Supplementary Methods

Topologically associated domain (TAD) identification in human ES cells

Human ES cell TADs were obtained from Dixon et al. 2012 (**Supplementary Table 3**). These TADs were identified using the DI-HMM method as reported in that study.

Nested subTAD level domain identification in human ES cell from Hi-C

To obtain subTAD level domain identification, we downloaded raw reads from Hi-C on human H1 ESC, mapped, filtered, binned, and bias-corrected the data, and ran the DI-HMM method using parameters tuned to detect smaller domains. Paired end reads from Hi-C on human H1 ES cell were obtained from Dixon et al. 2015 (**Supplementary Table 3**). Bowtie2 was used to independently align reads to the mm9 genome through the HiC-Pro software (Servant et al. 2015). The global parameters used were: *--very-sensitive –L 30 –score-min L,-0.6,-0.2 –end-to-end --reorder*. The local parameters used were: *--very-sensitive –L 20 –score-min L,-0.6,-0.2 –end-to-end --reorder)*. Only uniquely aligned reads were paired. Unmapped reads, non-uniquely mapped reads, and PCR duplicates were removed. Reads were binned at 40kb resolution and balanced using the iterative correction and eigenvector decomposition (ICED) technique to generate Hi-C maps. Matrix balanced counts were then quantile normalized. Counts were subject to square root transformation and a full hierarchy of TADs and subTADs across length scales were called by varying the DI parameter in the DI-HMM method, as previously described (Dixon et al. 2012).

In brief, the DI-HMM method computes a directionality index (DI) for genomic intervals which reflects the degree of upstream or downstream bias. A hidden Markov model (HMM) is used to determine “true” upstream or downstream biased states. A domain is initiated at the beginning of a single downstream biased HMM state and ends at the last of an upstream biased HMM state. TAD calls were obtained using a DI parameter of 50, corresponding to a genomic interval of 2 Mb (50 bins times 40 kb resolution). subTADs were called using DI parameters of 8, 10, 15, 20, and 35, corresponding to genomic intervals of 320, 400, 600, 800, and 1400 kb. Eight bins was determined to be the lower limit for subTAD calls due to the narrow distribution of DIs relative to the number of bins genome-wide resulting in rank deficiency during matrix inversion within

HMM. We define all domains as the concatenation of unique, non-redundant TADs and subTADs (**Supplementary Table 4**).

TAD and nested subTAD level domain identification in human cortical plate from Hi-C

Mapped, filtered, and bias-corrected Hi-C data binned at 40 kb resolution from human cortical plate neurons was obtained from Won et al 2016 (**Supplementary Table 3**). TADs and subTADs were called on this data as described above using the DI-HMM method, using a DI parameter of 50 for TADs and parameters of 7, 8, 20, and 35 for nested subTADs. TADs and all domain calls are listed in **Supplementary Table 4.**

Domain boundary identification for domains from hES cell and human cortical plate neurons

Starting from the domains called in each cell type as described above, the boundaries of those domains were obtained using the following method. Because our method of calling subTADs concatenates the results of multiple DI parameters, at some locations in the genome there is variation in where the edge coordinate of a domain lies, and thus where the domain boundary is. To properly account for this variation, we merge boundaries to consolidate redundant boundary calls (for an illustration, see **Supplementary Figure 20**). To do this, from a list of domains, boundaries were obtained by first taking the edge coordinate of each domain and adding 40kb to either side. All boundaries that overlapped were then merged and a new boundary was created. The midpoint of each merged boundary was determined, and final boundaries were generated from this list by adding 100kb to either side of the midpoint. Therefore, the final size of all boundaries is 200kb. The 200kb boundary size was chosen to address the resolution limit of Hi-C assays, which is binned at 40kb. Domain boundaries for all datasets are listed in **Supplementary Table 4.**

Cell type specific TAD boundary identification

TAD boundaries from H1 ES cell (Human H1) and H1 ESC-derived lineages—mesendoderm (MES), mesenchymal stem cells (MSC), neural progenitor cells (NPC), and trophoblast-like cells (TRO)—were called by the Insulation Score (IS) method (Crane et al. 2015) and obtained from Schmitt et al. 2016 **(Supplementary Table 3)**. The domains called by Dixon et al. 2015 using

DI-HMM are not publicly available. Underlying differences in the algorithms make comparing boundaries called by IS and DI-HMM challenging. Thus, while throughout this paper we use H1 ES cell domains called using DI-HMM, only for the cell-type specific boundary analysis (Figure 2) do we use H1 ES cell domains called using IS so that we can compare these boundaries to those in the H1 ESC-derived lineages (**Supplementary Figure 7)**. We note in the main text that the majority of the boundaries called in H1 ES cell using DI-HMM are also identified using IS.

Germline Hi-C map generation

Contact frequency maps from mouse sperm were generated from pre-processed Hi-C read count data obtained from Jung et al 2017, GEO Accession number GSE79230 (**Supplementary Table 3**).

Selection of disease-associated STR (daSTR) genes for analysis

A total of 29 genes associated with trinucleotide repeat expansion disorders were initially selected for analysis based on a literature search (La Spada et al. 2010, Iyer et al. 2015). Of these 29, one (BEAN1) was dropped because a repetitive sequence with the reported repeat unit was not found within the gene in hg19. One (TBP) was dropped because it is located at the far edge of a chromosome beyond where chromatin conformation capture can query domains. One final gene (CSTB) was queried for boundary localization but dropped from all analyses that compared daSTRs to matched null repeats because there are no other repeats matching its sequence, length, and gene body location requirements in hg19. Thus, we were able to generate contact frequency heatmaps around 27 genes, and use 26 genes for analyses in Figures 3 and 5 (see **Supplementary Figure 21**).

Identification of the sequence of disease-associated STRs (daSTRs) and matched repeats

For each daSTR gene included in this analysis, the location of the longest contiguous stretch of the relevant repeat unit in the reported region within the gene body was determined in hg19 (**Supplementary Table 5**). Coordinates of genes and exons, introns, and UTRs were obtained from USCS Table Browser based on RefSeq genes downloaded in April 2017. For each daSTR, we identified a set of matched repeats. Matched repeats for each daSTR are repeats in hg19 that share the same repeat unit, are located in the same region within the gene body, and are at least 3

repeat units in length. One gene (CSTB) was queried for boundary localization but dropped from all analyses that compared daSTRs to matched nulls because there were no matched repeats for it found in hg19. These thresholds were determined based on both clinical data and population data that suggests that limiting nulls to be exactly the repeat length in hg19 is not representative of the true repeats seen across populations, and thus would likely not be representative of the cells used by previous studies for Hi-C.

Classifying disease-associated STR (daSTR) genes into 4 key classes

The 26 daSTR genes were grouped into 4 classes based on the repeat unit and location of the repeat in the gene body (**Supplementary Table 1**). These groups are motivated by prior literature showing that daSTRs with similar sequence and gene body locations behave similarly with respect to number of repeats required for disease and expansion mechanisms (Iyer et al 2015). The classes are: 1) Repeat unit CAG in exons and 5'UTRs (9 daSTRs), 2) Repeat unit GCG/CGG in exons and 5'UTRs (8 daSTRs), 3) Unique repeat units in introns (5 daSTRs), and 4) CTG introns in 3'UTRs (3 daSTRs). One daSTR gene (COMP) was not included in these groups because its repeat unit, 'GAC', did not fall into one of these four categories (**Supplementary Figure 21, Supplementary Table 2**). The final four groups of daSTR genes, their properties, and the number of matched repeats for each are listed in **Supplementary Table 2**.

Determining repeat localization and distance to boundary

Distance from a repeat to the nearest domain boundary was calculated as the minimum distance from either end of the repeat to its closest boundary. Repeats contained within a boundary are given distance of 0 bp. A repeat is considered at a boundary if it overlaps the boundary by one or more base pairs.

Determining CpG island, CTCF, and H3K9me3 content across the genome

To determine CpG island, CTCF, and H3K9me3 content across the genome, the entire hg19 genome was first partitioned into sequential 200kb bins using bedtools makewindows. Then, any bin that overlapped a domain boundary present in either H1 ES cell (Human H1), H1 ES cell derived lineages (mesendoderm (MES), mesenchymal stem cells (MSC), neural progenitor cells

(NPC), and trophoblast-like cells (TRO), or human cortical plate neurons were removed. Bins overlapping centromeres, telomeres, gene deserts (areas of the genome > 2 Mb with no genes) and regions not able to be queried (ie, ends of chromosomes) by Hi-C were also removed. The remaining bins are considered non-boundary bins (for an illustration, see **Supplementary Figure 22**). This method ensures that both boundaries and non-boundary bins are the same size, and that non-boundary bins are not dominated by hard to query regions of the genome. The number of CpG islands, CTCF peaks, and H3K9me3 peaks per boundary and 200kb non-boundary bin was determined using bedtools intersect. Centromeres and telomeres were obtained from USCS Table Browser. Genes were from the RefSeq gene track on USCS Table Browser in May 2017. CpG islands in hg19 were obtained from USCS Table Browser (group: Regulation, track: CpG Islands, table: cpgIslandExt). H1 ES cell CTCF ChIP-seq peaks were obtained from ENCODE (UCSC Accession wgEncodeEH000085). H1 ES cell H3K9me3 was obtained from GEO Accession GSM605325 (**Supplementary Table 3**).

Bootstrapped intervals for boundary localization percentage and CpG island density

Bootstrapped intervals for percent of repeats at boundary and CpG island density were constructed by randomly selecting groups of 26 repeats from both the matched repeat set and the daSTRs 1,000 times with replacement and computing the percent of the randomly chosen repeats at a boundary each time. Each group drawn from the matched repeats matched the distribution of the daSTRs such that all four different classes of daSTRs are represented in equal proportion to what is found in the daSTRs group.

Randomization tests for boundary localization and CpG island density

The null distribution was created by taking 10,000 draws without replacement of n=26 matched repeats, where each draw is proportionally matched to the daSTR group with respect to repeat sequence and gene body location, and computing the test statistic each time. Two test statistics were used: 1) The percent of repeats that were located at a boundary (Figure 3), and 2) For repeats at boundaries, the number of CpG islands present in those boundaries (Figure 5). The distribution of the draws from the matched repeats set matched that of the daSTRs such that one iteration of a random draw from the null group draws from the different types of repeats and

their genomic locations with equal proportion to what is found in the daSTR genes group. The null distribution was compared to the test statistic to calculate an empirical p-value.

Pileups of CTCF/CpG density around domain boundaries

Pileup plots depicting the number of CTCF/CpG sites centered around the middle of boundaries were generated by counting the number of CTCF/CpG sites in 200kb genomic bins flanking the center of the boundary and averaging this number by the total number of domain boundaries.

Gene expression analysis

Gene expression datasets in post-mortem brain samples from Huntington's disease patients and normal controls were obtained from GEO (**Supplementary Table 3**). Normalized data was downloaded plotted for the genes of interest with no further manipulation.

Cell Culture for 5C analysis in B cells

B lymphocytes were grown at 37°C and 5% CO_2 with 15% FBS (Atlanta Biologicals) in RPMI-1640 media with 2mM glutamine (Sigma-Aldrich) and 1% antibiotic-antimycotic (Sigma-Aldrich). Cells were split every 3-4 days in accordance with repository protocol.

5C library generation and sequencing

The procedure for 5C begins in a similar fashion as that for *in situ* Hi-C[1]. Briefly, 4 million cells were collected per library and fixed with 11% formaldehyde to a final concentration of 1% at room temperature for 10 min. Formaldehyde was quenched by addition of 2.5M glycine to a final concentration of 0.2M for 5 min. After spinning at 300g at 4°C for 5 min, cells were washed in cold 1x PBS and spun again at 300g at 4°C for 5 min. Cell pellets were then flash-frozen in liquid nitrogen. Crosslinked pellets were lysed (10mM Tris-HCl pH 8.0, 10mM NaCl, 0.2% NP-40, protease inhibitors) for 15 min. After spinning at 2,500g for 5 min, supernatant was discarded and pelleted nuclei were washed again with lysis buffer. Chromatin was solubilized in 50µl 0.5% SDS and incubated at 62°C for 10 min. SDS was quenched with 145µl of water and 25µl of 10% Triton X-100 and incubated at 37°C for 15 min. Nuclei were digested with 100U of HindIII at 37°C overnight. HindIII was then inactivated at 62°C for 20 min. After adding 950µl ligation mix (100µl 10% Triton X-100, 120µl 10x NEB T4 DNA Ligation buffer, 12µl 10 mg/ml

BSA, 718µl water, and 2000U T4 DNA Ligase), fragments were ligated at 16°C for 2 h. After spinning at 2,500g for 5 min, supernatant was discarded and pellet was resuspended in 300µl buffer (10mM Tris-Hcl pH 8.0, 0.5M NaCl, 1.0% SDS). Crosslinks were reversed with addition of 25µl 20 mg/ml Proteinase K and incubated at 65°C for 4 h. Additional 25µl 20 mg/ml Proteinase K was added and incubated at 65°C overnight. Prior to DNA extraction, 5µl 10 mg/ml RNaseA was added and incubated at 37°C for 30 min. DNA was extracted with 350µl phenol:chloroform and precipitated with sodium acetate and ethanol. Excess salt was removed using Amicon Ultra centrifugal filter unit. Double-alternating 5C primers were designed to the *Fmr1* locus[2] and added at 0.5 fM after phosphorylation of reverse primers. Library and primers were denatured at 95°C for 5 min. Primers were annealed at 55°C for 16 h. Primers were ligated with addition of 10U of Taq Ligase and incubation at 55°C for 1 h. Taq Ligase was then inactivated at 75°C for 10 min. Ligation products were combined with PCR mix (5µl 5x HF buffer, 0.2µl 25mM dNTP, 1.5µl 80µM Emulsion forward primers, 1.5µl 80µM Emulsion phosphorylated reverse primers, 0.25µl Phusion polymerase (NEB), 10.55µl water) and amplified in 3 stages: 1 cycle - 95°C for 5 min, 30 cycles - 98°C for 10 s, 62°C for 30 s, 72°C for 30 s, 1 cycle - 72°C for 10 min, 4°C thereafter. Size selection was done using AMPure XP beads and library preparation was performed with NEBNext Ultra DNA Library Prep Kit according to manufacturer protocol (NEB #7370). Libraries were sequenced on Illumina NextSeq 500 according to the manufacturer's instruction.

5C data mapping and processing

Paired-end reads were mapped to the reference human genome (hg19) using bowtie (**Supplementary Table 6**). Counts were first quantile-normalized, conditioning on GC content, between conditions[3]. Next, matrices were normalized individually using the iterative correction method[4]. Finally, primer level counts were converted to fragment level counts by averaging all primers that measured the same fragment. Normalized matrices were then binned at 2kb with a 20kb smoothing window.

Code availability

All code and usage instructions to reproduce results were provided to reviewers.

References

1. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
2. Lajoie, B.R., van Berkum, N.L., Sanyal, A. & Dekker, J. My5C: web tools for chromosome conformation capture studies. *Nat Methods* **6**, 690-1 (2009).
3. Beagan, J.A. *et al.* YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* (2017).
4. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**, 999-1003 (2012).