flank the *CACNA1C* VNTR, which are statistically prioritized SNPs after accounting for haplotype structure (Benner et al. 2016). This result suggests that SNP-focused analyses have been unable to discriminate the effect of the VNTR from single nucleotide variants.

A key question is the cell and tissue types relevant to the GWAS signal. SNPs in *CACNA1C* intron 3 are associated with decreased *CACNA1C* expression, which was shown in superior temporal gyrus (Eckart et al. 2016) and cerebellum (Gershon et al. 2014), although the direction of effect is not consistently observed. Studies in dorsolateral prefrontal cortex (Bigos et al. 2010) and induced neurons (Yoshimizu et al. 2015) report an association with increased *CACNA1C* expression. Our results corroborate a decrease in *CACNA1C* expression in brain tissue, namely cerebellum. We show that the statistical significance of *CACNA1C* eQTLs and schizophrenia GWAS SNPs increases with proximity to the VNTR, particularly VR2. These data underscore that the expected outcome of the causal variant(s) is regulation of *CACNA1C* expression. Additionally, this nominates the sequence of the *CACNA1C* VNTR as a possibly causal variant at this GWAS locus with particular focus on VR2. However, schizophrenia is not thought to involve brain networks in cerebellum; instead, several cortical networks are implicated, many of which involve prefrontal brain regions (van den Heuvel and Fornito 2014). One plausible explanation for detecting expression changes in cerebellum is its high cell type uniformity. The cerebellum contains 69 billion neurons, or 80% of the neurons in the human brain, and 99% of these neurons are granule cells (Azevedo et al. 2009; Consalez et al. 2020). Thus, the power to detect *CACNA1C* eQTLs in the bulk tissue of the cortex is complicated by a high prevalence of glial cells that don't express *CACNA1C* and its neuron cell type heterogeneity (Li et al. 2023). Although our results show an association between a *CACNA1C* VNTR allele and gene expression, the mechanism may be

more complex than direct regulation. For example, VR2 is in high LD with VR1 and the two regions on average exist 1.44 kb apart. Both could be important for modulating *CACNA1C* expression or other genomic features like DNase I hypersensitive sites could facilitate the effect of the causal variant(s).

One limitation of this study is that the sequences of rare *CACNA1C* VNTR Types or infrequent repeat units are only as accurate as the long-read haplotype assemblies that they are extracted from. The error rates of the assemblies in this dataset were estimated to be low at < 1 in 10,000 for PacBio continuous long read assemblies (HGSVC2) (Ebert et al. 2021), < 1 in 100,000 for PacBio HiFi assemblies (HGSVC2, HPRC) (Ebert et al. 2021; Liao et al. 2023), and < 1 in 10 million for the haploid CHM13 assembly created with PacBio HiFi reads that didn't require haplotype phasing (Nurk et al. 2022). However, repeat units that occur in few *CACNA1C* VNTR Types may represent true repeat units or infrequent sequencing errors. Surprisingly, we observe 11 distinct repeat units of size other than 30 bp that occur 14 times across this set of *CACNA1C* VNTR sequences. It is unclear whether any of these are the result of a sequencing error. Most are found in Type 5 sequences and most are only 1 bp longer or shorter than a typical 30 bp repeat unit. Otherwise, evidence suggests that the *CACNA1C* VNTR sequences identified in each genome are high-fidelity. Exactly two alleles were extracted from each genome and each was found on a single contig flanked by unique sequence. No individual had two identical alleles, lowering the chance that a read was assigned to the incorrect haplotype. For the 5 individuals that had assemblies created by HGSVC2 and HPRC, four had identical *CACNA1C* VNTR sequences in the assemblies created by both projects. Additionally, for the 3 trios in this dataset, children had two parental *CACNA1C* VNTR sequences with one from their mother and one from their father. Lastly, 5 out of 7 Types were represented by multiple sequences. VR sequences were assigned to their respective alleles tolerant of multiple mismatches to the consensus VR sequence. Thus, the collated model of this VNTR is robust to sequencing errors present at the individual sequence level.

Our analysis discretizes *CACNA1C* VNTR VR sequences that exist among a spectrum of human variation, which advantageously summarizes complex genomic data yet deemphasizes similarities across categories. Particularly, partitioning VR2 sequences into alleles had a number of exceptions. First, some VR2C sequences partially resembled both VR2B and VR2C, but were representative of neither. These were partitioned to VR2C by an edit distance threshold that relies on the underlying alignment for which there may be multiple equally optimized solutions. Second and related, the number of nucleotide positions (including gaps) that were different between VR2B and VR2C was 44, but the average difference between a non-consensus VR2B sequence and its

consensus was 45.6 aligned nucleotides. Thus the intra-allele variability was higher than the inter-allele differences for VR2B, which suggests VR2B and VR2C could be more similar than different, particularly in their putative function. One stark difference between the sequences that contain VR2B and VR2C is their alleles at VR1. *CACNA1C* VNTR sequences with VR2B have VR1B, while those with VR2C have VR1A, which underscores the question whether variable structural alleles in high LD should be genotyped separately or together as a feature of the whole structural variant. Lastly, VR2C showed high LD with schizophrenia FINEMAP SNPs too. Though such an uncommon allele (frequency = 12% across Type 1, 2, and 3 sequences) is less likely to be the source of a common association, its relatedness to VR2B and assessment along with it should be considered for future analyses in disease cohorts or assessments of functional effects.

The striking structural diversity of the *CACNA1C* VNTR and our results indicating that several Types are the result of smaller and larger duplications invites speculation around its mechanism of expansion. In contrast with somatic variation at DNA repeats that is proposed to be induced by homologous recombination after erroneous DNA repair or replication slippage, variation within subtelomeric VNTRs may emerge through meiotic recombination events. An enrichment of VNTRs relative to short tandem repeats (STRs) is found near the end of chromosomes (Sulovari et al. 2019; Audano et al. 2019; Barton et al. 2008), and *CACNA1C* lies only 2 Mb from the start of chromosome 12. One possible mechanism for the expansion of the *CACNA1C* VNTR is unequal exchange at meiosis through homologous pairing-dependent events, which could change repeat unit copy number and repeat unit sequences but preserve the repeating frame of the sequence (Wolff et al. 1991; Smith 1976). Models of this mutational process show that over time such repetitive sequences would be unsTable in length. Changes through unequal crossover events between either homologous chromosomes or sister chromatids would inevitably occur that lengthen or shorten the repetitive sequence (Wolff et al. 1991; Smith 1976). Higher meiotic recombination rates and higher frequency of double-strand breaks correlate with higher VNTR density at subtelomeres (Audano et al. 2019). We speculate that evolutionary adaptations, such as highly variable gene families, occur more often at the subtelomeric ends of chromosomes in humans. Examples of subtelomeric gene families driving adaptation to environmental changes exist in human (e.g., the olfactory receptor gene family) (Buck and Axel 1991), yeast (e.g., MAL gene family) (Brown et al. 2010), and the parasite causing malaria in humans *P. falciparum* (e.g., antigen genes) (Corcoran et al. 1988). Furthermore, the number of distinct structural alleles of the *CACNA1C* VNTR provides hints about its patterns of divergence. Given the long amount of time

that the *CACNA1C* VNTR has existed, we expect that these large structural changes were rare and occurred early in the human lineages they are found in today.

Schizophrenia is a heterogenous and etiologically complex psychiatric disorder with an incidence of approximately 1% and about 70% heritability (Lichtenstein et al. 2009; Sullivan et al. 2003). One gene family that is robustly implicated in schizophrenia and other psychiatric disorders is calcium channel genes (Trubetskoy et al. 2022; Curtis et al. 2011), yet the genetic and molecular underpinnings of these associations are not clear. Here we characterize the genetic diversity of a human-specific VNTR at the *CACNA1C* schizophrenia locus and contextualize its effect with single nucleotide variants and eQTLs in brain tissues. We have shown that employing local patterns of LD between a known marker locus and an unknown trait locus can effectively prioritize an unknown variant among other fine-mapped SNPs and reveal a tissue-specific effect. Several unanswered questions remain. While any of the four known mutations causal for Timothy Syndrome result in a gain of CaV1.2 channel function (Barrett and Tsien 2008; Splawski et al. 2004), it's unclear what the consequences of decreased *CACNA1C* expression are on the affected cell types and brain networks. As of yet, functional properties have not been described within *CACNA1C* intron 3, but comparison of the *CACNA1C* VNTR with the 30 bp sequence in chimpanzees shows that substitution at an adenine in the middle of the chimp sequence led to gain of splice acceptor and splice donor sites in the human sequence (Song et al. 2018). Further functional investigation is needed.

The results described here address a tradeoff between conducting a surgical single-locus analysis and consulting a pangenome (Liao et al. 2023) for detailed information on complex genomic regions. While pangenome graphs can be referenced quickly, their accuracy is inherent to the quality of the underlying alignment and they are susceptible to simplify the allelic complexity (Lu et al. 2023). Our map of the *CACNA1C* VNTR marks a milestone in teasing apart the association at a complex GWAS locus. More work is needed to test the function of each candidate causal variant in disease cohorts and a relevant cell type.

MATERIALS AND METHODS

Analysis of VNTR in long and short-read sequencing data.

The CHM13 haploid genome assembly was downloaded from the Telomere-to-Telomere (T2T) GitHub site. Haplotype-resolved assemblies for 34 individuals were downloaded from the HGSVC2 (Ebert et al. 2021) FTP site. In cases where two assemblies exist for an individual, assemblies using PacBio HiFi reads were preferred over PacBio continuous long reads (CLR) . Haplotype-resolved assemblies for 43 individuals were downloaded from the HPRC (Liao et al. 2023) S3 bucket. 5 individuals were sequenced by both HGSVC2 and HPRC (HG00733, HG02818, HG03486, NA19240, NA24385). HGSVC2 assemblies were used for the individuals (HG00733, HG03486, NA19240, NA24385) for which both projects had identical *CACNA1C* VNTR sequences. The HGSVC2 assembly of HG02818 had four contigs containing the consensus *CACNA1C* repeat unit, one of which matched the HPRC HG02818_paternal assembly. None of the other three contigs contained a full *CACNA1C* VNTR with unique flanking sequence, and thus the HPRC assembly was chosen for HG02818.

Corresponding short-read sequencing data (n = 70 individuals) were downloaded from the 1000 Genomes Project (Byrska-Bishop et al. 2022). Download paths are in **Table S1**. Assembly coordinates of each VNTR sequence are listed in **Table S2**.

Assemblies were scanned to identify contigs containing the previously identified *CACNA1C* intron 3 consensus repeat unit (Song et al. 2018). Tandem Repeats Finder (Benson 1999) v4.09 was run on each resulting contig using parameters `2 7 7 80 10 50 32 -m -f -d` to detect patterns near the size of the *CACNA1C* repeat unit. With a matching weight of 2 and a minimum alignment score of 50, assuming perfect alignment at least 25 characters will need to be aligned, which is permissive for a repeat unit of 30 bp. For a targeted approach to find the *CACNA1C* VNTR, the maximum period size is 32 bp.

Consensus *CACNA1C* VNTR sequences by Type.

An initial multiple sequence alignment of all CACNA1C VNTR sequences was created using `MAFFT -text -globalpair -maxiterate 1000` (Katoh and Standley 2013). To align *CACNA1C* VNTR sequences in a unit boundary-aware manner, each VNTR sequence was encoded as a sequence of ASCII characters where each character represents one of the 37 repeat units. Sets of sequences were grouped with similar alignment patterns. Alignments of these sets

were created iteratively until no sequence appeared visually misplaced, evidenced by the presence of large alignment gaps. Final alignments for each Type were created using `MAFFT --op 4 --text --globalpair`.

A consensus sequence for each Type was defined from the most frequent unit at each alignment position. If the most frequent unit was an alignment gap with ≤ 65% frequency, the most frequent non-gap unit was used. In cases where two units were in equal frequency at a position, the more common unit among all 34,172 repeat units was chosen for the consensus.

Assessing duplication within *CACNA1C* VNTR sequences.

An exemplar sequence with no duplications was defined for each VNTR Type. Using a sliding window of N repeat units, every overlapping window along the exemplar VNTR sequence was quantified in each sequence by counting exact matches using `grep`.

Because Types 1, 2 and 3 had similar repeat unit compositions (**Fig. S1C**), the Type 1 consensus sequence was defined as an exemplar sequence for all three. The width of the sliding window for Types 1, 2, and 3 was N = 6 repeat units.

For Types 4, 6, and 7, their consensus sequences were defined as exemplar sequences and scanned for duplications within themselves. The width of the sliding window for Types 4, 6, and 7 was N = 6 repeat units.

Because the Type 5 consensus sequence contained duplications itself, the Type 5 multiple sequence alignment was used to define a duplication-free reference sequence with the highest amount of shared sequence across all four Type 5 sequences. The Type 5 exemplar sequence is an artificial sequence created from the HG00735_paternal *CACNA1C* VNTR sequence by excluding 5 smaller insertions and 3 single repeat unit insertions not found in other Type 5 sequences. The width of the sliding window for Type 5 was N = 8 repeat units.

Identifying variable repeat regions and alleles.

Variability at each position in the multiple sequence alignment was computed as Shannon's uncertainty $H(X) = -\sum_{i=1}^{K} p(x)_i \, log_2(p(x)_i)$ where p_i is the fraction of repeat units of unit type i and K is the number of different repeat units at position X. Gaps are included in the calculation. $H(X)$

was normalized as $H(X)_{normalized} = \frac{H(X)}{log_2(K)}$ and a smoothing filter applied using the R package ksmooth. Variable regions were defined where $H(X)_{normalized} > 0.25$ and length > 7 aligned repeat units. Variable region sequences were inspected manually and narrowed by a maximum of 2 repeat units on either end (**Fig. S2**).

For each variable region, aligned sequences were extracted from Type 1 sequences and partitioned into alleles. First, a consensus sequence for the major allele was defined as the most common VR sequence (**Table S4, Table S5**). Hamming edit distances in repeat units were calculated between the consensus sequence and each unique VR sequence using `StrDist` in the R package DescTools. A consensus sequence for the second major allele was defined as the next most common VR sequence with a large edit distance from the first major allele. Similarly, edit distances (in repeat units) were calculated between the second allele and each unique VR sequence. Partitioning into alleles was done using edit distance thresholds. For example, a threshold *t* was chosen such that a sequence belonging to allele the A allele had less than *t* mismatches (in repeat units, including gaps) to the consensus A sequence and greater than *t* mismatches to the consensus B sequence. For VR1, t = 7.5 and for VR2, t = 10 for VR2A and t = 12.5 for VR2B. If VR sequences clustered into more than two groups, the other groups were manually defined as minor VR alleles. Unclassified VR sequences fell outside the selected edit distance thresholds.

For nucleotide-level alignments (**Fig. S4B** and **Fig. S5B**), each aligned VNTR sequence was converted from encoded characters to nucleotides while preserving the positions of alignment gaps. Global alignments were adjusted using `pairwiseAlignment` from the R package Biostrings to account for re-incorporation of non-30-bp repeat units.

To identify VR alleles in Type 2 and Type 3 sequences, *CACNA1C* VNTR sequences were scanned for matches to each consensus VR allele allowing for maximum 2 mismatches (in repeat units) using `vmatchPattern` from the R package Biostrings.

Calculating linkage disequilibrium between variable regions and nearby SNPs.

We analyzed LD between variable regions and nearby SNPs in 35 individuals in HGSVC2 (Ebert et al. 2021). Phased genotypes of VR1 and VR2 were included for the 34 HGSVC2 individuals contributing long-read assemblies to our dataset and one additional individual (HG02818) se-

quenced by both HGSVC2 and HPRC but whose VR alleles were identified using the HPRC assemblies. Custom VCF entries were created for VR1 and VR2 and inserted into the sorted file with phased variant calls. For each VR, the A allele (VR1A and VR2A) was encoded as the reference allele. VR1 was entered as a single biallelic variant and the four alleles of VR2 were split into three biallelic entries, one for each alternate allele (VR2B, VR2C, and VR2D). Positions in the reference genome for each entry were chosen within the reference *CACNA1C* VNTR region (chr12:2255791-2256090, GRCh38/hg38) on chromosome 12, starting at the first position (2255791-2255795). Only haplotypes with a Type 1 *CACNA1C* VNTR (n = 61) were encoded as either 0 or 1, indicating the reference or alternate VR allele respectively. Haplotypes with a Type 2-7 sequence or an unclassified VR allele were indicated as missing (“.”).

SNP genotypes were taken from the HGSCV2 phased variant calls within the 1.4 Mb region surrounding the *CACNA1C* locus (chr12:1600001-3000000, GRCh38/hg38). Of these, we retained 2,749 SNPs having MAF > 0.1 in 1000 Genomes (or in HGSVC2 if the SNP was not present in 1000 Genomes). SNPs were annotated using `bcftools annotate` and dbSNP 151.

LD was calculated between each SNP-VR pair using `vcftools --hap-r2`.

Data retrieval: GWAS.

Summary statistics of the CLOZUK+PGC2 schizophrenia meta-analysis and the subset of high-quality imputed SNPs (Pardiñas et al. 2018) were obtained from the Psychiatric Genomics Consortium repository. SNP coordinates were converted from GRCh37/hg19 by looking up GRCh38/hg38 coordinates in dbSNP using the rsID. Odds ratio (OR) was used to identify risk (OR > 1) and protective (OR < 1) alleles. Fine-mapped SNPs were obtained from Supplementary Table 4 of the same study (Pardiñas et al. 2018) and similarly converted to GRCh38/hg38.

Analysis of eQTLs from GTEx.

eQTLs from the v8 analysis freeze were downloaded from the GTEx Portal on 09/20/2017. Analysis was limited to significant associations with ENSG00000151067.21 (Ensembl) in 13 brain tissues. For eQTLs colocalized with schizophrenia associations, eQTL effect sizes were defined as the effect of the risk allele relative to the protective allele by inverting the slope of the linear regression when the reference allele matched the risk allele.

To estimate the magnitude of *CACNA1C* expression change associated with SNPs in intron 3, software for calculating the log allelic fold change (aFC) was downloaded from the aFC GitHub site. aFC is equivalent to the expected log-fold expression ratio of the individuals homozygous for the alternate allele to those homozygous for the reference allele of an eQTL (Mohammadi et al. 2017). aFC was calculated using `aFC.py --min_samps 2 --min_alleles 1 --log_xform 1 --log_base 2` with GTEx normalized expression values and covariates in the GTEx brain tissue with the most eQTLs (cerebellar hemisphere), the set of 132 GTEx SNP eQTLs for *CACNA1C* (Ensembl ID ENSG00000151067.21) in cerebellar hemisphere, and GTEx phased variant calls included as input files.

Calculating *CACNA1C* VNTR length by WGS coverage.

Local sequencing coverage was used to estimate *CACNA1C* VNTR length. Average depth over three regions (left flanking 10 kb: chr12:2245791-2255790, right flanking 10 kb: chr12:2256091-2266090, *CACNA1C* VNTR region: chr12:2255791-2256090, GRCh38/hg38) was calculated using `samtools depth -a`. *CACNA1C* VNTR length was estimated by computing the average of the ratio between coverage depths at the VNTR and at each flanking sequence. This length was scaled in terms of the number of repeat units by multiplying by 300/30 to account for the collapsed repeat in the GRCh38/hg38 reference sequence.

Chromosome 12 BAM files for four archaic human individuals were downloaded. *CACNA1C* VNTR lengths were estimated as before, using UCSC liftOver to identify the corresponding regions in GRCh37/hg19.

REFERENCES

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**: 663-675.e19.

Azevedo FAC, Carvalho LRB, Grinberg LT, Farfel JM, Ferretti REL, Leite REP, Filho WJ, Lent R, Herculano-Houzel S. 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology* **513**: 532–541.

Babb PL, Lahens NF, Correa-Garhwal SM, Nicholson DN, Kim EJ, Hogenesch JB, Kuntner M, Higgins L, Hayashi CY, Agnarsson I, et al. 2017. The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression. *Nat Genet* **49**: 895–903.

Bakhtiari M, Park J, Ding Y-C, Shleizer-Burko S, Neuhausen SL, Halldórsson BV, Stefánsson K, Gymrek M, Bafna V. 2021. Variable number tandem repeats mediate the expression of proximal genes. *Nat Commun* **12**: 2075.

Bakhtiari M, Shleizer-Burko S, Gymrek M, Bansal V, Bafna V. 2018. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res* **28**: 1709–1719.

Barrett CF, Tsien RW. 2008. The Timothy syndrome mutation differentially affects voltage- and calcium-dependent inactivation of CaV1.2 L-type calcium channels. *PNAS* **105**: 2157–2162.

Barton AB, Pekosz MR, Kurvathi RS, Kaback DB. 2008. Meiotic recombination at the ends of chromosomes in Saccharomyces cerevisiae. *Genetics* **179**: 1221–1235.

Bartsch D, Casadio A, Karl KA, Serodio P, Kandel ER. 1998. CREB1 encodes a nuclear activator, a repressor, and a cytoplasmic modulator that form a regulatory unit critical for long-term facilitation. *Cell* **95**: 211–223.

Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. 2016. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**: 1493–1501.

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

Bigos KL, Mattay VS, Callicott JH, Straub RE, Vakkalanka R, Kolachana B, Hyde TM, Lipska BK, Kleinman JE, Weinberger DR. 2010. Genetic variation in CACNA1C affects brain circuitries related to mental illness. *Arch Gen Psychiatry* **67**: 939–945.

Brown CA, Murray AW, Verstrepen KJ. 2010. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr Biol* **20**: 895–903.

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**: 175–187.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**: 3426-3440.e19.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.

Consalez GG, Goldowitz D, Casoni F, Hawkes R. 2020. Origins, Development, and Compartmentation of the Granule Cells of the Cerebellum. *Front Neural Circuits* **14**: 611841.

Corcoran LM, Thompson JK, Walliker D, Kemp DJ. 1988. Homologous recombination within subtelomeric repeat sequences generates chromosome size polymorphisms in P. falciparum. *Cell* **53**: 807–813.

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.

Cross-Disorder Group of the Psychiatric Genomics Consortium. 2013. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**: 1371–1379.

Curtis D, Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, Lawrence J, Anjorin A, Choudhury K, Datta SR, et al. 2011. Case-case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. *Psychiatr Genet* **21**: 1–4.

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* **135**: 827–837.

Deisseroth K, Bito H, Tsien RW. 1996. Signaling from synapse to nucleus: postsynaptic CREB phosphorylation during multiple forms of hippocampal synaptic plasticity. *Neuron* **16**: 89–101.

Drögemüller C, Karlsson EK, Hytönen MK, Perloski M, Dolf G, Sainio K, Lohi H, Lindblad-Toh K, Leeb T. 2008. A mutation in hairless dogs implicates FOXI3 in ectodermal development. *Science* **321**: 1462.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**: eabf7117.

Eckart N, Song Q, Yang R, Wang R, Zhu H, McCallion AS, Avramopoulos D. 2016. Functional Characterization of Schizophrenia-Associated Variation in CACNA1C. *PLoS ONE* **11**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4898738/ (Accessed July 22, 2020).

Ferreira MAR, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK, et al. 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* **40**: 1056–1058.

Fidalgo M, Barrales RR, Ibeas JI, Jimenez J. 2006. Adaptive evolution by mutations in the FLO11 gene. *Proc Natl Acad Sci U S A* **103**: 11228–11233.

Fondon JW, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* **101**: 18058–18063.

Garg P, Jadhav B, Lee W, Rodriguez OL, Martin-Trujillo A, Sharp AJ. 2022. A phenome-wide association study identifies effects of copy-number variation of VNTRs and multicopy genes on multiple human traits. *Am J Hum Genet* **109**: 1065–1076.

Gershon ES, Grennan K, Busnello J, Badner JA, Ovsiew F, Memon S, Alliey-Rodriguez N, Cooper J, Romanos B, Liu C. 2014. A rare mutation of CACNA1C in a patient with bipolar disorder, and decreased gene expression associated with a bipolar-associated common SNP of CACNA1C in brain. *Mol Psychiatry* **19**: 890–894.

GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**: 1318–1330.

Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N, Penn O, et al. 2019. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**: 61–68.

Impey S, Mark M, Villacres EC, Poser S, Chavkin C, Storm DR. 1996. Induction of CRE-mediated gene expression by stimuli that generate long-lasting LTP in area CA1 of the hippocampus. *Neuron* **16**: 973–982.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

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.

Li B, Tadross MR, Tsien RW. 2016. Sequential ionic and conformational signaling by calcium channels drives neuronal gene expression. *Science* **351**: 863–867.

Li YE, Preissl S, Miller M, Johnson ND, Wang Z, Jiao H, Zhu C, Wang Z, Xie Y, Poirion O, et al. 2023. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382**: eadf7044.

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324.

Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, Hultman CM. 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**: 234–239.

Liu Y, Blackwood DH, Caesar S, de Geus EJC, Farmer A, Ferreira M a. R, Ferrier IN, Fraser C, Gordon-Smith K, Green EK, et al. 2011. Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Mol Psychiatry* **16**: 2–4.

Lohi H, Young EJ, Fitzmaurice SN, Rusbridge C, Chan EM, Vervoort M, Turnbull J, Zhao X-C, Ianzano L, Paterson AD, et al. 2005. Expanded repeat in canine epilepsy. *Science* **307**: 81.

Lu T-Y, Human Genome Structural Variation Consortium, Chaisson MJP. 2021. Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nat Commun* **12**: 4250.

Lu T-Y, Smaruj PN, Fudenberg G, Mancuso N, Chaisson MJP. 2023. The motif composition of variable number tandem repeats impacts gene expression. *Genome Res* **33**: 511–524.

Ma H, Groth RD, Cohen SM, Emery JF, Li B, Hoedt E, Zhang G, Neubert TA, Tsien RW. 2014. γCaMKII Shuttles Ca2+/CaM to the Nucleus to Trigger CREB Phosphorylation and Gene Expression. *Cell* **159**: 281–294.

Mafessoni F, Grote S, de Filippo C, Slon V, Kolobova KA, Viola B, Markin SV, Chintalapati M, Peyrégne S, Skov L, et al. 2020. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proceedings of the National Academy of Sciences* **117**: 15132–15136.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–1195.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**: 222–226.

Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, et al. 2014. Transcriptional landscape of the prenatal human brain. *Nature* **508**: 199–206.

Mohammadi P, Castel SE, Brown AA, Lappalainen T. 2017. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res* **27**: 1872–1884.

Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Hujoel MLA, McCarroll SA, Loh P-R. 2023. Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer. *Cell* **186**: 3659-3673.e23.

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. *Science* **373**: 1499–1505.

Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, Als TD, Bigdeli TB, Børte S, Bryois J, et al. 2021. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Genet* **53**: 817–829.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**: 44–53.

Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, Legge SE, Bishop S, Cameron D, Hamshere ML, et al. 2018. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* **50**: 381–389.

Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. 2017. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**: 655–658.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**: 43–49.

Psychiatric GWAS Consortium Bipolar Disorder Working Group. 2011. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* **43**: 977–983.

Ren J, Gu B, Chaisson MJP. 2023. vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol* **24**: 175.

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin D-Y, Duan J, Ophoff RA, Andreassen OA, et al. 2011. Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**: 969–976.

Ruderfer DM, Fanous AH, Ripke S, McQuillin A, Amdur RL, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Bipolar Disorder Working Group of the Psychiatric Genomics Consortium, Cross-Disorder Working Group of the Psychiatric Genomics Consortium, Gejman PV, O'Donovan MC, et al. 2014. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol Psychiatry* **19**: 1017–1024.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**: 421–427.

Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* **51**: 30–35.

Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. *Science* **191**: 528–535.

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.

Splawski I, Timothy KW, Decher N, Kumar P, Sachse FB, Beggs AH, Sanguinetti MC, Keating MT. 2005. Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. *Proc Natl Acad Sci U S A* **102**: 8089–8096; discussion 8086-8088.

Splawski I, Timothy KW, Sharpe LM, Decher N, Kumar P, Bloise R, Napolitano C, Schwartz PJ, Joseph RM, Condouris K, et al. 2004. Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**: 19–31.

Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, Mattheisen M, Wang Y, Coleman JRI, Gaspar HA, et al. 2019. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat Genet* **51**: 793–803.

Sullivan PF, Kendler KS, Neale MC. 2003. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* **60**: 1187–1192.

Sulovari A, Li R, Audano PA, Porubsky D, Vollger MR, Logsdon GA, Warren WC, Pollen AA, Chaisson MJP, Eichler EE. 2019. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc Natl Acad Sci U S A* **116**: 23243–23253.

Trevino AE, Müller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, Farh K, Chang HY, Pașca AM, Kundaje A, et al. 2021. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**: 5053-5069.e23.

Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, Bryois J, Chen C-Y, Dennison CA, Hall LS, et al. 2022. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**: 502–508.

van den Heuvel MP, Fornito A. 2014. Brain networks in schizophrenia. *Neuropsychol Rev* **24**: 32–48.

Vecsey CG, Hawk JD, Lattal KM, Stein JM, Fabian SA, Attner MA, Cabrera SM, McDonough CB, Brindle PK, Abel T, et al. 2007. Histone deacetylase inhibitors enhance memory and synaptic plasticity via CREB:CBP-dependent transcriptional activation. *J Neurosci* **27**: 6128–6140.

Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. *Nat Genet* **37**: 986–990.

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. UnsTable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**: 1213–1216.

Wheeler DG, Groth RD, Ma H, Barrett CF, Owen SF, Safa P, Tsien RW. 2012. Ca(V)1 and Ca(V)2 channels engage distinct modes of Ca(2+) signaling to control CREB-dependent gene expression. *Cell* **149**: 1112–1124.

Wolff R, Nakamura Y, Odelberg S, Shiang R, White R. 1991. Generation of variability at VNTR loci in human DNA. *EXS* **58**: 20–38.

Yin JC, Del Vecchio M, Zhou H, Tully T. 1995. CREB as a memory modulator: induced expression of a dCREB2 activator isoform enhances long-term memory in Drosophila. *Cell* **81**: 107–115.

Yoshimizu T, Pan JQ, Mungenast AE, Madison JM, Su S, Ketterman J, Ongur D, McPhie D, Cohen B, Perlis R, et al. 2015. Functional implications of a psychiatric risk variant within CACNA1C in induced human neurons. *Mol Psychiatry* **20**: 162–169.

FIGURE LEGENDS

Fig. 1. Structure of the *CACNA1C* VNTR.

(*A*) *CACNA1C* showing VNTR in intron 3, GWAS SNPs associated with schizophrenia (Pardiñas et al. 2018), bipolar disorder (Mullins et al. 2021), and *CACNA1C* expression in brain tissue (GTEx Consortium 2020). **(*B*)** Schematic of *CACNA1C* VNTR analysis strategy. *CACNA1C* VNTR sequences are identified in phased, long-read haplotype assemblies (n = 155 assemblies from 78 individuals) from the Human Genome Structural Variation Consortium (HGSVC2) (Ebert et al. 2021), the Human Pangenome Reference Consortium (HPRC) (Liao et al. 2023), and the Telomere-to-Telomere project (Nurk et al. 2022). **(*C*)** Sequence logo of the *CACNA1C* repeat unit. Letter height is proportional to base abundance across all sequences. **(*D*)** Repeat unit key for **(*D-E*)**. Each common repeat unit (frequency > 0.0005) is represented as a different color, collectively representing > 98% of analyzed *CACNA1C* VNTR sequences. Nucleotide differences from the consensus repeat unit are colored and underlined. Remaining low-frequency repeat units are represented by one of two colors depending on their length (either 30 bp or not). The single 30-bp repeat unit found in chimpanzees is listed below the repeat units found in humans. **(*E*)** Seven alleles (called Types) of the *CACNA1C* VNTR. Repeat unit frequencies (y-axis) per position (x-axis) are shown above a consensus sequence. Numbers in parentheses indicate VNTR Type frequency among the 155 sequences. Type 6 and Type 7, representing single *CACNA1C* VNTR sequences dissimilar to any other Type, are included separately.

Fig. 2. *CACNA1C* VNTR structural diversity introduced by repeat unit variation.

(*A*) Relative frequency of each repeat unit in the repeat unit key of **Fig. 1D** by Type. Stacked bars are ordered from bottom to top by decreasing repeat unit frequency in Type 1. **(*B-C*)** Cumulative proportions of distinct **(*B*)** and all **(*C*)** *CACNA1C* VNTR repeat units by the Types they are found in. “All Types” indicates a repeat unit is found in Types 1-7. There are a small number of repeat units that are shared between Types 1, 2, and 3, but not found in Types 4, 5, 6, nor 7. Types along the x-axes are ordered by decreasing proportion. “Rest” includes the remaining repeat units that are found in a mix of Types. **(*D-E*)** Count of identical repeat units across the first **(*D*)** and last **(*E*)** 18 repeat units of each Type’s consensus sequence. Colors correspond to the repeat unit key (**Fig. 1D**).

Fig. 3. Duplications within *CACNA1C* VNTR sequences.

Kilobase-scale duplications in Type 2, Type 3 (***A***,***B***), and Type 5 (***C***,***D***) *CACNA1C* VNTR sequences. (***A***,***C***) Dosage of a sliding window (y-axis) by its position (x-axis) along an exemplar sequence. The exemplar sequence for Type 1, 2, and 3 is the Type 1 consensus sequence, and for Type 5 is a duplication-free Type 5 sequence with the highest amount of shared sequence across all four Type 5 sequences. Dosage of each window is quantified as the number of exact matches in the sequences scanned. Width of the sliding window for Type 1, 2, and 3 is 6 repeat units and for Type 5 is 8 repeat units. A quantity above 1 indicates a duplication of the sliding window somewhere in the sequence. Quantities directly indicate the degree of duplication of VNTR segments (i.e., a value of 4 indicates a quadruplication). Vertical dashed lines (paired by color) indicate regions where the dosage of the sliding window identifies large tandem duplications of VNTR segments. Sequence shared with the exemplar sequence is aligned. Large duplications are extracted from their positions and aligned below their paralogous sequence using a yellow callout. Mismatched repeat units to the exemplar sequence are shown with red ticks, although mismatches involving a gap in either sequence are not indicated with a red tick. Colors correspond to the repeat unit key in **Fig. 1D**. (***C***,***D***) Dosage in HG00735_paternal is not shown due to its identity with the exemplar sequence. Arrows indicate a tiling path through a series of duplications. Deletions are indicated by an alignment gap (black).

Fig. 4. Characterization of variable region alleles in Type 1 sequences.

(*A*) Repeat unit variability is shown as normalized Shannon's uncertainty H(x). Variable regions are defined at H(x) > 0.25 and length > 7 aligned repeat units. **(*B*)** Multiple sequence alignment of Type 1 sequences. Variable region boundaries are labeled with vertical gray boxes. **(*C-F*)** Variable Region 1 and 2 (VR1 and VR2) alleles, their lengths, and their frequencies in 134 Type 1 sequences. "Unclassified" includes alleles that do not cluster near the consensus (**Fig. S4A** and **Fig. S5A**) and whose full VNTR sequences have a large aligned edit distance to exemplar *CACNA1C* VNTR sequences (**Fig. S4B** and **Fig. S5B**). **(*B,C,E*)** Colors correspond to the repeat unit key in **Fig. 1D**.

Fig. 5. Schizophrenia association and eQTLs are in tight linkage disequilibrium with Variable Region 2.

(*A*) Schizophrenia-associated SNPs surrounding the *CACNA1C* intron 3 VNTR (Song et al. 2018) colocalize with eQTLs in brain tissue. Shown are schizophrenia-associated SNPs and the fine-mapped causal set (PP>95%) (Pardiñas et al. 2018). eQTLs in brain (cerebellum n = 125, cerebellar hemisphere n = 125 with 117 shared between these two replicate tissues, putamen n = 1) from the GTEx project (GTEx Consortium 2020) are depicted as arcs. Also shown are scATAC-seq in glutamatergic neurons from fetal brain (Trevino et al. 2021) and DNase-seq in fetal brain from the Roadmap Epigenomics Mapping Consortium (Maurano et al. 2012). **(*B*,*D*)** Count of each VR allele considered for the LD calculation in 70 haplotypes from 35 individuals in HGSVC2. VR1 **(*B*)** and VR2 **(*D*)** allele frequencies are comparable to frequencies observed in all 155 *CACNA1C* VNTR sequences. "Excluded" indicates a haplotype with a Type 2-7 sequence or an unclassified VR allele. **(*C,E*)** Linkage disequilibrium (LD, y-axis) between variable regions and surrounding SNPs (x-axis). SNPs in the fine-mapped schizophrenia set are highlighted in green. LD is shown with Variable Region 1 (VR1, ***C***) for alleles VR1A and VR1B and with Variable Region 2 (VR2, ***E***) for alleles VR2A and VR2B. LD with VR2C and VR2D is shown in **Fig. S8**. Three top schizophrenia GWAS SNPs are in perfect LD with VR2. **(*F*)** Analysis of relationship between VR2 and eQTL signal. Shown are candidate eQTL SNPs in brain. Statistical significance (y-axis) is strongly correlated with LD with VR2 (x-axis). A second distinct eQTL signal is not in LD with VR2 and is not associated with schizophrenia. All significant eQTL SNPs that were associated with schizophrenia decreased expression of *CACNA1C* (blue).

Fig. 6. Evolutionary history of *CACNA1C* VNTR diversity.

(*A-C*) Ancestry frequency in 155 haplotypes **(*A*)** by *CACNA1C* VNTR Type **(*B*)** and VR2 allele **(*C*)**. Ancestry of each individual is defined by its 1000 Genomes super-population. **(*D*)** Pileup of reads over the *CACNA1C* VNTR region (GRCh37/hg19) after whole-genome sequencing of four ancient hominin genomes. **(*E*)** Estimated average *CACNA1C* VNTR length in four ancient hominin genomes. Values are listed from earliest to latest estimated date that the hominin individual lived.

SUPPLEMENTARY FIGURE LEGENDS

Fig. S1. Details of *CACNA1C* VNTR structure and sequencing.

(*A*) All 155 *CACNA1C* VNTR sequences by Type, aligned. An asterisk at the end of Type 5 indicates HG00735_paternal. **(*B*)** Counts of haplotype assemblies for each Type by long-read technology. **(*C*)** Cumulative distribution of *CACNA1C* VNTR lengths in bp (n = 155 sequences).

Fig. S2. Duplication scans for Types 4, 6, and 7.

Each consensus sequence of Type 4 **(*A*)**, 6 **(*B*)**, and 7 **(*C*)** was scanned for duplications using a sliding window of width 6 repeat units. Five regions across Type 4 and four regions across Type 7 registered a dosage of 2, indicating some smaller copied segments within these sequences. These regions were on the order of hundreds of nucleotides and were not contiguous; thus, they did not suggest the same pattern of kilobase-scale duplications like in Types 1, 2, 3, and 5. Type 6 had a dosage of 1 across its whole sequence.

Fig. S3. Expanded view of variable regions.

Enlargement of VR1 and VR2 in the alignment of Type 1 sequences (**Fig. 4B**). Constant regions of 5 repeat units flank each VR. Sequences are ordered by VR2 allele classification (right), revealing near-complete correspondence with VR1 allele classification (left); black asterisks (left column) denote those infrequent VR1 alleles that deviate from this pattern.

Fig. S4. VR1 allele definition.

(*A*) Scatterplot of edit distances (in repeat units) to the VR1B consensus vs. the VR1A consensus. Each point represents one of the 18 distinct VR1 sequences. Overlapping points are plotted with a small amount of jitter. Shaded areas define thresholds used to partition sequences into alleles. **(*B*)** Scatterplot of edit distances (at the nucleotide level) to two exemplar full *CACNA1C* VNTR sequences. Each point represents a full, aligned Type 1 sequence (n = 134) and its coordinates are edit distances to a VNTR sequence with the consensus VR1B allele (HG00171_h2, y-axis) and a VNTR sequence with the consensus VR1A allele (HG00096_h1, x-axis). VR1 alone does not explain the clustering of these sequences. However, an exceptional VR1A in two *CACNA1C* VNTR sequences with VR2B (HG01361_paternal, NA20509_h2) clustered according to their VR1 allele. **(*C*)** Consensus VR1 sequences and their forms (see **Table S4** for details) identified in Type 1 sequences. Consensus alleles are in bold. Non-consensus VR1 sequences classified to each allele are numbered and listed below the consensus. Two CACNA1C VNTR sequences (HG00732_h1, HG00733_h2, and NA19650_h1) contained large deletions that included VR1, resulting in the VR1 sequence VR1B-2. The unclassified VR1 sequence is named by the sample haplotype IDs it is identified in. **(*D*)** VR1A sequences identified in Type 2 and Type 3 sequences. The consensus VR1A and VR1B alleles were the only known VR1 sequences from Type 1; otherwise the VR1A sequences in Type 2 and Type 3 sequences were unique (asterisks).

Fig. S5. VR2 allele definition.

(*A*) Scatterplot of edit distances (in repeat units) to the VR2B consensus vs. the VR2A consensus. Each point represents one of the 71 distinct VR2 sequences. Overlapping points are plotted with a small amount of jitter. Shaded areas define thresholds used to partition sequences into two common alleles. The two rarer alleles (VR2C and VR2D) were defined manually. **(*B*)** Scatterplot of edit distances (at the nucleotide level) to two exemplar full *CACNA1C* VNTR sequences. Each point represents a full, aligned Type 1 sequence (n = 134) and its coordinates are edit distances to a VNTR sequence with the consensus VR1B allele (HG00171_h2, y-axis) and a VNTR sequence with the consensus VR1A allele (HG00096_h1, x-axis). Sequences cluster according to VR2 allele. However, two *CACNA1C* VNTR sequences (HG01361_paternal, NA20509_h2) with a VR2B allele (VR2B-6) inappropriately clustered near sequences with VR2D alleles because they had VR1A (VR1A-6) where most sequences with VR2B had VR1B. **(*C*)** Consensus VR2 sequences and their forms (see **Table S5** for details) identified in Type 1 sequences. Consensus alleles are in bold. Non-consensus VR2 sequences classified to each allele are numbered and listed below the consensus. One CACNA1C VNTR sequence (NA19650_h1) had a large deletion that included VR1 and part of VR2, resulting in the VR2 sequence VR2A-32. The unclassified VR2 sequence is named by the sample haplotype IDs it is identified in. **(*D*)** VR2 sequences identified in Type 2 and Type 3 sequences by scanning for matches to each consensus VR2 sequence (**Materials and Methods**). The VR2B-2 sequence was the only known VR2 sequence from Type 1; otherwise the VR2 sequences in Type 2 and Type 3 sequences were unique (asterisks). No Type 2 nor Type 3 sequences contained a consensus VR2 sequence. The unclassified VR2 sequence is named by the sample haplotype ID it is identified in.

Fig. S6. Previously-identified VR boundaries.

(*A*) Repeat unit variability, shown as normalized Shannon's uncertainty H(x), above its multiple sequence alignment of 27 *CACNA1C* VNTR sequences that were long-read sequenced following PCR amplification and size-selection (Song et al. 2018). Variable regions are defined at H(x) > 0.25 and length > 7 aligned repeat units. **(*B*)** Expanded view of variable regions comparing their boundaries defined previously to the boundaries described in this paper.

Fig. S7. Variable region alleles in Type 2 and Type 3 *CACNA1C* VNTR sequences.

(*A*,*C*) VR sequences in Type 2 sequences. VR1 **(*A*)** is duplicated in Type 2 and is found as VR1A and VR1B in equal proportion. VR2 **(*C*)** in Type 2 is found as VR2B, VR2C, and one unclassified VR2 sequence. **(*B*,*D*)** VR sequences in Type 3 sequences. VR1 in Type 3 **(*B*)** exists as VR1A only, which is either duplicated or triplicated. VR2 in Type 3 **(*D*)** exists as VR2C. Mismatched repeat units between each VR allele and the sequences identified in Type 2 and Type 3 are shown with red ticks, although mismatches involving a gap in either sequence are not indicated with a red tick. **(*E*)** Frequency of VR2 alleles by Type 1, 2, and 3 sequences. **(*F*,*G*)** Overall frequencies of VR alleles inclusive of Type 1, 2, and 3 *CACNA1C* VNTR sequences.

Fig. S8. Linkage disequilibrium between schizophrenia association and minor VR2 alleles. **(*A-B*)** Linkage disequilibrium (LD, y-axis) between VR2 minor alleles and surrounding SNPs (x-axis). SNPs in the fine-mapped schizophrenia set are highlighted in green. LD is shown with VR2C **(*A*)** and VR2D **(*B*)**. VR2C (frequency = 0.08) shows a similar LD pattern as VR2B with schizophrenia GWAS SNPs. Though observed infrequently, VR2D (frequency = 0.05) displays a different LD structure with nearby SNPs.

Fig. S9. VR2 linkage disequilibrium with GWAS and eQTL SNPs per brain tissue.
Analysis of the relationship between VR2, *CACNA1C* brain eQTLs, and schizophrenia GWAS results in 11 brain tissues. Each point represents a SNP, colored by eQTL slope **(*A***, blue**)** and GWAS **(*B***, yellow**)** significance if the SNP is identified by both analyses. Significant brain eQTLs overlapping GWAS SNPs are in strong LD with VR2. Cerebellum and cerebellar hemisphere, and cortex and frontal cortex represent the same anatomical regions in the same tissues but from different sampling sites.

Fig. S10. Inferring history of *CACNA1C* VNTR via ancestry and length estimates from short-read WGS data.
(*A*) Schematic for estimation of *CACNA1C* VNTR length (in repeat units). Estimate corresponds to the average of both alleles within a diploid genome. Average sequencing coverage depth is computed across three regions: the *CACNA1C* VNTR region (V) and its two flanking 10-kb segments (F_1 and F_2). Average VNTR length (in repeat units) is computed as the average of V/F_1 and V/F_2 scaled by a conversion factor. **(*B*)** Average VNTR length was estimated from sequencing coverage in short-read WGS data from 1000 Genomes Project 30X on GRCh38 (n = 70 individuals). Matching long-read assemblies were used to establish actual length; length was averaged across both alleles from the same individual. **(*C*)** Scatterplot of VNTR lengths estimated from WGS (averaged across each individual's two alleles, y-axis) vs. lengths measured directly from long-read haplotype assemblies (n = 140 VNTR sequences, n = 70 diploid individuals).

SUPPLEMENTARY TABLE LEGENDS

Table S1. Data sources for long-read haplotype assemblies and short-read WGS.

Sample, the seven-character ID assigned to each sample by 1000 Genomes with the exception of two Genome in a Bottle samples (NA24385 (HG002) and NA24631 (HG005)) and one HapMap sample (NA12878 (HG001)). Project, the creator of the haplotype assembly. Read type, PacBio long-read technology used to create initial phased contig assemblies. HiFi or CLR Coverage, reported as data yield of PacBio reads (Gbp) divided by estimated genome size (for HGSVC2 and HPRC samples). For T2T, HiFi coverage is reported as the mean coverage of all chromosomes. Values compiled from the supplementary information of the relevant publications: (Ebert et al. 2021) Table S6, (Liao et al. 2023) Table S1, (Nurk et al. 2022) main text). Haplotype 1 sample ID and Haplotype 2 sample ID, a unique name for each assembly shortened from their filename. Parentally-resolved haplotype assemblies from HPRC were assigned arbitrarily to haplotypes 1 and 2. Haplotype 1 download path and Haplotype 2 download path, link to download each assembly in fasta format. WGS download path, link to download each sample's high-coverage WGS data from 1000 Genomes, when available. 1000 Genomes Population Code, population membership of each sample when available, as defined by (1000 Genomes Project Consortium et al. 2015).

Table S2. Coordinates of VNTR in each assembly.

Haplotype sample ID, sample and haplotype. Filename, name of fasta file containing genomic assembly. Region (contig:start-end), coordinates of *CACNA1C* VNTR in each assembly.

Table S3. Repeat unit counts.

Repeat unit, the 158 unique repeat units identified from the *CACNA1C* VNTR sequences analyzed in this manuscript. Length, length in nucleotides of each repeat unit. Count, number of occurrences of each repeat unit in 155 *CACNA1C* VNTR sequences plus the VNTR region of GRCh38/hg38. Frequency, global frequency of each repeat unit. Samples, haplotype sample IDs of *CACNA1C* VNTR sequences containing at least 1 copy of a repeat unit. Num. samples, number of Samples. Maximum value is 156, which includes 155 *CACNA1C* VNTR sequences plus the VNTR region of GRCh38/hg38. Frac. samples, fraction of 156 samples containing at least 1 copy of a repeat unit.

Table S4. VR1 alleles.

Detail of VR1 alleles shown in **Fig. S4**. Allele, a unique identifier for each distinct VR1 sequence. Consensus alleles do not have a number appended to their identifier. Count, number of *CACNA1C* VNTR sequences with each VR1 sequence. Samples, haplotype sample IDs of *CACNA1C* VNTR sequences with each VR1 sequence. Num. mismatches, number of repeat unit mismatches between each VR1 sequence and the consensus sequence of its VR1 allele. The unclassified VR1 sequence has 10 and 9 mismatches (in repeat units) to VR1A and VR1B, respectively. Sequence, VR1 nucleotide sequences aligned to each other. Type, the *CACNA1C* VNTR Types each VR1 sequence is found in.

Table S5. VR2 alleles.

Detail of VR2 alleles shown in **Fig. S5**. Allele, a unique identifier for each distinct VR2 sequence. Consensus alleles do not have a number appended to their identifier. Count, number of *CACNA1C* VNTR sequences with each VR2 sequence. Samples, haplotype sample IDs of *CACNA1C* VNTR sequences with each VR2 sequence. Num. mismatches, number of repeat unit mismatches between each VR2 sequence and the consensus sequence of its VR2 allele. The unclassified VR2 sequence found in Type 1 has 12, 20, 21, and 30 mismatches (in repeat units) to VR2A, VR2B, VR2C, and VR2D, respectively. Sequence, VR2 nucleotide sequences aligned to each other. Type, the *CACNA1C* VNTR Types each VR1 sequence is found in.

Table S6. Number of nucleotide positions different between VR2 alleles.

Number of mismatched nucleotide positions (including gaps) between each pair of aligned VR2 consensus sequences. NA is used to avoid redundant entries of the symmetric table.

Table S7. Average number of nucleotide positions different between sequences belonging to each VR allele.

Average number of mismatched nucleotide positions (including gaps) between consensus and non-consensus sequences of a VR allele. NA is used to indicate VR alleles that are found in the dataset.

SUPPLEMENTARY DATA

Data S1. Alignment of all *CACNA1C* VNTR sequences.

All *CACNA1C* VNTR sequences analyzed in this manuscript, aligned within each Type and provided in fasta format. Sequences in this file are ordered by Type then by haplotype sample ID (**Table S1**).

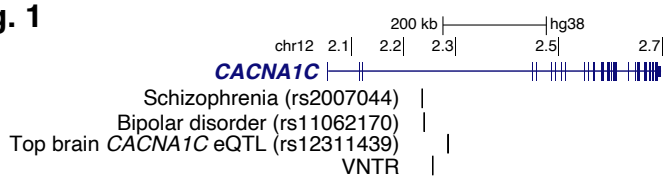
Data S2. *CACNA1C* VNTR Type consensus sequences.

Consensus sequences for *CACNA1C* VNTR Types described in this manuscript, provided in fasta format.

Consensus sequences for each Type were converted to nucleotide sequences by swapping each character for its corresponding repeat unit (**Data S2**). Types 2, 4, and 5 had characters representing grouped infrequent units in their consensus sequence at 1, 5, and 97 positions, respectively. These characters were converted to the most common infrequent unit at each position. Most of these positions (75%, 77/103) contained the same infrequent unit in all sequences without an alignment gap. One position in Type 5 had two infrequent units in equal proportion. In this case, the repeat unit with a higher global frequency was chosen for the consensus. The remaining positions (24%, 25/103) had one infrequent unit in a majority of sequences without an alignment gap. Only one repeat unit in any consensus sequence was a repeat unit of size other than 30 bp. It was shared by 3/4 *CACNA1C* VNTR sequences in Type 5.

Fig. 1

A



B Long-read haplotype assemblies (n = 155)

Project	Num. assemblies	Num. individuals
HGSVC2 (CLR)	42	21
HGSVC2 (HiFi)	26	13
HPRC (HiFi)	86	43
CHM13 (HiFi)	1	1

↓

Identify contig containing *CACNA1C* consensus repeat unit

↓

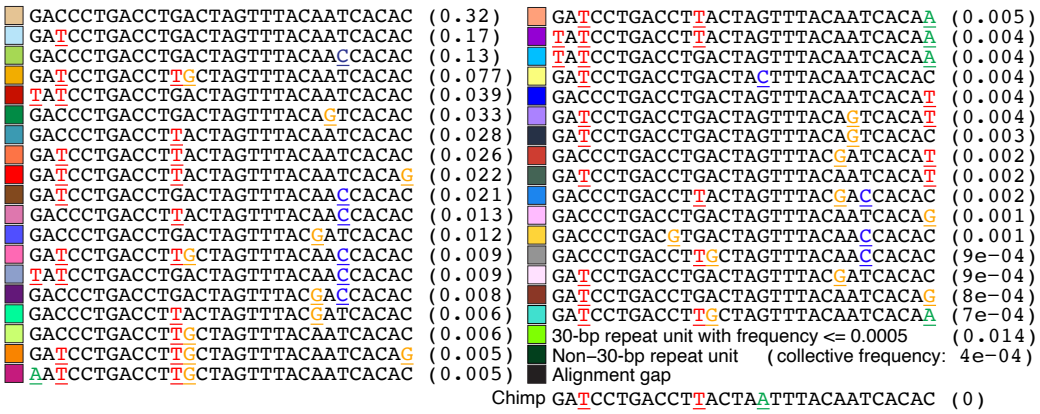
Parse VNTR into 30-bp repeat units (Tandem Repeats Finder)

↓

Multiple sequence alignment (MAFFT)

C

D Repeat units (frequency)



E

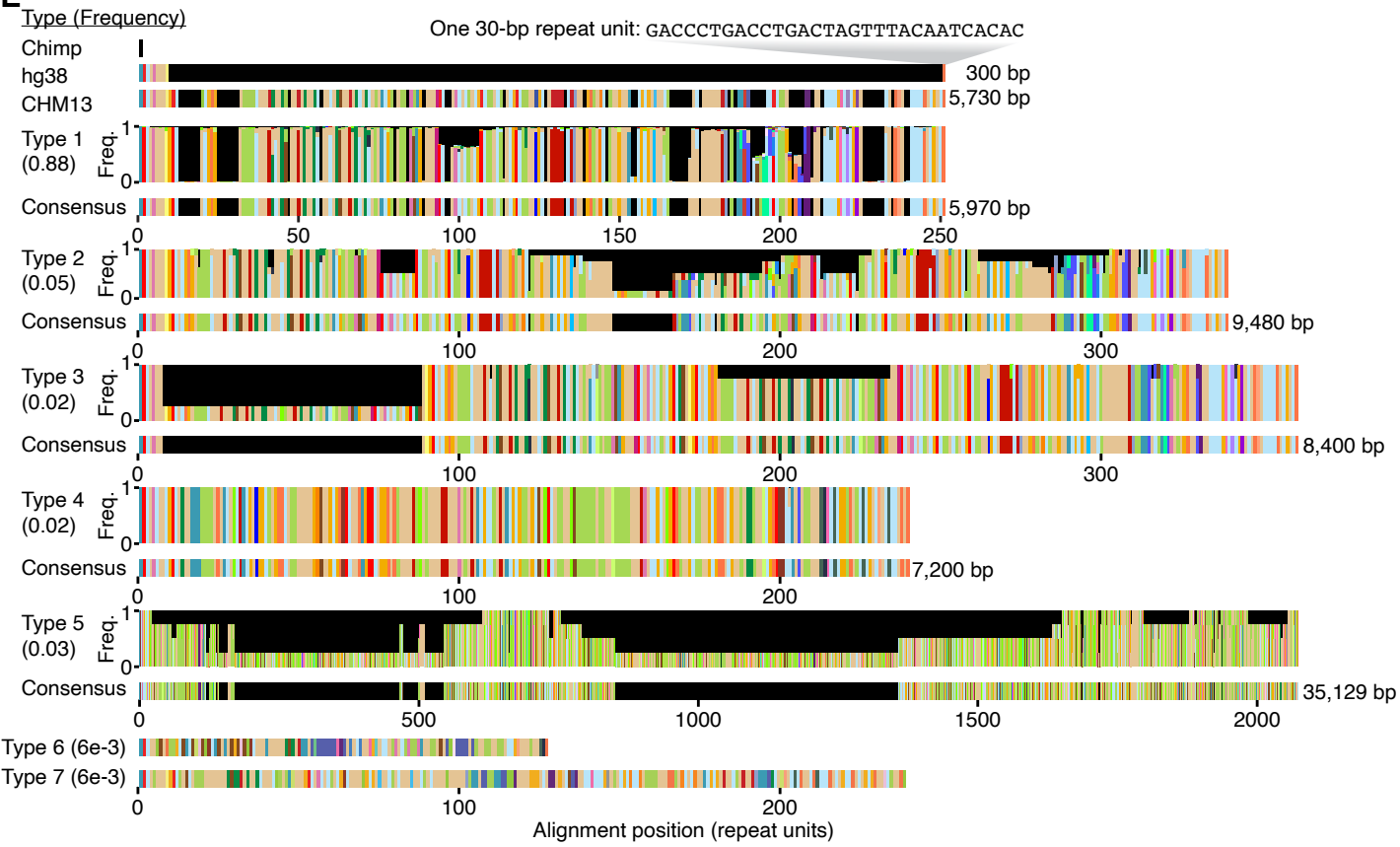
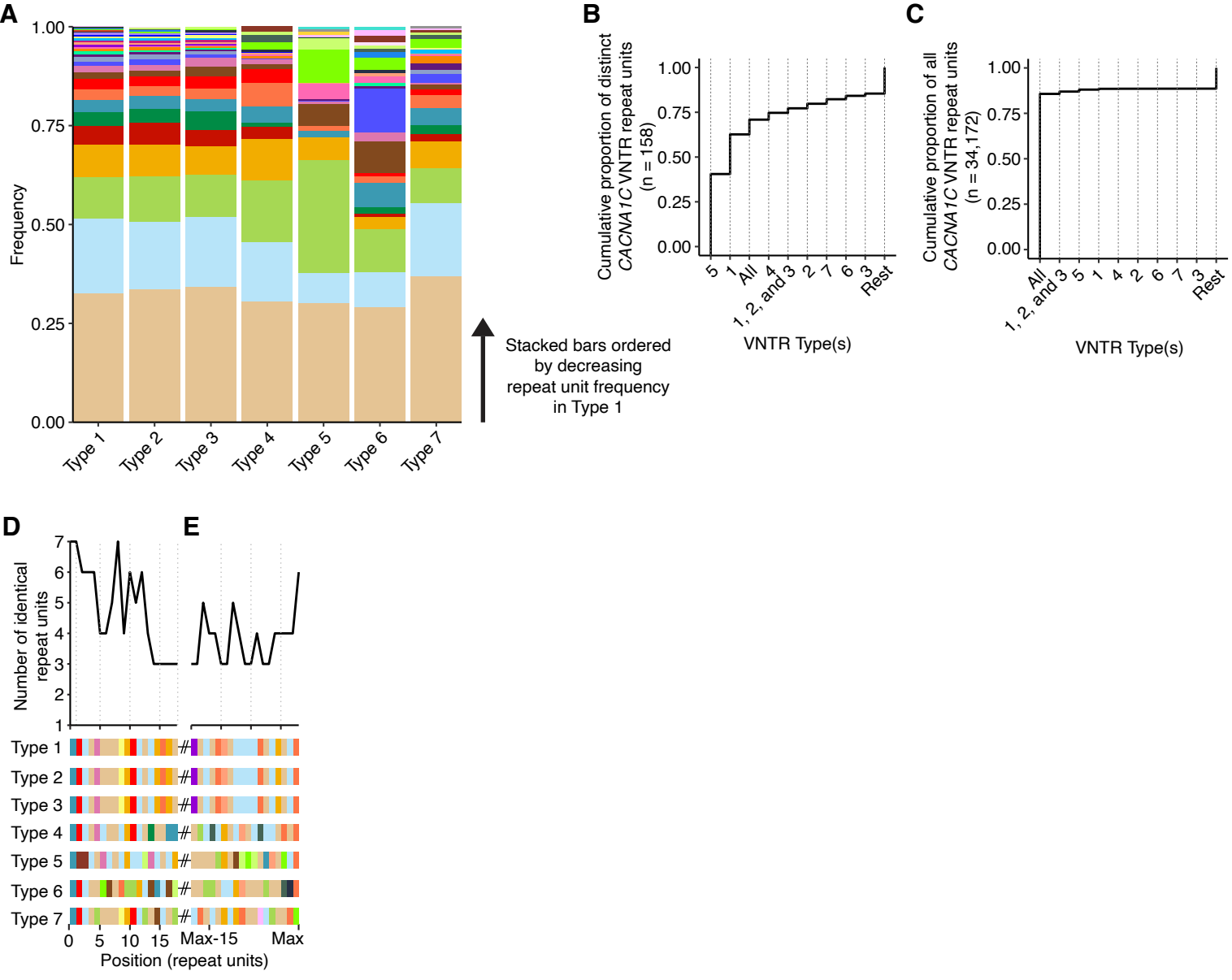


Fig. 2


A
Frequency
1.00
0.75
0.50
0.25
0.00
Type 1
Type 2
Type 3
Type 4
Type 5
Type 6
Type 7
Stacked bars ordered by decreasing repeat unit frequency in Type 1
B
Cumulative proportion of distinct CACNA1C VNTR repeat units (n = 158)
5
1
All
4
1, 2, and 3
2
7
6
3
Rest
VNTR Type(s)
C
Cumulative proportion of all CACNA1C VNTR repeat units (n = 34,172)
All
1, 2, and 3
5
1
4
2
6
7
3
Rest
VNTR Type(s)
D
E
Number of identical repeat units
Type 1
Type 2
Type 3
Type 4
Type 5
Type 6
Type 7
0
5
10
15
Max-15
Max
Position (repeat units)

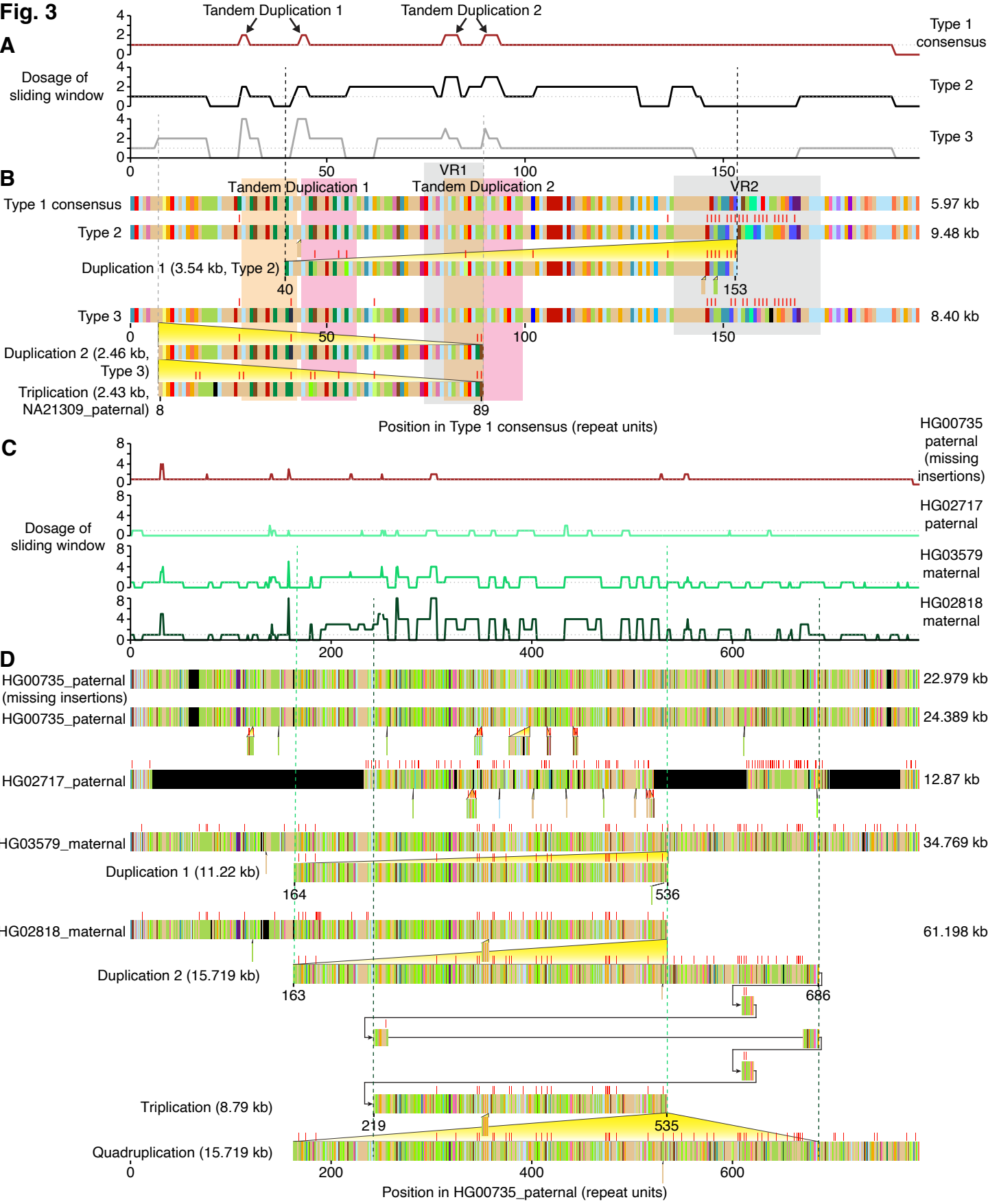
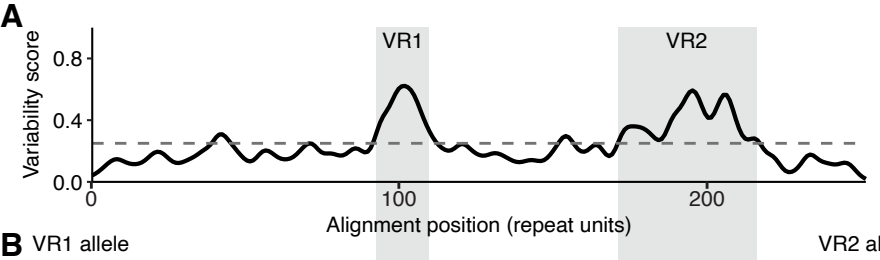


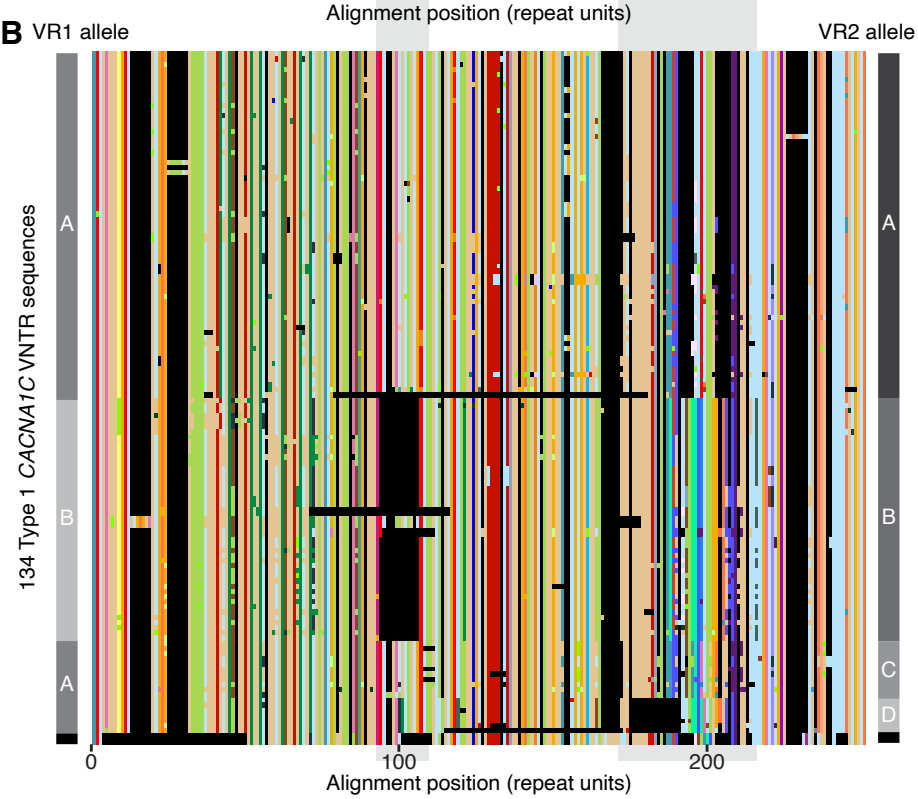
Fig. 3
A
Tandem Duplication 1
Tandem Duplication 2
Type 1 consensus
Dosage of sliding window
Type 2
Type 3
B
VR1
VR2
Tandem Duplication 1
Tandem Duplication 2
Type 1 consensus
5.97 kb
Type 2
9.48 kb
Duplication 1 (3.54 kb, Type 2)
40
153
Type 3
8.40 kb
Duplication 2 (2.46 kb, Type 3)
Triplication (2.43 kb, NA21309_paternal)
8
89
Position in Type 1 consensus (repeat units)
C
HG00735 paternal (missing insertions)
HG02717 paternal
HG03579 maternal
HG02818 maternal
Dosage of sliding window
D
HG00735_paternal (missing insertions)
22.979 kb
HG00735_paternal
24.389 kb
HG02717_paternal
12.87 kb
HG03579_maternal
34.769 kb
Duplication 1 (11.22 kb)
164
536
HG02818_maternal
61.198 kb
Duplication 2 (15.719 kb)
163
686
Triplication (8.79 kb)
219
535
Quadruplication (15.719 kb)
Position in HG00735_paternal (repeat units)

Fig. 4

A



B

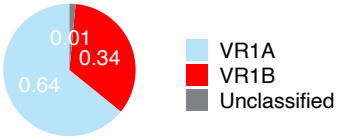


C Variable region (VR) 1

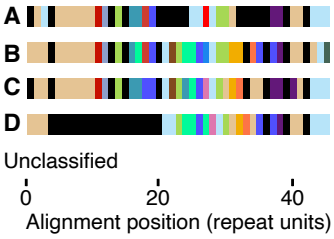
	Count	Sans-gap length
A	86	15
B	46	4
Unclassified	2	5

repeat units

D VR1 allele frequencies in Type 1 sequences



E Variable region (VR) 2



	Count	Sans-gap length
A	67	29
B	47	37
C	11	36
D	7	25
Unclassified	2	29

repeat units

F VR2 allele frequencies in Type 1 sequences

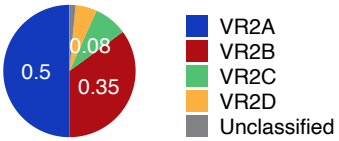


Fig. 5

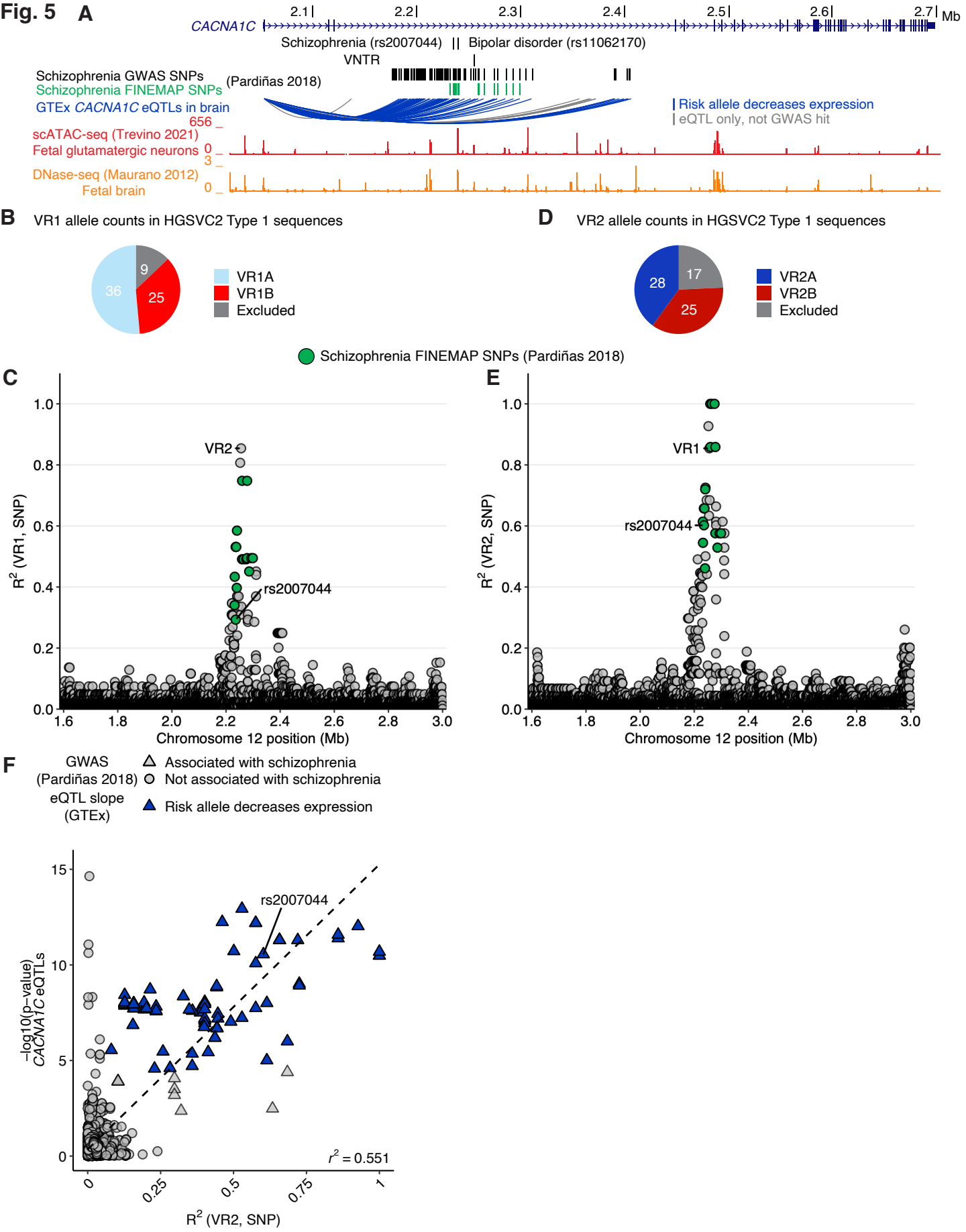
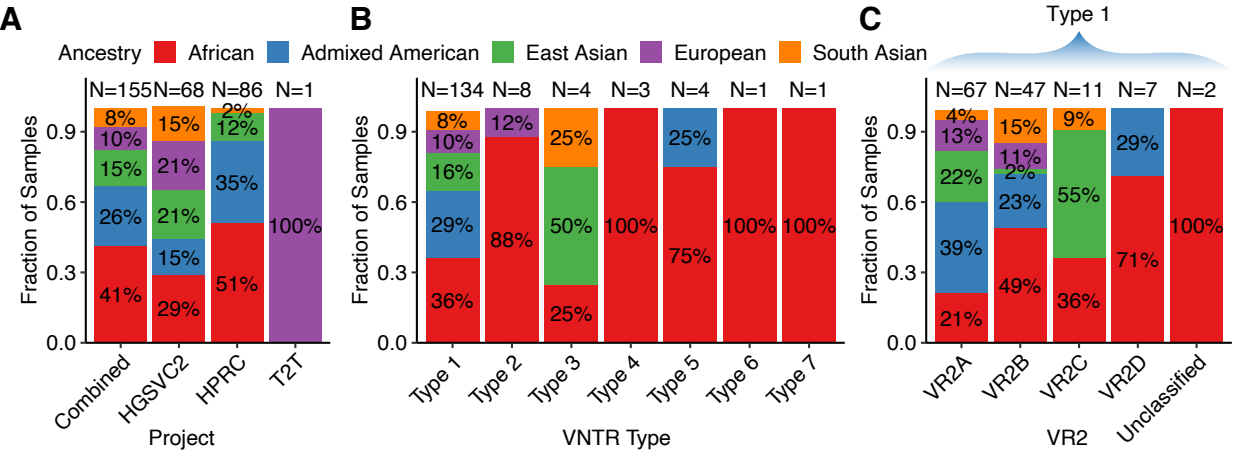
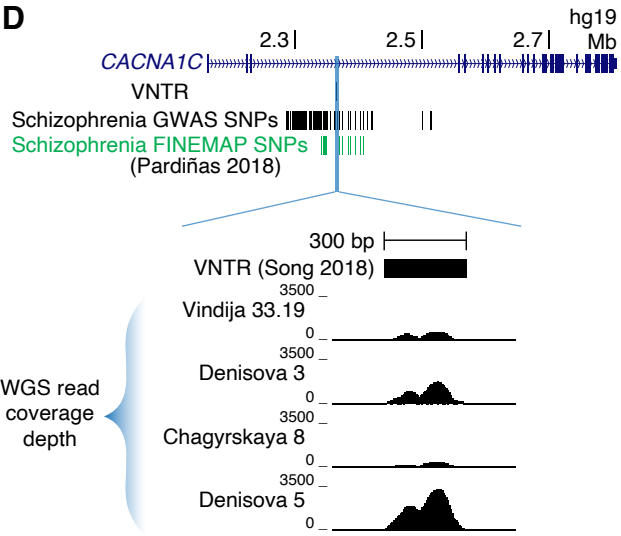

A
CACNA1C
2.1
2.2
2.3
2.4
2.5
2.6
2.7
Mb
Schizophrenia (rs2007044)
Bipolar disorder (rs11062170)
VNTR
Schizophrenia GWAS SNPs
Schizophrenia FINEMAP SNPs
(Pardiñas 2018)
GTEx CACNA1C eQTLs in brain
Risk allele decreases expression
eQTL only, not GWAS hit
scATAC-seq (Trevino 2021)
Fetal glutamatergic neurons
656
0
DNase-seq (Maurano 2012)
Fetal brain
3
0
B
VR1 allele counts in HGSVC2 Type 1 sequences
36
25
9
VR1A
VR1B
Excluded
D
VR2 allele counts in HGSVC2 Type 1 sequences
28
25
17
VR2A
VR2B
Excluded
Schizophrenia FINEMAP SNPs (Pardiñas 2018)
C
VR2
rs2007044
R2 (VR1, SNP)
Chromosome 12 position (Mb)
E
VR1
rs2007044
R2 (VR2, SNP)
Chromosome 12 position (Mb)
F
GWAS (Pardiñas 2018)
Associated with schizophrenia
Not associated with schizophrenia
eQTL slope (GTEx)
Risk allele decreases expression
rs2007044
−log10(p−value) CACNA1C eQTLs
r2 = 0.551
R2 (VR2, SNP)

Fig. 6





E

Sample	Avg. VNTR length (repeat units)	Avg. VNTR length (bp)	Date (kya)	Reference
Vindija 33.19	112	3,360	50	Prüfer 2017
Denisova 3	278.3	8,349	63-55	Meyer 2012
Chagyrskaya 8	71.8	2,154	80-60	Mafessoni 2020
Denisova 5	352	10,560	120	Prüfer 2014

Fig. S1

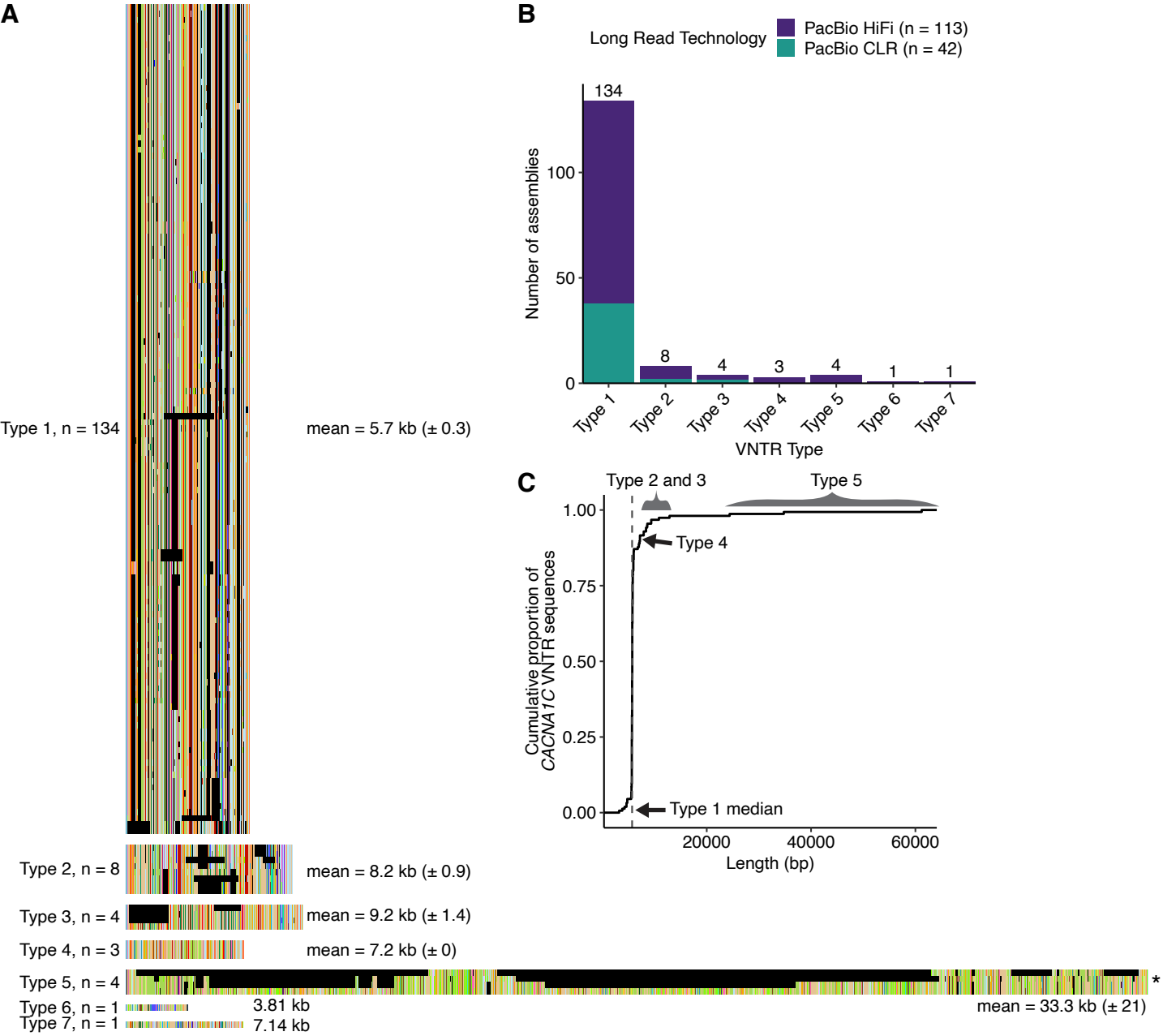

A
Type 1, n = 134
mean = 5.7 kb (± 0.3)
Type 2, n = 8
mean = 8.2 kb (± 0.9)
Type 3, n = 4
mean = 9.2 kb (± 1.4)
Type 4, n = 3
mean = 7.2 kb (± 0)
Type 5, n = 4
*
Type 6, n = 1
3.81 kb
Type 7, n = 1
7.14 kb
mean = 33.3 kb (± 21)
B
Long Read Technology
PacBio HiFi (n = 113)
PacBio CLR (n = 42)
Number of assemblies
134
8
4
3
4
1
1
Type 1
Type 2
Type 3
Type 4
Type 5
Type 6
Type 7
VNTR Type
C
Type 2 and 3
Type 5
Type 4
Type 1 median
Cumulative proportion of CACNA1C VNTR sequences
Length (bp)

Fig. S2

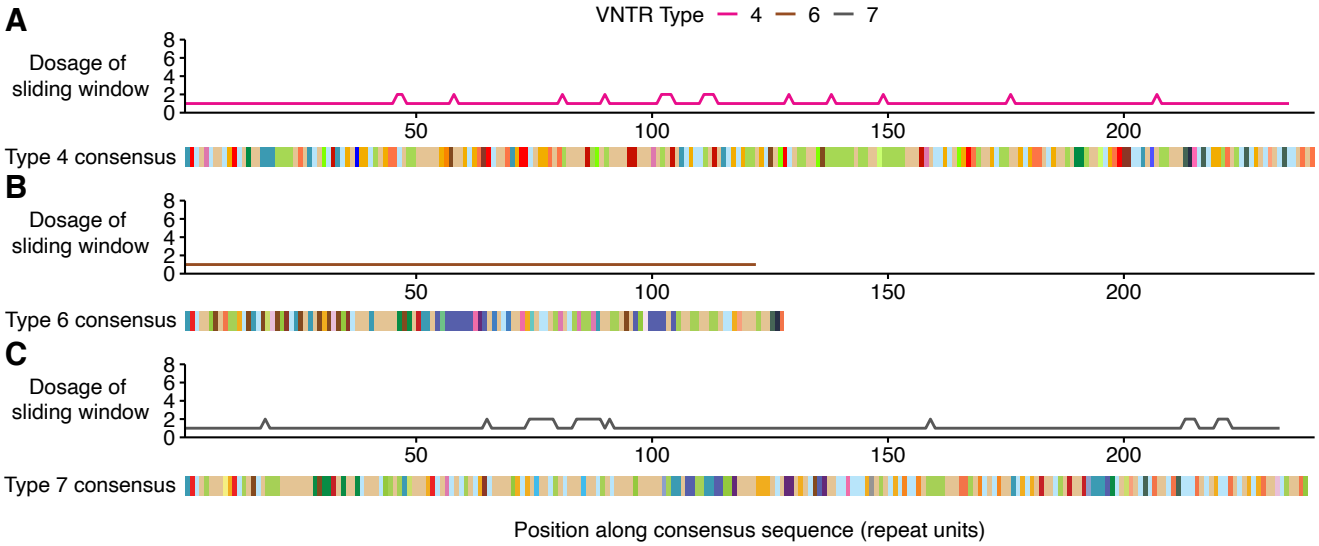

A
VNTR Type 4 6 7
Dosage of sliding window
8
6
4
2
0
50
100
150
200
Type 4 consensus
B
Dosage of sliding window
Type 6 consensus
C
Dosage of sliding window
Type 7 consensus
Position along consensus sequence (repeat units)

Fig. S3

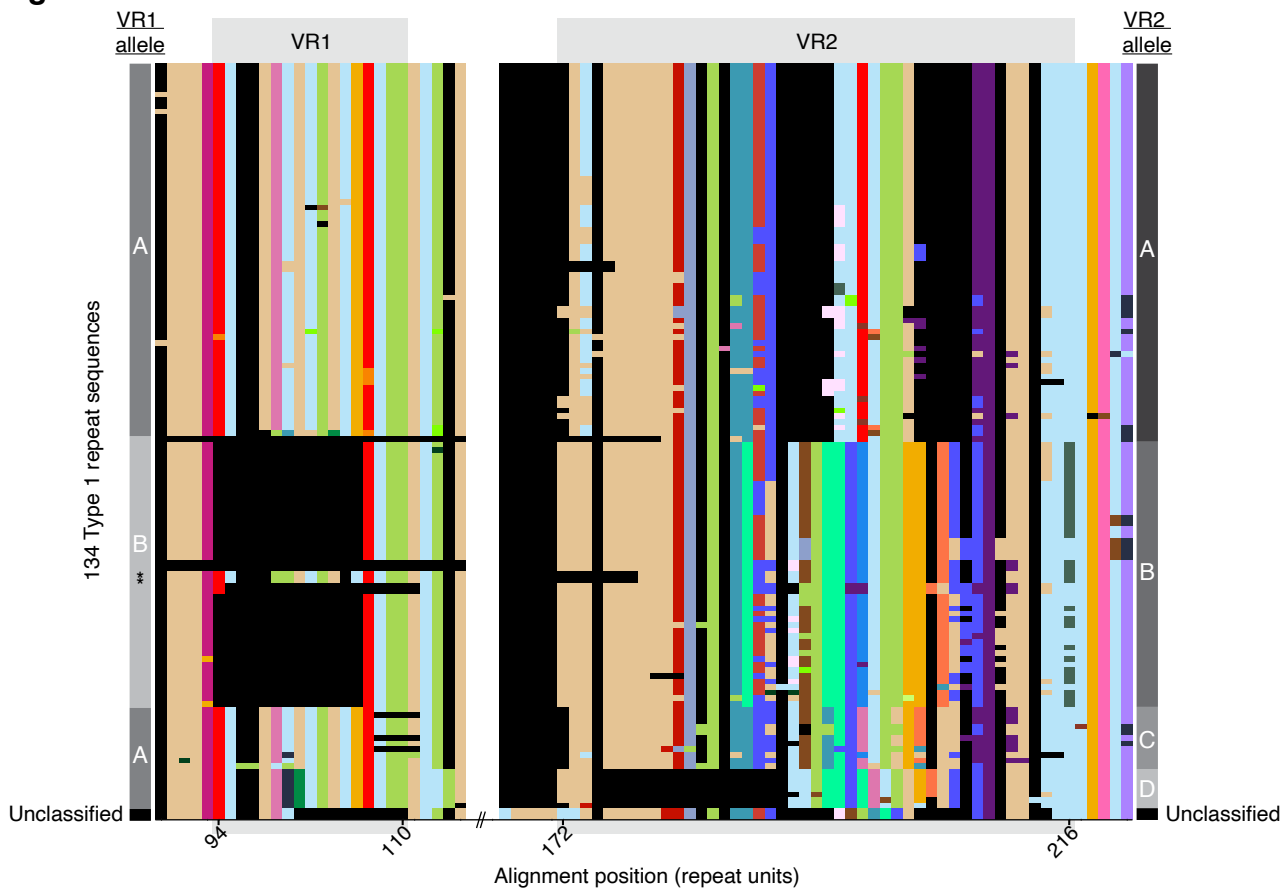

VR1 allele
VR1
VR2
VR2 allele
134 Type 1 repeat sequences
A
B
*
A
A
B
C
D
Unclassified
Unclassified
94
110
172
216
Alignment position (repeat units)

Fig. S4

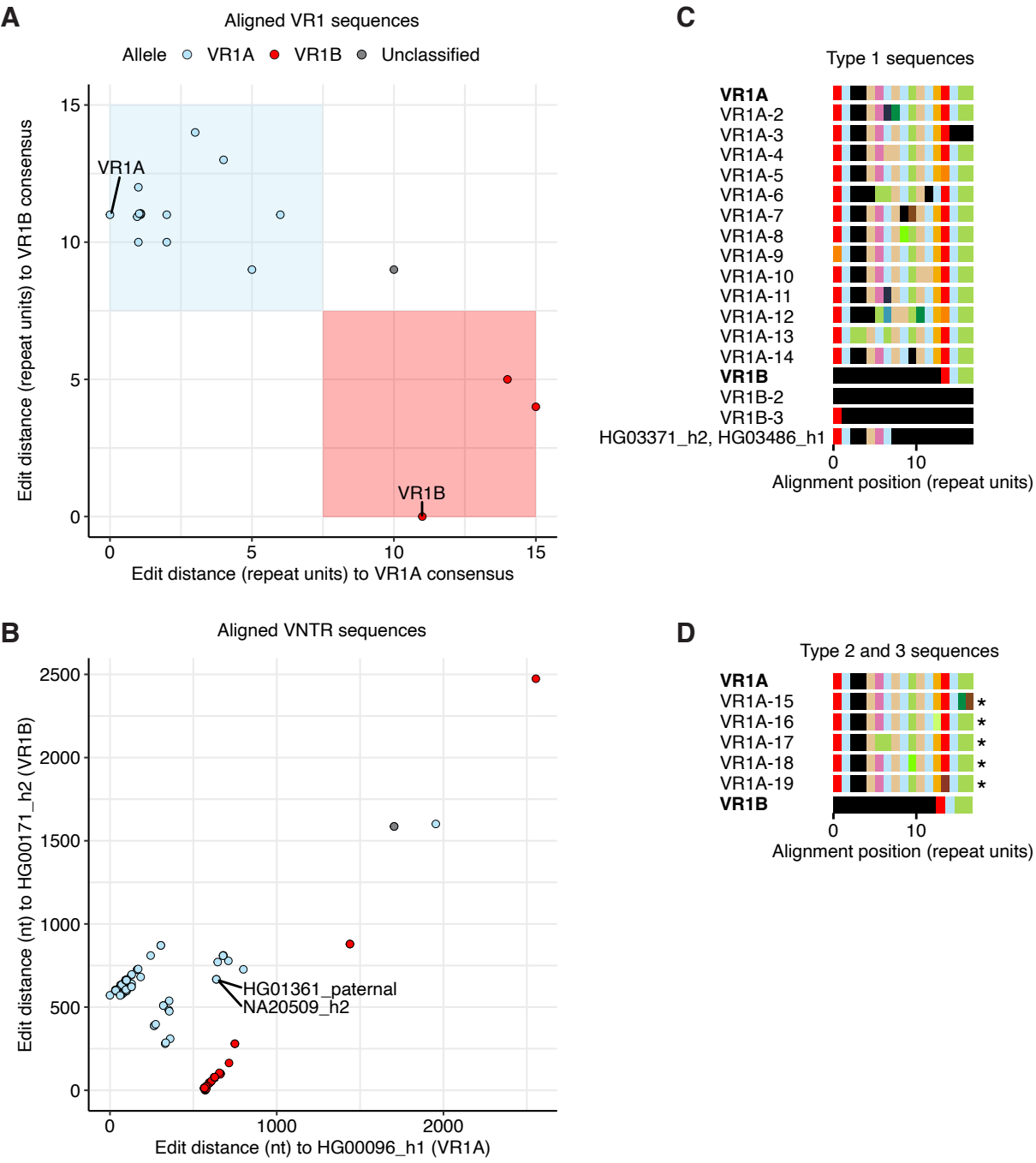

A
Aligned VR1 sequences
Allele VR1A VR1B Unclassified
VR1A
VR1B
Edit distance (repeat units) to VR1B consensus
Edit distance (repeat units) to VR1A consensus
B
Aligned VNTR sequences
HG01361_paternal
NA20509_h2
Edit distance (nt) to HG00171_h2 (VR1B)
Edit distance (nt) to HG00096_h1 (VR1A)
C
Type 1 sequences
VR1A
VR1A-2
VR1A-3
VR1A-4
VR1A-5
VR1A-6
VR1A-7
VR1A-8
VR1A-9
VR1A-10
VR1A-11
VR1A-12
VR1A-13
VR1A-14
VR1B
VR1B-2
VR1B-3
HG03371_h2, HG03486_h1
Alignment position (repeat units)
D
Type 2 and 3 sequences
VR1A
VR1A-15
VR1A-16
VR1A-17
VR1A-18
VR1A-19
VR1B
Alignment position (repeat units)

Fig. S5

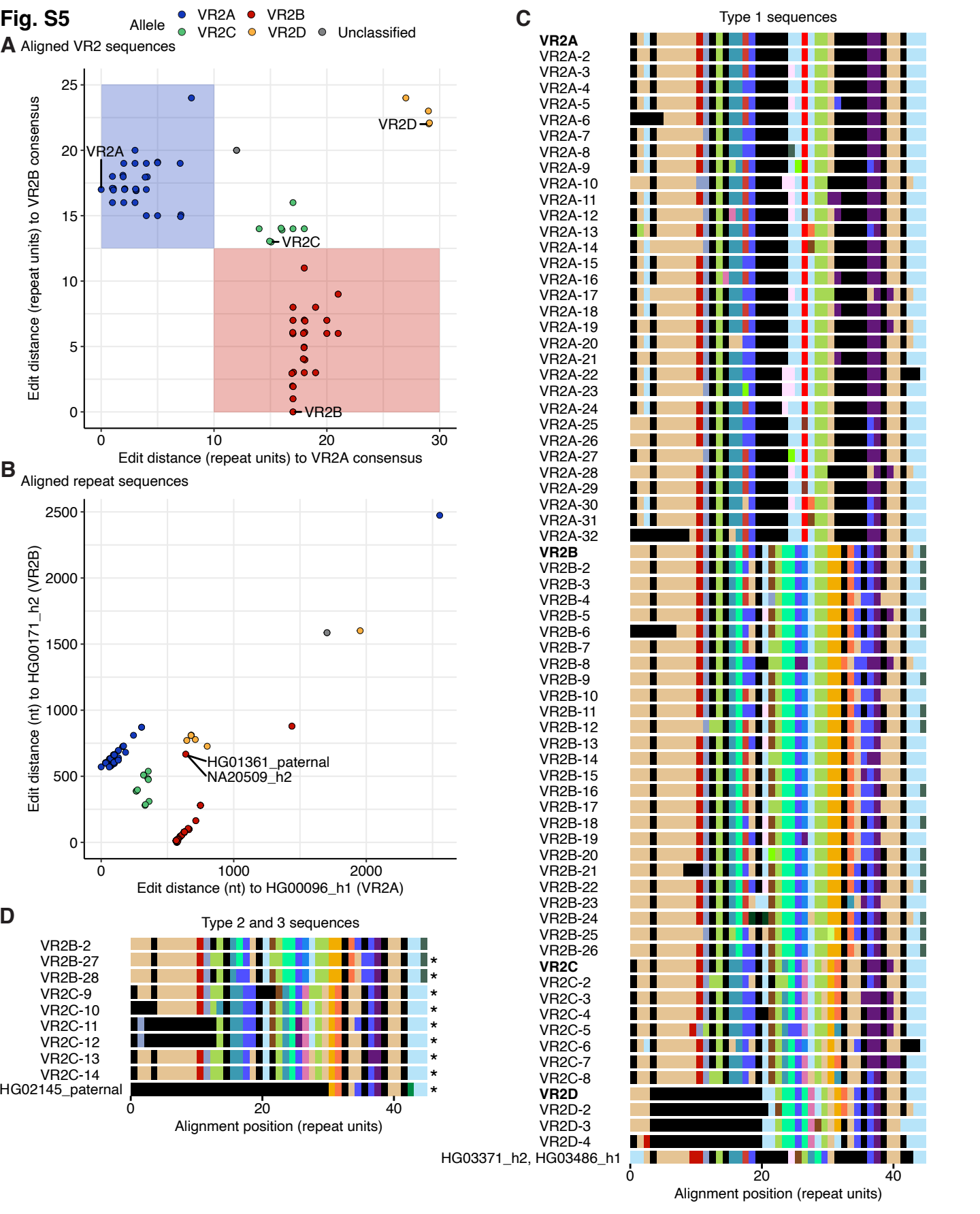
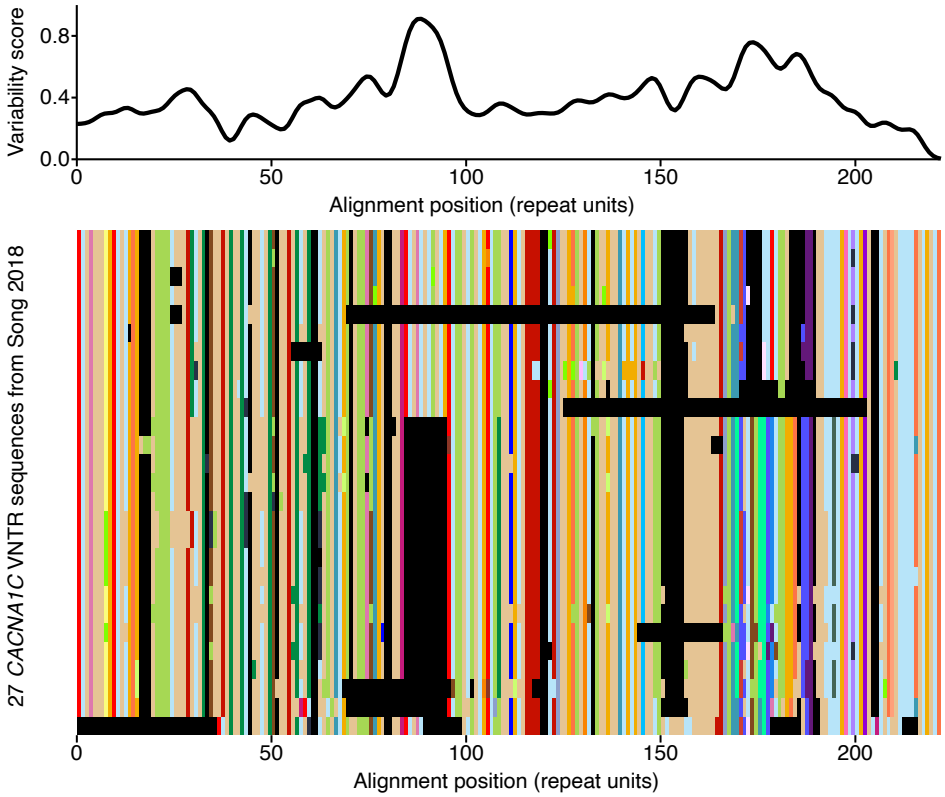

Allele
VR2A
VR2B
VR2C
VR2D
Unclassified
A
Aligned VR2 sequences
Edit distance (repeat units) to VR2B consensus
Edit distance (repeat units) to VR2A consensus
VR2A
VR2B
VR2C
VR2D
B
Aligned repeat sequences
Edit distance (nt) to HG00171_h2 (VR2B)
Edit distance (nt) to HG00096_h1 (VR2A)
HG01361_paternal
NA20509_h2
C
Type 1 sequences
VR2A
VR2A-2
VR2A-3
VR2A-4
VR2A-5
VR2A-6
VR2A-7
VR2A-8
VR2A-9
VR2A-10
VR2A-11
VR2A-12
VR2A-13
VR2A-14
VR2A-15
VR2A-16
VR2A-17
VR2A-18
VR2A-19
VR2A-20
VR2A-21
VR2A-22
VR2A-23
VR2A-24
VR2A-25
VR2A-26
VR2A-27
VR2A-28
VR2A-29
VR2A-30
VR2A-31
VR2A-32
VR2B
VR2B-2
VR2B-3
VR2B-4
VR2B-5
VR2B-6
VR2B-7
VR2B-8
VR2B-9
VR2B-10
VR2B-11
VR2B-12
VR2B-13
VR2B-14
VR2B-15
VR2B-16
VR2B-17
VR2B-18
VR2B-19
VR2B-20
VR2B-21
VR2B-22
VR2B-23
VR2B-24
VR2B-25
VR2B-26
VR2C
VR2C-2
VR2C-3
VR2C-4
VR2C-5
VR2C-6
VR2C-7
VR2C-8
VR2D
VR2D-2
VR2D-3
VR2D-4
HG03371_h2, HG03486_h1
Alignment position (repeat units)
D
Type 2 and 3 sequences
VR2B-2
VR2B-27
VR2B-28
VR2C-9
VR2C-10
VR2C-11
VR2C-12
VR2C-13
VR2C-14
HG02145_paternal
Alignment position (repeat units)

Fig. S6

A


Variability score
0.0
0.4
0.8
0
50
100
150
200
Alignment position (repeat units)
27 CACNA1C VNTR sequences from Song 2018
0
50
100
150
200
Alignment position (repeat units)

B

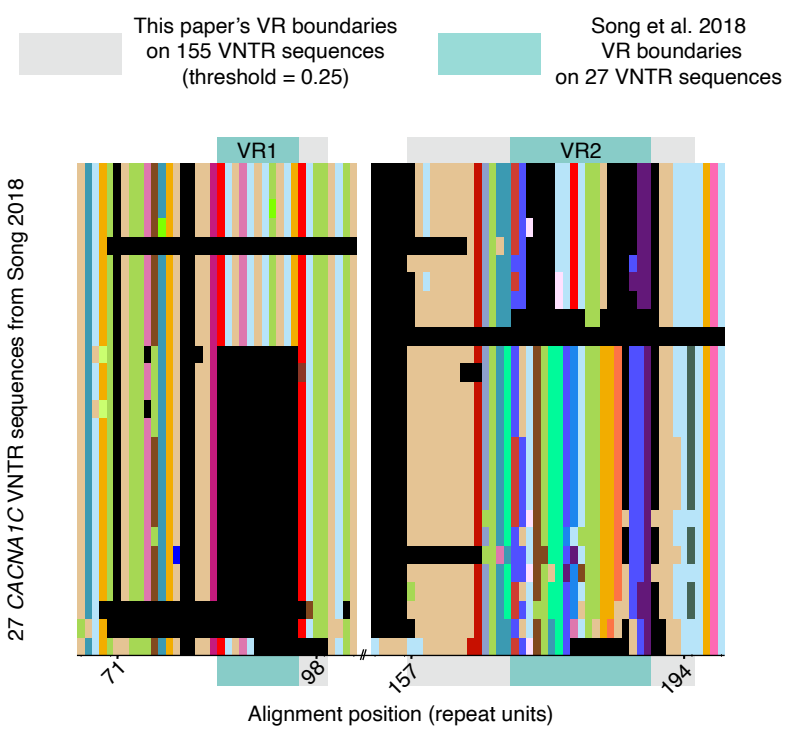

This paper's VR boundaries on 155 VNTR sequences (threshold = 0.25)
Song et al. 2018 VR boundaries on 27 VNTR sequences
VR1
VR2
27 CACNA1C VNTR sequences from Song 2018
71
98
157
194
Alignment position (repeat units)

Fig. S7

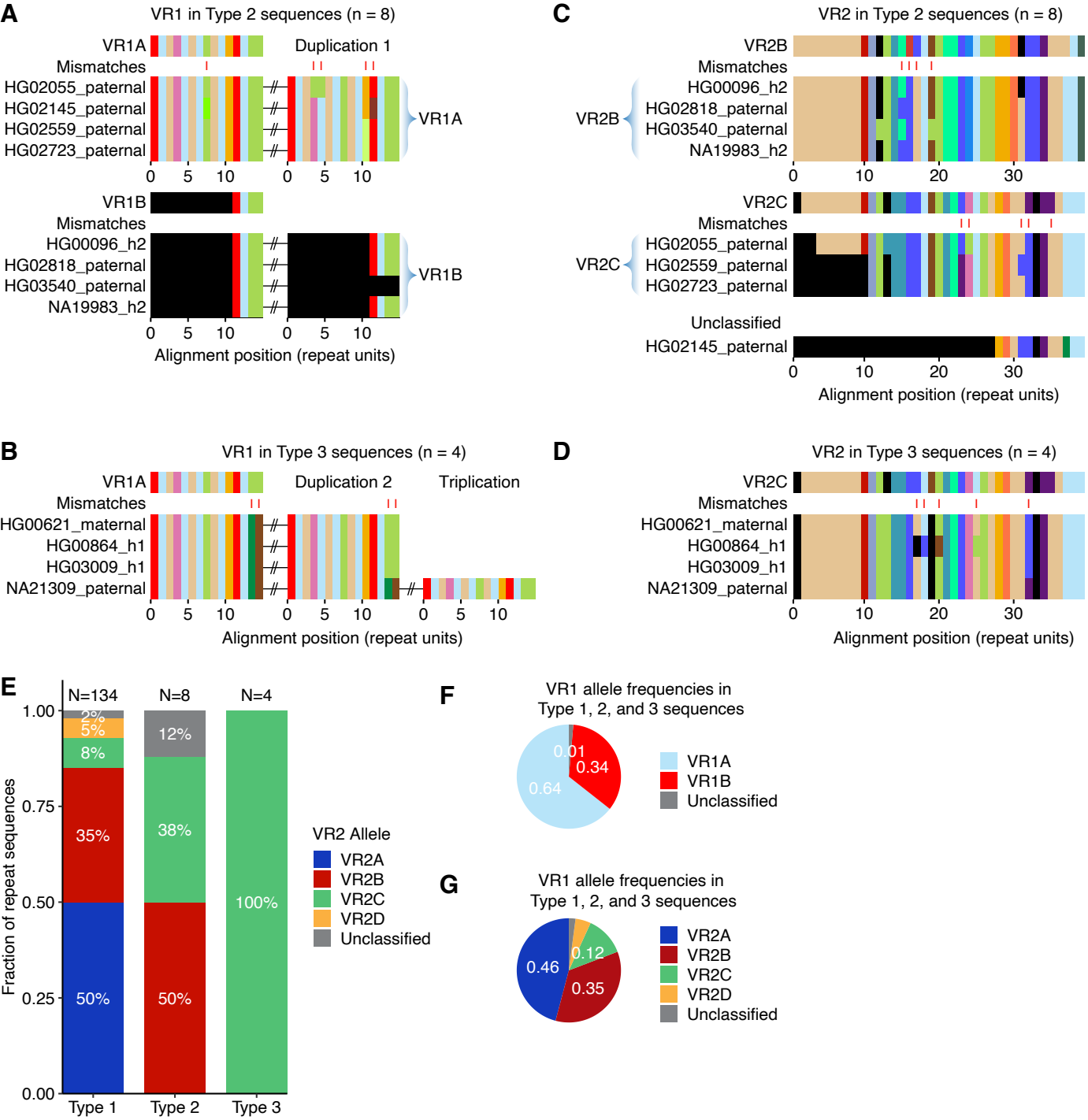
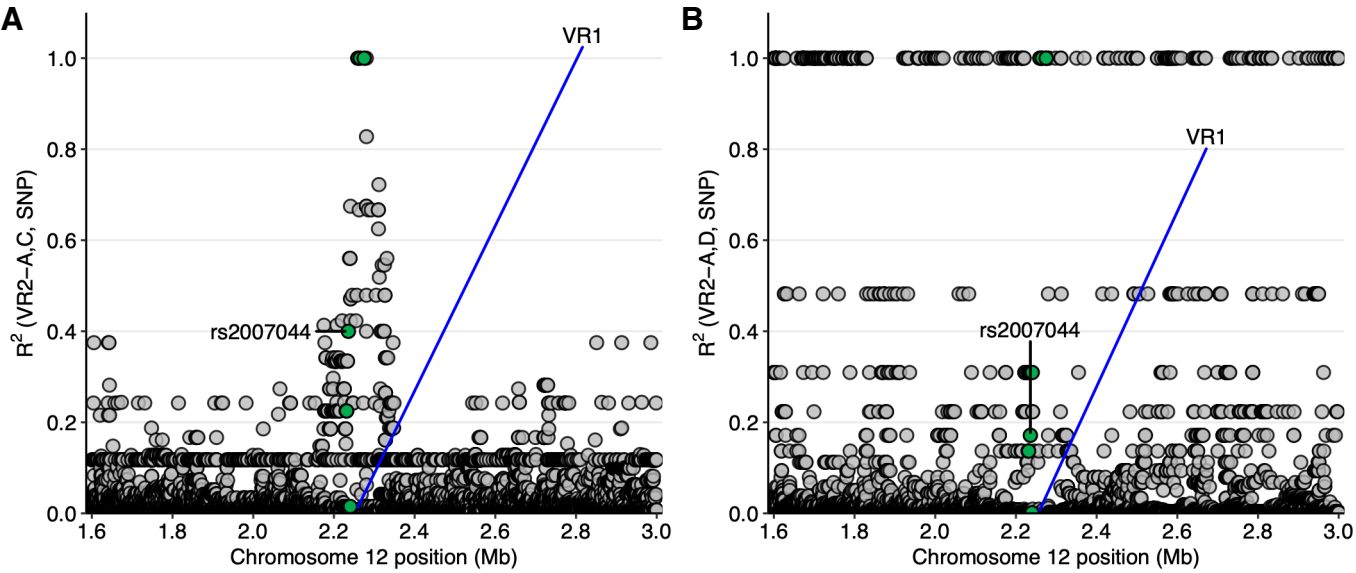
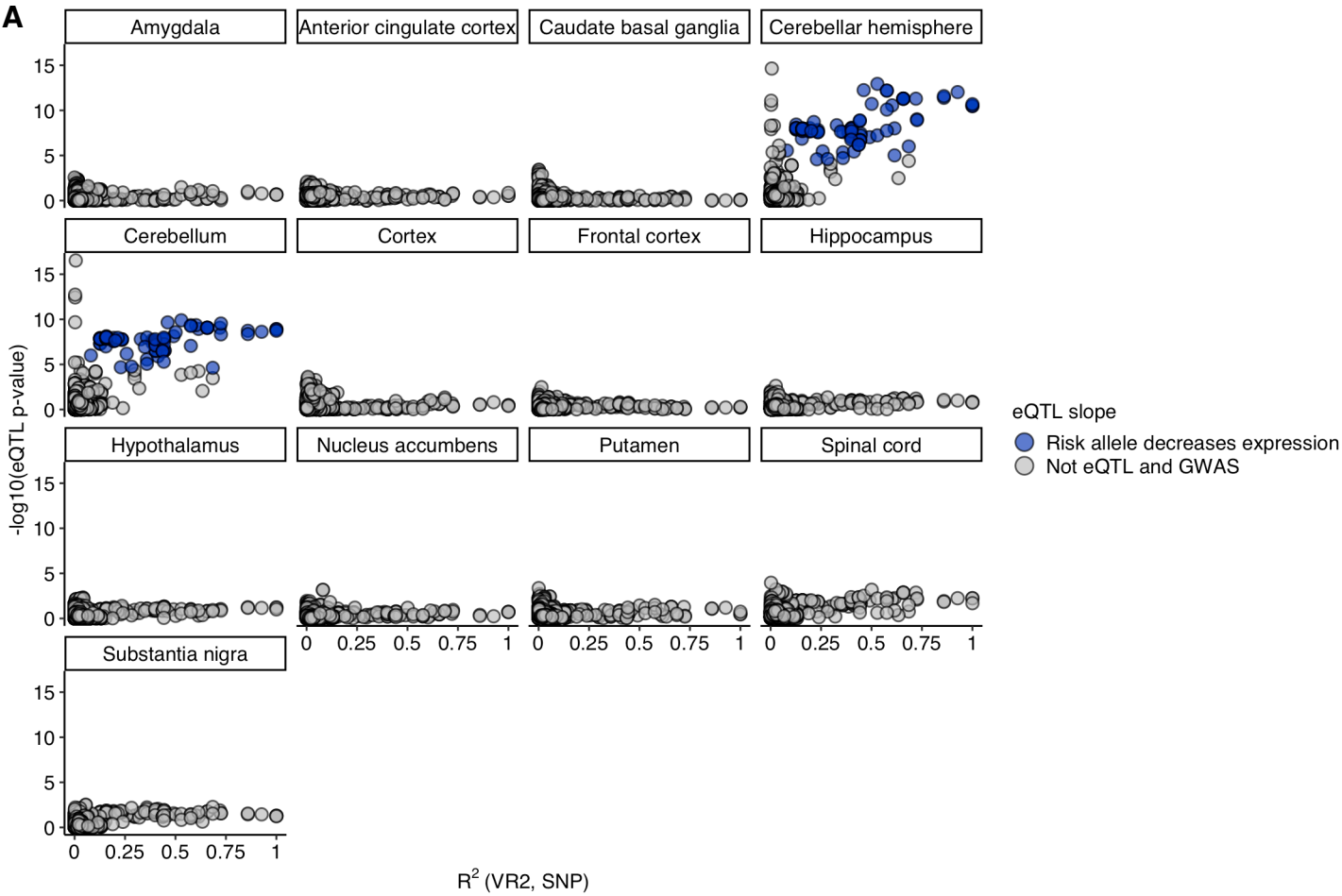

A
VR1 in Type 2 sequences (n = 8)
VR1A
Duplication 1
Mismatches
HG02055_paternal
HG02145_paternal
HG02559_paternal
HG02723_paternal
VR1A
VR1B
Mismatches
HG00096_h2
HG02818_paternal
HG03540_paternal
NA19983_h2
VR1B
Alignment position (repeat units)
B
VR1 in Type 3 sequences (n = 4)
VR1A
Duplication 2
Triplication
Mismatches
HG00621_maternal
HG00864_h1
HG03009_h1
NA21309_paternal
Alignment position (repeat units)
C
VR2 in Type 2 sequences (n = 8)
VR2B
Mismatches
HG00096_h2
HG02818_paternal
HG03540_paternal
NA19983_h2
VR2B
VR2C
Mismatches
HG02055_paternal
HG02559_paternal
HG02723_paternal
VR2C
Unclassified
HG02145_paternal
Alignment position (repeat units)
D
VR2 in Type 3 sequences (n = 4)
VR2C
Mismatches
HG00621_maternal
HG00864_h1
HG03009_h1
NA21309_paternal
Alignment position (repeat units)
E
N=134
N=8
N=4
Fraction of repeat sequences
2%
5%
8%
35%
50%
12%
38%
50%
100%
Type 1
Type 2
Type 3
VR2 Allele
VR2A
VR2B
VR2C
VR2D
Unclassified
F
VR1 allele frequencies in
Type 1, 2, and 3 sequences
0.01
0.34
0.64
VR1A
VR1B
Unclassified
G
VR1 allele frequencies in
Type 1, 2, and 3 sequences
0.46
0.12
0.35
VR2A
VR2B
VR2C
VR2D
Unclassified

Fig. S8


A
VR1
rs2007044
R^2 (VR2–A,C, SNP)
1.0
0.8
0.6
0.4
0.2
0.0
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
Chromosome 12 position (Mb)
B
VR1
rs2007044
R^2 (VR2–A,D, SNP)
1.0
0.8
0.6
0.4
0.2
0.0
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
Chromosome 12 position (Mb)


A
Amygdala
Anterior cingulate cortex
Caudate basal ganglia
Cerebellar hemisphere
Cerebellum
Cortex
Frontal cortex
Hippocampus
Hypothalamus
Nucleus accumbens
Putamen
Spinal cord
Substantia nigra
-log10(eQTL p-value)
R2 (VR2, SNP)
eQTL slope
Risk allele decreases expression
Not eQTL and GWAS

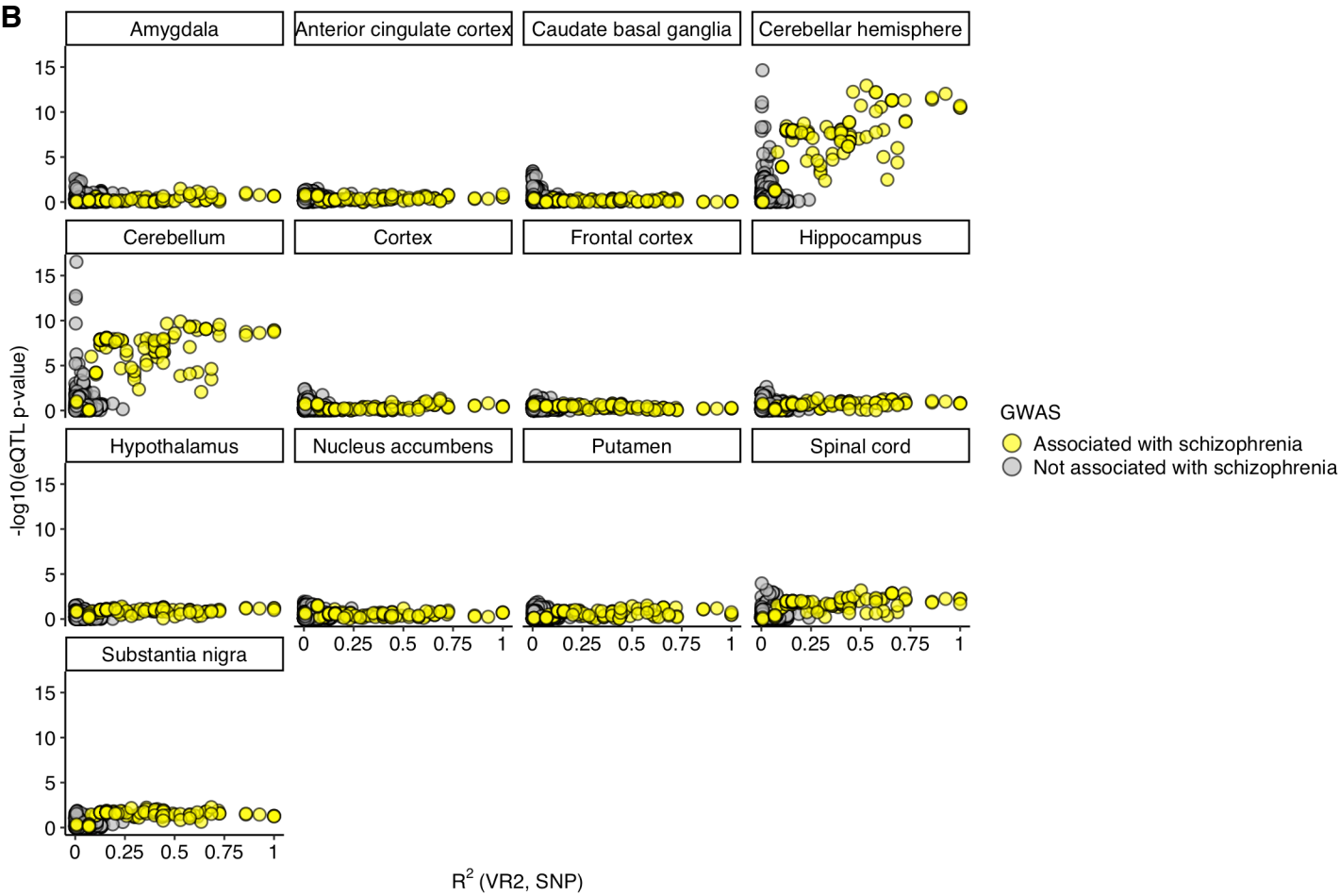
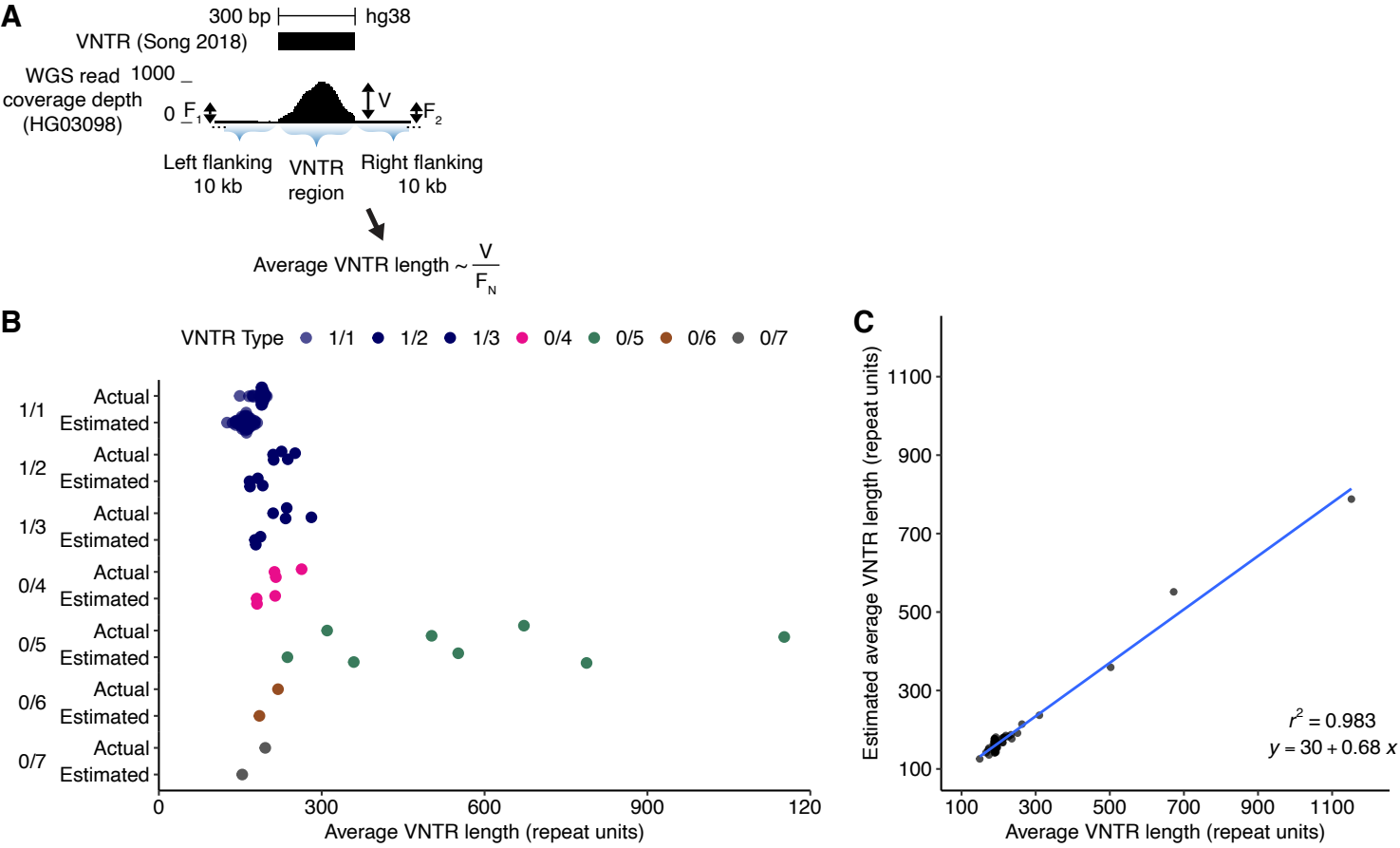

B
Amygdala
Anterior cingulate cortex
Caudate basal ganglia
Cerebellar hemisphere
Cerebellum
Cortex
Frontal cortex
Hippocampus
Hypothalamus
Nucleus accumbens
Putamen
Spinal cord
Substantia nigra
-log10(eQTL p-value)
R2 (VR2, SNP)
GWAS
Associated with schizophrenia
Not associated with schizophrenia

Fig. S10


A
300 bp
hg38
VNTR (Song 2018)
WGS read coverage depth (HG03098)
1000
0
F1
V
F2
Left flanking 10 kb
VNTR region
Right flanking 10 kb
Average VNTR length ~ V / FN
B
VNTR Type
1/1
1/2
1/3
0/4
0/5
0/6
0/7
Actual
Estimated
0
300
600
900
120
Average VNTR length (repeat units)
C
Estimated average VNTR length (repeat units)
1100
900
700
500
300
100
r2 = 0.983
y = 30 + 0.68 x
Average VNTR length (repeat units)