Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia

Janet H.T. Song,[1,2,4] Craig B. Lowe,[1,3,4] and David M. Kingsley[1,3,*]

Bipolar disorder (BD) and schizophrenia (SCZ) are highly heritable diseases that affect more than 3% of individuals worldwide. Genome-wide association studies have strongly and repeatedly linked risk for both of these neuropsychiatric diseases to a 100 kb interval in the third intron of the human calcium channel gene *CACNA1C*. However, the causative mutation is not yet known. We have identified a human-specific tandem repeat in this region that is composed of 30 bp units, often repeated hundreds of times. This large tandem repeat is unstable using standard polymerase chain reaction and bacterial cloning techniques, which may have resulted in its incorrect size in the human reference genome. The large 30-mer repeat region is polymorphic in both size and sequence in human populations. Particular sequence variants of the 30-mer are associated with risk status at several flanking single-nucleotide polymorphisms in the third intron of *CACNA1C* that have previously been linked to BD and SCZ. The tandem repeat arrays function as enhancers that increase reporter gene expression in a human neural progenitor cell line. Different human arrays vary in the magnitude of enhancer activity, and the 30-mer arrays associated with increased psychiatric disease risk status have decreased enhancer activity. Changes in the structure and sequence of these arrays likely contribute to changes in *CACNA1C* function during human evolution and may modulate neuropsychiatric disease risk in modern human populations.

More than 3% of the global population has bipolar disorder (BD) or schizophrenia (SCZ), and both diseases are among the top 25 causes of disability worldwide.[1–3] Along with the disability cost, both disorders are associated with an increased risk of suicide.[4,5] There are limited treatment options for BD and SCZ, and the burden of these diseases may be increasing.[6] Improved diagnosis and treatments may come from a better understanding of the molecular pathways that contribute to disease risk.

Both BD and SCZ are highly heritable. While they are classified as different diseases based on their clinical symptoms, they share a similar set of genomic risk variants.[7] Genome-wide association studies (GWASs) for BD and SCZ have consistently implicated risk variants in or near genes involved in calcium signaling.[8–14] Calcium signaling-related genes are also enriched for rare variants in families multiply affected by BD[15] and in individuals with SCZ,[16,17] suggesting that calcium signaling plays an important role in both BD and SCZ etiology.

Some of the strongest and best-replicated associations for BD and SCZ map within *CACNA1C*, which encodes the pore-forming subunit of the $Ca_V1.2$ calcium channel.[18] Disease-associated single-nucleotide polymorphisms (SNPs) are in strong linkage disequilibrium with each other and contained within a 100 kb region of the gene's third intron.[8–14,19] Underscoring the importance of this genomic region for psychiatric disease in humans, these disease-associated SNPs have also been associated with anxiety, depression-related symptoms, obsessive-compulsive symptoms, decreased performance in memory-related tasks, major depression, and autism.[13,20–28]

The causative variants at loci identified by GWASs could be the assayed SNPs themselves[29] or other variants tightly linked to the SNP markers.[30] Previous studies have investigated the functional consequences of the genotyped SNPs in *CACNA1C* and other closely linked SNPs.[31,32] However, the mutation responsible for the association between SNPs within *CACNA1C* and human neuropsychiatric diseases is still unknown.

Given the difficulty in identifying causal mutations at *CACNA1C* over the past 10 years,[8] we considered whether there might be additional structural variants at the locus that are not easily detected using current genotyping and sequencing methods. For example, copy-number variants and expansions and contractions of micro- and mini-satellite sequences can be difficult to identify with short-read or Sanger sequencing technologies. Nevertheless, these types of mutations have been implicated in a wide range of neurological diseases, including Huntington disease, spinocerebellar ataxia, fragile X syndrome, depression, and aggression and impulsivity behaviors,[33–40] and likely contribute to undetected variation and missing heritability for additional human traits.[41,42]

To search for unrecognized copy-number variants at the *CACNA1C* locus, we examined regions of the genome where no mutations were identified by large-scale sequencing projects such as the 1000 Genomes Project,[43] yet DNA sequencing reads consistently differed from the reference human assembly. We identified one such region (hg38; chr12:2255791–2256090) within the 100 kb interval associated with BD and SCZ. In the most recent human reference genome (hg38), this region is assembled as a tandem repeat composed of ten 30 bp units. Chimpanzees

[1]Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA; [2]Department of Genetics, Stanford University, Stanford, CA 94305, USA; [3]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA
[4]These authors contributed equally to this work
*Correspondence: kingsley@stanford.edu
https://doi.org/10.1016/j.ajhg.2018.07.011.

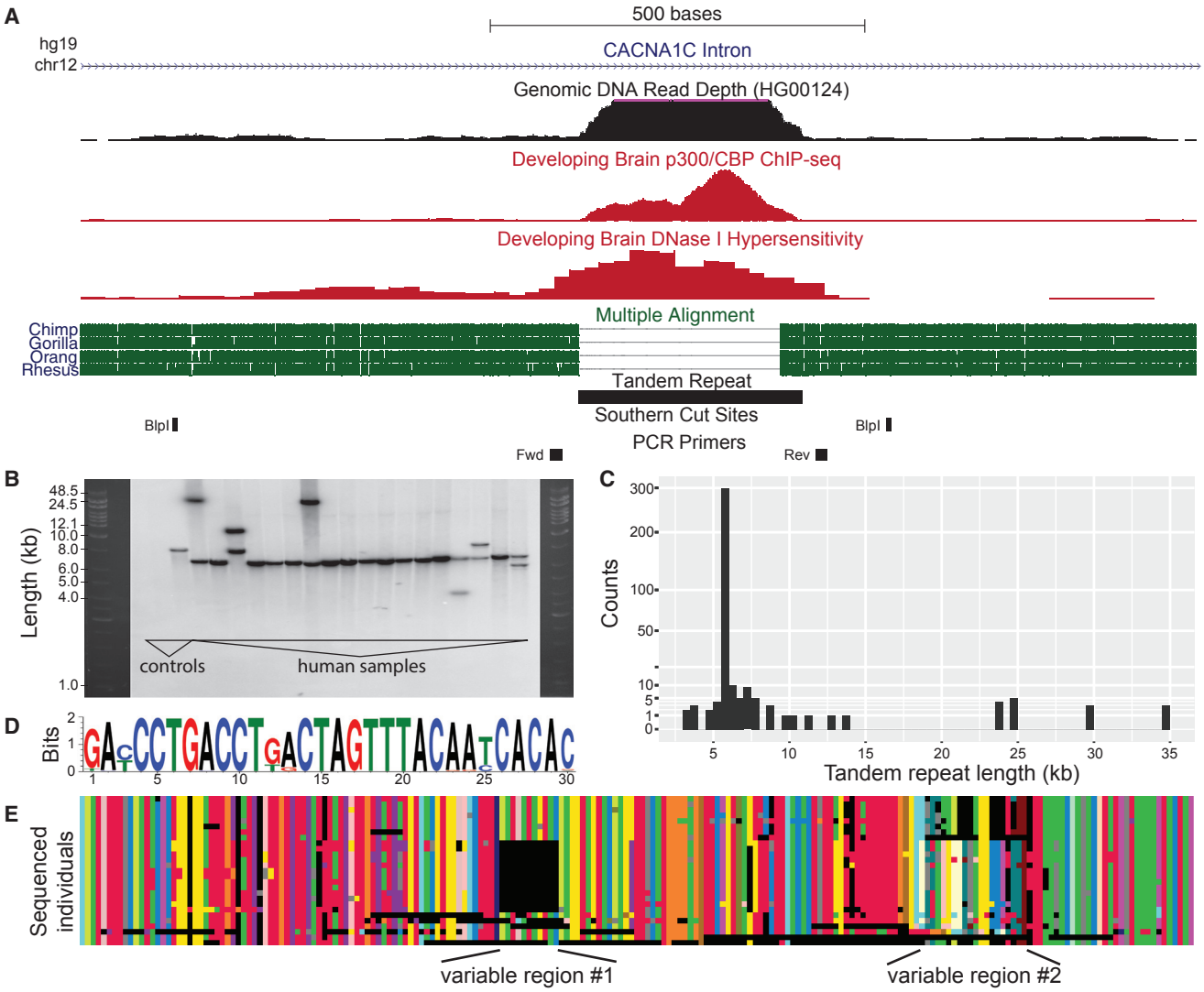


Figure 1. Human-Specific Tandem Repeat Region Is Composed of 30-mer Sequence Units Repeated Head-to-Tail in Multi-kilobase Arrays

(A) The tandem repeat is located in the third intron of *CACNA1C*. The human reference assembly predicts ten copies of the 30 bp segment while chimpanzees and other simians have a single copy of the 30 bp segment. More distantly related placental mammals, out to Afrotheria, have a region that aligns to the 30 bp segment, but with insertions or deletions. There is an abnormally large number of genomic DNA sequencing reads mapping to the tandem repeat region, consistent with this repeat being further expanded in human individuals. The repeat region also shows enrichment for p300/CBP binding and DNase I hypersensitivity in the developing human brain.

(B) We performed Southern blot analysis on 18 human individuals by probing for the 30 bp repeat after digesting with BlpI. We also included two controls: mouse DNA (no orthologous sequence) and the 8 kb vector from which the probe was transcribed. The human reference genome predicts a BlpI fragment of approximately 900 bp. In contrast, all humans tested show much larger BlpI fragment sizes (4,000 to 35,000 bp), and many individuals show dual bands indicating distinct alleles at the locus.

(C) Frequency distribution of 362 repeat allele lengths detected by Southern blot analysis.

(D) The 30-mer sequence logo calculated from the 30-mer variants present in human repeat arrays that were sequenced with long-read (PacBio) technology. Some positions are nearly invariant, whereas others vary from 30-mer to 30-mer.

(E) Structure and composition of tandem repeat arrays sequenced by PacBio long-read technology. Each row represents a different sequenced array, and each color represents a distinct 30-mer variant. Black regions indicate gaps that we have introduced to maximize repeat alignments between arrays. Many regions are organized similarly in all arrays, but common variable regions distinguish array subtypes.

and other non-human primates have a single instance of a homologous 30 bp sequence at this location (Figure 1A), suggesting that an ancestral 30-mer sequence has expanded in the *CACNA1C* intron during human evolution. Strikingly, the number of reads from individuals in the 1000 Genomes Project that map to this 300 bp segment is 3–379× greater than expected based on the reference assembly, and these reads contain multiple base substitutions. Read depth coverage and composition in Neanderthal and Denisovan genomes[44,45] fall within the

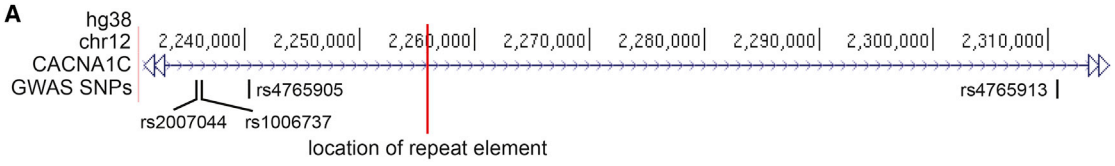
range observed among modern human populations (Figure S1). Further investigation identified a longer (3.3 kb) repeat array at this location in the Venter assembly (HuRef)[46] and an approximately 6 kb repeat array in the genome of a hydatidiform mole sequenced to 40× coverage with long-read technology.[47] Collectively, these data suggest that hominins have a large and variable tandem repeat in the neuropsychiatric risk-associated region of *CACNA1C*. The size of the tandem array is likely under-represented in the human reference genome by one or two orders of magnitude based on empirical estimates from read depth coverage (see Supplemental Methods).

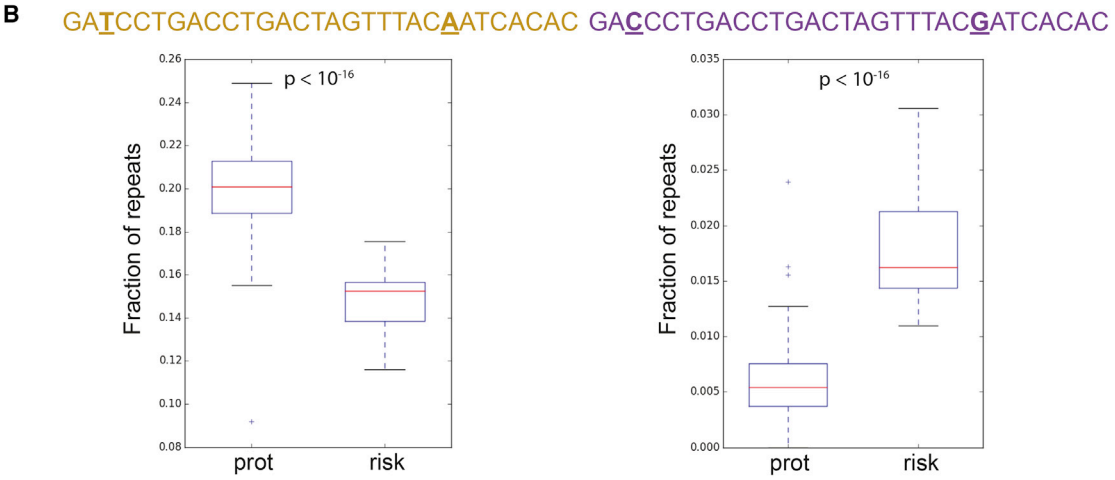
To further characterize the size of the tandem arrays using independent methods, we examined DNA from humans and our closest living relative, the chimpanzee. Polymerase chain reaction (PCR) amplification and sequencing from six chimpanzees confirmed a single instance of a 30-mer sequence, which exactly matches the chimpanzee reference genome. In contrast, when we performed Southern blots on human DNA (see Supplemental Methods), we found restriction fragment sizes consistent with repeat arrays of 3,000 to 30,000+ bp, with the majority of human repeat arrays showing sizes of approximately 6,000 bp (Figures 1B and 1C). We never observed a band size consistent with the human reference genome (hg38). The smallest band size seen in the 181 human samples we assayed (362 alleles) was 10 times larger than the repeat size annotated in the reference assembly (300 bp), while the largest was more than 100 times larger.

To understand why the human genome assembly appears to have a version of the tandem repeat that is not representative of the human population, we examined four bacterial artificial chromosome (BAC) clones derived from a single individual that were used in the sequencing and assembly of the human genome.[48] One BAC clone matched the length and sequence present in the assembly (300 bp). A BAC library made from a single individual should have at most two alleles; however, the four BACs all gave different tandem repeat lengths. Compounding this anomalous result, colonies picked from a single BAC clone are expected to be identical, but two of the four BACs produced subclones with varying tandem repeat lengths (Figure S2). In our experience, multi-kilobase tandem repeats, whose size was determined by Southern blot, reduced in length after amplification by PCR under routine conditions or when propagated using standard circular vectors in bacteria. We propose that the human reference assembly is based on a BAC clone that was correctly sequenced and assembled but that the sequence present in the BAC is an artifact of the instability of this tandem repeat when cloned and propagated using standard methods. Given the large size disparity between the version represented in the current human genome assembly and the alleles detected by Southern blot, as well as the instability of this region, we believe that the allele present in the current human genome assembly is not present in humans.

In addition to variation in the length of this repeat region in the human population, the 30-mer units that comprise each array also show sequence changes. For example, the array in the reference assembly is composed of four identical 30-mer units and six unique units that each contain a small number of SNPs. This variability in 30-mers is also seen in the large number of reads from the 1000 Genomes Project that map to this area. To better understand this variation in tandem repeat arrays, even for arrays of the same length, we performed long-read (PacBio) sequencing of repeat arrays amplified from 20 individuals using optimized PCR conditions (see Supplemental Methods). The size of the resulting PCR fragments using our optimized conditions matched the corresponding repeat lengths determined from Southern blots. The sequenced arrays were entirely composed of 30-mer units repeated head-to-tail. Some positions in the 30-mer unit appear to be largely invariant (e.g., position 2 is almost always an "A"), whereas other positions are more variable (Figures 1D and S3). For instance, the most common 30-mer unit (31%) is 5′-GAC̲CCTGACCTGACTAGTTTA CAATCACAC-3′ and the second most common (17%) is 5′-GAT̲CCTGACCTGACTAGTTTACAATCACAC-3′ (difference underlined). When aligning tandem repeat array variants, the structural organization of the 30-mer units within each array emerged (Figure 1E). Across all PacBio-sequenced repeat arrays, certain regions, such as the beginning and the end of each repeat array, contain the same 30-mer units organized almost identically. However, other regions (marked in Figure 1E) are more variable and contain specific patterns of 30-mer units that are consistently found in only a subset of the sequenced arrays.

The presence of a large and variable repeat region in the third intron of human *CACNA1C* raises the possibility that variation in the tandem array contributes to functional changes at the locus. To test whether the length or sequence of the tandem repeat region shows any association with genomic risk markers for BD and SCZ, we examined whole-genome sequence reads from individuals in the 1000 Genomes Project. We limited our analysis to individuals of European or East Asian descent, the two groups in which BD and SCZ risk status has been previously associated with four SNPs clustered in the third intron of the gene (rs2007044, rs1006737, rs4765905, and rs4765913; Figure 2A).[8–14,19] We first identified all sequencing reads from this repeat region. To infer the length of the repeat array (average of the individual's two alleles), we used the fraction of all sequencing reads for the individual that are from the repeat region (see Supplemental Methods). To estimate the sequence composition of the repeat array (averaged over the two alleles), we calculated the fraction of all 30-mers in the sequence reads identical to each observed sequence variant of the 30-mer unit (see Supplemental Methods). Since our length and composition statistics represent a mixture of the two alleles present in each person, we limited our analysis to individuals that are homozygous risk or protective at each of the four





C

	protective		risk		
	mean	stdev	mean	stdev	pval
GATCCTGACCTGACTAGTTTACAATCACAC	0.2002	0.0178	0.1474	0.0152	1.5E-17
GATCCTGACCTTACTAGTTTACAATCACAG	0.0241	0.0054	0.0146	0.0045	4.2E-13
GACCCTGACCTGACTAGTTTACGACCACAC	0.0090	0.0045	0.0044	0.0022	6.3E-09
GACCCTGACCTGACTAGTTTACGATCACAT	0.0031	0.0021	0.0010	0.0013	1.6E-08
GACCCTGACCTTACTAGTTTACAACCACAC	0.0111	0.0034	0.0068	0.0035	1.1E-07
GACCCTGACCTTACTAGTTTACGATCACAC	0.0002	0.0008	0.0109	0.0025	3.6E-18
GACCCTGACCTGACTAGTTTACGATCACAC	0.0058	0.0028	0.0180	0.0048	1.1E-17
GACCCTGACCTTACTAGTTTACGACCACAC	3.6E-05	0.0002	0.0029	0.0015	8.9E-17
GATCCTGACCTGACTAGTTTACAATCACAT	0.0002	0.0005	0.0054	0.0027	1.8E-16
GACCCTGACCTGACTAGTTTACAACCACAC	0.0930	0.0158	0.1068	0.0151	1.4E-05

Figure 2. 30-mer Repeat Variants Are Associated with Protective or Risk Status at GWAS SNPs Linked to Neuropsychiatric Disease
(A) Genome browser view of the third intron of *CACNA1C*. A red line marks the location of the repeat region. The human-specific 30-mer repeats are embedded in a region defined by four SNPs that are repeatedly associated with BD and SCZ.
(B) We identified individuals from the 1000 Genomes Project that have the protective genotype at all four GWAS SNPs (protective haplotype) and individuals with the risk genotype at all four GWAS SNPs (risk haplotype). We used only European and East Asian individuals because GWASs have only been done with these populations. For each possible 30-mer repeat unit, we determined what fraction of 30-mers in the reads that map to this locus in each individual exactly match that particular variant. The 30-mer sequence on the left is significantly associated with the protective haplotype ("prot"), whereas the 30-mer variant on the right is significantly associated with the risk haplotype ("risk"). Base pair differences between the two 30-mer variants presented here are underlined. Shown are standard box-and-whisker plots where the box represents the lower quartile, median, and upper quartile, and the whiskers represent the range of the measurements. Outliers ("+") are data points that are outside the nearest quartile + 1.5× the interquartile range.
(C) The table lists the mean and standard deviation of the fraction of reads that exactly match a given 30-mer for individuals with the protective or risk haplotype. Repeats enriched in the protective haplotype group are shown in yellow, and repeats enriched in the risk haplotype group are shown in purple. The p values were calculated using the Wilcoxon rank-sum test with Bonferroni correction (see Supplemental Methods).

SNPs commonly associated with BD and SCZ (Figure 2A). These SNPs are all tightly linked and define risk and protective haplotypes, making it possible to study repeat structures associated with risk or protective genotypes at *CACNA1C*.

We first tested whether repeat length is consistently associated with genotype status at the four GWAS SNPs. None of the four SNPs show a significant association with repeat length and the direction of effect is not consistent (Figure S4). It does not appear that repeat array length is associated with the protective or risk genotypes at the GWAS SNPs, at least not in a simple manner.

We then tested whether specific sequence variants of the 30-mer unit are associated with the risk or protective alleles at the four GWAS SNPs. For each sequence variant of the 30-mer unit, we tested whether its propensity to appear in reads from this repeat region differs between individuals that are homozygous for risk or protective genotypes at the four SNPs (Figure 2B). We identified a number of 30-mer units that are consistently associated with a genotypic class across all four SNPs (Figure S5). When considering only individuals that are homozygous protective at all four GWAS SNPs ("protective haplotype") or homozygous risk at all four GWAS SNPs ("risk haplotype"), five 30-mer variants

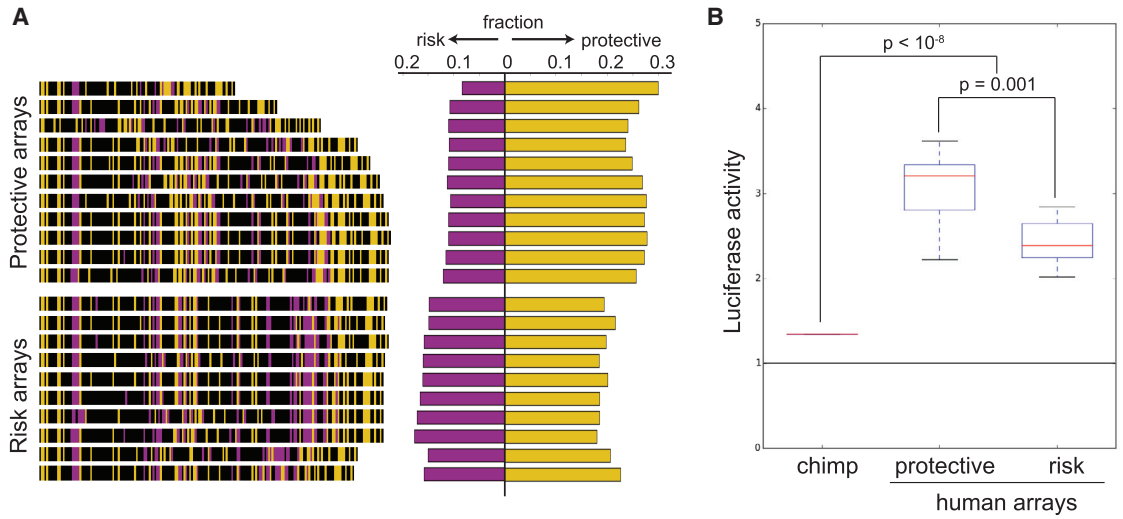


Figure 3. Human-Specific Repeat Arrays Act as Enhancers in Neural Cells
(A) The single 30-mer found in chimpanzees (30 bp) and 21 different human repeat arrays (3.5–6 kb) were cloned upstream of a minimal promoter driving expression of the luciferase reporter gene. 30-mer variants significantly associated with the protective haplotype are colored yellow, 30-mer variants significantly associated with the risk haplotype are purple, and non-significant variants are black. The fraction of total 30-mer variants associated with either the risk or the protective haplotype varies between the protective-associated and risk-associated repeat arrays as expected, although the differences are subtle.
(B) Constructs were assayed for luciferase activity in a human neural progenitor cell line (ReNcell Cx), as described in Supplemental Methods. Human repeat alleles drove significantly higher luciferase activity compared to the single 30-mer found in chimpanzees ($p < 10^{-8}$). Protective arrays drove significantly higher luciferase activity than risk arrays ($p = 0.001$). The p values were calculated using the Wilcoxon rank-sum test.

are associated with the protective haplotype, and five 30-mer variants are associated with the risk haplotype (Figure 2C). These particular 30-mer units tend to be located in the variable regions observed when aligning the PacBio-sequenced repeat arrays (marked in Figure 1E). Thus, while there is no straightforward association between the overall length of repeat arrays and the risk or protective haplotype, the abundance of particular 30-mer units is significantly and consistently associated with markers for psychiatric disease risk at *CACNA1C*.

The repeat region gives a significant signal for p300 enrichment in ChIP-seq experiments performed on tissue from the developing human brain[49] and also shows an open chromatin signal during human brain development (Figures 1A and S6).[50] Both of these results are consistent with the repeat region acting as a distal enhancer element during brain development. To experimentally test whether the 30-mer repeat arrays show enhancer activity in developing neural cells, we cloned the single 30-mer sequence found in chimpanzees (30 bp), as well as 21 different human repeat arrays (3.2–6 kb), upstream of a basal promoter and a luciferase reporter gene (Figure 3A). We used a linear cloning vector that greatly improved repeat stability, and we confirmed clone stability via comparison to the expected size and sequence as determined from Southern blot analysis and PacBio sequencing, respectively (see Supplemental Methods). We then transfected each construct into a human neural progenitor cell line (ReNcell Cx) and measured luciferase activity (see Supplemental Methods). The chimpanzee construct, containing a single 30 bp unit, weakly enhanced luciferase activity relative to the empty vector ($p = 0.01$), while the much larger repeat arrays found in humans significantly enhanced luciferase activity compared to both the empty vector ($p < 10^{-8}$) and the chimpanzee construct ($p < 10^{-8}$, Figure 3B). We additionally tested three individual 30-mer sequence variants that are commonly observed in humans. Like the single 30 bp unit found in chimpanzees, these 30-mer variants acted as weak enhancers in the luciferase assay (Figure S7). These results suggest that the expansion of a single 30 bp unit to hundreds of tandem repeats at the *CACNA1C* locus during human evolution has strengthened an existing enhancer element.

Although all of the tested human repeat arrays consistently acted as enhancers, there was substantial quantitative variation in enhancer strength among human repeat arrays (Figure S8). To test whether enhancer strength varied for arrays linked to protective or risk GWAS SNPs for neuropsychiatric disease, we determined the genotypes of the individuals from which these human repeat arrays were cloned. Repeat arrays derived from individuals with the protective haplotype were classified as "protective," and repeat arrays derived from individuals with the risk haplotype were classified as "risk." For repeat arrays derived from individuals who are heterozygous at the GWAS SNPs, we determined the proportion of protective- and risk-associated 30-mer variants from PacBio sequencing. We then asked whether these proportions most closely resembled individuals with the protective haplotype or individuals with the risk haplotype in the 1000 Genomes Project and

designated the ambiguous repeat arrays accordingly (Figure S9, see Supplemental Methods). The repeat arrays characteristic of the protective haplotype drove significantly higher luciferase activity than repeat arrays characteristic of the risk haplotype ($p = 0.001$, Figure 3). In contrast, we did not observe an association between repeat length in the human repeat arrays we tested (3.2–6 kb) and luciferase activity (Figure S10). These data show that compositional differences between human repeat arrays lead to functional differences in enhancer activity and suggest that differences in the repeat arrays may be causative genomic changes underlying the association between linked *CACNA1C* markers and susceptibility to neuropsychiatric disease.

While the transcriptional enhancer is located within the *CACNA1C* locus, the enhancer might also affect the expression of other linked genes.[51] In human brain samples, the enhancer is present in a topologically associating domain (TAD) that contains both *CACNA1C* and seven downstream genes.[52,53] In human dorsolateral prefrontal cortex, more than 90% of Hi-C associations found within 5 kb of the enhancer map to other locations within *CACNA1C*, including locations near *CACNA1C* transcription start sites.[52] These results are consistent with the enhancer regulating the expression of *CACNA1C*.

Previous studies have tested whether risk and protective genotypes at the *CACNA1C* locus lead to higher or lower *CACNA1C* expression in the brain. Studies in the dorsolateral prefrontal cortex and cerebellum reported decreased *CACNA1C* expression in individuals with risk variants at human GWAS SNPs.[20,54] In contrast, studies in the superior temporal gyrus and fibroblast-derived induced human neurons reported increased *CACNA1C* expression in individuals with risk variants at human GWAS SNPs.[32,55] Our data show that risk-associated repeat arrays have reduced enhancer activity in the particular human neural progenitor cell line we tested. We note that differences in human repeat arrays could also underlie more complex expression differences at other tissues or developmental time points. The base pair changes seen in particular 30-mer motifs that are associated with risk or protective genotypes alter the predicted binding sites for a number of potential trans-regulatory factors (Table S1). These factors themselves vary in expression and abundance in different brain regions,[56–60] which could in turn lead to differential effects of repeat variants at different times or places *in vivo*.

Previous studies of coding region mutations suggest that both loss-of-function and gain-of-function alterations in *CACNA1C* can lead to behavioral changes in mice and humans with similarities to BD and SCZ. For example, in mouse models where *CACNA1C* expression levels are either globally reduced or ablated only in specific brain regions, mice display increased anxiety and depression in behavioral tests such as the elevated plus maze, light-dark box, and learned helplessness test.[18,61–63] Conversely, gain-of-function mutations in *CACNA1C* lead to Timothy syndrome (TS) in humans, an autosomal-dominant disease where afflicted individuals display autism-like symptoms in addition to a host of non-neurological pathologies.[64,65] Although TS is normally lethal in young children, a rare individual with TS who survived into his late teens developed BD.[54] These studies suggest that modulating *CACNA1C* expression levels, such as through human variation at the repeat arrays we report in this study, could result in behavioral changes associated with BD and SCZ.

We note that the 30-mer repeat arrays might have additional functional effects beyond the enhancer activities we characterize here. For example, the most common 30-mer sequences have open reading frames in both directions (Figure S11A). Previous studies have shown that some tandem repeats are transcribed and translated even in the absence of conventional ATG start codons.[66–68] The tandem repeat also contains canonical splice site consensus sequences, including a donor site, an acceptor site, branch sites, and a polypyrimidine tract (Figure S11B). Intriguingly, the single 30-mer found in chimpanzees has an "A" at the 17th position, whereas the vast majority of human 30-mers (99.94%) have a "G" at that position. This single base pair difference means that chimpanzees do not have canonical splice donor or acceptor sites at this locus. Finally, when organized in head-to-tail fashion, the 30-mers also form a CpG site located between the "C" that ends most 30-mers and the "G" that begins the next 30-mer (Figure S11B). The tandem repeat arrays may affect translation, splicing, or methylation, in addition to forming a functional enhancer sequence within *CACNA1C*.

Tandem repeats have previously been proposed as a possible causal basis for the evolution of both species-specific traits and individual-to-individual variation in complex phenotypes such as neurological functions in humans.[69] Our studies have identified a dramatic expansion of a 30-mer sequence that generates human-specific tandem arrays in a key gene related to calcium signaling, gene expression, and behavior. The human-specific repeat arrays show enhancer activity in human neural progenitor cells, and risk-associated versions of the tandem repeat have less enhancer activity than protective-associated versions. We hypothesize that generation of these repeat arrays has modified $Ca_V1.2$ function during human evolution and that structural and compositional differences of the 30-mer repeats among humans represent causal genomic changes that modify risk of neuropsychiatric disease in modern populations.

Many diseases that are particularly common in human populations occur at body locations that have also undergone dramatic and relatively recent evolution in the human lineage. For example, humans have a high incidence of lower back, knee, and foot problems, likely due to the recent evolutionary transition to upright bipedal walking.[70] More than 70% of young adults develop impacted third molars (wisdom teeth), likely due to evolutionary reduction of jaw size in the human lineage and modern changes in diet.[71,72] Similarly, the high prevalence

of neurological diseases in modern humans may be, in part, due to recent evolutionary changes in genes controlling brain size, connectivity, and function in humans compared to other primates.[73,74] Tandem repeat expansions provide a particularly interesting class of genomic variants in evolution and disease studies because the generation of new tandem repeats can not only alter gene functions between species, but also make the same genes prone to variation and diversity among individuals of the same species.[69]

Producing new cellular and animal models that carry either chimpanzee or various human 30-mer repeat arrays at the *CACNA1C* locus should make it possible to further characterize both the evolutionary and disease effects of this repeat region. In addition, the sequence differences in the 30-mer repeats can now be used as a feature to group affected individuals into distinct genetic subtypes. Further stratification of individuals based on *CACNA1C* repeat genotypes may prove useful for refined disease association studies or for identifying affected individuals who are likely to show favorable responses to drugs targeting calcium channel activity. These drugs have long been available but have produced mixed results as treatments for psychiatric diseases.[75,76]

Finally, our research illustrates how characterizing hidden variation in the human genome can uncover variants associated with both human evolution and disease. SNPs are still the most commonly studied type of variant in most genotyping and trait association studies. However, structural variants and repeat sequences make up a substantial fraction of the human genome, show abundant variation both within and between species, and may contribute to key phenotypic traits and disease susceptibilities in humans and other organisms.[77]

Accession Numbers

The accession numbers for the repeat array sequences reported in this paper are GenBank: MH645925–MH645951.

Supplemental Data

Supplemental Data include Supplemental Methods, 11 figures, and 1 table and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.07.011.

Acknowledgments

We wish to thank members of the Kingsley Lab for useful discussions and comments on the manuscript. Research reported in this publication was supported in part by the NIDCR of the National Institutes of Health (K25DE025316 to C.B.L.), and by a National Science Foundation Graduate Research Fellowship and a Stanford Graduate Fellowship (J.H.T.S.). D.M.K. is an Investigator of the Howard Hughes Medical Institute. DNA or tissue samples were obtained through the NIH Neurobiobank from the Human Brain and Spinal Fluid Resource Center (VA West Los Angeles Healthcare Center), the University of Maryland Brain and Tissue Bank, the Harvard Brain Tissue Resource Center, the University of Miami Brain Endowment Bank, the Mt. Sinai Brain Bank, and the Brain Tissue Donation Program at the University of Pittsburgh. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of Interests

The authors declare no competing interests.

Received: May 2, 2018
Accepted: July 13, 2018
Published: August 9, 2018

References

1. Saha, S., Chant, D., Welham, J., and McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. PLoS Med. *2*, e141.
2. Merikangas, K.R., Jin, R., He, J.-P., Kessler, R.C., Lee, S., Sampson, N.A., Viana, M.C., Andrade, L.H., Hu, C., Karam, E.G., et al. (2011). Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch. Gen. Psychiatry *68*, 241–251.
3. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet *388*, 1545–1602.
4. Krishnan, K.R. (2005). Psychiatric and medical comorbidities of bipolar disorder. Psychosom. Med. *67*, 1–8.
5. Baldessarini, R.J., Pompili, M., and Tondo, L. (2006). Suicide in bipolar disorder: Risks and management. CNS Spectr. *11*, 465–471.
6. Saha, S., Chant, D., and McGrath, J. (2007). A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time? Arch. Gen. Psychiatry *64*, 1123–1131.
7. Forstner, A.J., Hecker, J., Hofmann, A., Maaser, A., Reinbold, C.S., Mühleisen, T.W., Leber, M., Strohmaier, J., Degenhardt, F., Treutlein, J., et al. (2017). Identification of shared risk loci and pathways for bipolar disorder and schizophrenia. PLoS ONE *12*, e0171595.
8. Ferreira, M.A., O'Donovan, M.C., Meng, Y.A., Jones, I.R., Ruderfer, D.M., Jones, L., Fan, J., Kirov, G., Perlis, R.H., Green, E.K., et al.; Wellcome Trust Case Control Consortium (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat. Genet. *40*, 1056–1058.
9. Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.Y., Duan, J., Ophoff, R.A., Andreassen, O.A., et al.; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.
10. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al.; Multicenter Genetic Studies of Schizophrenia Consortium; Psychosis Endophenotypes International Consortium; and Wellcome Trust Case Control

Consortium 2 (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. *45*, 1150–1159.
11. Ripke, S., Neale, B.M., Corvin, A., Walters, J.T., Farh, K.H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
12. Sklar, P., Ripke, S., Scott, L.J., Andreassen, O.A., Cichon, S., Craddock, N., Edenberg, H.J., Nurnberger, J.I., Rietschel, M., Blackwood, D., et al.; 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.
13. Smoller, J.W., Ripke, S., Lee, P.H., Neale, B., Nurnberger, J.I., Santangelo, S., Sullivan, P.F., Perlis, R.H., Purcell, S.M., Fanous, A., et al.; 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.
14. Ruderfer, D.M., Fanous, A.H., Ripke, S., McQuillin, A., Amdur, R.L., Gejman, P.V., O'Donovan, M.C., Andreassen, O.A., Djurovic, S., Hultman, C.M., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Bipolar Disorder Working Group of the Psychiatric Genomics Consortium; and Cross-Disorder Working Group of the Psychiatric Genomics Consortium (2014). Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. Mol. Psychiatry *19*, 1017–1024.
15. Ament, S.A., Szelinger, S., Glusman, G., Ashworth, J., Hou, L., Akula, N., Shekhtman, T., Badner, J.A., Brunkow, M.E., Mauldin, D.E., et al.; Bipolar Genome Study (2015). Rare variants in neuronal excitability genes influence risk for bipolar disorder. Proc. Natl. Acad. Sci. USA *112*, 3576–3581.
16. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185–190.
17. Andrade, A., Hope, J., Allen, A., Yorgan, V., Lipscombe, D., and Pan, J.Q. (2016). A rare schizophrenia risk variant of CACNA1I disrupts $Ca_V3.3$ channel activity. Sci. Rep. *6*, 34233.
18. Dedic, N., Pöhlmann, M.L., Richter, J.S., Mehta, D., Czamara, D., Metzger, M.W., Dine, J., Bedenk, B.T., Hartmann, J., Wagner, K.V., et al. (2018). Cross-disorder risk gene CACNA1C differentially modulates susceptibility to psychiatric disorders during development and adulthood. Mol. Psychiatry *23*, 533–543.
19. Nie, F., Wang, X., Zhao, P., Yang, H., Zhu, W., Zhao, Y., Chen, B., Valenzuela, R.K., Zhang, R., Gallitano, A.L., and Ma, J. (2015). Genetic analysis of SNPs in CACNA1C and ANK3 gene with schizophrenia: A comprehensive meta-analysis. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *168*, 637–648.
20. Bigos, K.L., Mattay, V.S., Callicott, J.H., Straub, R.E., Vakkalanka, R., Kolachana, B., Hyde, T.M., Lipska, B.K., Kleinman, J.E., and Weinberger, D.R. (2010). Genetic variation in CACNA1C affects brain circuitries related to mental illness. Arch. Gen. Psychiatry *67*, 939–945.
21. Casamassima, F., Hay, A.C., Benedetti, A., Lattanzi, L., Cassano, G.B., and Perlis, R.H. (2010). L-type calcium channels and psychiatric disorders: A brief review. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *153B*, 1373–1390.
22. Erk, S., Meyer-Lindenberg, A., Schnell, K., Opitz von Boberfeld, C., Esslinger, C., Kirsch, P., Grimm, O., Arnold, C., Haddad, L., Witt, S.H., et al. (2010). Brain function in carriers of a genome-wide supported bipolar disorder variant. Arch. Gen. Psychiatry *67*, 803–811.
23. Green, E.K., Grozeva, D., Jones, I., Jones, L., Kirov, G., Caesar, S., Gordon-Smith, K., Fraser, C., Forty, L., Russell, E., et al.; Wellcome Trust Case Control Consortium (2010). The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. Mol. Psychiatry *15*, 1016–1022.
24. Liu, Y., Blackwood, D.H., Caesar, S., de Geus, E.J., Farmer, A., Ferreira, M.A., Ferrier, I.N., Fraser, C., Gordon-Smith, K., Green, E.K., et al.; Wellcome Trust Case-Control Consortium (2011). Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. Mol. Psychiatry *16*, 2–4.
25. Hori, H., Yamamoto, N., Fujii, T., Teraishi, T., Sasayama, D., Matsuo, J., Kawamoto, Y., Kinoshita, Y., Ota, M., Hattori, K., et al. (2012). Effects of the CACNA1C risk allele on neurocognition in patients with schizophrenia and healthy individuals. Sci. Rep. *2*, 634.
26. Zhang, Q., Shen, Q., Xu, Z., Chen, M., Cheng, L., Zhai, J., Gu, H., Bao, X., Chen, X., Wang, K., et al. (2012). The effects of CACNA1C gene polymorphism on spatial working memory in both healthy controls and patients with schizophrenia or bipolar disorder. Neuropsychopharmacology *37*, 677–684.
27. He, K., An, Z., Wang, Q., Li, T., Li, Z., Chen, J., Li, W., Wang, T., Ji, J., Feng, G., et al. (2014). CACNA1C, schizophrenia and major depressive disorder in the Han Chinese population. Br. J. Psychiatry *204*, 36–39.
28. Li, J., Zhao, L., You, Y., Lu, T., Jia, M., Yu, H., Ruan, Y., Yue, W., Liu, J., Lu, L., et al. (2015). Schizophrenia related variants in CACNA1C also confer risk of autism. PLoS ONE *10*, e0133247.
29. Guenther, C.A., Tasic, B., Luo, L., Bedell, M.A., and Kingsley, D.M. (2014). A molecular basis for classic blond hair color in Europeans. Nat. Genet. *46*, 748–752.
30. Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med. *373*, 895–907.
31. Roussos, P., Mitchell, A.C., Voloudakis, G., Fullard, J.F., Pothula, V.M., Tsang, J., Stahl, E.A., Georgakopoulos, A., Ruderfer, D.M., Charney, A., et al. (2014). A role for noncoding variation in schizophrenia. Cell Rep. *9*, 1417–1429.
32. Eckart, N., Song, Q., Yang, R., Wang, R., Zhu, H., McCallion, A.S., and Avramopoulos, D. (2016). Functional characterization of schizophrenia-associated variation in CACNA1C. PLoS ONE *11*, e0157086.
33. DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.A., Flynn, H., Adamson, J., et al. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron *72*, 245–256.
34. Gardiner, S.L., van Belzen, M.J., Boogaard, M.W., van Roon-Mom, W.M.C., Rozing, M.P., van Hemert, A.M., Smit, J.H., Beekman, A.T.F., van Grootheest, G., Schoevers, R.A., et al.

(2017). Huntingtin gene repeat size variations affect risk of lifetime depression. Transl. Psychiatry *7*, 1277.
35. Gardiner, S.L., van Belzen, M.J., Boogaard, M.W., van Roon-Mom, W.M.C., Rozing, M.P., van Hemert, A.M., Smit, J.H., Beekman, A.T.F., van Grootheest, G., Schoevers, R.A., et al. (2017). Large normal-range TBP and ATXN7 CAG repeat lengths are associated with increased lifetime risk of depression. Transl. Psychiatry *7*, e1143.
36. Landefeld, C.C., Hodgkinson, C.A., Spagnolo, P.A., Marietta, C.A., Shen, P.-H., Sun, H., Zhou, Z., Lipska, B.K., and Goldman, D. (2018). Effects on gene expression and behavior of untagged short tandem repeats: the case of arginine vasopressin receptor 1a (AVPR1a) and externalizing behaviors. Transl. Psychiatry *8*, 72.
37. Lindblad, K., Savontaus, M.L., Stevanin, G., Holmberg, M., Digre, K., Zander, C., Ehrsson, H., David, G., Benomar, A., Nikoskelainen, E., et al. (1996). An expanded CAG repeat sequence in spinocerebellar ataxia type 7. Genome Res. *6*, 965–971.
38. Renton, A.E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L., et al.; ITALSGEN Consortium (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron *72*, 257–268.
39. The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell *72*, 971–983.
40. Verkerk, A.J., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F.P., et al. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell *65*, 905–914.
41. Hannan, A.J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet. *26*, 59–65.
42. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298.
43. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
44. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.
45. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.
46. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.
47. Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.
48. Osoegawa, K., Mammoser, A.G., Wu, C., Frengen, E., Zeng, C., Catanese, J.J., and de Jong, P.J. (2001). A bacterial artificial chromosome library for sequencing the complete human genome. Genome Res. *11*, 483–496.
49. Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R.V., McKinsey, G.L., Pattabiraman, K., Silberberg, S.N., Blow, M.J., et al. (2013). A high-resolution enhancer atlas of the developing telencephalon. Cell *152*, 895–908.
50. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
51. Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am. J. Hum. Genet. *76*, 8–32.
52. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., and Ren, B. (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. *17*, 2042–2059.
53. Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.
54. Gershon, E.S., Grennan, K., Busnello, J., Badner, J.A., Ovsiew, F., Memon, S., Alliey-Rodriguez, N., Cooper, J., Romanos, B., and 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.
55. Yoshimizu, T., Pan, J.Q., Mungenast, A.E., Madison, J.M., Su, S., Ketterman, J., Ongur, D., McPhie, D., Cohen, B., Perlis, R., and Tsai, L.H. (2015). Functional implications of a psychiatric risk variant within CACNA1C in induced human neurons. Mol. Psychiatry *20*, 162–169.
56. Bulfone, A., Martinez, S., Marigo, V., Campanella, M., Basile, A., Quaderi, N., Gattuso, C., Rubenstein, J.L., and Ballabio, A. (1999). Expression pattern of the Tbr2 (Eomesodermin) gene during mouse and chick brain development. Mech. Dev. *84*, 133–138.
57. Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176.
58. Liu, X., Bates, R., Yin, D.M., Shen, C., Wang, F., Su, N., Kirov, S.A., Luo, Y., Wang, J.Z., Xiong, W.C., and Mei, L. (2011). Specific regulation of NRG1 isoform expression by neuronal activity. J. Neurosci. *31*, 8491–8501.
59. Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., van de Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature *489*, 391–399.
60. Miller, J.A., Ding, S.L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199–206.
61. Dao, D.T., Mahon, P.B., Cai, X., Kovacsics, C.E., Blackwell, R.A., Arad, M., Shi, J., Zandi, P.P., O'Donnell, P., Knowles, J.A., et al.; Bipolar Genome Study (BiGS) Consortium (2010). Mood disorder susceptibility gene CACNA1C modifies mood-related behaviors in mice and interacts with sex to

influence behavior in mice and diagnosis in humans. Biol. Psychiatry *68*, 801–810.
62. Jeon, D., Kim, S., Chetana, M., Jo, D., Ruley, H.E., Lin, S.Y., Rabah, D., Kinet, J.P., and Shin, H.S. (2010). Observational fear learning involves affective pain system and Cav1.2 Ca2+ channels in ACC. Nat. Neurosci. *13*, 482–488.
63. Lee, A.S., Ra, S., Rajadhyaksha, A.M., Britt, J.K., De Jesus-Cortes, H., Gonzales, K.L., Lee, A., Moosmang, S., Hofmann, F., Pieper, A.A., and Rajadhyaksha, A.M. (2012). Forebrain elimination of cacna1c mediates anxiety-like behavior in mice. Mol. Psychiatry *17*, 1054–1055.
64. Splawski, I., Timothy, K.W., Sharpe, L.M., Decher, N., Kumar, P., Bloise, R., Napolitano, C., Schwartz, P.J., Joseph, R.M., Condouris, K., et al. (2004). Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. Cell *119*, 19–31.
65. Splawski, I., Timothy, K.W., Decher, N., Kumar, P., Sachse, F.B., Beggs, A.H., Sanguinetti, M.C., and Keating, M.T. (2005). Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. Proc. Natl. Acad. Sci. USA *102*, 8089–8096, discussion 8086–8088.
66. Zu, T., Gibbens, B., Doty, N.S., Gomes-Pereira, M., Huguet, A., Stone, M.D., Margolis, J., Peterson, M., Markowski, T.W., Ingram, M.A., et al. (2011). Non-ATG-initiated translation directed by microsatellite expansions. Proc. Natl. Acad. Sci. USA *108*, 260–265.
67. Cleary, J.D., and Ranum, L.P. (2013). Repeat-associated non-ATG (RAN) translation in neurological disease. Hum. Mol. Genet. *22* (R1), R45–R51.
68. Bañez-Coronel, M., Ayhan, F., Tarabochia, A.D., Zu, T., Perez, B.A., Tusi, S.K., Pletnikova, O., Borchelt, D.R., Ross, C.A., Margolis, R.L., et al. (2015). RAN translation in Huntington disease. Neuron *88*, 667–677.
69. Nithianantharajah, J., and Hannan, A.J. (2007). Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. BioEssays *29*, 525–535.
70. Pennisi, E. (2012). Evolutionary biology. The burdens of being a biped. Science *336*, 974.
71. Stedman, H.H., Kozyak, B.W., Nelson, A., Thesier, D.M., Su, L.T., Low, D.W., Bridges, C.R., Shrager, J.B., Minugh-Purvis, N., and Mitchell, M.A. (2004). Myosin gene mutation correlates with anatomical changes in the human lineage. Nature *428*, 415–418.
72. Lieberman, D. (2013). The Story of the Human Body: Evolution, Health, and Disease (New York: Pantheon Books).
73. Oksenberg, N., Stevison, L., Wall, J.D., and Ahituv, N. (2013). Function and regulation of AUTS2, a gene implicated in autism and human evolution. PLoS Genet. *9*, e1003221.
74. Srinivasan, S., Bettella, F., Mattingsdal, M., Wang, Y., Witoelar, A., Schork, A.J., Thompson, W.K., Zuber, V., Winsvold, B.S., Zwart, J.A., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, The International Headache Genetics Consortium (2016). Genetic markers of human evolution are enriched in schizophrenia. Biol. Psychiatry *80*, 284–292.
75. Hollister, L.E., and Trevino, E.S. (1999). Calcium channel blockers in psychiatric disorders: a review of the literature. Can. J. Psychiatry *44*, 658–664.
76. Zamponi, G.W. (2016). Targeting voltage-gated calcium channels in neurological and psychiatric diseases. Nat. Rev. Drug Discov. *15*, 19–34.
77. Kronenberg, Z.N., Fiddes, I.T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O.S., Underwood, J.G., Nelson, B.J., Chaisson, M.J.P., Dougherty, M.L., et al. (2018). High-resolution comparative analysis of great ape genomes. Science *360*. https://doi.org/10.1126/science.aar6343.

The American Journal of Human Genetics, Volume *103*

Supplemental Data

Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia

Janet H.T. Song, Craig B. Lowe, and David M. Kingsley

Supplemental Figures

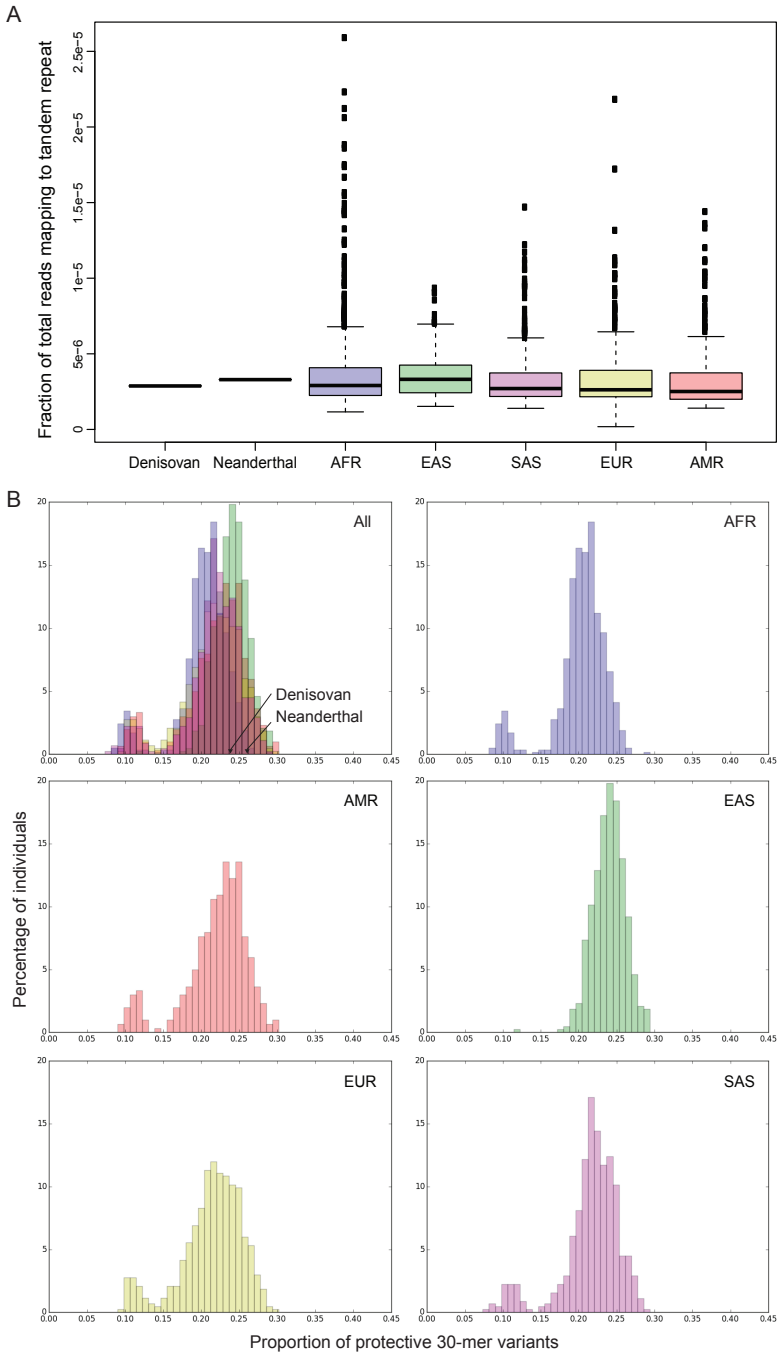


Figure S1: **Comparison of repeat arrays in archaic hominins and modern human populations.** We analyzed both read depth and the type of 30-mer repeat sequences in genomes from one Denisovan,[1] one Neanderthal,[2] and modern humans from the 1000 Genomes Project[3] subdivided by super population code into Africans (AFR), Ad-Mixed Americans (AMR), East Asians (EAS), Europeans (EUR), and South Asians (SAS). For both (A) mean repeat array length and (B) the proportion of 30-mer variants significantly associated with the protective haplotype (see Supplemental Methods), the Denisovan and Neanderthal genomes fall within the range of modern humans. Repeat length and composition vary among modern human populations ($p < 10^{-50}$ for both repeat length and composition by the k-sample Anderson-Darling test). However, this region of *CACNA1C* is not one of the loci that shows strong evidence for positive selection in modern humans.[4–8]

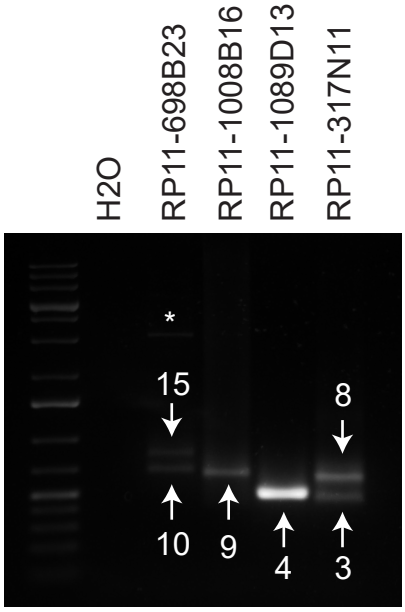


Figure S2: **Bacterial artificial chromosomes (BACs) used in assembling the human reference genome contain highly reduced tandem repeat arrays.** Four BACs made from the same individual have variable copy number at the tandem repeat. The numbers and arrows indicate the number of 30-mer units in each PCR product, as determined by sequencing. The asterisk indicates a non-specific PCR artifact.

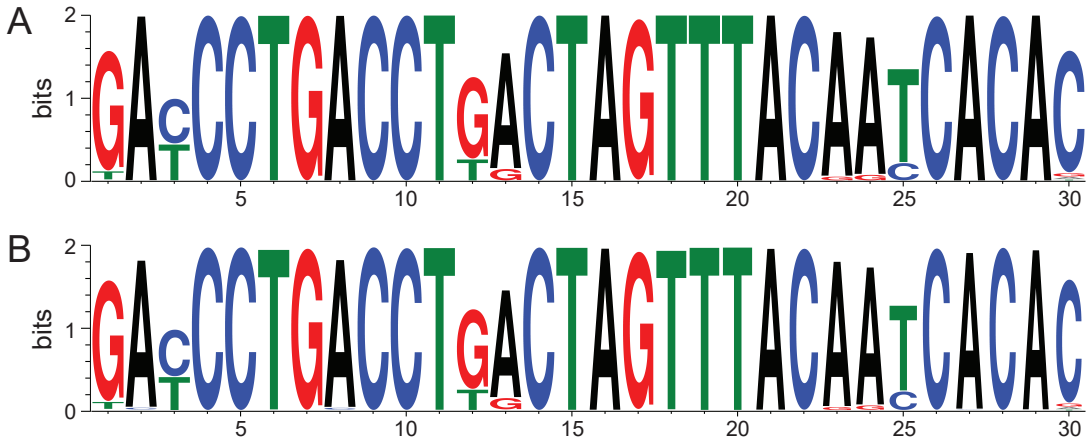


Figure S3: **Motifs of the 30-mer units that comprise the tandem repeat array.** Consensus motifs determined from PacBio-sequenced human repeat arrays (A) or whole genome DNA sequencing reads that map to this region in individuals of European and East Asian descent in the 1000 Genomes Project (B) are very similar. Some positions in the motif are largely invariant, whereas other positions vary from 30-mer unit to 30-mer unit.

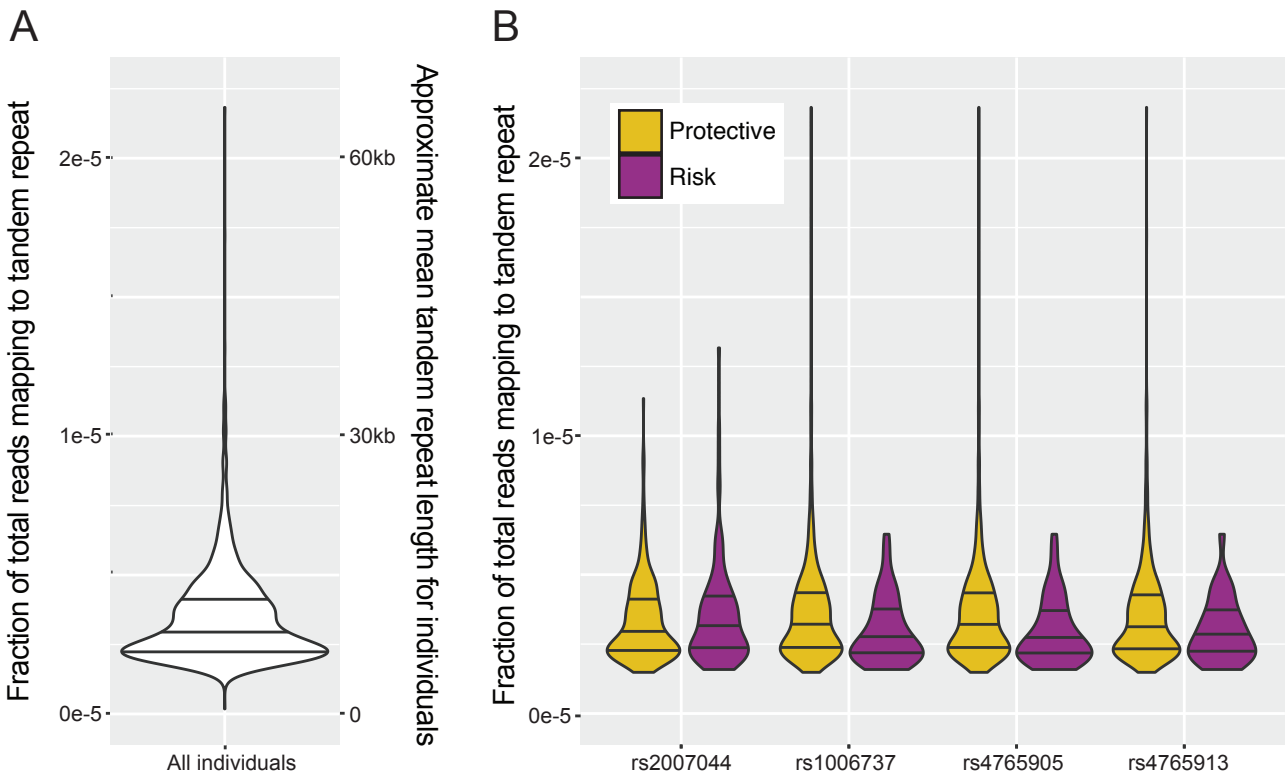


Figure S4: **Inferred repeat array length does not show a simple association with GWAS SNPs.** (A) We infer the mean repeat array length for an individual sequenced as part of the 1000 Genomes Project by calculating the fraction of total reads that map to the repeat region (see Supplemental Methods). (B) To understand if the repeat array length may be correlated with either the risk or protective allele at the four GWAS SNPs, we visualized the distribution of allele sizes (average of two alleles) present in those individuals homozygous for either the risk or protective alleles at that SNP (see Supplemental Methods). After correcting for multiple hypothesis testing, none of the four SNPs had a significant difference between the allele sizes in the risk or protected individuals (Wilcoxon rank-sum test; p-value threshold of 0.01).

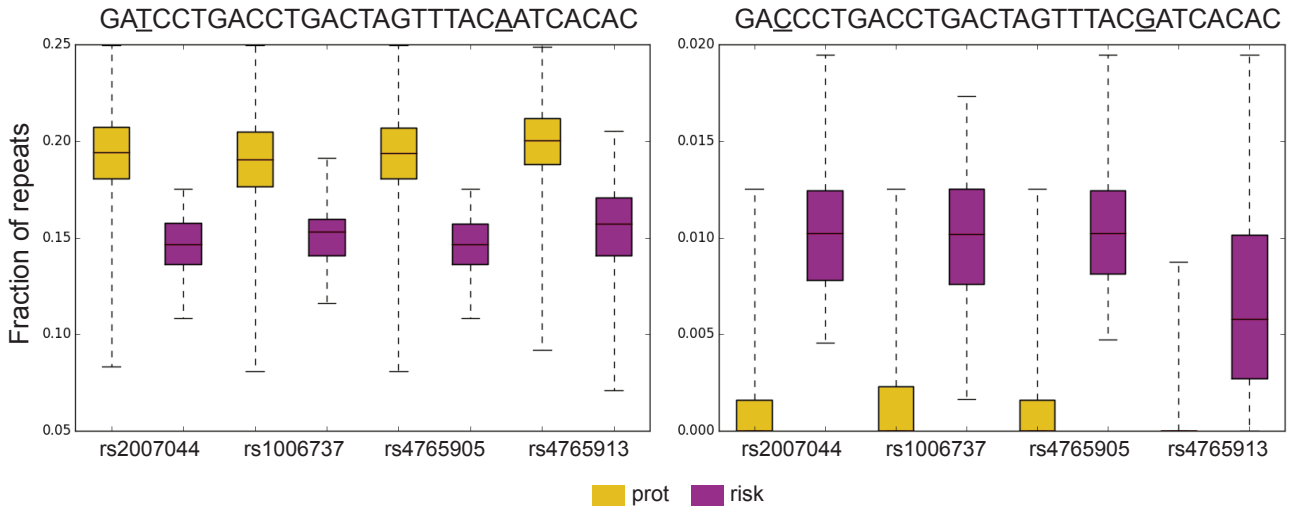


Figure S5: **Particular 30-mer sequence variants are associated with the protective or risk genotype at GWAS SNPs.** For each possible 30-mer repeat sequence, we determined what fraction of 30-mers found in each individual of a given cohort exactly match that particular variant. Two examples of significantly associated 30-mer units are plotted here. The 30-mer variant on the left is significantly associated with the protective genotype at all four GWAS SNPs, whereas the 30-mer variant on the right is significantly associated with the risk genotype. Sequence differences between the two 30-mer units are underlined.

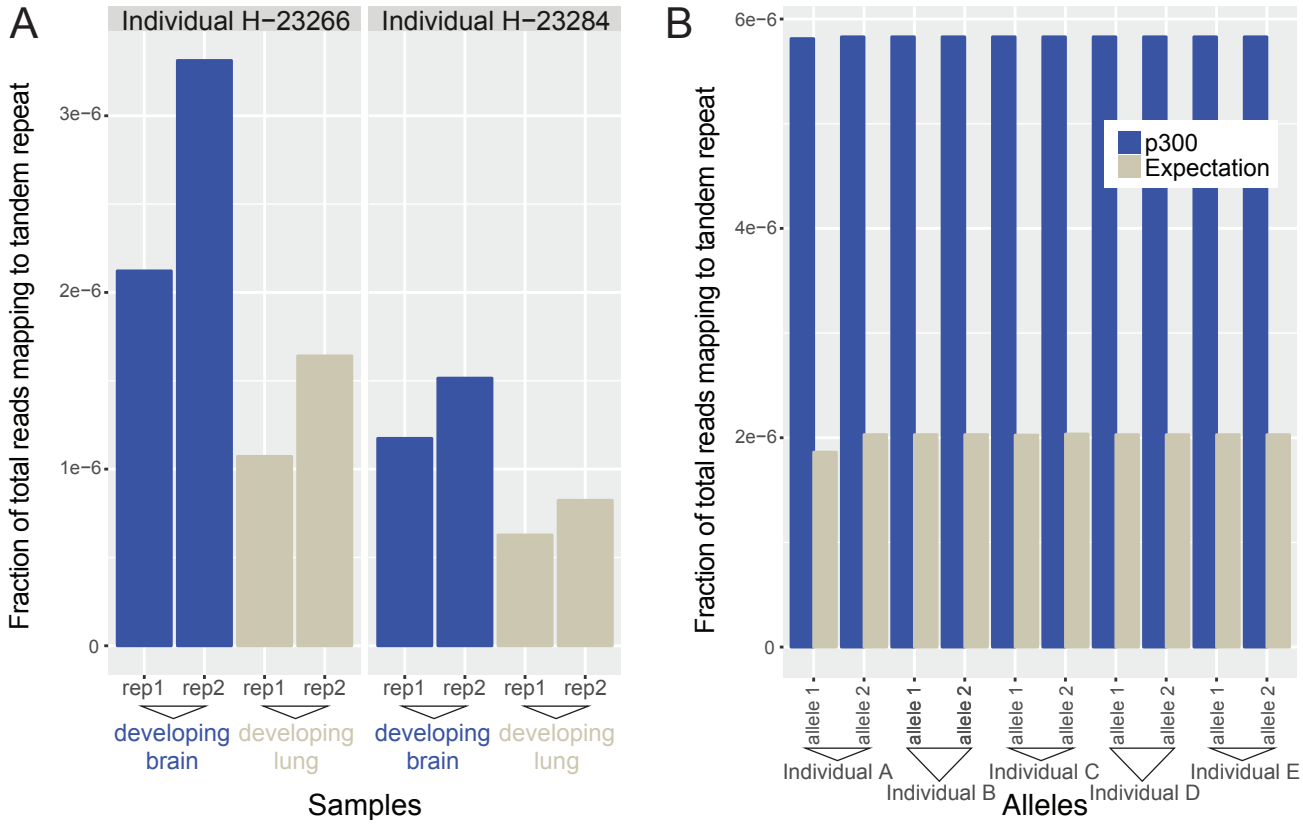


Figure S6: **Signatures of enhancer function in the developing human brain.** Both DNase I hypersensitivity and p300 ChIP-seq experiments rely on creating sequencing libraries enriched for genomic locations where the DNA is either open or associated with p300, respectively. The analysis of whether significant enrichment exists is more straightforward when the reference genome matches the individual sequenced; however, in the case of this repeat region, we expect a large number of reads to map back to this location even with no enrichment from the assay since the repeat expansion is much larger in human tissue than in the human assembly (Fig. 1). (A) There are two individuals from the Roadmap Epigenomics[9] data set where DNase I hypersensitivity experiments were done on both the developing brain (two replicates) and the developing lung (two replicates). While we do not know what the repeat array lengths are for these two individuals, and therefore how to normalize the read depth, we do know that all experiments on the same individual should be normalized to the same degree. For both individuals, the DNase signal is stronger in the two replicates from the developing brain than it is for the two replicates in the developing lung (see Supplemental Methods). These results are consistent with the repeat array being in open chromatin in the developing brain. (B) We do not have other sequencing libraries from the same individual for comparison to the p300 ChIP-seq assay performed on the developing human brain.[10] However, based on SNPs seen in the p300 ChIP-seq reads, this individual appears to most closely match five individuals from which we have long-read sequencing and know the sequence of their repeat arrays (see Supplemental Methods). When we map the p300 data to human assemblies modified to have one of these 10 alleles instead of the one present in the assembly reference, there is still an approximately 3x enrichment for the reads from the p300 data set over what we would expect for these allele sizes. For the p300 enrichment to be solely due to the individual having large repeat alleles, the individual would need to have one or more alleles over 24kb, which is only seen in approximately 4% of alleles (Fig. 1).

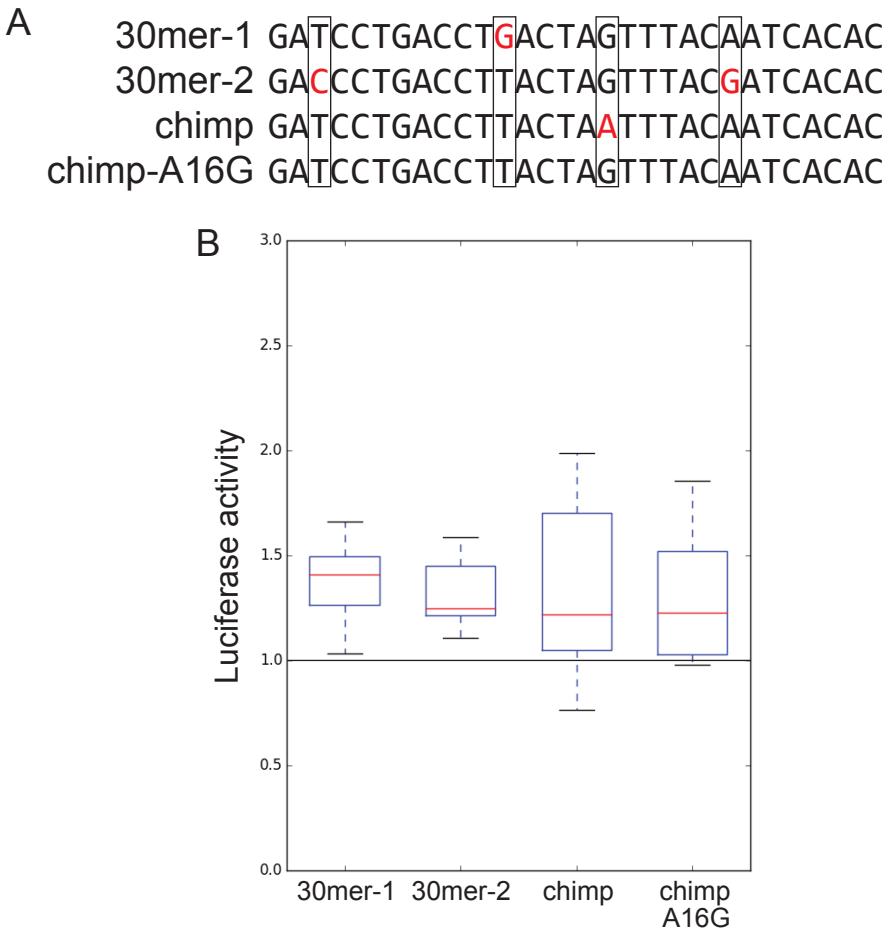


Figure S7: **Enhancer assays with single 30-mer units.** Four different 30-mer units were cloned upstream of a minimal promoter driving the luciferase gene and assayed individually for luciferase activity relative to the empty vector (horizontal line at 1.0) as described in Supplemental Methods. 30mer-1 is a 30-mer significantly associated with the protective haplotype, and 30mer-2 is a 30-mer significantly associated with the risk haplotype. Chimp is the 30-mer unit found in chimpanzees, while chimp-A16G has been engineered to have a G instead of an A at the 16th position. 30mer-1, 30mer-2, and chimp drove mean luciferase activities that were higher than empty vector controls (p = 0.0008, 0.002, and 0.01, respectively by the Wilcoxon rank-sum test, based on n = 8, 7, and 14). Chimp-A16G trended in the same direction but was not statistically significant (p = 0.25 based on n = 4). None were statistically different from the others. Note that the chimp sequence is very rare in humans, chimp-A16G is the seventh most common 30-mer unit in humans, and neither is significantly associated with either the protective or risk haplotype in humans.

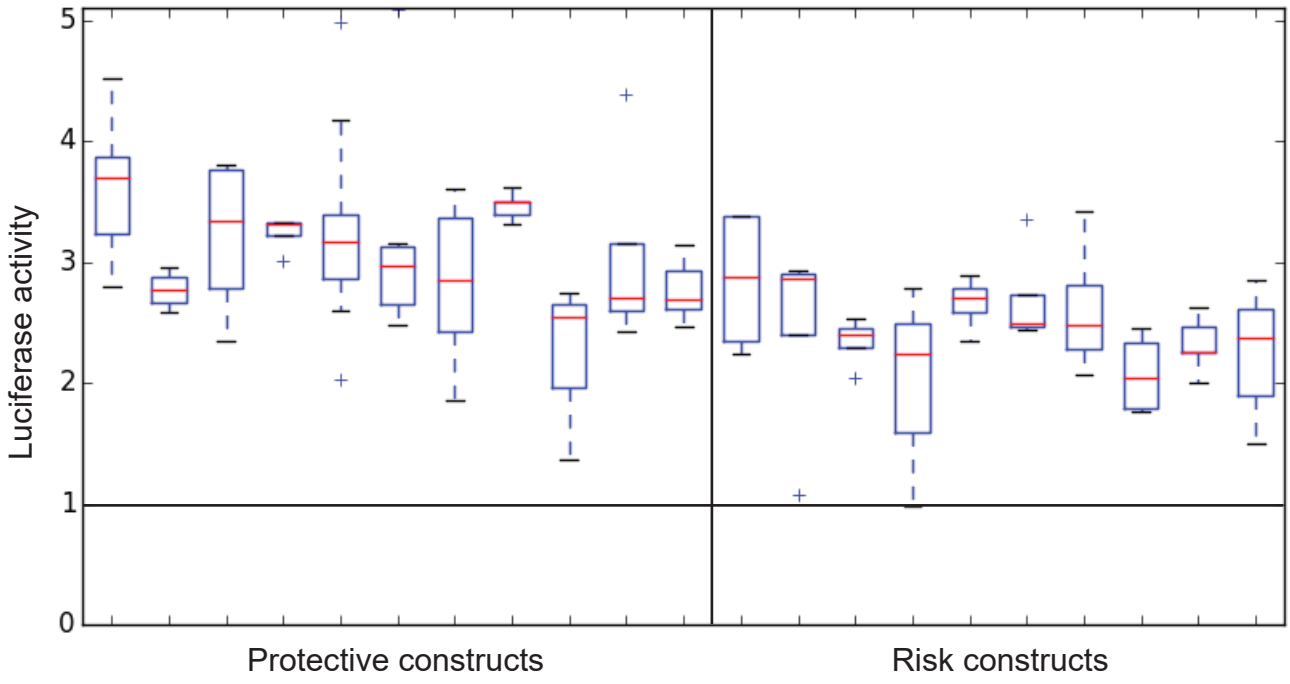


Figure S8: **Protective human repeat arrays drive higher luciferase activity than risk human repeat arrays.** 11 human repeat arrays characteristic of the protective haplotype and 10 repeat arrays characteristic of the risk haplotype were cloned upstream of a minimal promoter driving the luciferase gene. These constructs were then assayed for luciferase activity, as described in Supplemental Methods. Protective arrays drove significantly higher luciferase activity than risk arrays ($p = 0.001$, Wilcoxon rank-sum test).

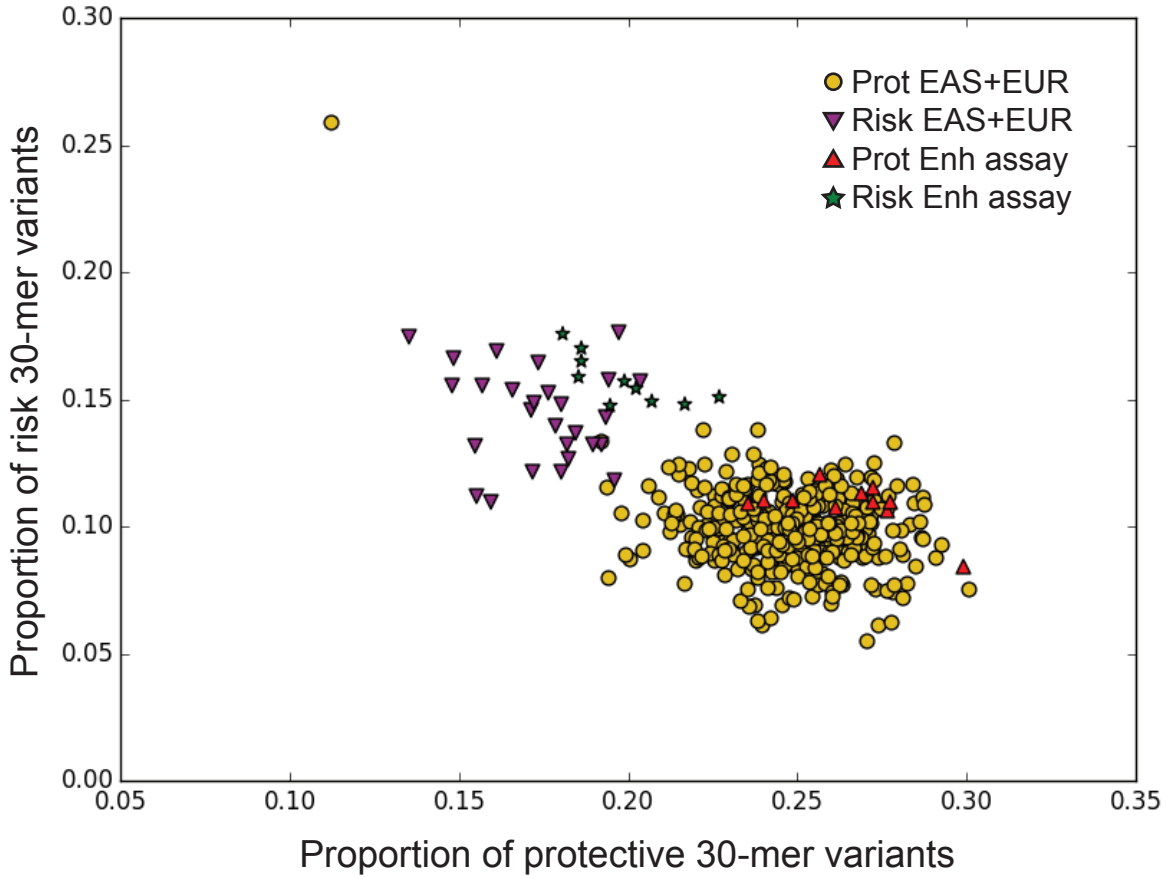


Figure S9: **Repeat arrays tested in enhancer assays cluster based on their proportion of protective- and risk-associated 30-mer variants.** The proportion of 30-mers that exactly match a 30-mer variant significantly associated with the protective haplotype or the risk haplotype are plotted for East Asian and European individuals in the 1000 Genomes Project who are homozygous protective at all four GWAS SNPs (Prot EAS+EUR, yellow) or homozygous risk at all four GWAS SNPs (Risk EAS+EUR, purple). Since these two groups of individuals were themselves used to identify the 30-mer variants plotted here, they separate as expected. PacBio-sequenced repeat arrays that were tested in the enhancer assay cluster either with Prot EAS+EUR (Prot Enh assay, red) or Risk EAS+EUR (Risk Enh assay, green).

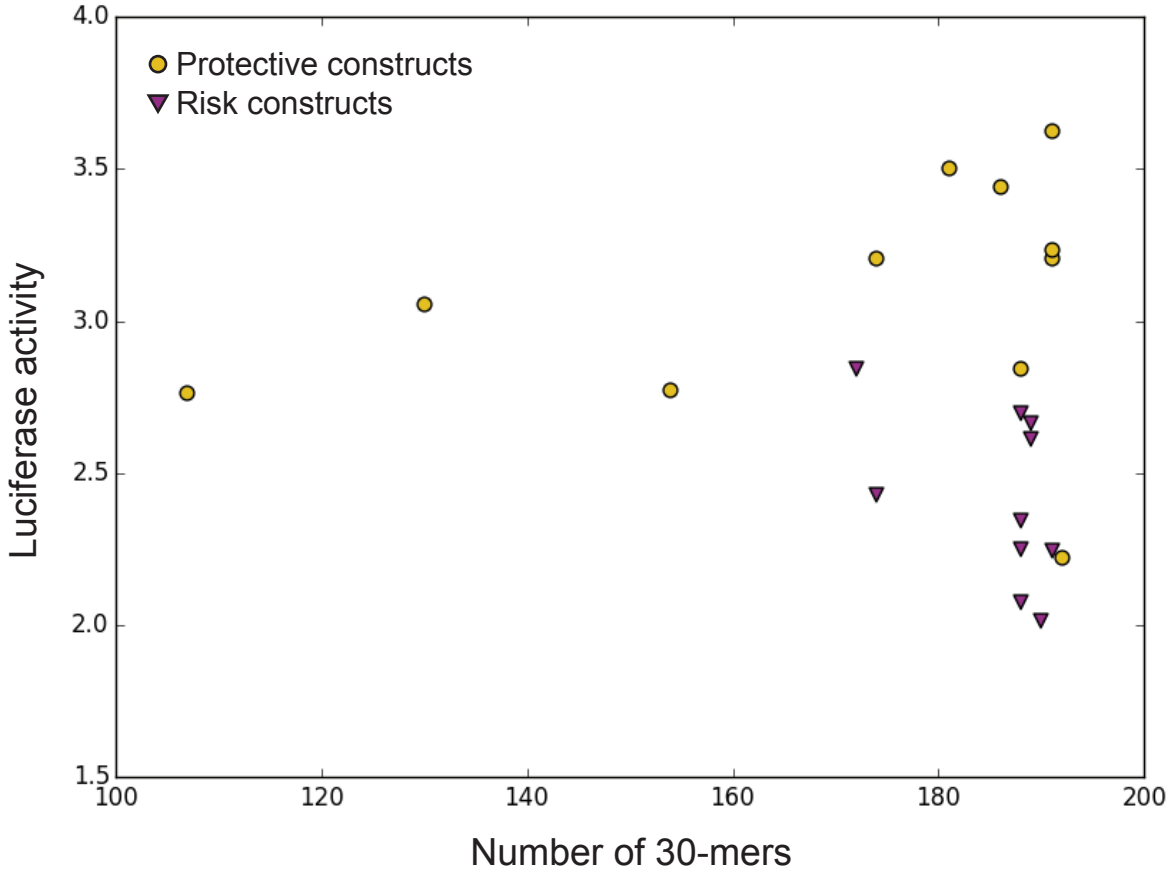


Figure S10: **Enhancer activity is not associated with human repeat array length.** 21 human repeat arrays were cloned upstream of a minimal promoter driving the luciferase gene and assayed for luciferase activity, as described in Supplemental Methods. These arrays contained between 107 and 192 30-mers. The number of 30-mers in a repeat array was not significantly associated with luciferase activity in this length range ($R = -0.21$, $p = 0.37$ using the Spearman correlation).

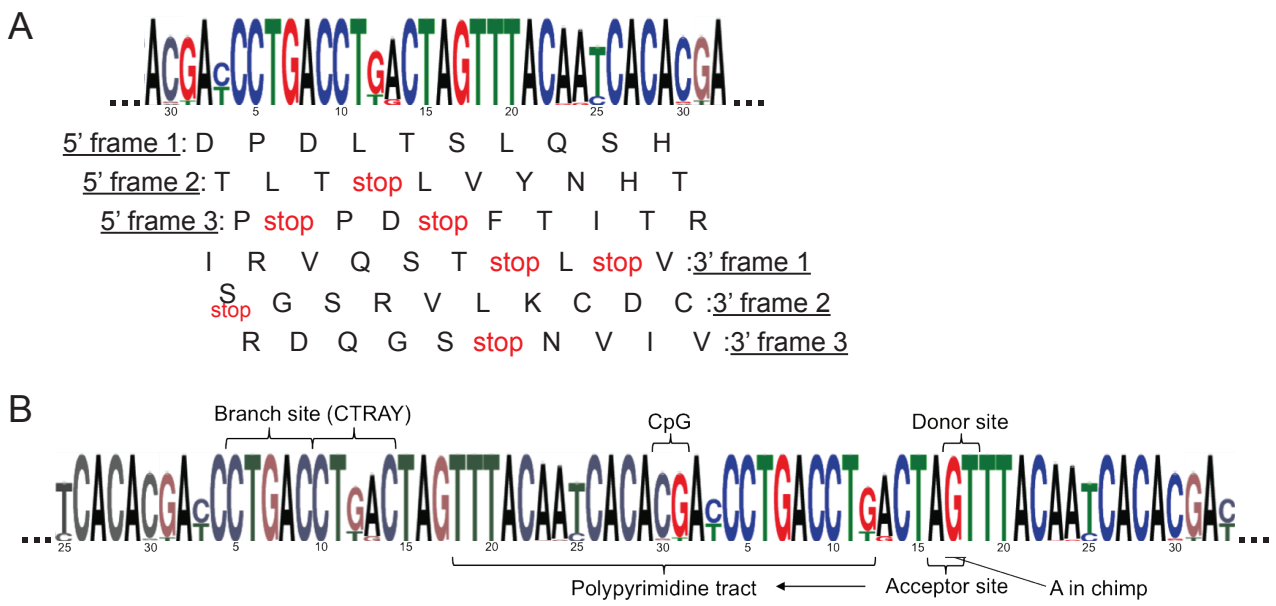


Figure S11: **30-mer repeat sequences contain open reading frames and putative CpG methylation and splicing sites.** (A) Predicted amino acid sequences are shown for each potential reading frame in the 5 and 3 direction. The first reading frame in the 5' direction is open in most 30-mer sequence variants. The second reading frame in the 3' direction is also open in the most common 30-mer sequence variant; however, all of the PacBio-sequenced repeat arrays also contain multiple 30-mer variants that have a stop codon in this frame. (B) A tandem doublet of 30-mer repeat units contains a putative CpG site at the junction between repeats, as well as canonical splicing sequences, including a putative donor site, acceptor site, polypyrimidine tract, and branch sites. The single 30-mer repeat found in chimpanzees has an A at position 17, which removes both the putative donor and acceptor sites.

Supplemental Tables

Table S1: Transcription factors with motif that matches part of 30-mer

Transcription Factor	Motif	Position	Protective Count	Risk Count	Difference		Transcription Factor	Motif	Position	Protective Count	Risk Count	Difference
Eomes	GATCACAC	23	0	2	0.4		Cdx1	TTTACAAC	18	1	1	0
Nrg1	GACCCTGA	1	2	4	0.4		Cdx2	TTTACGAC	18	1	1	0
PmTbr	GATCACAC	23	0	2	0.4		Cdx2	TTTACAAC	18	1	1	0
Rxrg	TGACCTTA	6	2	0	0.4		E4f1	TTTACGAC	18	1	1	0
SpTbr	GATCACAC	23	0	2	0.4		Ecm22	TTTACGAC	18	1	1	0
Tbr1	GATCACAC	23	0	2	0.4		Esrra	TGACCTTA	6	2	2	0
Abd-B	TTTACGAT	18	1	2	0.2		Esrrb	TGACCTTA	6	2	2	0
BCL11A	TCCTGACC	3	2	1	0.2		Gli1	GACCACAC	23	1	1	0
BCL11A	CTGACCTT	5	1	2	0.2		Gli2	GACCACAC	23	1	1	0
BCL11B	TCCTGACC	3	2	1	0.2		Gli3	GACCACAC	23	1	1	0
BCL11B	CTGACCTT	5	1	2	0.2		HLH-25	ACCACACG	24	2	2	0
Bhlhb2	TCACACGA	25	1	2	0.2		Hmbox1	TTACTAGT	11	2	2	0
Cdx1	TTTACGAT	18	1	2	0.2		Hmbox1	TGACTAGT	11	3	3	0
Cdx2	TTTACGAT	18	1	2	0.2		Hnf4a	TGACCTTA	6	2	2	0
Cdx2	TTTACAAT	18	2	1	0.2		Hoxa10	TTTACGAC	18	1	1	0
Cphx	CAATCACA	22	2	1	0.2		Hoxa11	TTTACGAC	18	1	1	0
Dux1	TACAATCA	20	2	1	0.2		Hoxa13	TTTACGAC	18	1	1	0
Ecm22	GTTTACGA	17	1	2	0.2		Hoxa9	TTTACGAC	18	1	1	0
Eomes	AATCACAC	23	1	0	0.2		Hoxb13	TTTACGAC	18	1	1	0
Esrra	TGACCCTG	30	1	0	0.2		Hoxb9	TTTACGAC	18	1	1	0
Esrra	CCTGACCT	4	2	3	0.2		Hoxc10	TTTACGAC	18	1	1	0
GFI1	CAATCACA	22	2	1	0.2		Hoxc11	TTTACGAC	18	1	1	0
GFI1B	AATCACAG	23	1	0	0.2		Hoxc12	TTTACGAC	18	1	1	0
Hdx	TACGATCA	20	1	2	0.2		Hoxc12	TTTACAAC	18	1	1	0
Hdx	TACAATCA	20	2	1	0.2		Hoxc13	TTTACGAC	18	1	1	0
HLH-1	CACATGAC	26	1	0	0.2		Hoxc9	TTTACGAC	18	1	1	0
Hnf4a	TGACCCTG	30	1	0	0.2		Hoxd10	TTTACGAC	18	1	1	0
Hoxa10	TTTACGAT	18	1	2	0.2		Hoxd11	TTTACGAC	18	1	1	0
Hoxa11	TTTACGAT	18	1	2	0.2		Hoxd12	TTTACGAC	18	1	1	0
Hoxa13	TTTACGAT	18	1	2	0.2		Hoxd12	TTTACAAC	18	1	1	0
Hoxa13	GTTTACAA	17	3	2	0.2		Hoxd13	TTTACGAC	18	1	1	0
Hoxa9	TTTACGAT	18	1	2	0.2		HOXD13	TTTACGAC	18	1	1	0
Hoxb13	TTTACGAT	18	1	2	0.2		HOXD13	TTTACAAC	18	1	1	0
Hoxb9	TTTACGAT	18	1	2	0.2		Lhx6	CAATCACA	22	2	2	0
Hoxc10	TTTACGAT	18	1	2	0.2		Max	ATCACATG	24	1	1	0
Hoxc11	TTTACGAT	18	1	2	0.2		Mlx	TCACATGA	25	1	1	0
Hoxc12	TTTACGAT	18	1	2	0.2		NR1H4	TGACCTTA	6	2	2	0
Hoxc13	TTTACGAT	18	1	2	0.2		NR1H4	TGACCTGA	6	3	3	0
Hoxc13	TTTACAAT	18	2	1	0.2		Nr2e1	CTGACCTT	5	2	2	0
Hoxd11	TTTACGAT	18	1	2	0.2		Nr2e1	CCTGACCT	4	3	3	0
Hoxd12	TTTACGAT	18	1	2	0.2		Nr2f1	TGACCTTA	6	2	2	0
Hoxd13	TTTACGAT	18	1	2	0.2		Nr2f1	TGACCTGA	6	3	3	0
HOXD13	TTTACGAT	18	1	2	0.2		Nr2f2	TGACCTTA	6	2	2	0
HOXD13	TTTACAAT	18	2	1	0.2		Nr2f2	TGACCTGA	6	3	3	0
Irx5	ACATGATC	27	0	1	0.2		Nr2f6	TGACCTTA	6	2	2	0
Lhx8	TACAATCA	20	2	1	0.2		Nr2f6	TGACCTGA	6	3	3	0
Nrg1	TGACCCTG	30	1	0	0.2		Nr5a1	TGACCTTA	6	2	2	0
PmTbr	AATCACAC	23	1	0	0.2		NSY-7	TGACCTTA	6	2	2	0
Rxra	TGACCCTG	30	1	0	0.2		PF14_79	TACAACCA	20	1	1	0
Rxrb	TGACCCTG	30	1	0	0.2		Rara	TGACCTTA	6	2	2	0
Rxrb	CCTGACCT	4	2	3	0.2		Rara	TGACCTGA	6	3	3	0
Rxrg	TGACCCTG	30	1	2	0.2		Rarb	TGACCTTA	6	2	2	0
Spdef	GGATCCTG	30	1	0	0.2		Rarb	TGACCTGA	6	3	3	0
SpTbr	AATCACAC	23	1	0	0.2		Rarg	TGACCTTA	6	2	2	0
Tbf1	ACCCTGAC	2	3	4	0.2		Rarg	TGACCTGA	6	3	3	0
Tbr1	AATCACAC	23	1	0	0.2		Rxra	TGACCTTA	6	2	2	0
Upc2	GTTTACGA	17	1	2	0.2		Rxrb	TGACCTTA	6	2	2	0
Usv1	GACCCTGA	1	3	4	0.2		Tcf2	TTACTAGT	11	2	2	0
Abd-B	TTTACGAC	18	1	1	0		Tcfec	TCACATGA	25	1	1	0
BCL11A	CCCTGACC	3	2	2	0		Tye7	ATCACATG	24	1	1	0
BCL11B	CCCTGACC	3	2	2	0		Upc2	TTTACGAC	18	1	1	0
Bhlhb2	TCACATGA	25	1	1	0		Upc2	TTTACGAC	18	1	1	0
Cbf1	TCACATGA	25	1	1	0		Vhr2	TGACTAGT	11	3	3	0
Cdx1	TTTACGAC	18	1	1	0							

Protective count: number of the five protective 30-mer variants with motif
Risk count: number of the five risk 30-mer variants with motif
Difference: absolute value of $\frac{protective}{5} - \frac{risk}{5}$

Supplemental Methods

Analysis of 1000 Genomes Project Data

We analyzed individuals sequenced as part of the 1000 Genomes Project.[3] For each individual we remotely accessed the read mappings already performed by the consortium (ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/*/*/alignment/*.cram) and extracted reads that overlapped either the repeat region directly (hg38; chr12:2255791-2256090) or the decoy region that holds a similar sequence (chrUn_KN707670v1_decoy). We counted the number of unique reads, using the read name as a unique identifier, and ensured that the alignment was not marked as a secondary placement. We used the "*.bas" summary files provided by the consortium for each individual to extract the total number of mapped reads without exhaustively counting the reads. We also examined whole-genome DNA sequencing of one Denisovan and one Neanderthal (http://cdna.eva.mpg.de/neandertal/altai/).[1,2]

For each of the four GWAS SNPs (rs2007044, rs1006737, rs4765905, and rs4765913), we used the phase three integrated genotype calls (v5a.20130502) to extract identifiers for individuals who are homozygous for the risk or protective alleles at the given GWAS SNP. To test for associations between the repeat array and the GWAS SNPs, we only considered individuals in the 1000 genomes of either European or East Asian ancestry (population codes: CEU, TSI, FIN, GBR, IBS, CDX, CHB, CHS, JPT, and KHV). We performed the analysis both with and without the small number of related individuals in the 1000 Genomes Project, and both produced similar results.

To infer the mean allele length, we assume that a random read is equally likely to begin at every base in the genome. Therefore, the fraction of reads that overlap the repeat region should be approximately equal to the fraction of the genome that is the repeat region. For a read length of X and a region length of Y, there are $(X-1)+Y$ coordinates where a read could have its left-most coordinate and overlap the region by at least one base position. The total coordinates where a read could begin is approximately equal to the size of the genome assembly. This allows us to infer the repeat length using the equation:

$$\frac{overlapReads}{totalReads} = \frac{(readLen-1)+regionLen}{genomeLen} \quad \text{(S1)}$$

$$regionLen = \frac{overlapReads}{totalReads} * genomeLen - readLen + 1 \quad \text{(S2)}$$

To test for a significant association between inferred length and SNP genotype, we used the Wilcoxon rank-sum test with a Bonferroni correction for the four tests done (one for each SNP). No association between inferred length and SNP genotype was significant using an adjusted p-value threshold of 0.01.

To test for associations between sequence variants of the 30-mer unit and genotypes at the four GWAS SNPs, we considered all 30-mer sequence variants that were found at least 500 times among all 2688 individuals in the 1000 Genomes Project (> 0.2 average times per individual). There are 292 30-mer sequence variants that fit this criterion. For each European or East Asian individual, we then calculated the fraction of that individual's 30-mers that exactly match each of the 292 30-mer sequence variants being analyzed. To test for an association at each of the four GWAS SNPs, we use the Wilcoxon rank-sum test to compare the prevalence of a 30-mer sequence variant between those individuals homozygous for the risk allele and those homozygous for the protective allele. We use a significance threshold of 0.01 after correcting for the 1168 tests performed. At this threshold, 16 sequence variants are associated with the genotype at one or more GWAS SNPs and 6 sequence variants show a consistent association at all four GWAS SNPs. We performed a second association test where we considered only individuals that are homozygous protective or risk at all four GWAS SNPs (designated as the protective or risk haplotype). Statistical significance between the protective and risk groups was again assessed with the Wilcoxon rank-sum test with a Bonferroni correction for 292 tests and a p-value threshold of 0.01. There are 10 sequence variants that strongly associate with the protective or risk haplotype at this locus (Fig. 2). Plots are standard box-and-whisker plots where the box represents the lower quartile, median, and upper quartile, and the whiskers represent the range of the measurements. Outliers (+) are data points that are outside the nearest quartile + 1.5x the interquartile range.

We also performed the same analysis for Europeans and East Asians separately. For Europeans, 13 sequence variants are associated with the genotype at one or more GWAS SNPs and 7 sequence variants show a consistent association at all four GWAS SNPs. 10/13 and 6/7 sequence variants are also associated when Europeans and East Asians are considered together. When we only consider Europeans with the protective or risk haplotype, 10 sequence variants are significantly associated, 9 of which are also associated when Europeans and East Asians are considered together. For East Asians, 9 sequence variants are associated with the genotype at rs2007044; 8/9 are associated with one or more GWAS SNPs when Europeans and East Asians are considered together. There are no associations with the other SNPs, most likely due a lack of power arising from the low number of East Asians that are homozygous risk at those SNPs in the 1000 Genomes Project.

DNase I hypersensitivity and p300 ChIP-Seq datasets

The Roadmap Epigenomics Consortium has produced a large number of chromatin-related assays on primary human tissue.[9] We analyzed the DNase I hypersensitivity data from individuals H-23266 and H-23284 because assays were performed on both developing brain tissue and developing lung tissue from the same individual. For both individuals, there are two replicates performed on the developing brain and two replicates for the developing lung. We downloaded the location of mapped reads in the hg19 genome assembly and calculated the fraction of reads that overlapped the tandem repeat of interest (hg19; chr12:2364957-2365256), relative to the total number of mapped reads for that experiment. The assays consistently give a stronger signal in the developing brain compared to the developing lung within the same individual (Fig. S6).

Another research group previously performed a p300 ChIP-seq experiment on developing human brain tissue.[10] We down-

loaded the raw reads (SRR630871) and re-mapped them to the hg19 genome assembly using BWA [11] since the previous analysis had filtered many repetitive regions of the genome. Reads from the p300 ChIP-seq experiment overlap rs1006737 and rs4765905 and all four reads report the risk allele. We have sequenced 10 alleles from five individuals that share this genotype. These 10 alleles are all in the most common size range of approximately 6 kb (Fig. 1). For each of the 10 alleles we created a modified genome assembly where we replaced the repeat array present in the hg19 assembly with the array we had sequenced using long-read technology. When we mapped the original p300 ChIP-seq reads to these modified genome assemblies, there was still a 3x enrichment of reads over this tandem repeat (Fig. S6).

Bacterial artificial chromosome (BAC) analysis

We performed PCR on the BACs RP11-698B23, RP11-1008B16, RP11-1089D13, and RP11-317N11 with primers 5'-AGGAGGTGGTGGCTACAGAT-3' and 5'-CCATCCCTGAGTTGTGTGCA-3' (Fig. 1). These BACs are from the RPCI human BAC library 11. [12]

Southern blot of repeat array length in healthy individuals

DNA from presumed healthy individuals were obtained from Coriell and from the NIH Neurobiobank. 5-10 μg of human DNA was digested with the restriction enzyme BlpI (New England Biolabs), run on a 0.5% agarose gel, and transferred using the TurboBlotter Kit (GE Healthcare Life Sciences). Following cross-linking, the membrane was pre-hybridized for 6+ hours at 60C in QuikHyb Hybridization Solution (Agilent) supplemented with 1 mg of denatured UltraPure Salmon Sperm DNA Solution (ThermoFisher). The membrane was then hybridized overnight at 60C with radio-labeled probe made using the Random Primers DNA Labeling System (Invitrogen) from a DNA template with 10 30-mer repeats and ≈500 bp of flanking unique sequence (primer set: 5'-AGGAAAGCACCATTCCCCAG-3' and 5'-CCATCCCTGAGTTGTGTGCA-3'). The next day, the membrane was washed twice in 2X SSC at room temperature and then 3 times for 30 minutes each in 2X SSC, 1% SDS at 60C before exposure and subsequent imaging. BlpI restriction enzyme sites were sequenced for a subset of samples, including all samples with alleles > 20 kb, to ensure that they were intact.

PacBio sequencing of repeat arrays

The repeat array was amplified using primers 5'-TGGCCCTACGGATATCACAT-3' and 5'-TGAGTTGTGTGCAAGTGGC-3' with barcoded tags. The PCR was performed using LA Taq DNA Polymerase (ClonTech) and the Perfect Match PCR Enhancer (Agilent) using the Mg2+ plus buffer provided by ClonTech, dNTPs at a final concentration of 400 μM, an annealing temperature of 56C for 30 seconds, and an extension temperature of 68C for 4 minutes. The expected size of the PCR product was selected with BluePippin (Sage Science). We prepared libraries following the protocol "Preparing Amplicon Libraries using PacBio Barcoded Adapters for Multiplex SMRT Sequencing" (https://www.pacb.com/wp-content/uploads/2015/09/Procedure-Checklist-Preparing-Amplicon-Libraries-using-PacBio-Barcoded-Adapters-for-Multiplex-SMRT-Sequencing.pdf) and sequenced on the PacBio RS II. Data were analyzed using the Long Amplicon Analysis protocol in a SMRT Portal on Amazon Web Services. Identified alleles have at least 30 supporting reads.

Enhancer Assays

We cloned the minimal promoter and the luc2 luciferase gene from pGL4.23 (Promega) using primers 5'-CAAGCTTAGACACTAGAGGGTATATAATGGA-3' and 5'-GGATCCTTATCGATTTTACCACATTT-3' into the linear pJAZZ vector (Lucigen). We then amplified the repeat array from human DNA using the primers and conditions described above. The repeat array was then cloned upstream of the minimal promoter. Proper insertion of the repeat arrays was confirmed by restriction digests and Sanger sequencing.

400 ng of each construct was transfected with 10 ng of pRL-TK (Promega) into a human neural progenitor cell line, ReNcell Cx, (maintained as per vendors instructions, Millipore) using the 96-well shuttle system nucleofector with solution P3 and program 96-DC-104 (Lonza). 48 hours after transfection, cells were assayed for luciferase activity using the Dual-Luciferase Reporter Assay System (Promega) and run on a GloMax-Multi+ Detection System (Promega). Four replicate transfections were performed per construct for each experiment. Mean background readings from untransfected wells were subtracted from all measurements. Luciferase measurements from the pGL4.23 luc2 gene were normalized to measurements of Rluc from pRL-TK. Each construct was tested in a minimum of four different experiments. The normalized measurements for each construct are plotted in Fig. S6, and the means for each construct are plotted in Fig. 2B. All plots are standard box-and-whisker plots where the box represents the lower quartile, median, and upper quartile, and the whiskers represent the range of the measurements. Outliers (+) are data points that are outside the nearest quartile + 1.5x the interquartile range. Statistical significance was assessed with the Wilcoxon rank-sum test.

Repeat arrays were classified as protective or risk as follows: If the individual from which the repeat array was cloned has the protective haplotype (homozygous protective at all four GWAS SNPs), repeat arrays derived from that individual were considered protective. Likewise, if the individual has the risk haplotype, repeat arrays from that individual were considered risk. If an individual was heterozygous at the GWAS SNPs, we then determined the proportion of protective-associated and risk-associated 30-mer variants (variants listed in Fig. 2C) for each repeat array and asked whether its proportion of 30-mer variants was more similar to the 30-mer composition of individuals from the 1000 Genomes Project with the protective haplotype or risk haplotype (Fig. S9). There was never any discrepancy between an individuals SNPs and the designation of the repeat array as protective or risk. For instance, if an individual

was heterozygous at these SNPs, one tandem repeat allele always grouped with the 30-mer composition of individuals with the protective haplotype, and one tandem repeat allele always grouped with individuals with the risk haplotype.

Transcription Factor Motifs

We used the motifs generated from universal protein binding microarrays in the Uniprobe repository.[13,14] We searched this repository against each 30-mer significantly associated with the protective or risk haplotype using a score threshold of 0.45, and counted the number of protective-associated 30-mers and risk-associated 30-mers for each identified motif (Table S1). We included motifs from all available species because transcription factor motifs tend to be highly conserved between species.[15]

Supplemental References

[1] Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prufer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.

[2] Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

[3] 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

[4] Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.

[5] Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

[6] Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

[7] Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D. J., Griesemer, D., Karlsson, E. K., Wong, S. H., et al. (2013). Identifying recent adaptations in large-scale genomic data. Cell *152*, 703–713.

[8] Li, M. J., Wang, L. Y., Xia, Z., Wong, M. P., Sham, P. C., and Wang, J. (2014). dbPSHP: a database of recent positive selection across human populations. Nucleic Acids Res. *42*, D910–916.

[9] Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

[10] Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R. V., McKinsey, G. L., Pattabiraman, K., Silberberg, S. N., Blow, M. J., et al. (2013). A high-resolution enhancer atlas of the developing telencephalon. Cell *152*, 895–908.

[11] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

[12] Osoegawa, K., Mammoser, A. G., Wu, C., Frengen, E., Zeng, C., Catanese, J. J., and de Jong, P. J. (2001). A bacterial artificial chromosome library for sequencing the complete human genome. Genome Res. *11*, 483–496.

[13] Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

[14] Barrera, L. A., Vedenko, A., Kurland, J. V., Rogers, J. M., Gisselbrecht, S. S., Rossin, E. J., Woodard, J., Mariani, L., Kock, K. H., Inukai, S., et al. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science *351*, 1450–1454.

[15] Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E. E., et al. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife *4*, e04837.