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

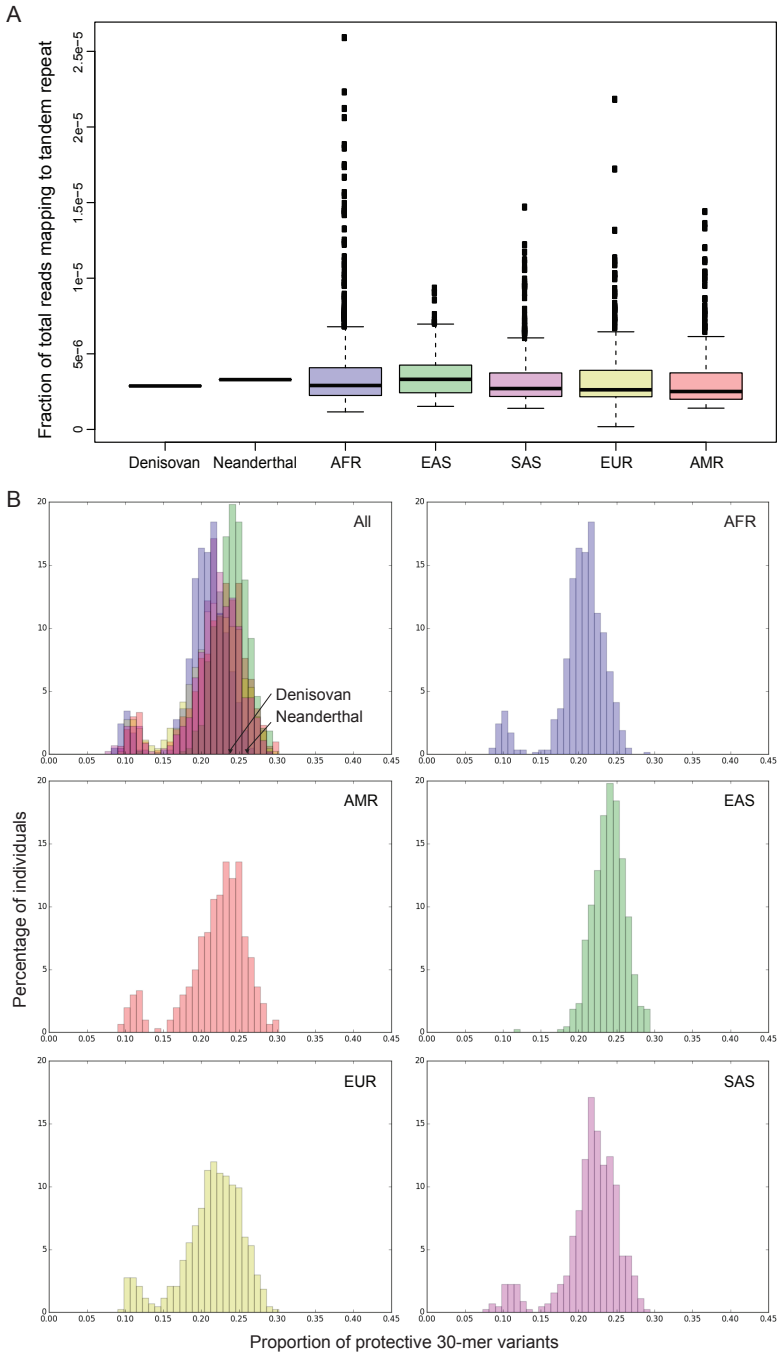

Figure S1: **Comparison of repeat arrays in archaic hominins and modern human populations.** We analyzed both read depth and the type of 30-mer repeat sequences in genomes from one Denisovan,[1] one Neanderthal,[2] and modern humans from the 1000 Genomes Project[3] subdivided by super population code 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]

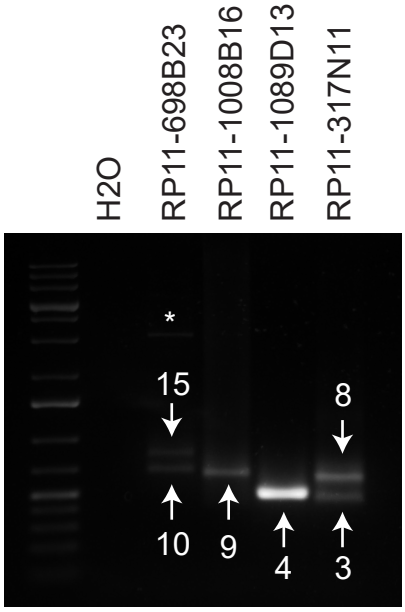

Figure S2: **Bacterial artificial chromosomes (BACs) used in assembling the human reference genome contain highly reduced tandem repeat arrays.** Four BACs made from the same individual have variable copy number at the tandem repeat. The numbers and arrows indicate the number of 30-mer units in each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

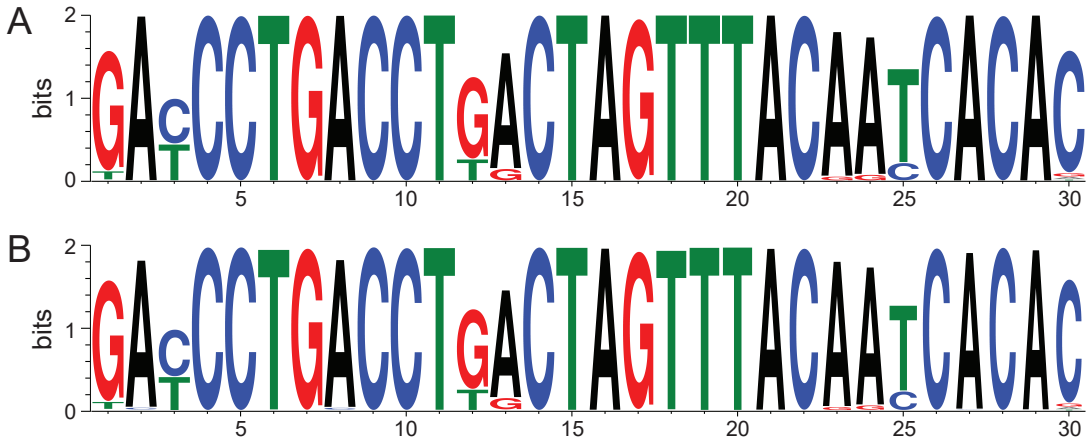

Figure S3: **Motifs of the 30-mer units that comprise the tandem repeat array.** Consensus motifs determined from PacBio-sequenced human repeat arrays (A) or whole genome DNA sequencing reads that map to this region in individuals of European and East Asian descent in the 1000 Genomes Project (B) are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

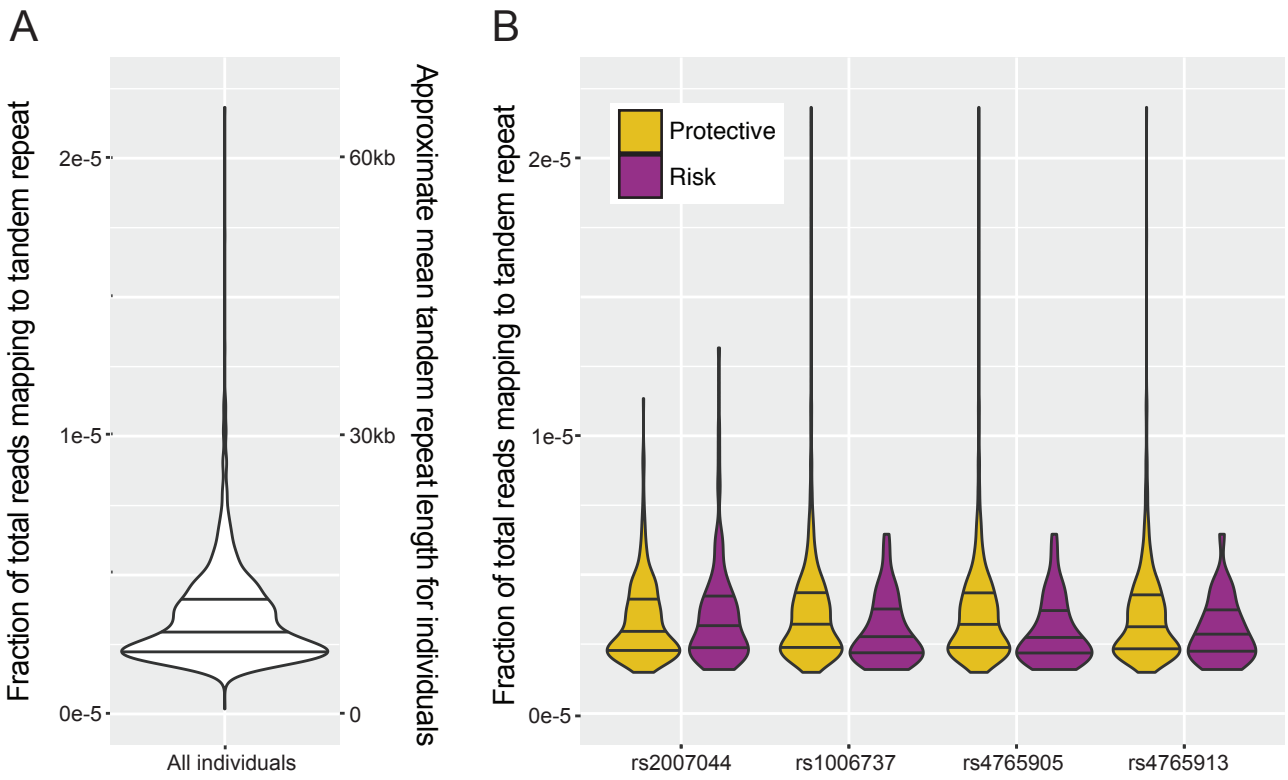

Figure S4: **Inferred repeat array length does not show a simple association with GWAS SNPs.** (A) We infer the mean repeat array length for an individual sequenced as part of the 1000 Genomes Project by calculating the fraction of total reads that map to the repeat region (see Supplemental Methods). (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).

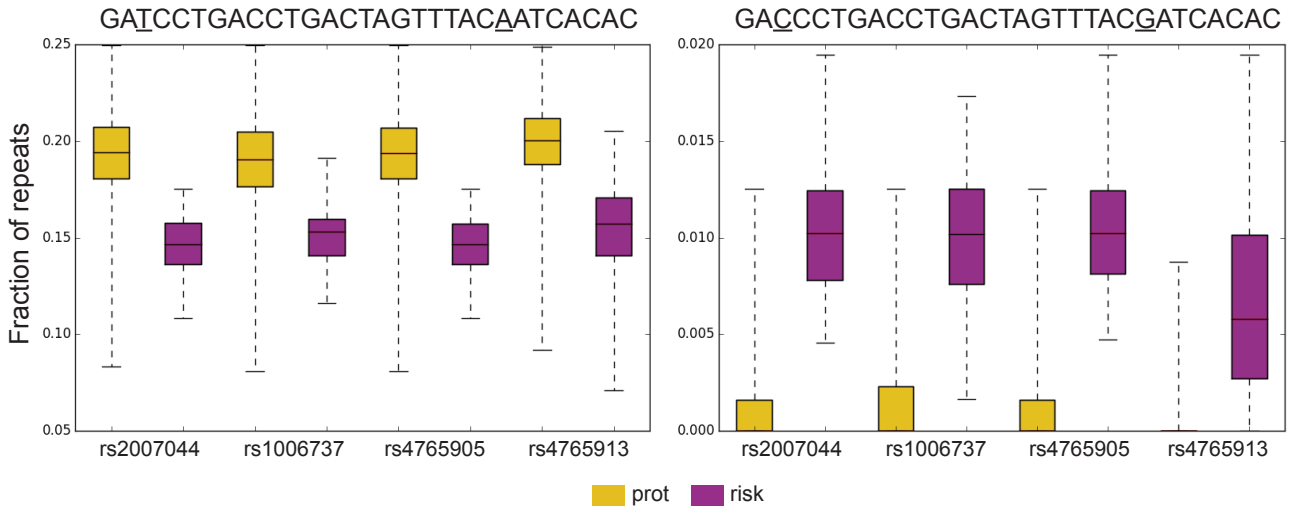

Figure S5: **Particular 30-mer sequence variants are associated with the protective or risk genotype at GWAS SNPs.** For each possible 30-mer repeat sequence, we determined what fraction of 30-mers found in each individual of a given cohort exactly match that particular variant. Two examples of significantly 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.

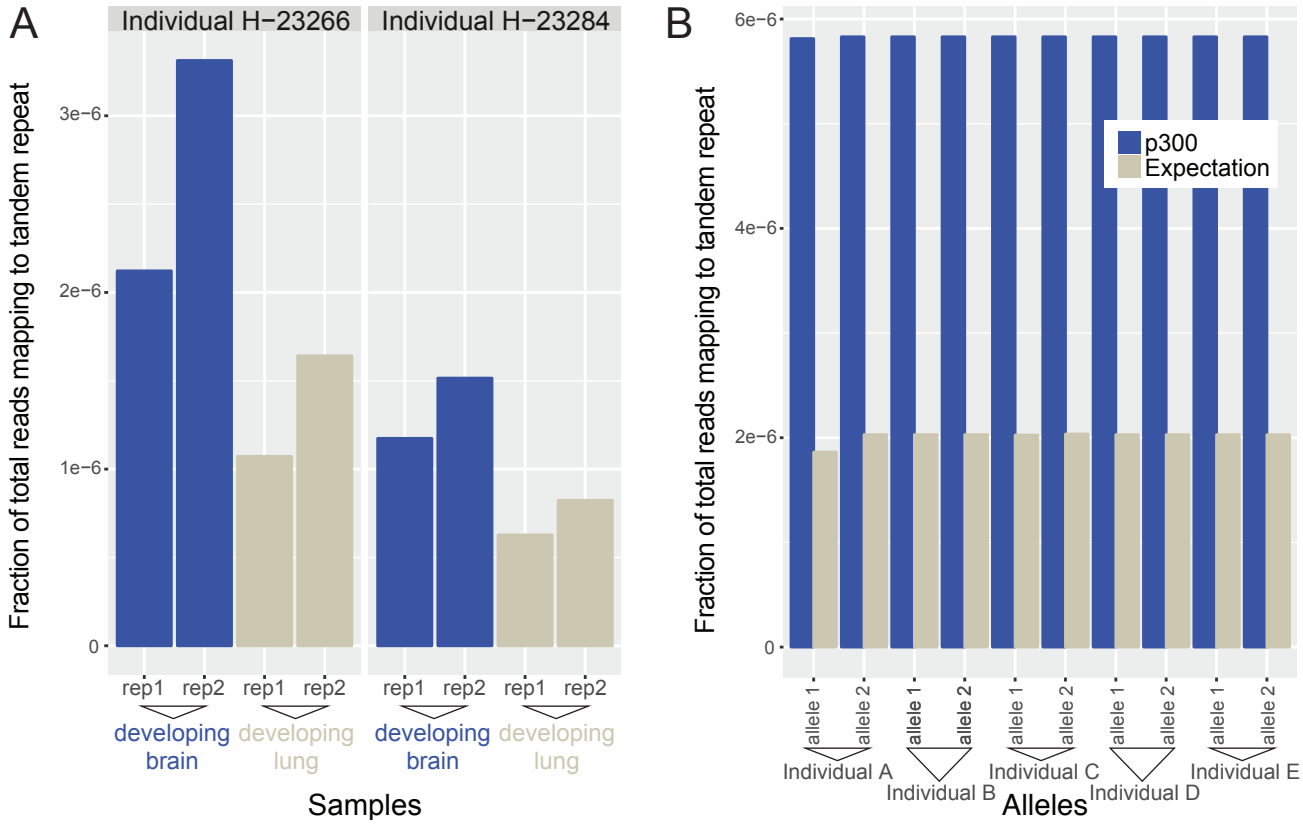

Figure S6: **Signatures of enhancer function in the developing human brain.** Both DNase I hypersensitivity and p300 ChIP-seq experiments rely on creating sequencing libraries enriched for genomic locations where the DNA is either open or associated with p300, respectively. The analysis of whether significant 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).

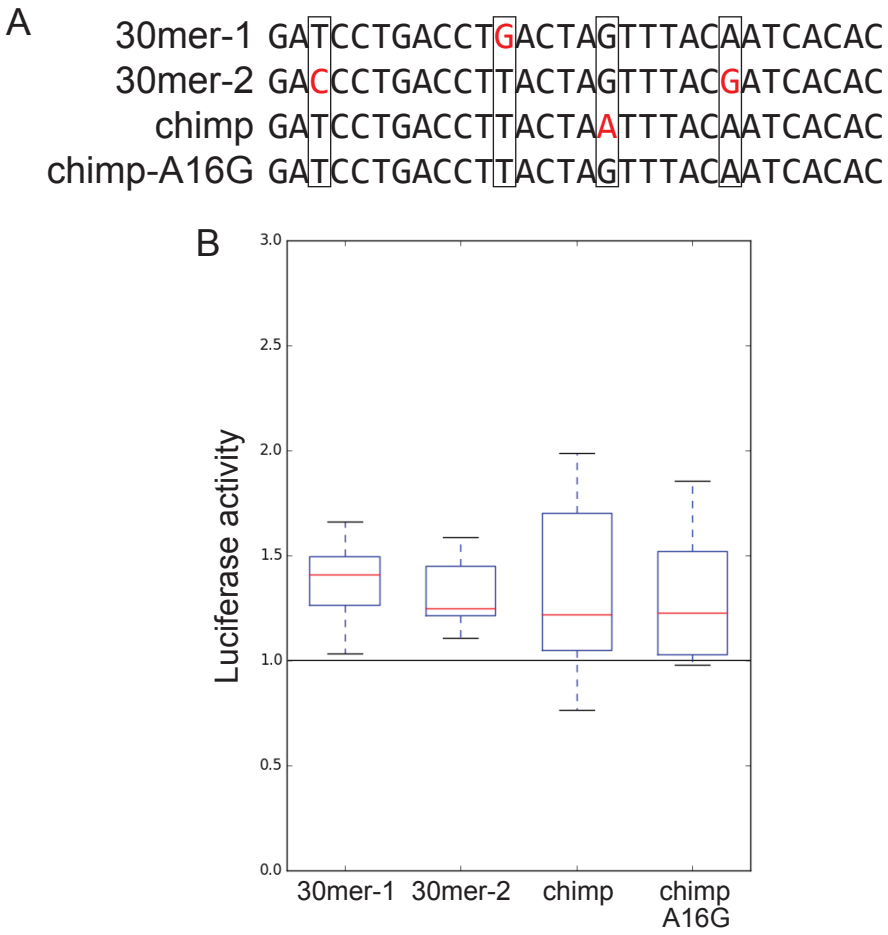

Figure S7: **Enhancer assays with single 30-mer units.** Four different 30-mer units were cloned upstream of a minimal promoter driving the luciferase gene and assayed individually for luciferase activity relative to the empty vector (horizontal line at 1.0) as described in Supplemental Methods. 30mer-1 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.

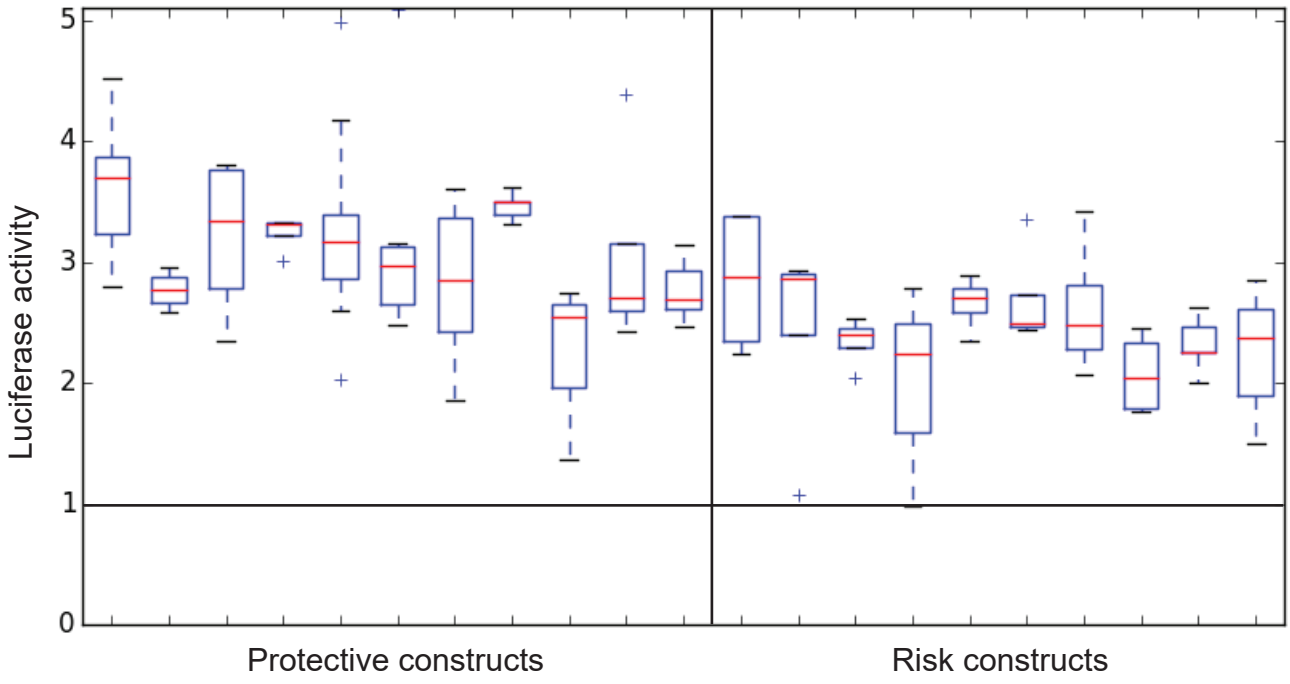

Figure S8: **Protective human repeat arrays drive higher luciferase activity than risk human repeat arrays.** 11 human repeat arrays characteristic of the protective haplotype and 10 repeat arrays characteristic of the risk haplotype were cloned upstream of a minimal promoter driving the luciferase gene. 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).

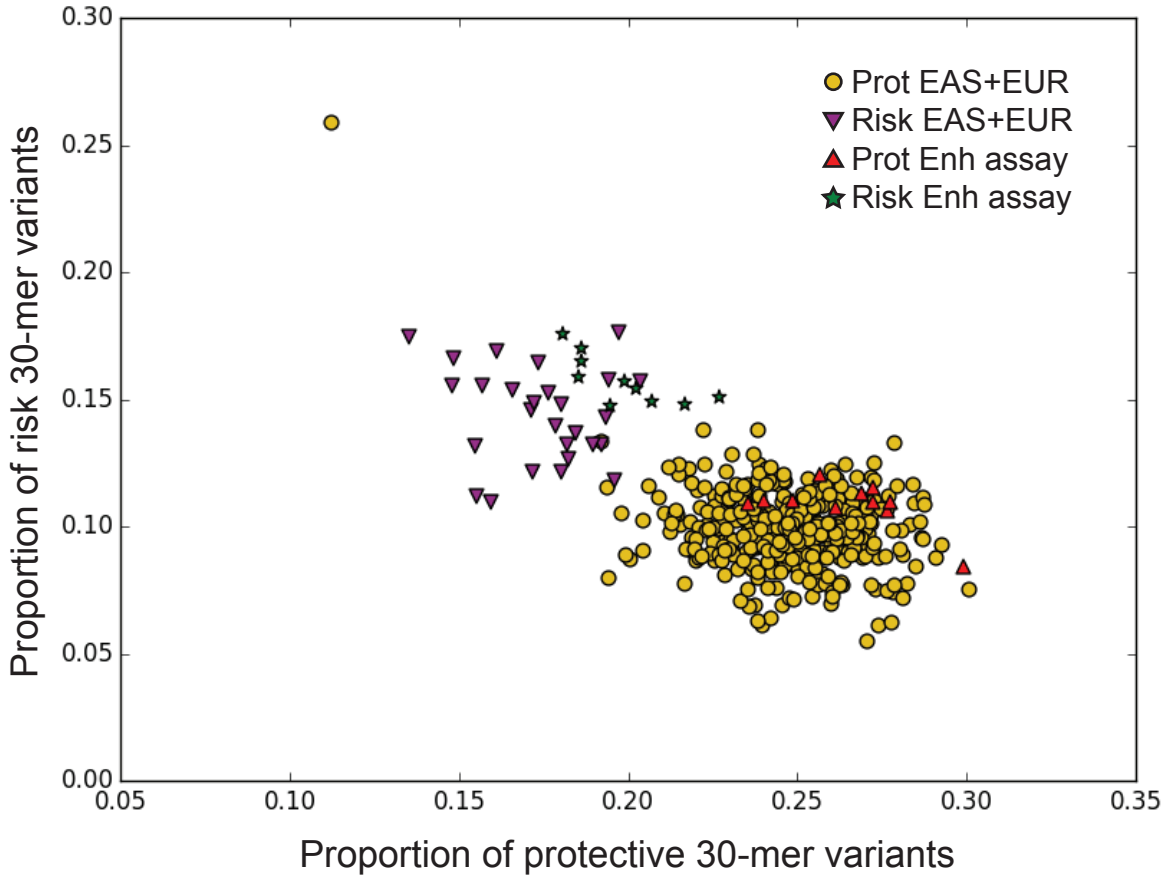

Figure S9: **Repeat arrays tested in enhancer assays cluster based on their proportion of protective- and risk-associated 30-mer variants.** The proportion of 30-mers that exactly match a 30-mer variant significantly associated with the protective haplotype or the risk haplotype are plotted for East 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).

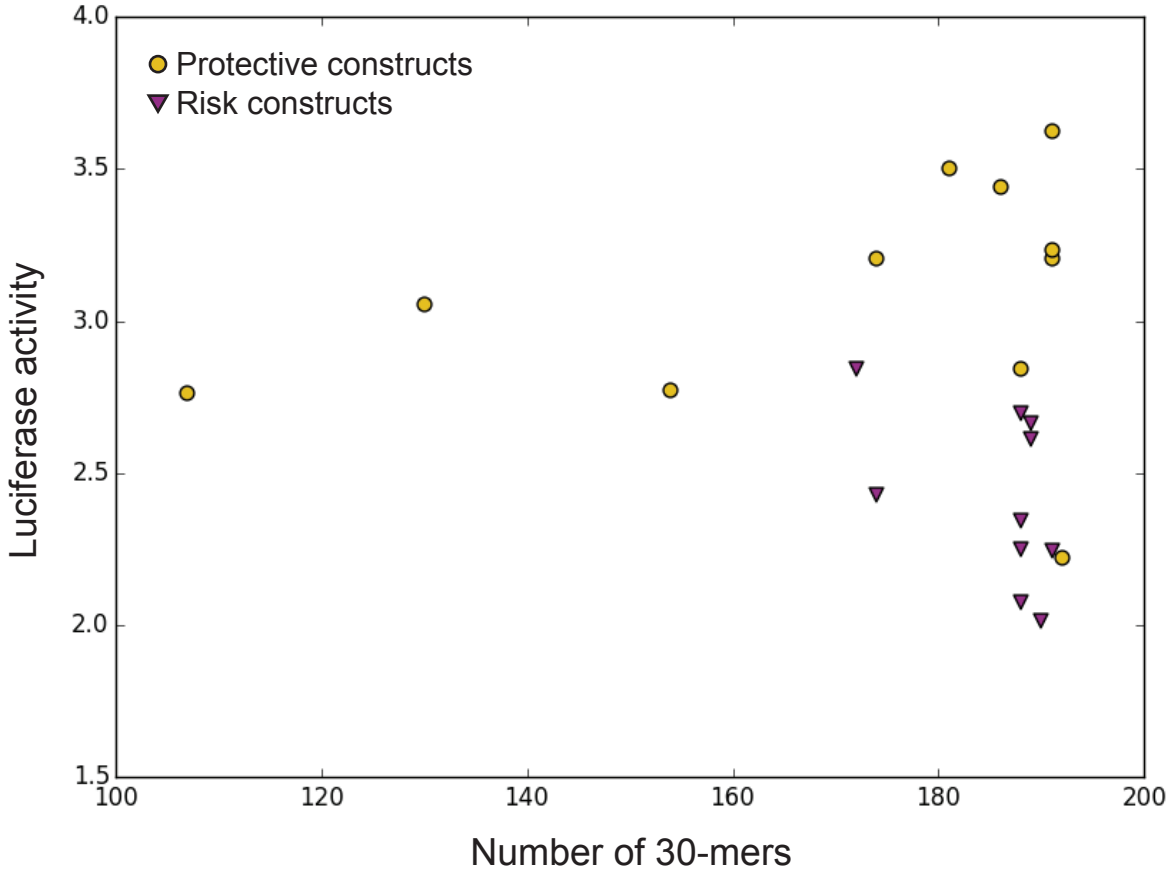

Figure S10: **Enhancer activity is not associated with human repeat array length.** 21 human repeat arrays were cloned upstream of a minimal promoter driving the luciferase gene and assayed for luciferase activity, as described in Supplemental Methods. These arrays contained between 107 and 192 30-mers. 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).

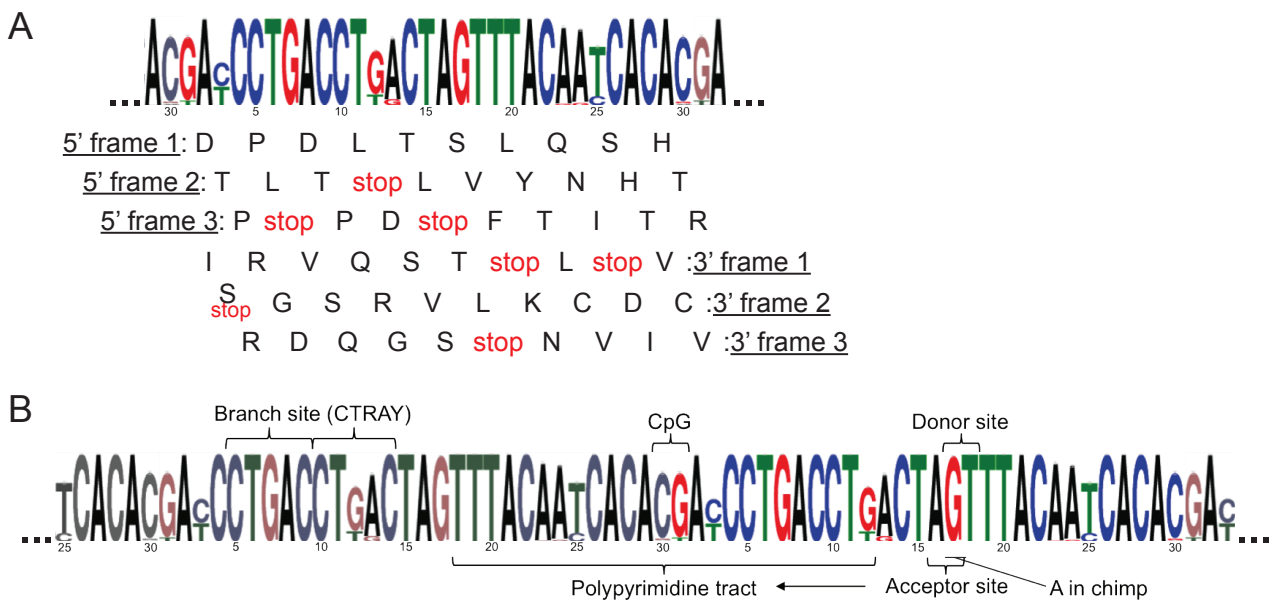

Figure S11: **30-mer repeat sequences contain open reading frames and putative CpG methylation and splicing sites.** (A) Predicted amino acid sequences are shown for each potential reading frame in the 5 and 3 direction. The first reading frame in the 5' direction is open in most 30-mer sequence variants. 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.