PNAS
www.pnas.org

Supplementary Information for

Human-specific tandem repeat expansion and differential gene expression during primate evolution

Arvis Sulovari, Ruiyang Li, Peter A. Audano, David Porubsky, Mitchell R. Vollger, Glennis A. Logsdon, Human Genome Structural Variation Consortium, Wesley C. Warren, Alex A. Pollen, Mark J.P. Chaisson, Evan E. Eichler

Corresponding author: Evan E. Eichler, Ph.D
Email: eee@gs.washington.edu

This PDF file includes:

Supplementary methods
Figures S1 to S11
Legends for Datasets S1 to S14
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S14

www.pnas.org/cgi/doi/10.1073/pnas.1912175116

METHODS

Phasing of the tandem repeat loci in NHP genomes. The SMRT PacBio data from the three NHP were aligned against GRCh38 using BLASR(25) and the same alignment parameters as those used by the HGSVC:

```
--bam --insertion 8 --deletion 8 --mismatch 4 --indelRate 3 --
advanceExactMatches 10 --maxMatch 25 --sdpTupleSize 13 --out --nproc 8
```

The generated BAM and the 10X phased variants were processed through PhasedSV (https://github.com/mchaisso/phasedsv). Although PhasedSV was designed to identify and sequence-resolve SVs at a whole-genome scale, we used it in a targeted fashion, such that it assembled NHP alleles only for our genomic regions of interest. Additionally, no stitching was performed to join the haplotype-resolved contigs. Although stitching improves the contiguity of the assemblies, we do not anticipate it to affect our ability to resolve tandem repeat regions since these are relatively small in size (the vast majority measure ≤10 kbp in length). Importantly, when PhasedSV cannot haplotype-partition PacBio reads for a given locus (e.g., when we have a paucity of phased heterozygous SNVs from 10X), it will utilize all mapped reads for that region and assemble the same contig in both alleles.

Sequence annotation and visualization. The sequences from each of the six NHPs were analyzed with RepeatMasker v.4.0.3(1) and TRF v.4.07b(2) using the same parameter settings as those used for annotation of HGSVC tandem repeat sequences. To visualize the structure of all tandem repeats, we constructed all pairwise dotplots for a total of 18 sequence sources: GRCh38, the eight human samples (six HGSVC haplotypes, haploid hydatidiform cell line CHM13, and the haplotype unaware Yoruban individual assembly), and nine NHP samples (six haplotypes resolved in this study and three haplotype unaware assemblies from Clint (chimpanzee)(3), Susie (gorilla)(4) and Susie (orangutan)(3)). A subset of expanded and disease-associated loci were selected for detailed visualizations using dotplots, multiple sequence alignments (MSAs) of the assembled human and NHP haplotypes, and Miropeats(5). We developed the visualization tool Dottedpython (https://github.com/ruiyangl/Dottedpython), which uses a variety of sequence analyses methods to visualize the structure of tandem repeat sequences across different samples. It incorporates dot matrix analyses (first described by Sonnhammer and Durbin(6)), repeat length statistics from TRF(2), MSAs of the repeat motif sequence using Clustal Omega(7), and gene annotation based on RefSeq gene annotations from UCSC GRCh38. Tandem repeats sequences are analyzed using pairwise dot matrix, which are generated for each unique pair of samples (i.e., ${}^{18}C_2 + 18 = 173$ unique pairs including self-comparisons). To generate a dot matrix, we use prefix doubling to construct a suffix array(8). Our program searches for identical 10-mers in two sequences. It has a time complexity of *O(nlog(n))* and a linear memory usage.

We also created a new dynamic visualization framework, Sequence Composition Viewer, specifically designed for tandem repeats. To produce a sequence composition view, we require 1) the sequence of the tandem repeat with some flanking sequence, and 2) the periodicity (i.e., motif length) of the STR/VNTR. Flanking sequences for STRs and VNTRs were set to ~500 bp and ~1.5 kbp, respectively. Next, we apply KAnalyze(9) to all the available haplotype-resolved sequences from human and NHP haplotypes to establish the abundance of each available k-mer that has the same length as the motif of the tandem repeat. Next, the k-mers are ordered by decreasing order of abundance, and each k-mer sequence is paired with a specific color from "hot red" to "cold blue", such that the most abundant k-mer corresponds to the hottest color. Our color wheel contains 50 distinct heat colors. The k-mer-to-color pairing information is used to recursively replace exact matches in the input sequence with specific colors. This is done to control for the redundant nature of the k-mer sequences and to avoid assigning multiple colors to the same k-mer. For example, consider a 4 bp motif STR, where the two most abundant k-mers are AAGA followed by AAAG. The hottest color will be assigned to each exact match of AAGA in the input sequence; if after all copies of AAGA are assigned a color, our input sequence no longer contains AAAG motifs, then the second hottest color is assigned to the next most abundant 4-mer with non-zero abundance. After all 50 unique colors have been assigned to available k-mers, the rest of the input sequence is colored in gray. To properly align these multicolor tracts across multiple samples, we force a left and right alignment of the unique flanking sequence by breaking the tract in the last occurrence of two consecutive "hot colors"; most often the break will occur at the end of the tandem repeat region. The longest pure tract length (i.e., longest tract of uninterrupted perfectly repeating motif) is reported next to each sample, along with the motif sequence and the GRCh38 coordinates.

We used the start and end coordinates of the expanded region to search for gene annotations and report the distance to the nearest gene up and downstream of the repeat based on NCBI RefSeq data. Tandem repeats located <10 kbp away from the nearest gene were classified as upstream or downstream from a gene; otherwise they were classified as intergenic. Gene enhancer annotations were extracted from the nonredundant database GeneHancer(10), which aggregates enhancer information from sources including ENCODE, FANTOM and VISTA Enhancer Browser.

Sequence motif analysis was performed using TRF(2) with the following command-line options: `trf <Input file> 2 7 7 80 10 50 2000 -h -d -ngs > <output file>`. MSAs were generated using the sequence of the motifs corresponding to the longest total length in each sample. Each STR and VNTR repeat motif was analyzed through Multiple EM for Motif Elicitation (MEME)(11).

Identification of HSE STRs/VNTRs. We identified HSE tandem repeats according the following four criteria. First, to control for any technical biases that may have affected the sequence or mapping accuracy, we employed the following filters: removed loci overlapping with segmental duplications or containing satellite DNA, required that the haplotype-resolved tandem repeat sequence was flanked by ≥100 bp of unique GRCh38 sequence, required ≥3 samples with contiguously assembled sequence in each primate cohort, and tandem repeats had to have a motif length ≥2 bp. Second, we tested for differences in the lengths of tandem repeats between the two primate groups using the nonparametric two-tailed Wilcoxon rank-sum test. We adjusted for multiple hypothesis testing using the false discovery rate (FDR)(12) and required each HSE to have a q-value ≤ 0.05. Third, we quantified the variance ratio of the copy number values using the V_{ST} statistic in the two primate cohorts using the following formula at each tandem repeat locus:

$$V_{ST} = (Var_{Total} – (N_{NHP} \times Var_{NHP} + N_{HS} \times Var_{HS}))/Var_{Total}$$

where Var_{Total} is the total variance of all copy number values across the human and NHP haplotypes, and Var_{NHP} and Var_{HS} represent the variance for the specific primate groups. To establish a data-driven V_{ST} threshold for the most copy number different tandem repeats, we determined the first inflection point in the probability distribution function of V_{ST} using the extreme distance estimator approach with the Chebyshev confidence interval. This inflection point was found to be at $V_{ST} = 0.45$ [C.I. = 0.36 – 0.53]; hence, we required all candidates for human-specific repeat expansions to have a $V_{ST} \geq 0.45$. Lastly, to account for the copy number variability of a tandem repeat in each primate cohort, we required that the longest NHP tandem repeat and the shortest human tandem repeat (i.e., the minimal copy number difference) differ by >5 tandem repeat copies. This value was determined based on the trimodal distribution of minimal human-specific versus NHP copy number differences. We observe three distinct peaks in the probability distribution function, each of which corresponds to the ≤1 copy number difference (left-most peak), between >1 and ≤ 5 copies (middle peak), and >10 tandem copies (right-most peak). Since we aimed to capture all tandem repeat loci represented by the right-most distribution, we required copy number differences larger than 5 tandem copies.

Differential splicing analysis. To identify splicing differences, we obtained multi-tissue RNA-seq data from five chimpanzee brains(13) (NCBI BioProject accession number: PRJNA236446) and ten human brains from the GTEx(14) project. The FASTQ files were aligned to the transcriptome in GRCh38 coordinates using the STAR (2.6.1d) aligner(15) and gene transfer format file for the full transcriptome from GENCODE v29. The following human–chimpanzee brain tissue pairings were considered for the identification of differentially spliced striatum and cortex genes, respectively: striatum versus nucleus accumbens basal ganglia and orbital prefrontal cortex versus frontal cortex BA9. Next, the LeafCutter pipeline(16) was used to identify intron clusters (defined by the presence of ≥50 RNA split reads and introns of length ≤500 kbp) followed by differential splicing analysis between the human and the chimpanzee samples. The χ^2 test p-values associated with differential splicing were conservatively adjusted using the family-wise error rate (FWER) procedure across 79,322 intron clusters of the human transcriptome, resulting in adjusted threshold $\alpha = 0.05/79,322 = 6.3\times10^{-7}$. This analysis yielded 1,397 and 977 differentially spliced cortex and striatum genes, respectively. A local R Shiny app was used for visualizing the LeafCutter results (more details on running LeafCutter can be found here: https://davidaknowles.github.io/leafcutter). R version 3.4.0 was used for all statistical analyses.

Multiple regression models. The overlap of our tandem repeats with a gene (GR), density of tandem repeats at the subtelomeres, and the overlap with differential expression gene sets were regressed against a combination of the following: the average length of the tandem repeat in human samples (HL), the distance to the nearest gene (GD), the length of the nearest gene (GL), and a few categorical variables including subtelomere overlap (SS, i.e., ≤5 Mbp from the chromosome arms), which we have recently shown to be significantly enriched for tandem repeats(17), *ab initio* status of the repeat (AB), HSE status (HE), cells-type associated genes (CT), STR/VNTR repeat type (RT), and overlap of the tandem repeat loci with interspersed repeat elements as defined by RepeatMasker (RM, i.e., Alu, LINE, ERVK, SVA, and transposons). Two-way interactions of the explanatory variables were considered assuming sufficient degrees of freedom. The following model represents the general framework we used for the explanatory variables tested for association with each specific response variable:

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \mathrm{HL} + \beta_2 \mathrm{GD} + \beta_3 \mathrm{GD} + \beta_4 \mathrm{GL} + \beta_5 \mathrm{RT} + \beta_6 \mathrm{RM} + (\mathrm{HL} + \mathrm{HE} + \mathrm{GD} + \mathrm{GL} + \mathrm{RT} + \mathrm{RM})^2 + \varepsilon$$

Where $\ln \frac{p_i}{1-p_i}$ represents the natural log of the likelihood that each event *i* occurs (i.e., the natural log of odds), and ε represents the random effects, which are modelled as a normal distribution with $\mu = 0$ and $\sigma = 1$. We pursued this approach for modelling our annotation data because we can test for association while controlling for potential confounders. For example, the odds of a tandem repeat overlapping with a gene might be affected by the tandem repeat's size, or the size of the gene, or the repeat's location in the genome; hence, in the generalized linear model framework above, we control for all of these confounders.

Validation experiments. We performed the following three independent sequence validation experiments. BAC sequencing: We aligned 199 NHP BAC insert sequences(21) to haplotype-resolved tandem repeat sequences from our three NHP genomes. Importantly, these BAC genomic libraries were generated using DNA from the same three NHP individuals as those used in our study: Clint (chimpanzee), Kamilah (gorilla), and Susie (orangutan). We directly compared the length and percent sequence identity of the tandem repeat sequences from BACs to the corresponding individuals' haplotype-resolved sequences generated by PacBio and 10X. We used minimap2 (v2.16)(30) to map all 199 BACs against GRCh38; the mapping coordinates were used to identify all overlapping STR/VNTR loci. Next, the overlapping tandem repeat sequences were mapped against the BACs. The cigar strings were processed using three different methods to produce percent sequence identity, as previously reported in the segmental duplication assembly pipeline(31). In order for a sequence to be considered validated, we required ≥99% length concordance and ≥99% sequence identity between our haplotype-resolved assemblies and BACs. The sequence identity metric we used considers each base pair in an indel.

Macaque assembly comparisons: We used the most recent macaque genome long-read assembly (*Macaca mulatta*, GenBank accession: GCA_003339765.3) to extract sequence from the regions homologous to the human *ab initio* and expanded tandem repeats. This assembly was chosen for two reasons: 1) macaque represents a more distant outgroup to the human lineage than our current outgroup (i.e., orangutan), allowing us to evaluate the human specificity of our tandem repeats and identify potential false positives in this set; and 2) importantly, this assembly was constructed using PacBio long reads, which unlike the older short-read assemblies, does not contain the inherent sequencing bias present over repetitive regions.

Since a LiftOver chain was not available to convert GRCh38 coordinates of tandem repeats to macaque reference coordinates, we extracted 2 kbp sequences from both flanking regions of each tandem repeat locus in GRCh38 and mapped these to the macaque primary contigs with minimap2. The resulting unique mapping positions (i.e., MAPQ ≥ 40) of the upstream and downstream flanks in the macaque contigs were used to guide the extraction of the homologous human tandem repeat sequences. This mapping approach is advantageous over mapping of the complete human tandem repeat locus, as it controls for potential fracturing of the alignment over the tandem repeat, or other misalignment artifacts, due to the vast structural genomic differences that may exist between the macaque and human genomes over tandem repeats. After characterizing each extracted macaque sequence with TRF, we employed separate validation criteria for the human *ab initio* and the HSE loci. In the case of *ab initio* loci, the homologous macaque sequence had to contain no tandem repeats for it to be considered validated. For each HSE locus, we required every homologous macaque sequence to contain a tandem repeat that is at most as long as the largest allele of that locus in the three NHP genomes.

Validation with orthogonal long-read sequence datasets: We carried out an additional orthogonal validation for STRs and VNTRs from the CHM13 CLR (continuous long-read) assembly, using orthogonal HiFi (high-fidelity) circular consensus sequencing data and ultra-long Oxford Nanopore Technologies (UL-ONT) sequence reads generated from the same source cell line (CHM13)(32). We considered a repeat validated if there was ≥99% length concordance as well as ≥99% sequence concordance between the CLR and the HiFi assemblies of STRs/VNTRs. For regions where HiFi and CLR were in disagreement (i.e., >1% discordance), size concordance for each of the two technologies were compared to ONT (allowing ≤5% discordance), assuming ONT more accurately captured the true length. To access the accuracy of the CLR sequence, we further analyzed the regions that were size-concordant between HiFi and CLR using the underlying circular consensus sequence data from HiFi restricting the analysis to the highest quality of data QV > 30 (i.e., < 0.1% sequencing error).

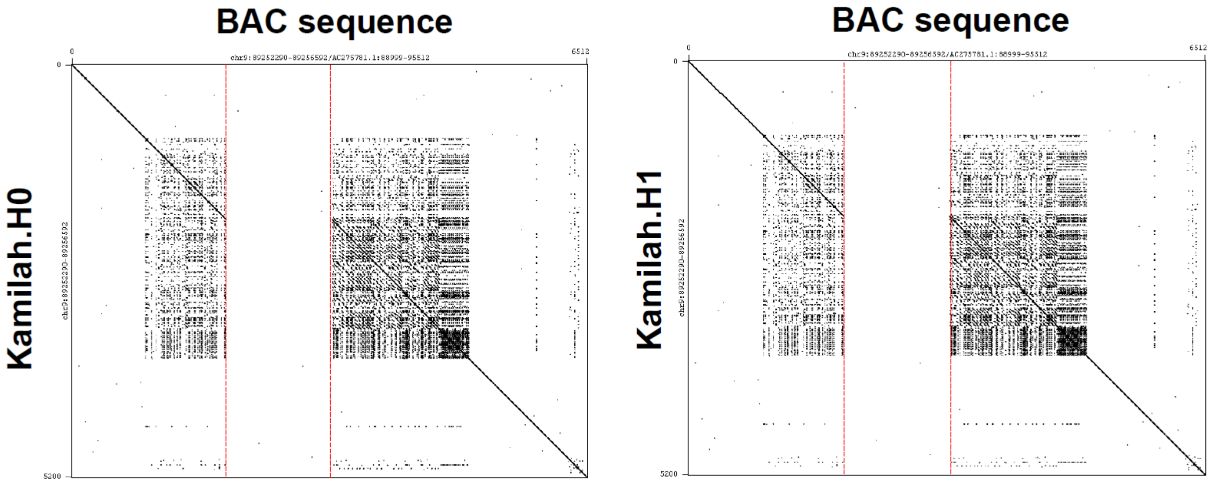


Fig. S1. BAC propagation artefact. An *E. coli* sequence insertion occurs in the gorilla BAC sequence and leads to a false negative validation result. Gepard(18) was used for making the dotplots. The inserted sequence is shown in between the dotted red vertical lines, representing a ~1.25 kbp insertion corresponding to the transposase *E. coli* gene.

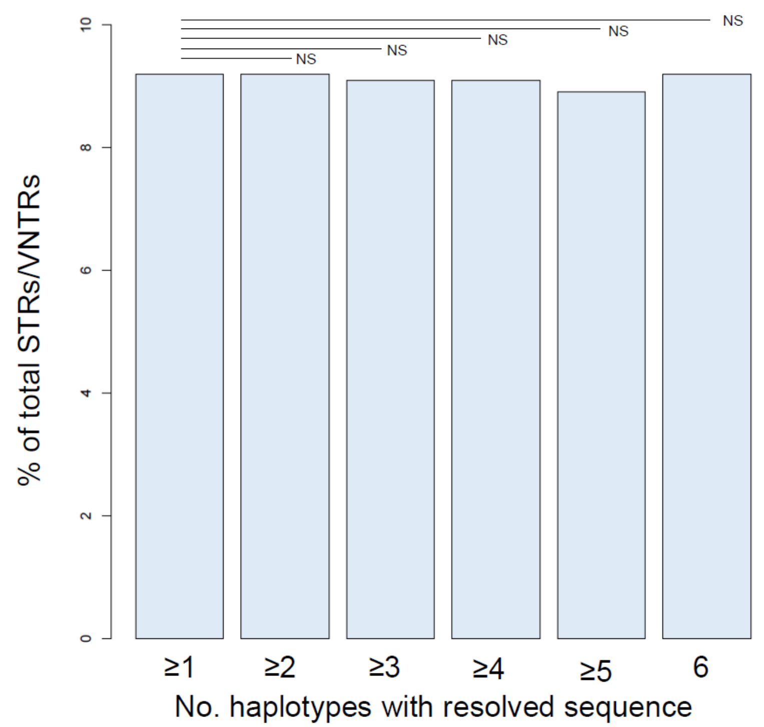


Fig. S2. Missing data and discovery of human-specific STR/VNTR expansions. The proportion of missing haplotype-resolved sequence is consistent for different thresholds. The x-axis indicates the number of haplotypes (from human and NHPs) that have been successfully assembled across all of the HSE and *ab initio* loci (n = 1,584). The label "NS" refers to the lack of statistical significance (χ^2 test p-value > 0.05) for tests that were conducted between different pairs of missing thresholds.

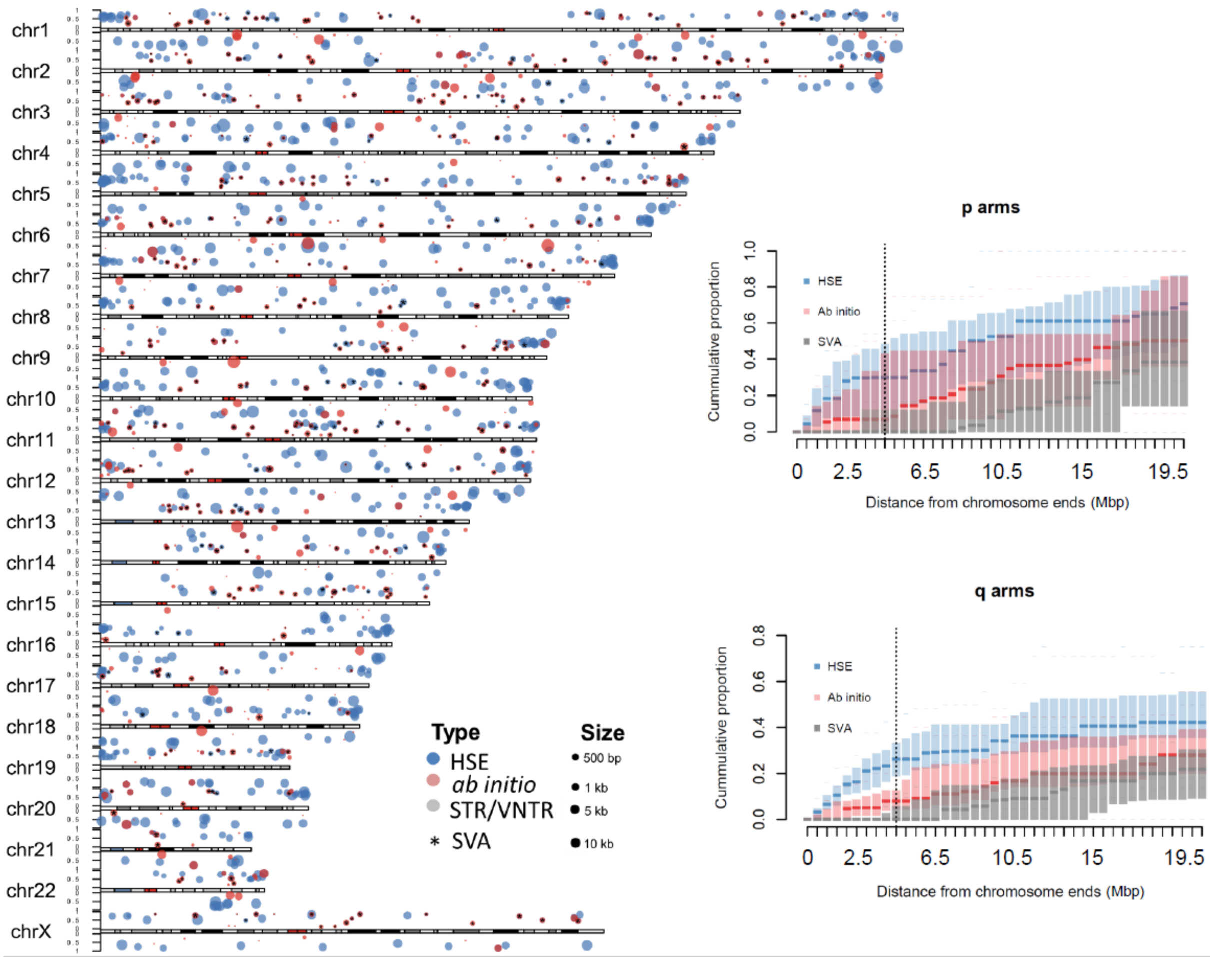


Fig. S3. The genome-wide distribution of HSE and *ab initio* STRs/VNTRs. The karyoplot to the left depicts the genome-wide distribution of HSE (blue) and *ab initio* (red) STRs (bottom panel) and VNTRs (top panel). The size of the filled circles corresponds to the tandem repeat size, while the star "*" indicates loci that are SVA-associated. The y-axes of both top and bottom panels in each chromosome correspond to the V_{ST} value for each locus ranging from 0 to 1 (**Methods**). The colored boxplots to the right summarize the cumulative proportion of STRs and VNTRs observed as we move away from the telomere, towards the centromere. Three types of loci are represented for each chromosome arm, including HSE (blue), *ab initio* (red), and SV-associated (gray). The latter represent negative controls for the subtelomeric enrichment, since SVA-associated tandem repeats do not display such an enrichment. The dotted vertical line represents the boundary for the subtelomere region at 5 Mbp position away from the chromosome ends.

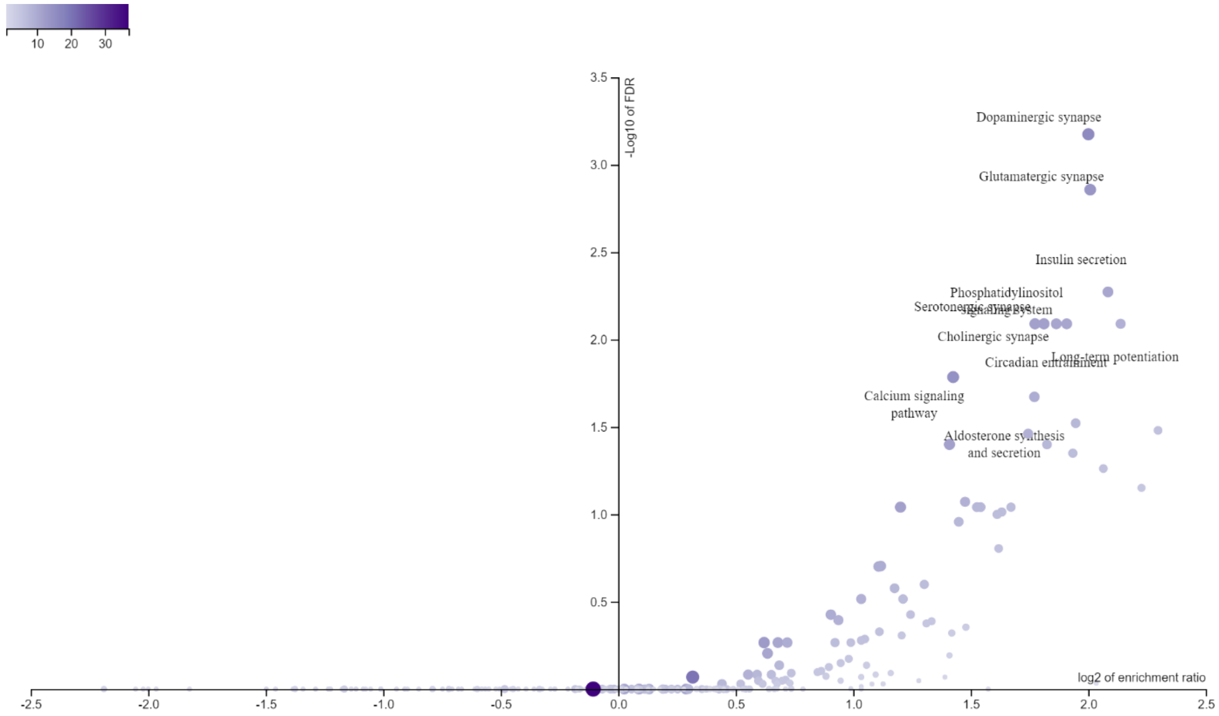


Fig. S4. KEGG overrepresentation analysis. A list of 814 genes overlapping with *ab initio* and HSEs was used as input. The top three enriched pathways were dopaminergic synapse (hsa04728) pathway (enrichment ratio = 4.0, FDR p-value = 6.7×10^{-4}), glutamatergic synapse (hsa04724, enrichment ratio = 4.02, FDR p-value = 0.0013), and insulin secretion (hsa04911, enrichment ratio = 4.24, FDR p-value = 0.005). WebGestalt was used for this analysis.

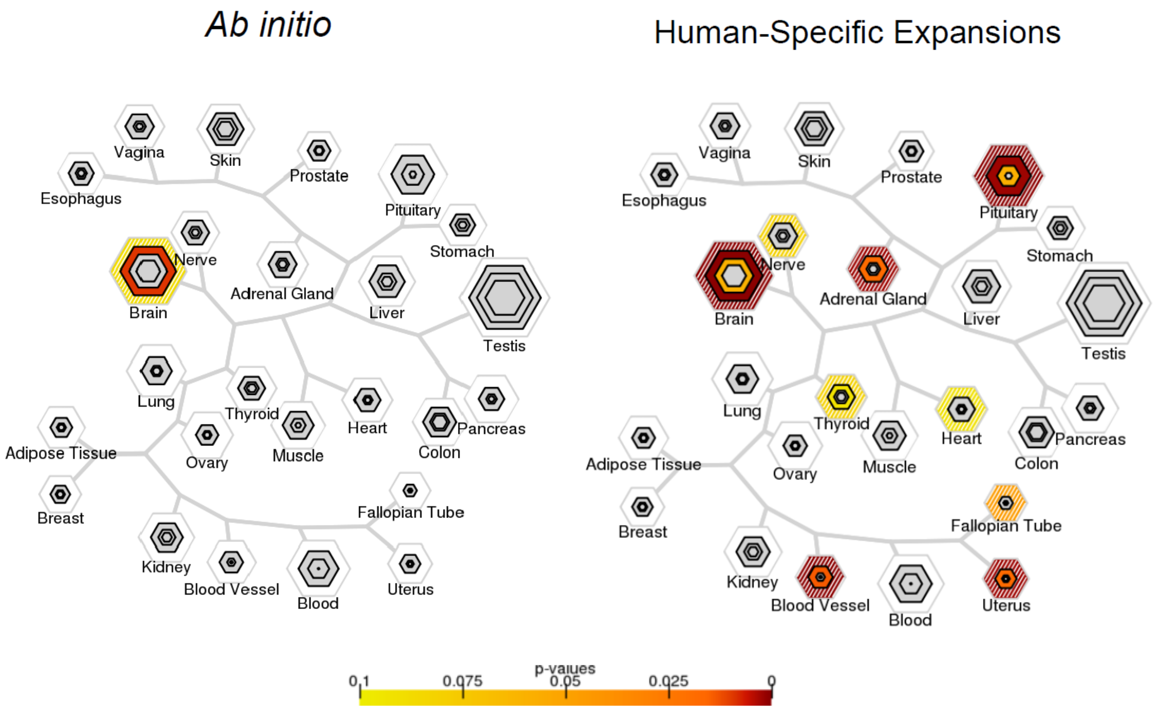
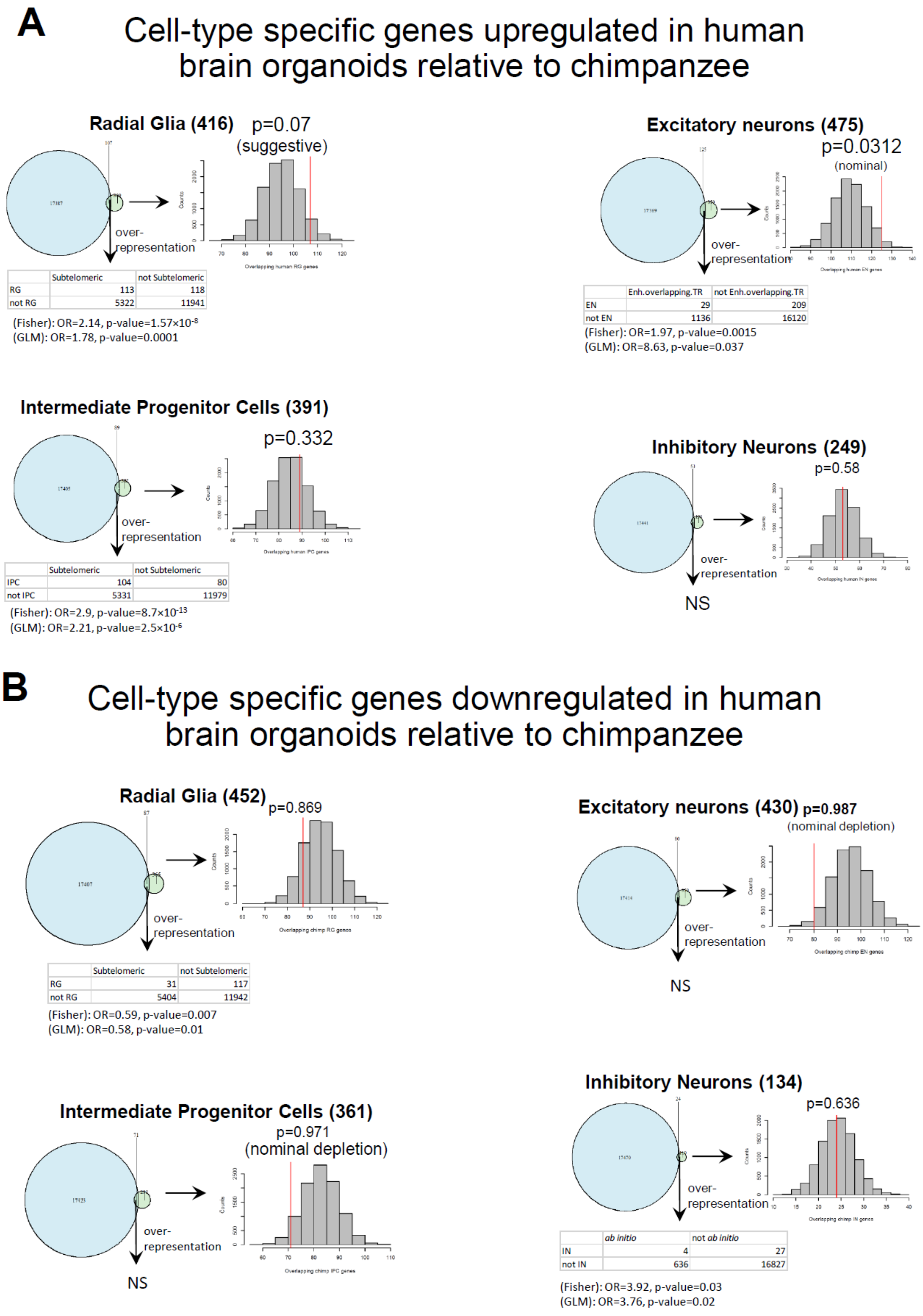


Fig. S5. Tissue-specific enrichment analysis. Both *ab initio* and HSE intronic events are enriched for brain-specific genes. The cell-specific enrichment tool (CSEA(19)) was used for the generation of both plots. For both bullseye plots, different stringency thresholds are shown by up to four concentric hexagons, where the “hotter” colors correspond to more significant the p-values.



A

Cell-type specific genes upregulated in human brain organoids relative to chimpanzee

Radial Glia (416) p=0.07 (suggestive)

	Subtelomeric	not Subtelomeric
RG	113	118
not RG	5322	11941

(Fisher): OR=2.14, p-value=1.57×10^{-8}
(GLM): OR=1.78, p-value=0.0001

Excitatory neurons (475) p=0.0312 (nominal)

	Enh.overlapping.TR	not Enh.overlapping.TR
EN	29	209
not EN	1136	16120

(Fisher): OR=1.97, p-value=0.0015
(GLM): OR=8.63, p-value=0.037

Intermediate Progenitor Cells (391) p=0.332

	Subtelomeric	not Subtelomeric
IPC	104	80
not IPC	5331	11979

(Fisher): OR=2.9, p-value=8.7×10^{-13}
(GLM): OR=2.21, p-value=2.5×10^{-6}

Inhibitory Neurons (249) p=0.58

B

Cell-type specific genes downregulated in human brain organoids relative to chimpanzee

Radial Glia (452) p=0.869

	Subtelomeric	not Subtelomeric
RG	31	117
not RG	5404	11942

(Fisher): OR=0.59, p-value=0.007
(GLM): OR=0.58, p-value=0.01

Excitatory neurons (430) p=0.987 (nominal depletion)

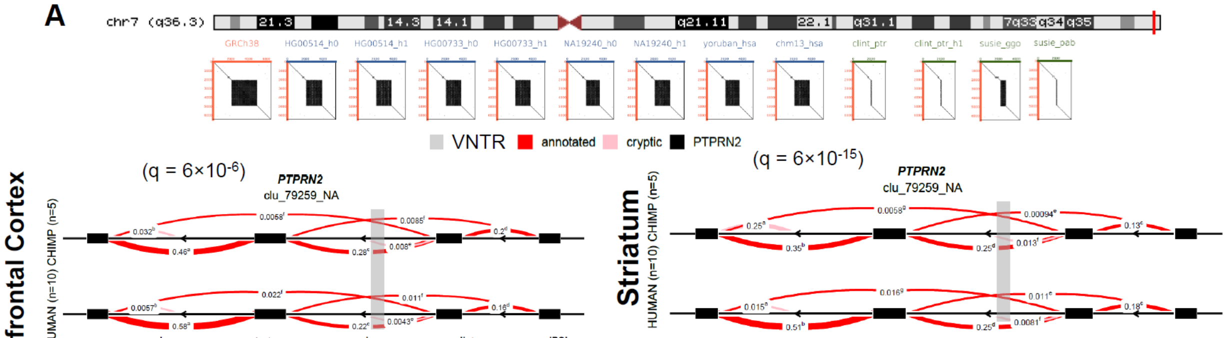
Intermediate Progenitor Cells (361) p=0.971 (nominal depletion)

Inhibitory Neurons (134) p=0.636

	ab initio	not *ab initio*
IN	4	27
not IN	636	16827

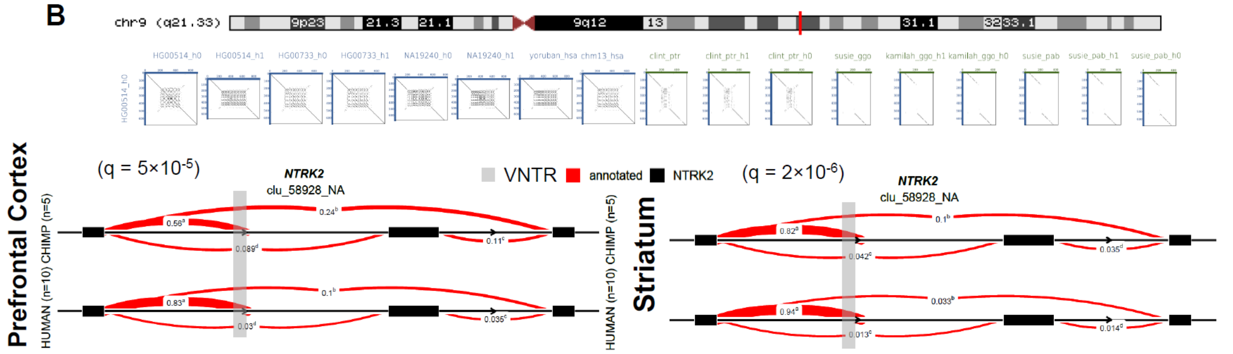
(Fisher): OR=3.92, p-value=0.03
(GLM): OR=3.76, p-value=0.02

Fig. S6. Association between STRs/VNTRs and cell-specific gene expression. A visual summary of the cell-type-specific enrichment analyses. The two plots represent the **A**) human brain organoid upregulated genes and **B**) human brain organoid downregulated genes, both with respect to the chimpanzee brain organoid gene levels. In each category we tested for enrichment with four different cell types: RG, EN, IPC and IN. The intersection of our STRs/VNTRs (n=17,494) and these gene lists were evaluated from overrepresentation using both a generalized linear model (i.e., GLM; **Methods**) as well as a Fisher's exact test for count data. Additionally, each histogram represents the null distribution of the expected overlaps between our set of tandem repeats and each respective gene list, determined by 100,000 genome-wide shuffles. After multiple-testing correction, we observe a significant overrepresentation of STRs/VNTRs for genes that are human upregulated in radial glia and intermediate progenitor cells, as supported by at least one of the three statistical models.



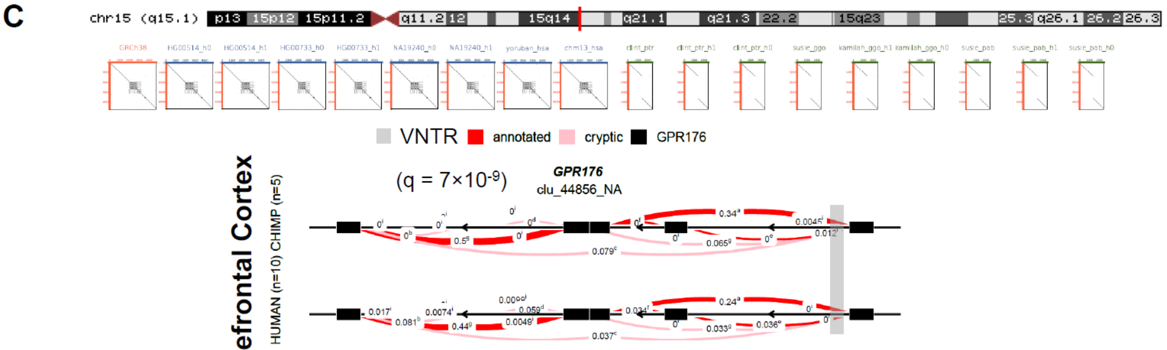
		chr	start	end	verdict	dPSI
	a	chr7	158205273	158316819	annotated	0.091
	b	chr7	158205273	158249607	cryptic_fiveprime	−0.050
→	c	chr7	158316932	158489735	annotated	−0.048
	d	chr7	158489785	158587558	annotated	0.009
→	e	chr7	158449067	158489735	cryptic_threeprime	−0.003
	f	chr7	158205273	158489735	annotated	0.000
	g	chr7	158316932	158587558	annotated	0.000

		chr	start	end	verdict	dPSI
	a	chr7	158205273	158249607	cryptic_fiveprime	−0.240
	b	chr7	158205273	158316819	annotated	0.199
	c	chr7	158489785	158587558	annotated	0.063
→	d	chr7	158316932	158489735	annotated	−0.021
	e	chr7	158316932	158587558	annotated	0.010
→	f	chr7	158449067	158489735	cryptic_threeprime	−0.007
	g	chr7	158205273	158489735	annotated	−0.005

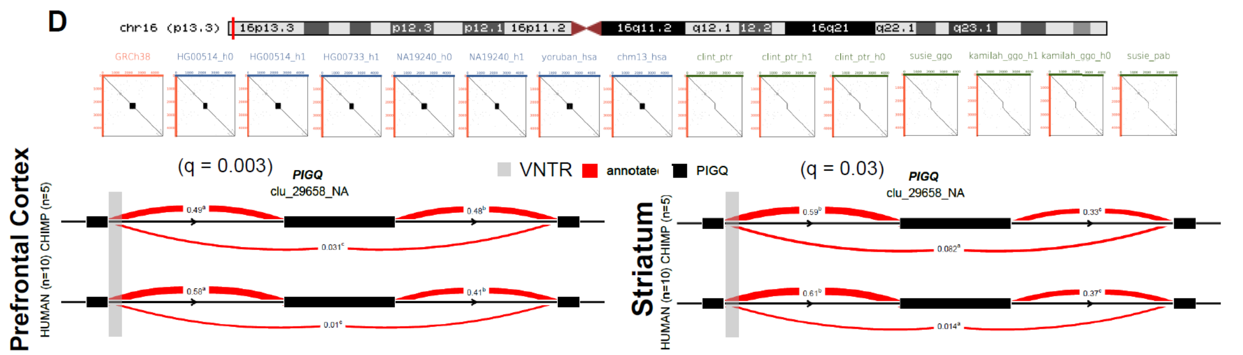


		chr	start	end	verdict	dPSI
→	a	chr9	84752085	84810542	annotated	0.290
	b	chr9	84752085	84867243	annotated	−0.153
	c	chr9	84861087	84867243	annotated	−0.073
→	d	chr9	84752085	84861040	annotated	−0.065

		chr	start	end	verdict	dPSI
→	a	chr9	84752085	84810542	annotated	0.123
	b	chr9	84752085	84867243	annotated	−0.070
→	c	chr9	84752085	84861040	annotated	−0.027
	d	chr9	84861087	84867243	annotated	−0.026



		chr	start	end	verdict	dPSI
→	a	chr15	39807258	39919855	annotated	−0.073
	b	chr15	39802254	39805220	cryptic_fiveprime	0.061
→	c	chr15	39802254	39919855	novel annotated pair	−0.061
	d	chr15	39805404	39807006	cryptic_threeprime	0.058
→	e	chr15	39860922	39919855	annotated	0.041
	f	chr15	39807258	39860820	annotated	0.036
→	g	chr15	39802254	39807006	annotated	−0.031
	h	chr15	39807197	39919855	novel annotated pair	−0.031



		chr	start	end	verdict	dPSI
→	*a*	chr16	579180	580183	annotated	0.110
	b	chr16	580263	580858	annotated	−0.093
→	*c*	chr16	579180	580858	annotated	−0.018

		chr	start	end	verdict	dPSI
→	*a*	chr16	579180	580858	annotated	−0.053
	b	chr16	579180	580183	annotated	0.031
→	*c*	chr16	580263	580858	annotated	0.022

Fig. S7. Differential splicing analysis. RNA-seq analysis identifies differential isoform usage between human and chimpanzee brains tissues over human-specific tandem repeat expansions. These four VNTRs were predicted to overlap with splice variants according to SpliceAI predictions(20). HSE tandem repeats harboring potential splice variants were located in the introns of *PTPRN2* (**A**) and *NTRK2* (**B**), while two *ab initio* VNTRs contained splice variants for *GPR176* (**C**) and *PIGQ* (**D**). The corresponding donor gain SpliceAI scores were consistently the highest of all four possible splicing variants categories: *PTPRN2* (delta score = 0.6951), *NTRK2* (delta score = 0.8607), *GPR176* (delta score = 0.6936), and *PIGQ* (delta score = 0.8567). For each gene we show the results from LeafCutter(16), which include dPSI (i.e., delta PSI, which is the difference in the proportions of reads spliced in at a given splice junction between human and chimpanzee samples) and FDR-corrected differential splicing significance (i.e., q-value). A positive dPSI suggests a higher PSI value for a given splice junction in human tissues compared to chimpanzee, and vice versa. The black horizontal arrows to the left of the table indicate the exon-exon junction that overlaps with the tandem repeat displayed at the top of each plot. *GPR176* is differentially spliced in cortex but not in striatum.

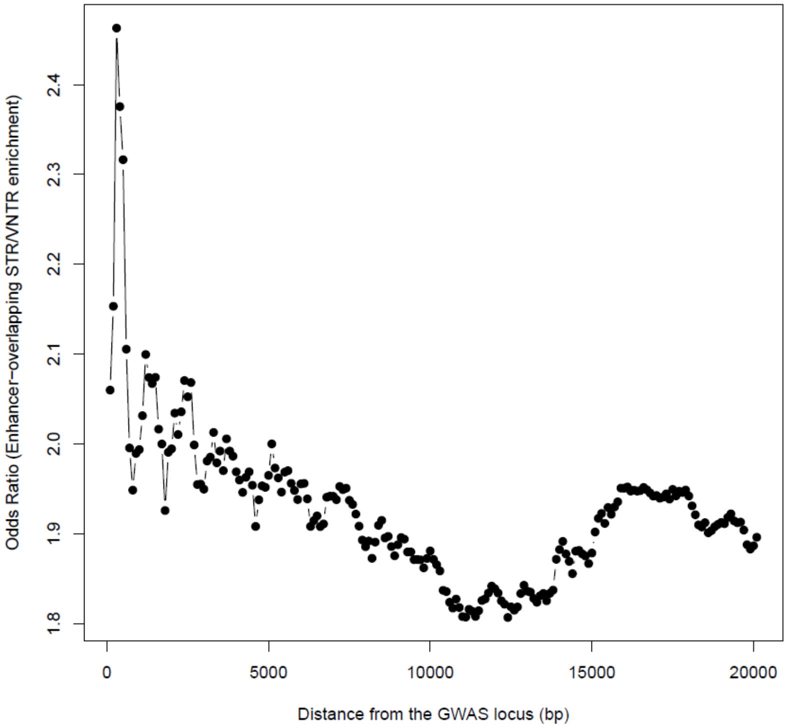
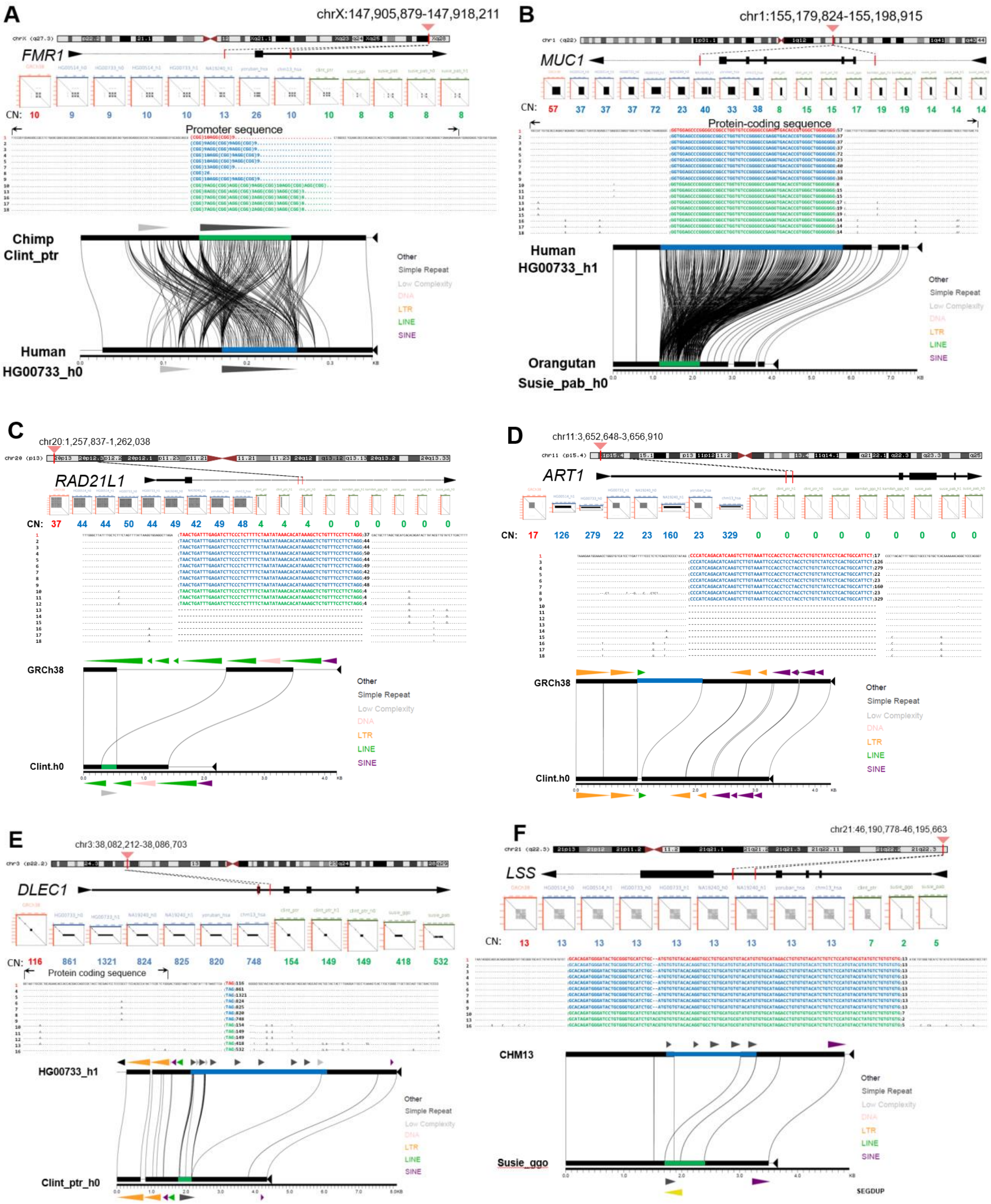
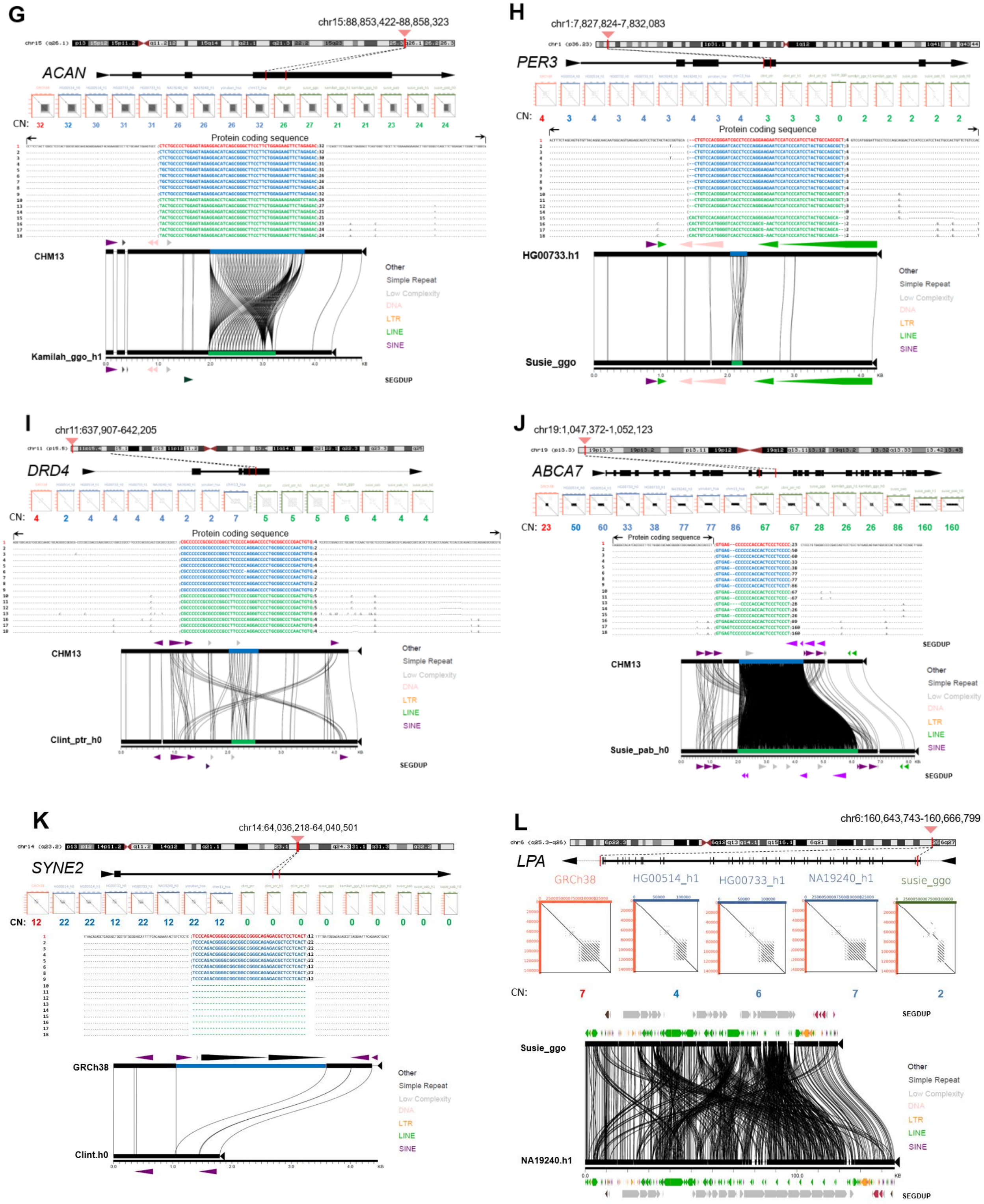


Fig. S8. Proximity of enhancer-overlapping STRs/VNTRs to GWAS signals. Enhancer-overlapping tandem repeats are enriched for GWAS SNPs, compared to non-enhancer-overlapping tandem repeats. The y-axis indicates the odds ratio of the Fisher's exact test conducted in every window around the GWAS SNP. We used 100 bp windows and increased the window size by 100 bp each time until we reached 20 kbp from the GWAS SNP. The null observation for the GWAS-enhancer overlap was 9.17%, based on 1,000 permutations of the enhancer region coordinates, corresponding to a significant enrichment (OR=1.47 [1.42-1.52], p-value < 2.2×10^{-16}).


A
chrX:147,905,879-147,918,211
FMR1
CN: 10 9 9 10 10 13 26 10 10 8 8 8 8
Promoter sequence
Chimp
Clint_ptr
Human
HG00733_h0
Other
Simple Repeat
Low Complexity
DNA
LTR
LINE
SINE
B
chr1:155,179,824-155,198,915
MUC1
CN: 57 37 37 37 72 23 40 33 38 8 15 15 17 19 19 14 14 14
Protein-coding sequence
Human
HG00733_h1
Orangutan
Susie_pab_h0
Other
Simple Repeat
Low Complexity
DNA
LTR
LINE
SINE
C
chr20:1,257,837-1,262,038
RAD21L1
CN: 37 44 44 50 44 49 42 49 48 4 4 4 0 0 0 0 0 0
GRCh38
Clint.h0
Other
Simple Repeat
Low Complexity
DNA
LTR
LINE
SINE
D
chr11:3,652,648-3,656,910
ART1
CN: 17 126 279 22 23 160 23 329 0 0 0 0 0 0 0 0 0
GRCh38
Clint.h0
Other
Simple Repeat
Low Complexity
DNA
LTR
LINE
SINE
E
chr3:38,082,212-38,086,703
DLEC1
CN: 116 861 1321 824 825 820 748 154 149 149 418 532
Protein coding sequence
HG00733_h1
Clint_ptr_h0
Other
Simple Repeat
Low Complexity
DNA
LTR
LINE
SINE
F
chr21:46,190,778-46,195,663
LSS
CN: 13 13 13 13 13 13 13 13 13 7 2 5
CHM13
Susie_ggo
Other
Simple Repeat
Low Complexity
DNA
LTR
LINE
SINE
SEGDUP


G
chr15:88,853,422-88,858,323
ACAN
CN: 32 32 30 31 31 26 26 26 32 26 27 21 21 23 24 24
Protein coding sequence
CHM13
Kamilah_ggo_h1
SEGDUP
H
chr1:7,827,824-7,832,083
PER3
CN: 4 3 4 3 4 3 4 3 4 3 3 3 0 2 2 2 2 2
Protein coding sequence
HG00733.h1
Susie_ggo
I
chr11:637,907-642,205
DRD4
CN: 4 2 4 4 4 4 2 2 7 5 5 5 6 4 4 4
Protein coding sequence
CHM13
Clint_ptr_h0
SEGDUP
J
chr19:1,047,372-1,052,123
ABCA7
CN: 23 50 60 33 38 77 77 86 67 67 28 26 26 86 160 160
Protein coding sequence
SEGDUP
CHM13
Susie_pab_h0
SEGDUP
K
chr14:64,036,218-64,040,501
SYNE2
CN: 12 22 22 12 22 12 22 12 0 0 0 0 0 0 0 0 0
GRCh38
Clint.h0
L
chr6:160,643,743-160,666,799
LPA
GRCh38
HG00514_h1
HG00733_h1
NA19240_h1
susie_ggo
CN: 7 4 6 7 2
SEGDUP
Susie_ggo
NA19240.h1
SEGDUP
Other
Simple Repeat
Low Complexity
DNA
LTR
LINE
SINE

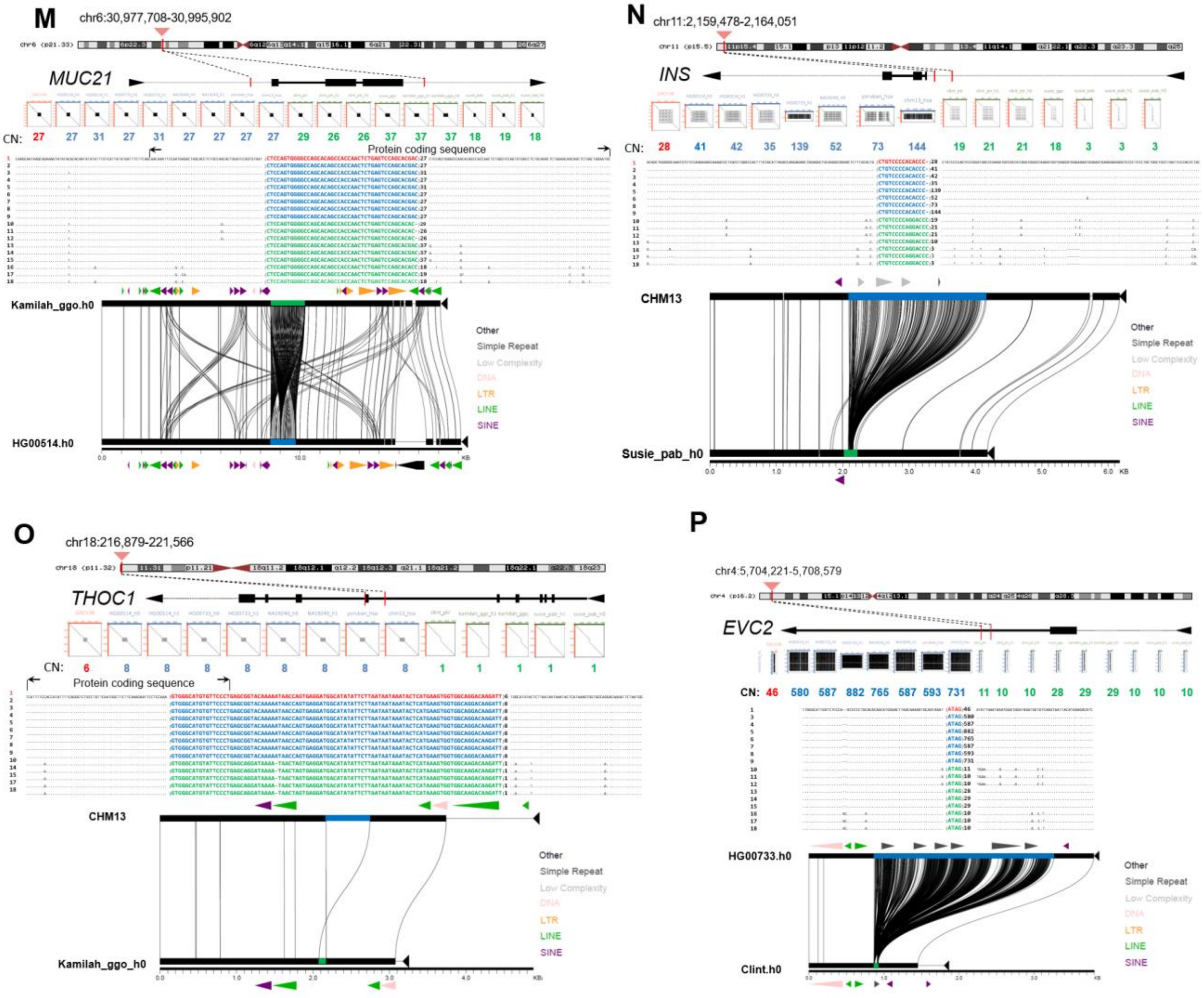
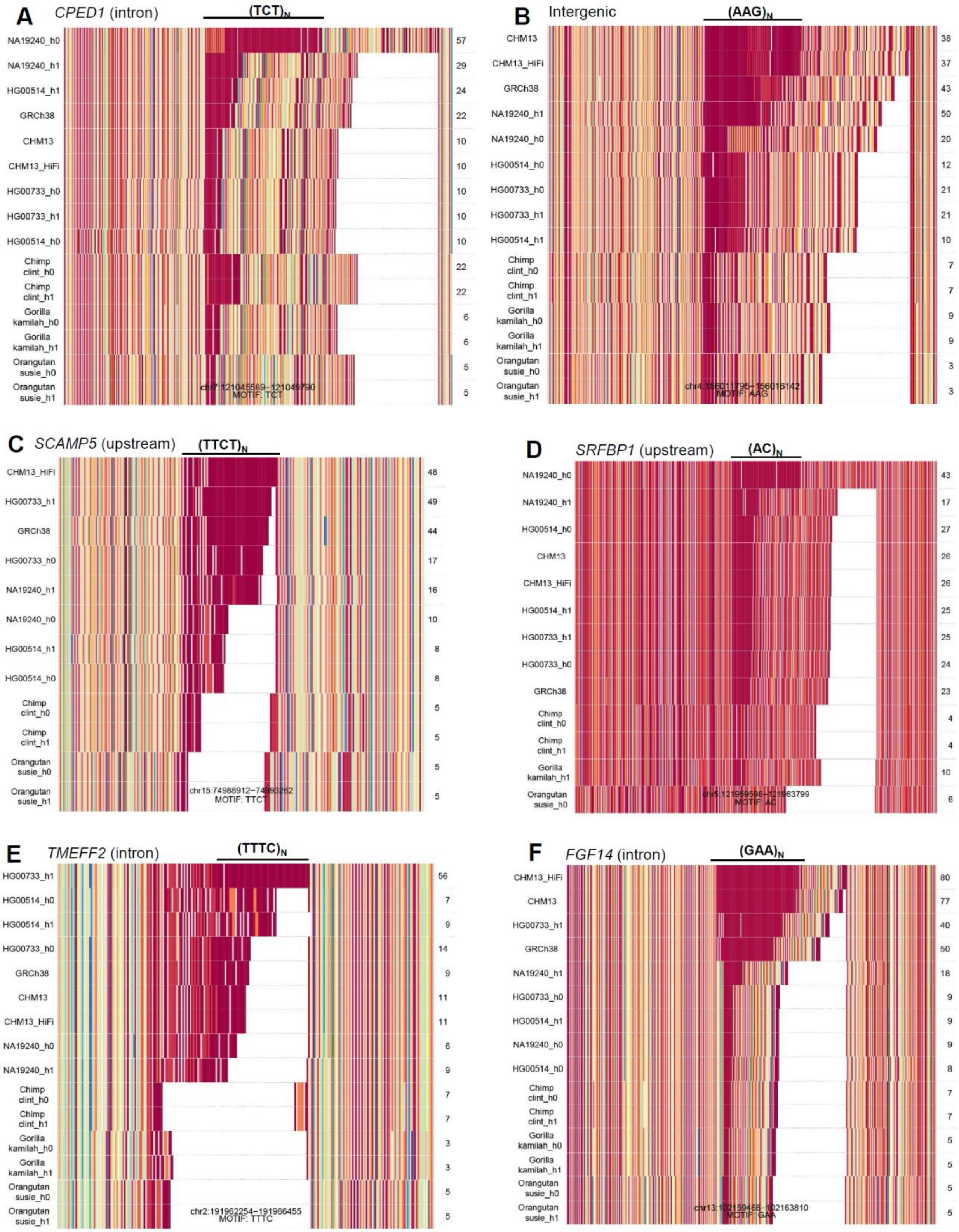
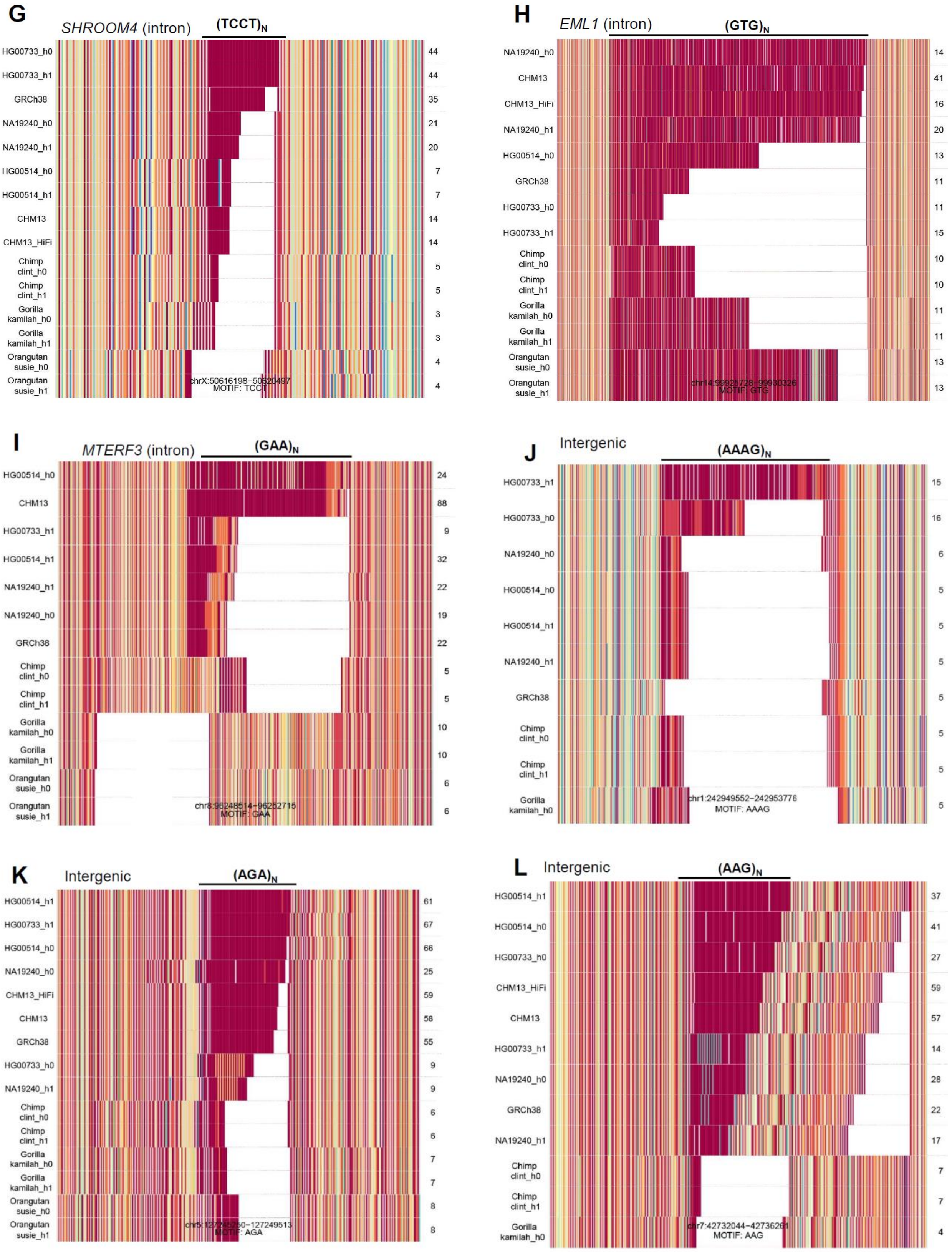
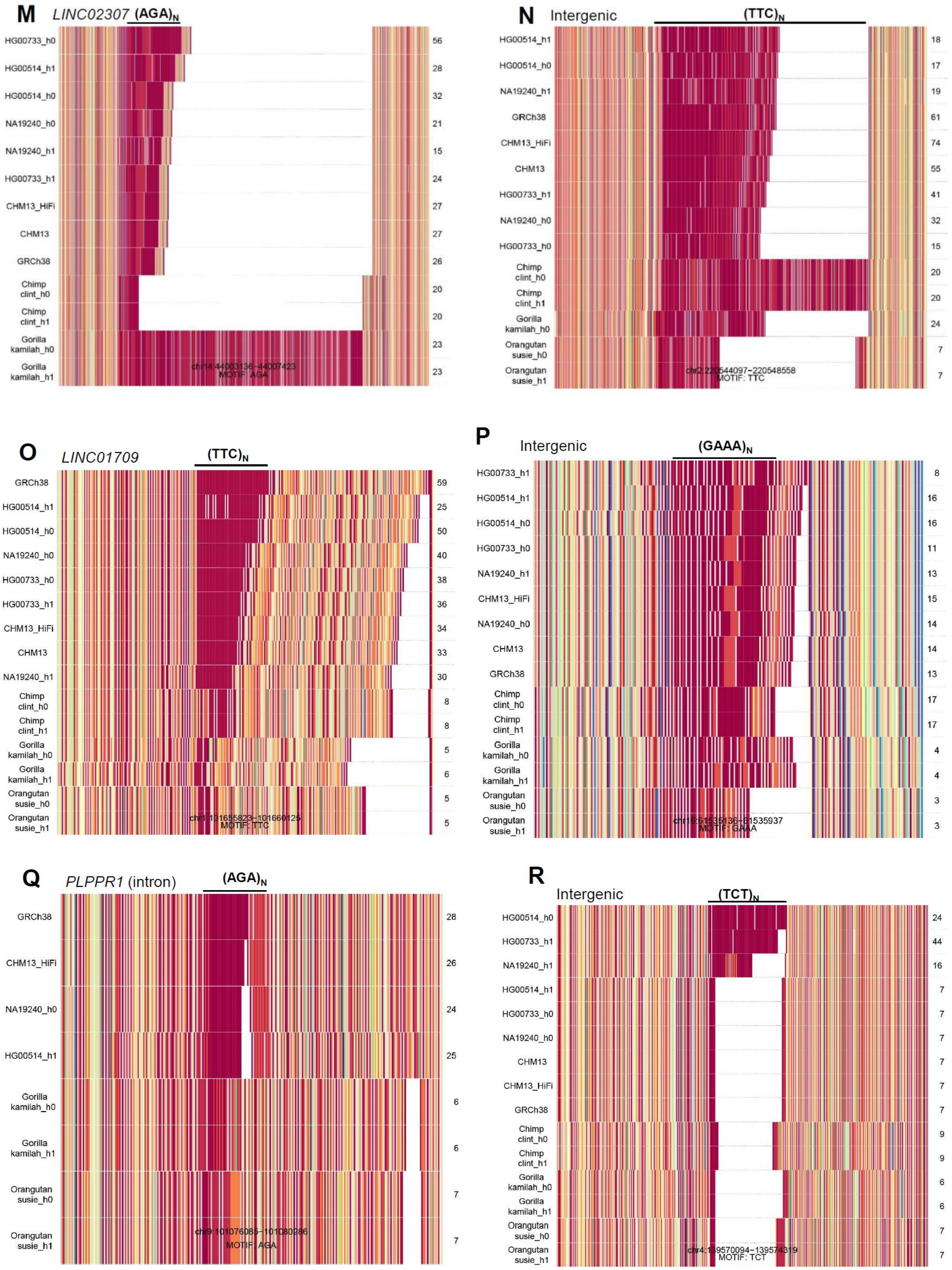

M
chr6:30,977,708-30,995,902
MUC21
CN: 27 27 31 27 31 27 27 27 27 29 26 26 37 37 37 18 19 18
Protein coding sequence
Kamilah_ggo.h0
HG00514.h0
N
chr11:2,159,478-2,164,051
INS
CN: 28 41 42 35 139 52 73 144 19 21 21 18 3 3 3
CHM13
Susie_pab_h0
O
chr18:216,879-221,566
THOC1
CN: 6 8 8 8 8 8 8 8 8 1 1 1 1 1
Protein coding sequence
CHM13
Kamilah_ggo_h0
P
chr4:5,704,221-5,708,579
EVC2
CN: 46 580 587 882 765 587 593 731 11 10 10 28 29 29 10 10 10
HG00733.h0
Clint.h0
Other
Simple Repeat
Low Complexity
DNA
LTR
LINE
SINE

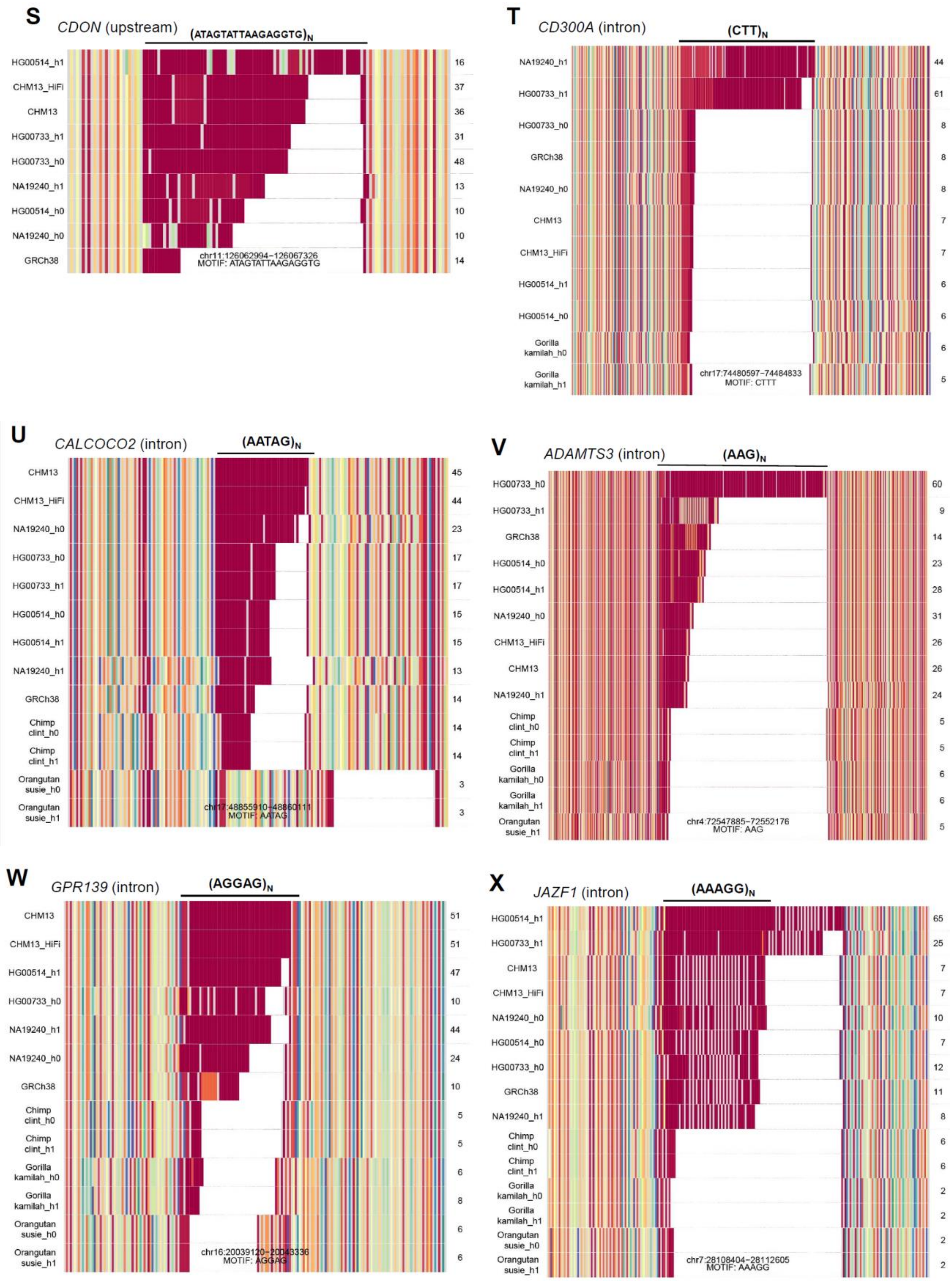
Fig. S9. Sequence annotation of human-expanded and disease-associated STRs/VNTRs. A collection of visualizations and MSAs for *ab initio*, HSE, and disease-associated tandem repeats. The left to right order of samples in the dotplots corresponds to the top to bottom order in the MSA. The numbers to the left of the MSA correspond to the following samples: 1 - GRCh38, 2 - HG00514 h0, 3 - HG00514 h1, 4 - HG00733 h0, 5 - HG00733 h1, 6 - NA19240 h0, 7 - NA19240 h1, 8 - Yoruban human assembly (collapsed), 9 - CHM13, 10 - chimpanzee collapsed, 11 - chimpanzee h0, 12 - chimpanzee h1, 13 - gorilla collapsed, 14 - gorilla h0, 15 - gorilla h1, 16 - orangutan collapsed, 17 - orangutan h0, and 18 - orangutan h1. "CN" represents the copy number of the tandem repeat. **A)** The trinucleotide repeat for the fragile X locus, which is located in the promoter of *FMR1*; the longest uninterrupted tract of CGG repeats is 13 copies in humans and 10 copies in the NHPs. **B)** An HSE VNTR overlaps directly with the protein-coding portion of *MUC1*; the largest human haplotype contains 72 tandem repeat copies of a 60-mer, and the largest NHP haplotype contains 19 copies of the same motif. A 1 bp insertion in one of the copies of this 60-mer has been shown to cause medullary cystic kidney disease type I(11). **C)** An HSE in the intron of *RAD21L1*, with the largest human allele measuring 50 copies of a 63-mer, and the largest NHP allele containing 4 tandem copies. **D)** A large *ab initio* VNTR in the intron

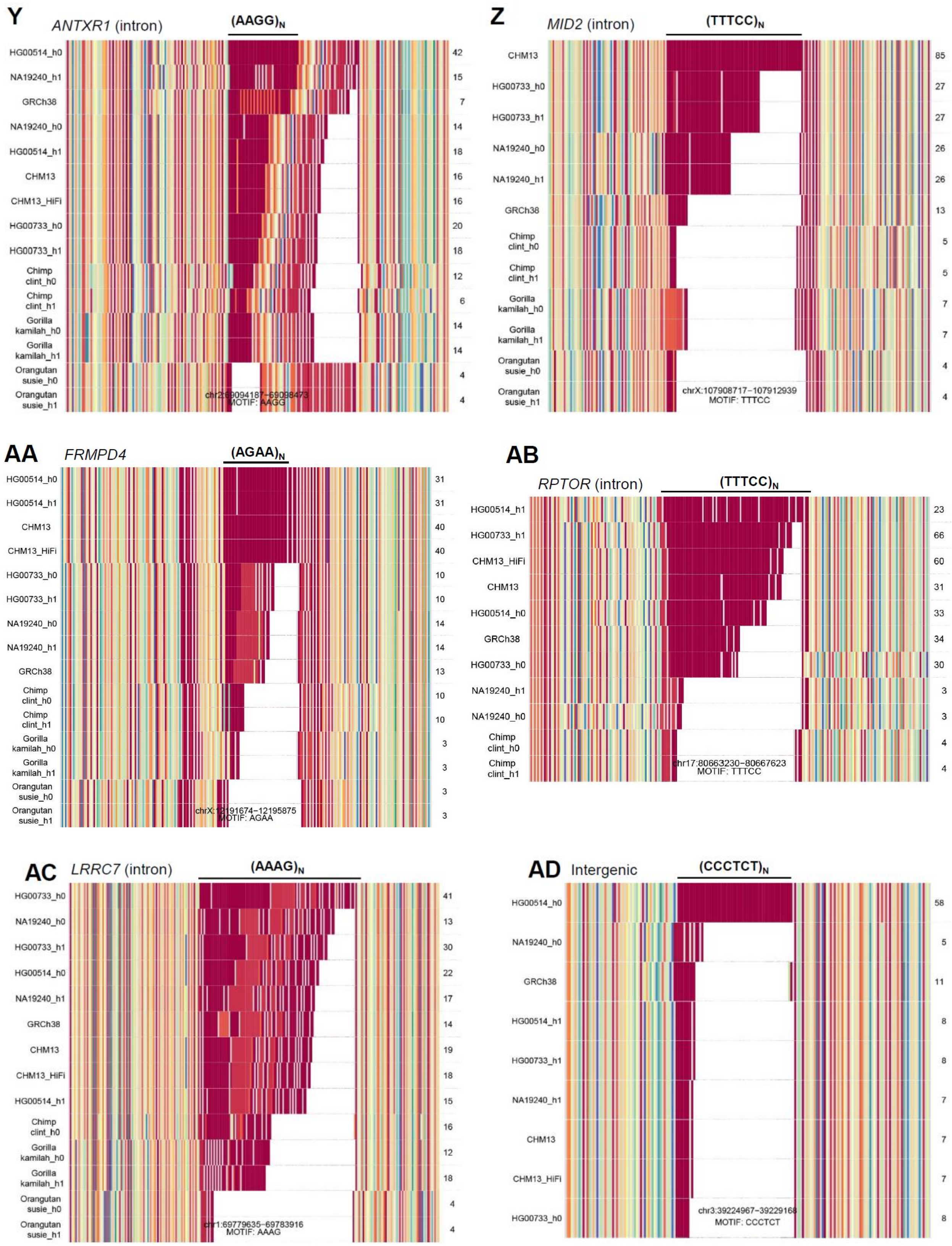
of *ART1* is not associated with retrotransposable elements, and it contains 329 tandem copies of a 63-mer in the largest human haplotype. **E)** A large HSE STR is located near the splice site of *DLEC1*; it is composed of 1,310 tandem repeat copies of a trinucleotide in the largest human haplotype and 415 copies of the same trinucleotide in the largest NHP haplotype. **F)** An HSE VNTR located in the intron of *LSS* contains 13 copies of a 117-mer in the largest human haplotype and 7 copies of a 119-mer in the largest NHP haplotype. **G)** A disease-associated VNTR is located in the exon of *ACAN* and contains 26-32 copies of a 57-mer in the human haplotypes and 21-24 copies in the NHP haplotypes. A contraction of this VNTR down to 13 copies has been associated with osteochondritis dissecans risk in humans(21). **H)** A disease-associated VNTR is located in the protein-coding portion of *PER3* and is composed of 3-4 copies of a 54-mer in the human haplotypes and 2-3 copies of a similar but not identical motif in the NHP haplotypes. The disease-associated event is an additional copy gain of the 54-mer motif, which is associated with an earlier age of onset for bipolar disorder(22). **I)** A disease-associated VNTR is located in the exon of *DRD4* and contains 2-7 copies of a 48-mer in the human haplotypes and 4-6 copies in the NHP haplotypes. The seven copy allele of this repeat has been associated with a higher risk of ADHD(23) and OCD(24) onset. **J)** An Alzheimer's disease risk VNTR is located directly in a splice site of *ABCA7*(25). The largest representation of this locus in human haplotypes is an 86-copy 25-mer, while the largest NHPs representation is 160 copies of a 27-mer. **K)** An *ab initio* VNTR in the intron of *SYNE2* is part of an SVA element, and all human alleles contain an invariable 12 copies of a 40-mer. **L)** The *LPA* Kringle-IV locus is a ~5.6 kbp motif, where each copy contains multiple exons of the *LPA* gene; it is likely collapsed in our assemblies, with the largest human haplotype containing 7 copies, which is consistent with the GRCh38 representation, and the collapsed reference of the gorilla genome, which contains 2 copies of the same motif. **M)** A disease-associated VNTR is located in the protein-coding portion of *MUC21* and is composed of 27-31 copies of a 45-mer in the human haplotypes and 18-37 copies of the same 45-mer or a 1 bp shorter motif in NHP haplotypes. A 4 bp deletion in this VNTR has been shown to increase disease risk for diffuse panbronchiolitis(26). **N)** The first VNTR reported in the literature(27) is also associated with type I diabetes(28). This repeat is located in the promoter region of *INS* and consists of 35-144 copies of a 14 bp motif in human haplotypes and 3-21 copies of a 15 bp motifs in the NHP haplotypes. **O)** An HSE and *ab initio* VNTR overlapping directly with the splice site of *THOC1* contains 8 and 1 copies of a 95- and 94-mer in the largest human and NHP haplotypes, respectively. **P)** A large HSE located in the intron of *EVC2* contains 882 copies of a tetranucleotide repeat in the largest human haplotype, compared to only 29 copies of the same motif in the largest NHP haplotype.


A
CPED1 (intron)
(TCT)N
NA19240_h0 57
NA19240_h1 29
HG00514_h1 24
GRCh38 22
CHM13 10
CHM13_HiFi 10
HG00733_h0 10
HG00733_h1 10
HG00514_h0 10
Chimp clint_h0 22
Chimp clint_h1 22
Gorilla kamilah_h0 6
Gorilla kamilah_h1 6
Orangutan susie_h0 5
Orangutan susie_h1 5
chr7:121045589–121049790
MOTIF: TCT
B
Intergenic
(AAG)N
CHM13 38
CHM13_HiFi 37
GRCh38 43
NA19240_h1 50
NA19240_h0 20
HG00514_h0 12
HG00733_h0 21
HG00733_h1 21
HG00514_h1 10
Chimp clint_h0 7
Chimp clint_h1 7
Gorilla kamilah_h0 9
Gorilla kamilah_h1 9
Orangutan susie_h0 3
Orangutan susie_h1 3
chr4:156011795–156016142
MOTIF: AAG
C
SCAMP5 (upstream)
(TTCT)N
CHM13_HiFi 48
HG00733_h1 49
GRCh38 44
HG00733_h0 17
NA19240_h1 16
NA19240_h0 10
HG00514_h1 8
HG00514_h0 8
Chimp clint_h0 5
Chimp clint_h1 5
Orangutan susie_h0 5
Orangutan susie_h1 5
chr15:74988912–74993262
MOTIF: TTCT
D
SRFBP1 (upstream)
(AC)N
NA19240_h0 43
NA19240_h1 17
HG00514_h0 27
CHM13 26
CHM13_HiFi 26
HG00514_h1 25
HG00733_h1 25
HG00733_h0 24
GRCh38 23
Chimp clint_h0 4
Chimp clint_h1 4
Gorilla kamilah_h1 10
Orangutan susie_h0 6
chr5:121959598–121963799
MOTIF: AC
E
TMEFF2 (intron)
(TTTC)N
HG00733_h1 56
HG00514_h0 7
HG00514_h1 9
HG00733_h0 14
GRCh38 9
CHM13 11
CHM13_HiFi 11
NA19240_h0 6
NA19240_h1 9
Chimp clint_h0 7
Chimp clint_h1 7
Gorilla kamilah_h0 3
Gorilla kamilah_h1 3
Orangutan susie_h0 5
Orangutan susie_h1 5
chr2:191962254–191966455
MOTIF: TTTC
F
FGF14 (intron)
(GAA)N
CHM13_HiFi 80
CHM13 77
HG00733_h1 40
GRCh38 50
NA19240_h1 18
HG00733_h0 9
HG00514_h1 9
NA19240_h0 9
HG00514_h0 8
Chimp clint_h0 7
Chimp clint_h1 7
Gorilla kamilah_h0 5
Gorilla kamilah_h1 5
Orangutan susie_h0 5
Orangutan susie_h1 5
chr13:102159466–102163810
MOTIF: GAA

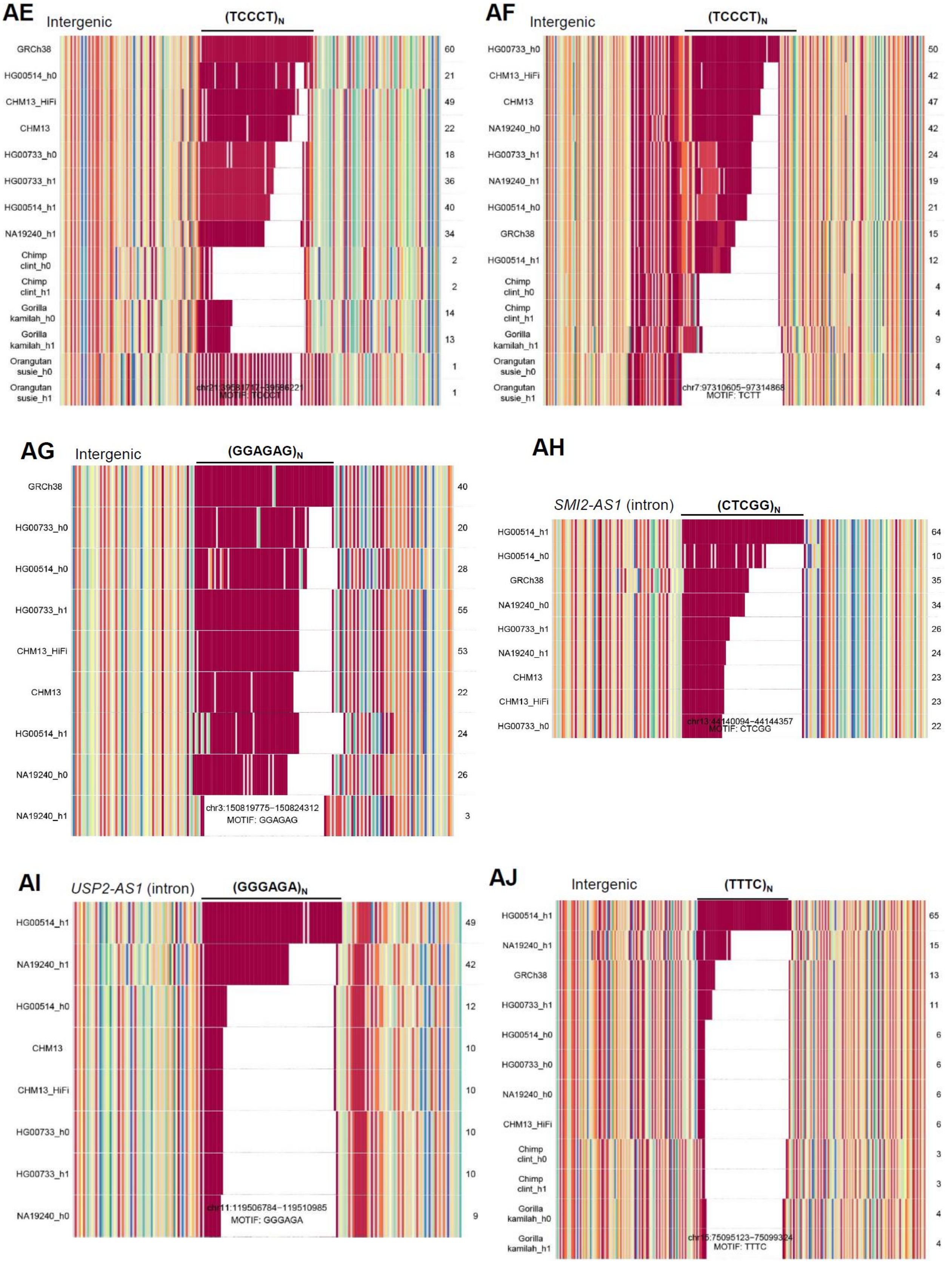

G
SHROOM4 (intron)
(TCCT)N
HG00733_h0 44
HG00733_h1 44
GRCh38 35
NA19240_h0 21
NA19240_h1 20
HG00514_h0 7
HG00514_h1 7
CHM13 14
CHM13_HiFi 14
Chimp clint_h0 5
Chimp clint_h1 5
Gorilla kamilah_h0 3
Gorilla kamilah_h1 3
Orangutan susie_h0 4
Orangutan susie_h1 4
chrX:50616198−50620497
MOTIF: TCCT
H
EML1 (intron)
(GTG)N
NA19240_h0 14
CHM13 41
CHM13_HiFi 16
NA19240_h1 20
HG00514_h0 13
GRCh38 11
HG00733_h0 11
HG00733_h1 15
Chimp clint_h0 10
Chimp clint_h1 10
Gorilla kamilah_h0 11
Gorilla kamilah_h1 11
Orangutan susie_h0 13
Orangutan susie_h1 13
chr14:99925728−99930326
MOTIF: GTG
I
MTERF3 (intron)
(GAA)N
HG00514_h0 24
CHM13 88
HG00733_h1 9
HG00514_h1 32
NA19240_h1 22
NA19240_h0 19
GRCh38 22
Chimp clint_h0 5
Chimp clint_h1 5
Gorilla kamilah_h0 10
Gorilla kamilah_h1 10
Orangutan susie_h0 6
Orangutan susie_h1 6
chr8:96248514−96252715
MOTIF: GAA
J
Intergenic
(AAAG)N
HG00733_h1 15
HG00733_h0 16
NA19240_h0 6
HG00514_h0 5
HG00514_h1 5
NA19240_h1 5
GRCh38 5
Chimp clint_h0 5
Chimp clint_h1 5
Gorilla kamilah_h0 5
chr1:242949552−242953776
MOTIF: AAAG
K
Intergenic
(AGA)N
HG00514_h1 61
HG00733_h1 67
HG00514_h0 66
NA19240_h0 25
CHM13_HiFi 59
CHM13 58
GRCh38 55
HG00733_h0 9
NA19240_h1 9
Chimp clint_h0 6
Chimp clint_h1 6
Gorilla kamilah_h0 7
Gorilla kamilah_h1 7
Orangutan susie_h0 8
Orangutan susie_h1 8
chr5:127245260−127249513
MOTIF: AGA
L
Intergenic
(AAG)N
HG00514_h1 37
HG00514_h0 41
HG00733_h0 27
CHM13_HiFi 59
CHM13 57
HG00733_h1 14
NA19240_h0 28
GRCh38 22
NA19240_h1 17
Chimp clint_h0 7
Chimp clint_h1 7
Gorilla kamilah_h0 4
chr7:42732044−42736261
MOTIF: AAG

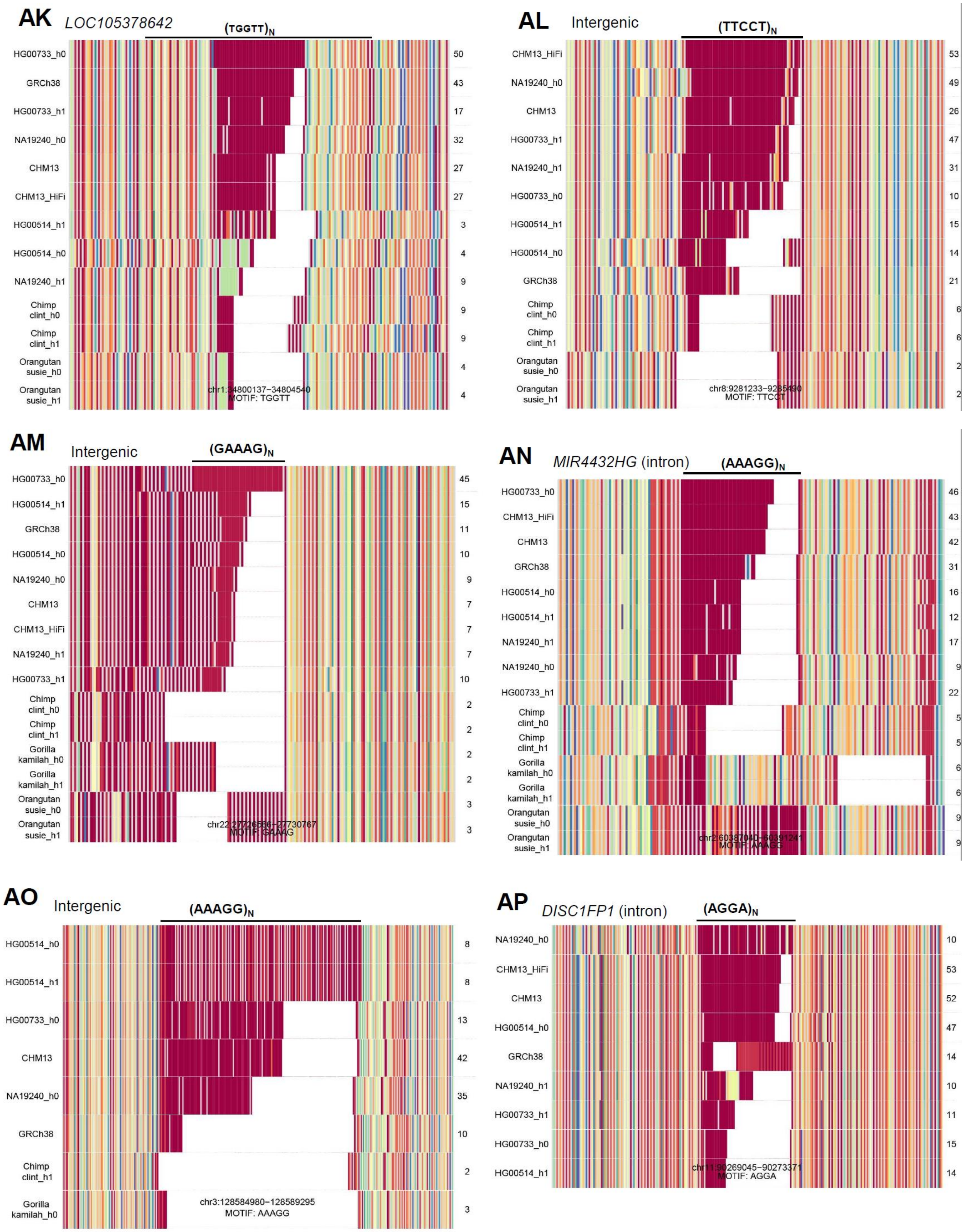

M
LINC02307
(AGA)$_N$
HG00733_h0 56
HG00514_h1 28
HG00514_h0 32
NA19240_h0 21
NA19240_h1 15
HG00733_h1 24
CHM13_HiFi 27
CHM13 27
GRCh38 26
Chimp clint_h0 20
Chimp clint_h1 20
Gorilla kamilah_h0 23
Gorilla kamilah_h1 23
chr14:44003136–44007423
MOTIF: AGA
N
Intergenic
(TTC)$_N$
HG00514_h1 18
HG00514_h0 17
NA19240_h1 19
GRCh38 61
CHM13_HiFi 74
CHM13 55
HG00733_h1 41
NA19240_h0 32
HG00733_h0 15
Chimp clint_h0 20
Chimp clint_h1 20
Gorilla kamilah_h0 24
Orangutan susie_h0 7
Orangutan susie_h1 7
chr2:220544097–220548558
MOTIF: TTC
O
LINC01709
(TTC)$_N$
GRCh38 59
HG00514_h1 25
HG00514_h0 50
NA19240_h0 40
HG00733_h0 38
HG00733_h1 36
CHM13_HiFi 34
CHM13 33
NA19240_h1 30
Chimp clint_h0 8
Chimp clint_h1 8
Gorilla kamilah_h0 5
Gorilla kamilah_h1 6
Orangutan susie_h0 5
Orangutan susie_h1 5
chr1:101655823–101660125
MOTIF: TTC
P
Intergenic
(GAAA)$_N$
HG00733_h1 8
HG00514_h1 16
HG00514_h0 16
HG00733_h0 11
NA19240_h1 13
CHM13_HiFi 15
NA19240_h0 14
CHM13 14
GRCh38 13
Chimp clint_h0 17
Chimp clint_h1 17
Gorilla kamilah_h0 4
Gorilla kamilah_h1 4
Orangutan susie_h0 3
Orangutan susie_h1 3
chr16:61535136–61535937
MOTIF: GAAA
Q
PLPPR1 (intron)
(AGA)$_N$
GRCh38 28
CHM13_HiFi 26
NA19240_h0 24
HG00514_h1 25
Gorilla kamilah_h0 6
Gorilla kamilah_h1 6
Orangutan susie_h0 7
Orangutan susie_h1 7
chr9:101076085–101080286
MOTIF: AGA
R
Intergenic
(TCT)$_N$
HG00514_h0 24
HG00733_h1 44
NA19240_h1 16
HG00514_h1 7
HG00733_h0 7
NA19240_h0 7
CHM13 7
CHM13_HiFi 7
GRCh38 7
Chimp clint_h0 9
Chimp clint_h1 9
Gorilla kamilah_h0 6
Gorilla kamilah_h1 6
Orangutan susie_h0 7
Orangutan susie_h1 7
chr4:139570094–139574319
MOTIF: TCT

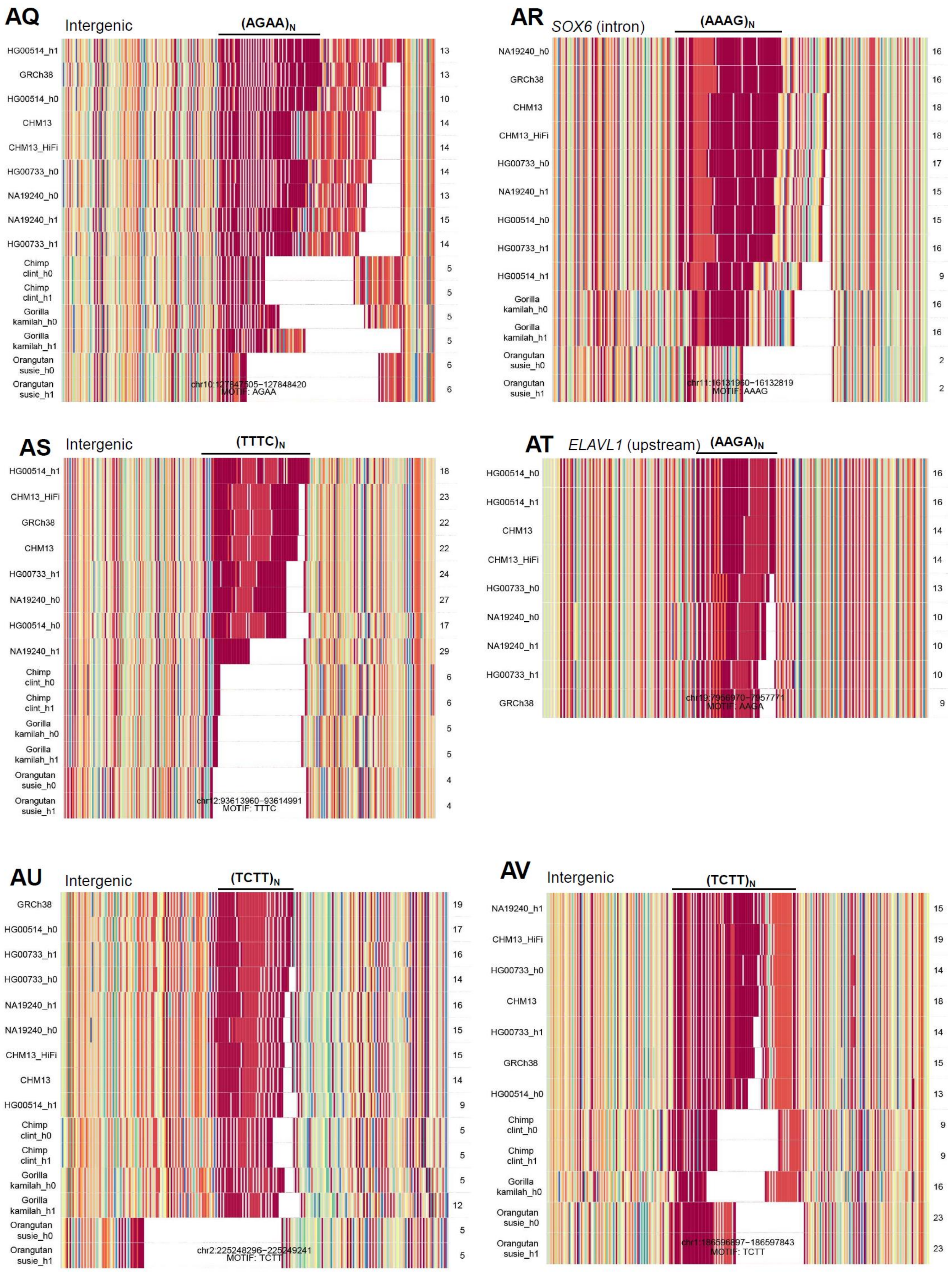

S
CDON (upstream)
(ATAGTATTAAGAGGTG)N
HG00514_h1 16
CHM13_HiFi 37
CHM13 36
HG00733_h1 31
HG00733_h0 48
NA19240_h1 13
HG00514_h0 10
NA19240_h0 10
GRCh38 14
chr11:126062994−126067326
MOTIF: ATAGTATTAAGAGGTG
T
CD300A (intron)
(CTT)N
NA19240_h1 44
HG00733_h1 61
HG00733_h0 8
GRCh38 8
NA19240_h0 8
CHM13 7
CHM13_HiFi 7
HG00514_h1 6
HG00514_h0 6
Gorilla kamilah_h0 6
Gorilla kamilah_h1 5
chr17:74480597−74484833
MOTIF: CTTT
U
CALCOCO2 (intron)
(AATAG)N
CHM13 45
CHM13_HiFi 44
NA19240_h0 23
HG00733_h0 17
HG00733_h1 17
HG00514_h0 15
HG00514_h1 15
NA19240_h1 13
GRCh38 14
Chimp clint_h0 14
Chimp clint_h1 14
Orangutan susie_h0 3
Orangutan susie_h1 3
chr17:48855910−48860111
MOTIF: AATAG
V
ADAMTS3 (intron)
(AAG)N
HG00733_h0 60
HG00733_h1 9
GRCh38 14
HG00514_h0 23
HG00514_h1 28
NA19240_h0 31
CHM13_HiFi 26
CHM13 26
NA19240_h1 24
Chimp clint_h0 5
Chimp clint_h1 5
Gorilla kamilah_h0 6
Gorilla kamilah_h1 6
Orangutan susie_h1 5
chr4:72547885−72552176
MOTIF: AAG
W
GPR139 (intron)
(AGGAG)N
CHM13 51
CHM13_HiFi 51
HG00514_h1 47
HG00733_h0 10
NA19240_h1 44
NA19240_h0 24
GRCh38 10
Chimp clint_h0 5
Chimp clint_h1 5
Gorilla kamilah_h0 6
Gorilla kamilah_h1 8
Orangutan susie_h0 6
Orangutan susie_h1 6
chr16:20039120−20043336
MOTIF: AGGAG
X
JAZF1 (intron)
(AAAGG)N
HG00514_h1 65
HG00733_h1 25
CHM13 7
CHM13_HiFi 7
NA19240_h0 10
HG00514_h0 7
HG00733_h0 12
GRCh38 11
NA19240_h1 8
Chimp clint_h0 6
Chimp clint_h1 6
Gorilla kamilah_h0 2
Gorilla kamilah_h1 2
Orangutan susie_h0 2
Orangutan susie_h1 2
chr7:28108404−28112605
MOTIF: AAAGG

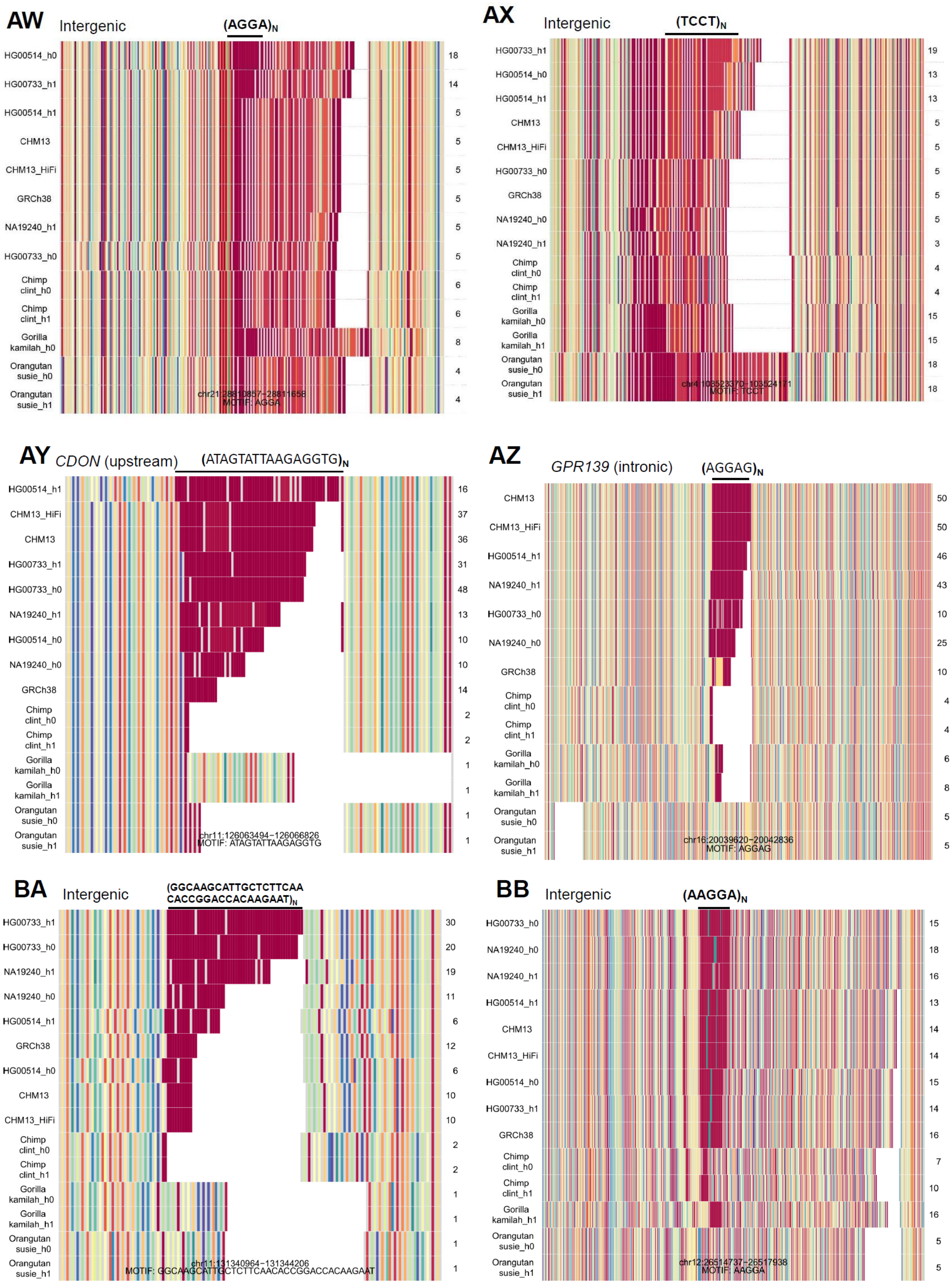


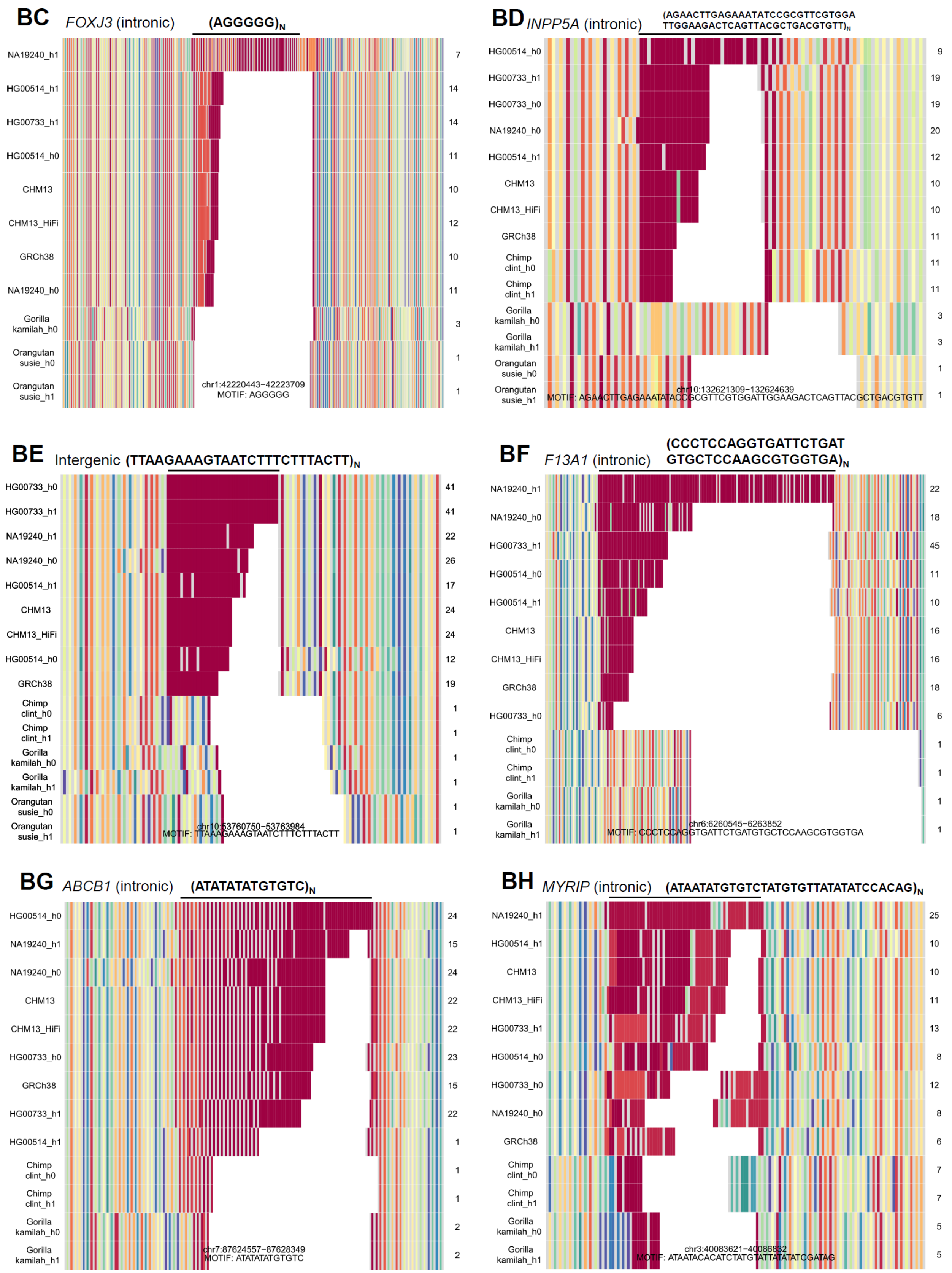
Y
ANTXR1 (intron)
$(AAGG)_N$
HG00514_h0 42
NA19240_h1 15
GRCh38 7
NA19240_h0 14
HG00514_h1 18
CHM13 16
CHM13_HiFi 16
HG00733_h0 20
HG00733_h1 18
Chimp clint_h0 12
Chimp clint_h1 6
Gorilla kamilah_h0 14
Gorilla kamilah_h1 14
Orangutan susie_h0 4
Orangutan susie_h1 4
chr2:69094187-69098473
MOTIF: AAGG
Z
MID2 (intron)
$(TTTCC)_N$
CHM13 85
HG00733_h0 27
HG00733_h1 27
NA19240_h0 26
NA19240_h1 26
GRCh38 13
Chimp clint_h0 5
Chimp clint_h1 5
Gorilla kamilah_h0 7
Gorilla kamilah_h1 7
Orangutan susie_h0 4
Orangutan susie_h1 4
chrX:107908717-107912939
MOTIF: TTTCC
AA
FRMPD4
$(AGAA)_N$
HG00514_h0 31
HG00514_h1 31
CHM13 40
CHM13_HiFi 40
HG00733_h0 10
HG00733_h1 10
NA19240_h0 14
NA19240_h1 14
GRCh38 13
Chimp clint_h0 10
Chimp clint_h1 10
Gorilla kamilah_h0 3
Gorilla kamilah_h1 3
Orangutan susie_h0 3
Orangutan susie_h1 3
chrX:12191674-12195875
MOTIF: AGAA
AB
RPTOR (intron)
$(TTTCC)_N$
HG00514_h1 23
HG00733_h1 66
CHM13_HiFi 60
CHM13 31
HG00514_h0 33
GRCh38 34
HG00733_h0 30
NA19240_h1 3
NA19240_h0 3
Chimp clint_h0 4
Chimp clint_h1 4
chr17:80663230-80667623
MOTIF: TTTCC
AC
LRRC7 (intron)
$(AAAG)_N$
HG00733_h0 41
NA19240_h0 13
HG00733_h1 30
HG00514_h0 22
NA19240_h1 17
GRCh38 14
CHM13 19
CHM13_HiFi 18
HG00514_h1 15
Chimp clint_h0 16
Gorilla kamilah_h0 12
Gorilla kamilah_h1 18
Orangutan susie_h0 4
Orangutan susie_h1 4
chr1:69779635-69783916
MOTIF: AAAG
AD
Intergenic
$(CCCTCT)_N$
HG00514_h0 58
NA19240_h0 5
GRCh38 11
HG00514_h1 8
HG00733_h1 8
NA19240_h1 7
CHM13 7
CHM13_HiFi 7
HG00733_h0 8
chr3:39224967-39229168
MOTIF: CCCTCT

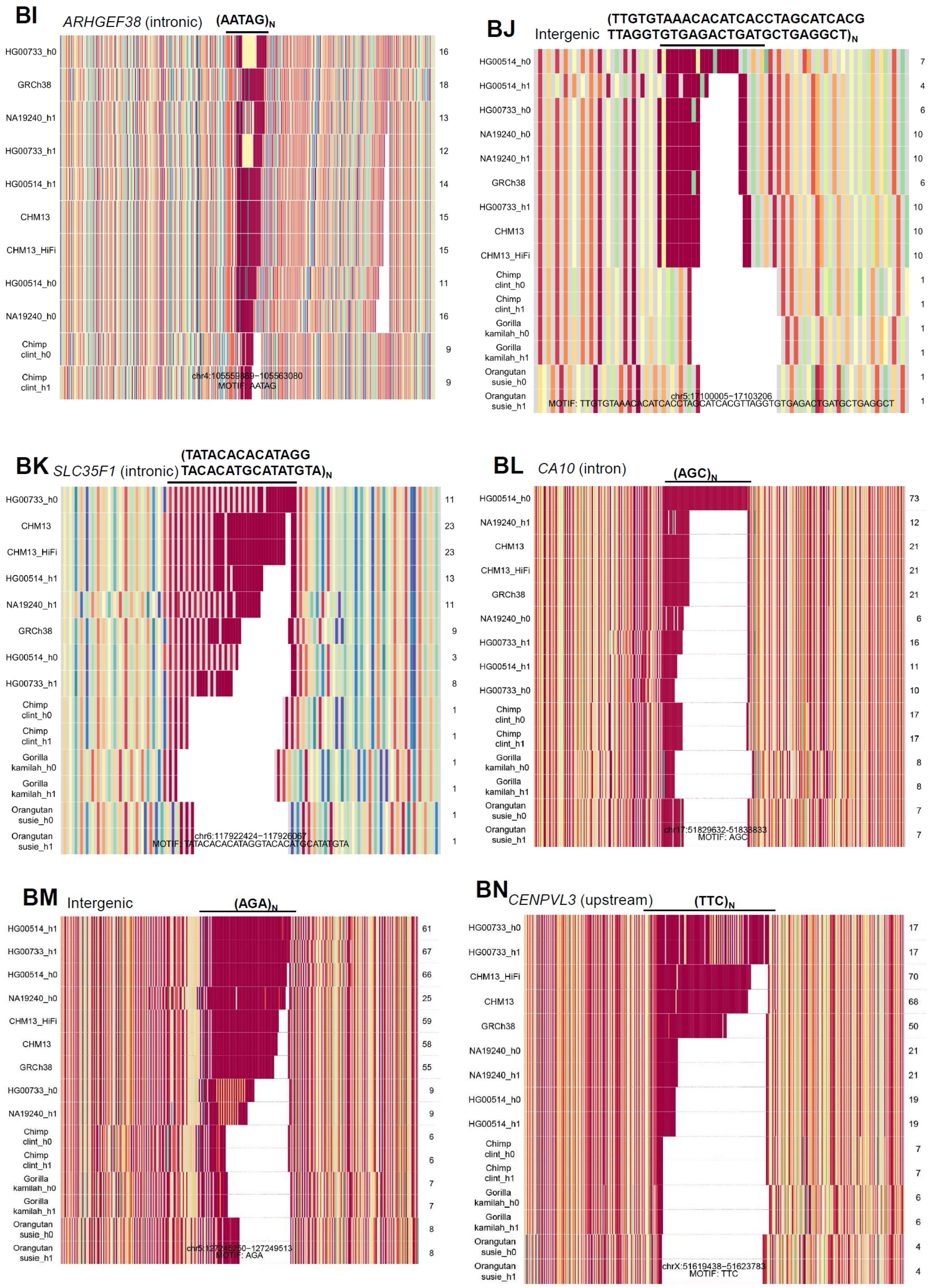

AE
Intergenic
(TCCCT)N
GRCh38
HG00514_h0
CHM13_HiFi
CHM13
HG00733_h0
HG00733_h1
HG00514_h1
NA19240_h1
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
60
21
49
22
18
36
40
34
2
2
14
13
1
1
chr21:39581717-39586221
MOTIF: TCCCT
AF
Intergenic
(TCCCT)N
HG00733_h0
CHM13_HiFi
CHM13
NA19240_h0
HG00733_h1
NA19240_h1
HG00514_h0
GRCh38
HG00514_h1
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
50
42
47
42
24
19
21
15
12
4
4
9
4
4
chr7:97310605-97314868
MOTIF: TCTT
AG
Intergenic
(GGAGAG)N
GRCh38
HG00733_h0
HG00514_h0
HG00733_h1
CHM13_HiFi
CHM13
HG00514_h1
NA19240_h0
NA19240_h1
40
20
28
55
53
22
24
26
3
chr3:150819775-150824312
MOTIF: GGAGAG
AH
SMI2-AS1 (intron)
(CTCGG)N
HG00514_h1
HG00514_h0
GRCh38
NA19240_h0
HG00733_h1
NA19240_h1
CHM13
CHM13_HiFi
HG00733_h0
64
10
35
34
26
24
23
23
22
chr13:44140094-44144357
MOTIF: CTCGG
AI
USP2-AS1 (intron)
(GGGAGA)N
HG00514_h1
NA19240_h1
HG00514_h0
CHM13
CHM13_HiFi
HG00733_h0
HG00733_h1
NA19240_h0
49
42
12
10
10
10
10
9
chr11:119506784-119510985
MOTIF: GGGAGA
AJ
Intergenic
(TTTC)N
HG00514_h1
NA19240_h1
GRCh38
HG00733_h1
HG00514_h0
HG00733_h0
NA19240_h0
CHM13_HiFi
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
65
15
13
11
6
6
6
6
3
3
4
4
chr15:75095123-75099324
MOTIF: TTTC


AK
LOC105378642
(TGGTT)N
HG00733_h0
GRCh38
HG00733_h1
NA19240_h0
CHM13
CHM13_HiFi
HG00514_h1
HG00514_h0
NA19240_h1
Chimp clint_h0
Chimp clint_h1
Orangutan susie_h0
Orangutan susie_h1
chr1:34800137−34804540
MOTIF: TGGTT
50
43
17
32
27
27
3
4
9
9
9
4
4
AL
Intergenic
(TTCCT)N
CHM13_HiFi
NA19240_h0
CHM13
HG00733_h1
NA19240_h1
HG00733_h0
HG00514_h1
HG00514_h0
GRCh38
Chimp clint_h0
Chimp clint_h1
Orangutan susie_h0
Orangutan susie_h1
chr8:9281233−9285490
MOTIF: TTCCT
53
49
26
47
31
10
15
14
21
6
6
2
2
AM
Intergenic
(GAAAG)N
HG00733_h0
HG00514_h1
GRCh38
HG00514_h0
NA19240_h0
CHM13
CHM13_HiFi
NA19240_h1
HG00733_h1
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
chr22:27726566−27730767
MOTIF: GAAAG
45
15
11
10
9
7
7
7
10
2
2
2
2
3
3
AN
MIR4432HG (intron)
(AAAGG)N
HG00733_h0
CHM13_HiFi
CHM13
GRCh38
HG00514_h0
HG00514_h1
NA19240_h1
NA19240_h0
HG00733_h1
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
chr2:60387040−60391241
MOTIF: AAAGG
46
43
42
31
16
12
17
9
22
5
5
6
6
9
9
AO
Intergenic
(AAAGG)N
HG00514_h0
HG00514_h1
HG00733_h0
CHM13
NA19240_h0
GRCh38
Chimp clint_h1
Gorilla kamilah_h0
chr3:128584980−128589295
MOTIF: AAAGG
8
8
13
42
35
10
2
3
AP
DISC1FP1 (intron)
(AGGA)N
NA19240_h0
CHM13_HiFi
CHM13
HG00514_h0
GRCh38
NA19240_h1
HG00733_h1
HG00733_h0
HG00514_h1
chr11:90269045−90273371
MOTIF: AGGA
10
53
52
47
14
10
11
15
14


AQ
Intergenic
$(AGAA)_N$
HG00514_h1 13
GRCh38 13
HG00514_h0 10
CHM13 14
CHM13_HiFi 14
HG00733_h0 14
NA19240_h0 13
NA19240_h1 15
HG00733_h1 14
Chimp clint_h0 5
Chimp clint_h1 5
Gorilla kamilah_h0 5
Gorilla kamilah_h1 5
Orangutan susie_h0 6
Orangutan susie_h1 6
chr10:127847505−127848420
MOTIF: AGAA
AR
SOX6 (intron)
$(AAAG)_N$
NA19240_h0 16
GRCh38 16
CHM13 18
CHM13_HiFi 18
HG00733_h0 17
NA19240_h1 15
HG00514_h0 15
HG00733_h1 16
HG00514_h1 9
Gorilla kamilah_h0 16
Gorilla kamilah_h1 16
Orangutan susie_h0 2
Orangutan susie_h1 2
chr11:16131960−16132819
MOTIF: AAAG
AS
Intergenic
$(TTTC)_N$
HG00514_h1 18
CHM13_HiFi 23
GRCh38 22
CHM13 22
HG00733_h1 24
NA19240_h0 27
HG00514_h0 17
NA19240_h1 29
Chimp clint_h0 6
Chimp clint_h1 6
Gorilla kamilah_h0 5
Gorilla kamilah_h1 5
Orangutan susie_h0 4
Orangutan susie_h1 4
chr12:93613960−93614991
MOTIF: TTTC
AT
ELAVL1 (upstream)
$(AAGA)_N$
HG00514_h0 16
HG00514_h1 16
CHM13 14
CHM13_HiFi 14
HG00733_h0 13
NA19240_h0 10
NA19240_h1 10
HG00733_h1 10
GRCh38 9
chr19:7956970−7957771
MOTIF: AAGA
AU
Intergenic
$(TCTT)_N$
GRCh38 19
HG00514_h0 17
HG00733_h1 16
HG00733_h0 14
NA19240_h1 16
NA19240_h0 15
CHM13_HiFi 15
CHM13 14
HG00514_h1 9
Chimp clint_h0 5
Chimp clint_h1 5
Gorilla kamilah_h0 5
Gorilla kamilah_h1 12
Orangutan susie_h0 5
Orangutan susie_h1 5
chr2:225248296−225249241
MOTIF: TCTT
AV
Intergenic
$(TCTT)_N$
NA19240_h1 15
CHM13_HiFi 19
HG00733_h0 14
CHM13 18
HG00733_h1 14
GRCh38 15
HG00514_h0 13
Chimp clint_h0 9
Chimp clint_h1 9
Gorilla kamilah_h0 16
Orangutan susie_h0 23
Orangutan susie_h1 23
chr1:186596897−186597843
MOTIF: TCTT


AW
Intergenic
(AGGA)N
HG00514_h0 18
HG00733_h1 14
HG00514_h1 5
CHM13 5
CHM13_HiFi 5
GRCh38 5
NA19240_h1 5
HG00733_h0 5
Chimp clint_h0 6
Chimp clint_h1 6
Gorilla kamilah_h0 8
Orangutan susie_h0 4
Orangutan susie_h1 4
chr21:28810857−28811658
MOTIF: AGGA
AX
Intergenic
(TCCT)N
HG00733_h1 19
HG00514_h0 13
HG00514_h1 13
CHM13 5
CHM13_HiFi 5
HG00733_h0 5
GRCh38 5
NA19240_h0 5
NA19240_h1 3
Chimp clint_h0 4
Chimp clint_h1 4
Gorilla kamilah_h0 15
Gorilla kamilah_h1 15
Orangutan susie_h0 18
Orangutan susie_h1 18
chr4:103523370−103524171
MOTIF: TCCT
AY
CDON (upstream)
(ATAGTATTAAGAGGTG)N
HG00514_h1 16
CHM13_HiFi 37
CHM13 36
HG00733_h1 31
HG00733_h0 48
NA19240_h1 13
HG00514_h0 10
NA19240_h0 10
GRCh38 14
Chimp clint_h0 2
Chimp clint_h1 2
Gorilla kamilah_h0 1
Gorilla kamilah_h1 1
Orangutan susie_h0 1
Orangutan susie_h1 1
chr11:126063494−126066826
MOTIF: ATAGTATTAAGAGGTG
AZ
GPR139 (intronic)
(AGGAG)N
CHM13 50
CHM13_HiFi 50
HG00514_h1 46
NA19240_h1 43
HG00733_h0 10
NA19240_h0 25
GRCh38 10
Chimp clint_h0 4
Chimp clint_h1 4
Gorilla kamilah_h0 6
Gorilla kamilah_h1 8
Orangutan susie_h0 5
Orangutan susie_h1 5
chr16:20039620−20042836
MOTIF: AGGAG
BA
Intergenic
(GGCAAGCATTGCTCTTCAA CACCGGACCACAAGAAT)N
HG00733_h1 30
HG00733_h0 20
NA19240_h1 19
NA19240_h0 11
HG00514_h1 6
GRCh38 12
HG00514_h0 6
CHM13 10
CHM13_HiFi 10
Chimp clint_h0 2
Chimp clint_h1 2
Gorilla kamilah_h0 1
Gorilla kamilah_h1 1
Orangutan susie_h0 1
Orangutan susie_h1 1
chr11:131340964−131344206
MOTIF: GGCAAGCATTGCTCTTCAACACCGGACCACAAGAAT
BB
Intergenic
(AAGGA)N
HG00733_h0 15
NA19240_h0 18
NA19240_h1 16
HG00514_h1 13
CHM13 14
CHM13_HiFi 14
HG00514_h0 15
HG00733_h1 14
GRCh38 16
Chimp clint_h0 7
Chimp clint_h1 10
Gorilla kamilah_h1 16
Orangutan susie_h0 5
Orangutan susie_h1 5
chr12:26514737−26517938
MOTIF: AAGGA


BC
FOXJ3 (intronic)
(AGGGGG)$_N$
NA19240_h1
HG00514_h1
HG00733_h1
HG00514_h0
CHM13
CHM13_HiFi
GRCh38
NA19240_h0
Gorilla kamilah_h0
Orangutan susie_h0
Orangutan susie_h1
chr1:42220443−42223709
MOTIF: AGGGGG
7
14
14
11
10
12
10
11
3
1
1
BD
INPP5A (intronic)
(AGAACTTGAGAAATATCCGCGTTCGTGGA
TTGGAAGACTCAGTTACGCTGACGTGTT)$_N$
HG00514_h0
HG00733_h1
HG00733_h0
NA19240_h0
HG00514_h1
CHM13
CHM13_HiFi
GRCh38
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
chr10:132621309−132624639
MOTIF: AGAACTTGAGAAATATACCGCGTTCGTGGATTGGAAGACTCAGTTACGCTGACGTGTT
9
19
19
20
12
10
10
11
11
11
3
3
1
1
BE
Intergenic (TTAAGAAAGTAATCTTTCTTTACTT)$_N$
HG00733_h0
HG00733_h1
NA19240_h1
NA19240_h0
HG00514_h1
CHM13
CHM13_HiFi
HG00514_h0
GRCh38
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
chr10:53760750−53763984
MOTIF: TTAAAGAAAGTAATCTTTCTTTACTT
41
41
22
26
17
24
24
12
19
1
1
1
1
1
1
BF
F13A1 (intronic)
(CCCTCCAGGTGATTCTGAT
GTGCTCCAAGCGTGGTGA)$_N$
NA19240_h1
NA19240_h0
HG00733_h1
HG00514_h0
HG00514_h1
CHM13
CHM13_HiFi
GRCh38
HG00733_h0
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
chr6:6260545−6263852
MOTIF: CCCTCCAGGTGATTCTGATGTGCTCCAAGCGTGGTGA
22
18
45
11
10
16
16
18
6
1
1
1
1
BG
ABCB1 (intronic)
(ATATATATGTGTC)$_N$
HG00514_h0
NA19240_h1
NA19240_h0
CHM13
CHM13_HiFi
HG00733_h0
GRCh38
HG00733_h1
HG00514_h1
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
chr7:87624557−87628349
MOTIF: ATATATATGTGTC
24
15
24
22
22
23
15
22
1
1
1
2
2
BH
MYRIP (intronic)
(ATAATATGTGTCTATGTGTTATATATCCACAG)$_N$
NA19240_h1
HG00514_h1
CHM13
CHM13_HiFi
HG00733_h1
HG00514_h0
HG00733_h0
NA19240_h0
GRCh38
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
chr3:40083621−40086832
MOTIF: ATAATACACATCTATGTATTATATATCGATAG
25
10
10
11
13
8
12
8
6
7
7
5
5


BI
ARHGEF38 (intronic)
(AATAG)$_N$
HG00733_h0
GRCh38
NA19240_h1
HG00733_h1
HG00514_h1
CHM13
CHM13_HiFi
HG00514_h0
NA19240_h0
Chimp clint_h0
Chimp clint_h1
16
18
13
12
14
15
15
11
16
9
9
chr4:105559869−105563080
MOTIF: AATAG
BJ
Intergenic
(TTGTGTAAACACATCACCTAGCATCACG TTAGGTGTGAGACTGATGCTGAGGCT)$_N$
HG00514_h0
HG00514_h1
HG00733_h0
NA19240_h0
NA19240_h1
GRCh38
HG00733_h1
CHM13
CHM13_HiFi
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
7
4
6
10
10
6
10
10
10
1
1
1
1
1
1
chr5:17100005−17103206
MOTIF: TTGTGTAAACACATCACCTAGCATCACGTTAGGTGTGAGACTGATGCTGAGGCT
BK
SLC35F1 (intronic)
(TATACACACATAGG TACACATGCATATGTA)$_N$
HG00733_h0
CHM13
CHM13_HiFi
HG00514_h1
NA19240_h1
GRCh38
HG00514_h0
HG00733_h1
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
11
23
23
13
11
9
3
8
1
1
1
1
1
1
chr6:117922424−117926067
MOTIF: TATACACACATAGGTACACATGCATATGTA
BL
CA10 (intron)
(AGC)$_N$
HG00514_h0
NA19240_h1
CHM13
CHM13_HiFi
GRCh38
NA19240_h0
HG00733_h1
HG00514_h1
HG00733_h0
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
73
12
21
21
21
6
16
11
10
17
17
8
8
7
7
chr17:51829632-51833833
MOTIF: AGC
BM
Intergenic
(AGA)$_N$
HG00514_h1
HG00733_h1
HG00514_h0
NA19240_h0
CHM13_HiFi
CHM13
GRCh38
HG00733_h0
NA19240_h1
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
61
67
66
25
59
58
55
9
9
6
6
7
7
8
8
chr5:127245250−127249513
MOTIF: AGA
BN
CENPVL3 (upstream)
(TTC)$_N$
HG00733_h0
HG00733_h1
CHM13_HiFi
CHM13
GRCh38
NA19240_h0
NA19240_h1
HG00514_h0
HG00514_h1
Chimp clint_h0
Chimp clint_h1
Gorilla kamilah_h0
Gorilla kamilah_h1
Orangutan susie_h0
Orangutan susie_h1
17
17
70
68
50
21
21
19
19
7
7
6
6
4
4
chrX:51619438−51623783
MOTIF: TTC

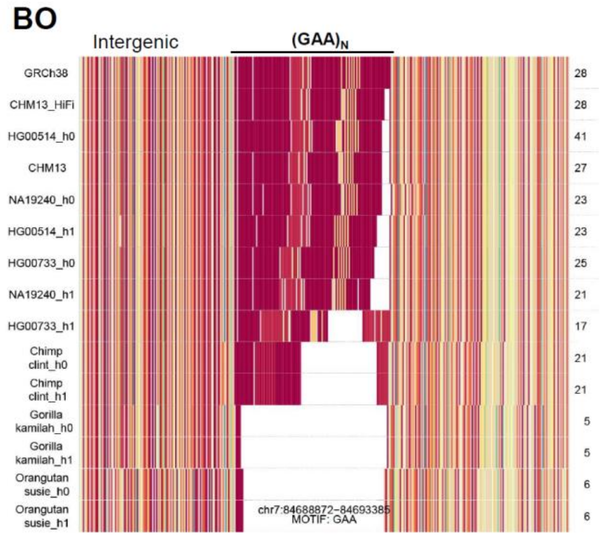


Fig. S10. STR/VNTR sequence composition plots. The sequences from each human and NHP haplotype were colored according to their k-mer abundance, from hot to cold colors for the most to least abundant k-mers, respectively. The order of the tracts reflects the order of overall sequence size from largest to smallest human samples, followed by the NHP samples ordered by their evolutionary distance to human. See Methods for more details. The number to the right of each row represents the maximal LPT (i.e., longest pure tract) copy number. For the CHM13 sample, sequence from both CLR and HiFi assemblies have been included as technical replicates. Some of the clustered STR/VNTR interruptions include: 11 tandem copies of (GTAGTATTAAGAGGTG) in CHM13 (**AY**), (AGGG)n interruptions in GRCh38 (**AZ**), (GAA)n interruptions in multiple human samples (**BC**), 13-17 copies of (AATGG) interruptions in both haplotypes of HG00733 (**BI**), and 4-8 copies of (ATAATACACATCTATGTATTATCTATCGATAG) interruptions in CHM13 and h1 of HG00733 (**BH**). An STR located in the intron of *CA10* is composed of 6-73 uninterrupted tandem repeat copies of AGC in human haplotypes. Interspersed interruptions of AGCC exist in both Yoruban haplotypes, while single interruptions of AG and AGG occur in both orangutan haplotypes (**BL**). An intergenic STR is composed of 9-67 uninterrupted tandem repeat copies of AGA in humans, while periodic interruptions of AGG occur in two unrelated human haplotypes. Interestingly, a comparably sized AT-rich region is located upstream of the variable AGA region of all human and NHP haplotypes with the exception of the orangutan haplotypes (**BM**). An STR located upstream of *CENPVL3* is composed of 17-70 uninterrupted tandem repeat copies of TTC in humans (**BN**). Interestingly, interruptions of CTC exist in CHM13, while interspersed interruptions of both TCC and TC exist in both haplotypes of the Puerto Rican sample. An intergenic STR is composed of 17-41 uninterrupted tandem repeat copies of GAA in humans. All human haplotypes contain two distinct clustered interruptions of GGA and GCA repeats, and the largest expansion of these interruptions exists in the same Puerto Rican haplotype (**BO**).

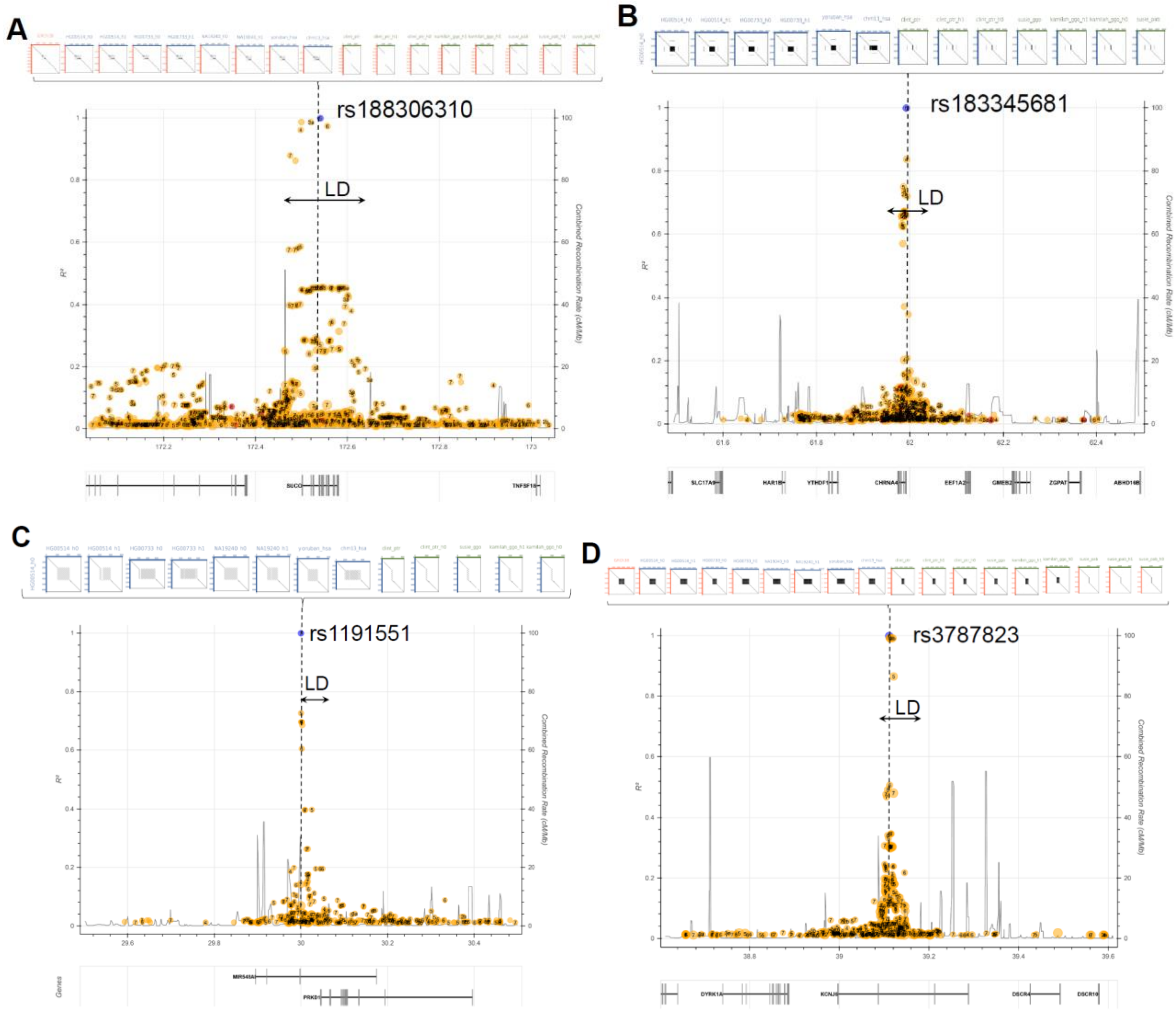


Fig. S11. Linkage between STRs/VNTRs and GWAS SNPs. Each circle in the plot corresponds to the 1000 Genomes Project (phase III) SNPs, and their linkage disequilibrium (LD) is measured by applying the R^2 statistic to the genotype frequencies from all 26 world populations. The GWAS SNPs correspond to the following phenotypes: (**A**) daytime sleep phenotypes(29), (**B**) emphysema imaging phenotypes(30), (**C**) schizophrenia(31), and (**D**) math ability(32). The position of the tandem repeat is shown with the dotted vertical line, while the two-sided horizontal arrow indicates the high-LD region, which was identified using R^2 values as guides. The vertical peaks correspond to the estimated recombination rate, and the corresponding cM/Mb values are shown on the right-hand vertical axis. The numbers inside each of the circles correspond to the RegulomeDB score, which predicts *in silico* the potential for involvement in gene regulation(33). The dotplots at the top of each plot compare the GRCh38 sequence (y-axis) that that assembled in all of the human and NHP haplotypes (x-axis), unless otherwise stated.

Dataset S1 (separate file). Summary of the total STRs/VNTRs identified in the human assemblies.

Dataset S2 (separate file). Summary of the human tandem repeats stratified by motif length.

Dataset S3 (separate file). 281 STRs/VNTRs missing from the GRCh38.p12 assembly.

Dataset S4 (separate file). Per-sample statistics of STR and VNTR motif size, variability, and overall repeat purity.

Dataset S5 (separate file). Summary of samples used in our study.

Dataset S6 (separate file). Number of STRs/VNTRs successfully phased in human and NHP haplotypes.

Dataset S7 (separate file). Sequence identity between assembled haplotypes and BAC sequences in NHP samples.

Dataset S8 (separate file). Ultra-long ONT validation of STRs/VNTRs with inconsistent representations in CLR and HiFi assemblies of CHM13.

Dataset S9 (separate file). Full list of *ab initio* and HSE loci and their tandem repeat annotation in human and NHP haplotypes.

Dataset S10 (separate file). Association results using generalized linear models.

Dataset S11 (separate file). Genes associated with the largest differential expression between human and chimpanzee cerebral organoids.

Dataset S12 (separate file). Summary of 1,719 GWAS signals located ≤2 kbp from the haplotype-resolved STRs/VNTRs.

Dataset S13 (separate file). Known disease-associated STRs/VNTRs.

Dataset S14 (separate file). STRs/VNTRs with ≥40 uninterrupted tandem repeat copies.

REFERENCES

1. Tempel S (2012) Using and understanding RepeatMasker. *Methods Mol Biol* 859:29-51.
2. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573-580.
3. Kronenberg ZN*, et al.* (2018) High-resolution comparative analysis of great ape genomes. *Science* 360(6393).
4. Gordon D*, et al.* (2016) Long-read sequence assembly of the gorilla genome. *Science* 352(6281):aae0344.
5. Parsons JD (1995) Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* 11(6):615-619.
6. Sonnhammer EL & Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167(1-2):GC1-10.
7. Sievers F & Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol* 1079:105-116.
8. Larsson NJ & Sadakane K (2007) Faster suffix sorting. *Theor Comput Sci* 387(3):258-272.
9. Audano P & Vannberg F (2014) KAnalyze: a fast versatile pipelined k-mer toolkit. *Bioinformatics* 30(14):2070-2072.
10. Fishilevich S*, et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* 2017.
11. Bailey TL, Williams N, Misleh C, & Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34(Web Server issue):W369-373.
12. Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J R Stat Soc B* 57(1):289-300.
13. Sousa AMM*, et al.* (2017) Molecular and cellular reorganization of neural circuits in the human lineage. *Science* 358(6366):1027-1032.
14. Mele M*, et al.* (2015) The human transcriptome across tissues and individuals. *Science* 348(6235):660-665.
15. Dobin A*, et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15-21.
16. Li YI*, et al.* (2018) Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* 50(1):151-158.
17. Audano PA*, et al.* (2019) Characterizing the major structural variant alleles of the human genome. *Cell* 176(3):663-675 e619.
18. Krumsiek J, Arnold R, & Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 23(8):1026-1028.
19. Dougherty JD, Schmidt EF, Nakajima M, & Heintz N (2010) Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res* 38(13):4218-4230.
20. Jaganathan K*, et al.* (2019) Predicting splicing from primary sequence with deep learning. *Cell* 176(3):535-548 e524.
21. Eser O*, et al.* (2011) Short aggrecan gene repetitive alleles associated with lumbar degenerative disc disease in Turkish patients. *Genet Mol Res* 10(3):1923-1930.
22. Benedetti F*, et al.* (2008) A length polymorphism in the circadian clock gene Per3 influences age at onset of bipolar disorder. *Neurosci Lett* 445(2):184-187.
23. LaHoste GJ*, et al.* (1996) Dopamine D4 receptor gene polymorphism is associated with attention deficit hyperactivity disorder. *Mol Psychiatry* 1(2):121-124.

24. Taj MJR*, et al.* (2013) DRD4 gene and obsessive compulsive disorder: do symptom dimensions have specific genetic correlates? *Prog Neuropsychopharmacol Biol Psychiatry* 41:18-23.

25. De Roeck A*, et al.* (2018) An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta Neuropathol* 135(6):827-837.

26. Hijikata M*, et al.* (2011) Molecular cloning of two novel mucin-like genes in the disease-susceptibility locus for diffuse panbronchiolitis. *Hum Genet* 129(2):117-128.

27. Jeffreys AJ, Wilson V, & Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature* 314(6006):67-73.

28. Pugliese A*, et al.* (1997) The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. *Nat Genet* 15(3):293-297.

29. Spada J*, et al.* (2016) Genome-wide association analysis of actigraphic sleep phenotypes in the LIFE Adult Study. *J Sleep Res* 25(6):690-701.

30. Cho MH*, et al.* (2015) A genome-wide association study of emphysema and airway quantitative imaging phenotypes. *Am J Respir Crit Care Med* 192(5):559-569.

31. Pardinas AF*, et al.* (2018) Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* 50(3):381-389.

32. Lee JJ*, et al.* (2018) Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* 50(8):1112-1121.

33. Boyle AP*, et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22(9):1790-1797.

The Human Genome Structural Variation Consortium

Mark J.P. Chaisson[1,2], Ashley D. Sanders[3], Xuefang Zhao[4,5], Ankit Malhotra[6], David Porubsky[7,8], Tobias Rausch[3], Eugene J. Gardner[9], Oscar L. Rodriguez[10], Li Guo[11,12,13], Ryan L. Collins[5,14], Xian Fan[15], Jia Wen[16], Robert E. Handsaker[17,18,19], Susan Fairley[20], Zev N. Kronenberg[1], Xiangmeng Kong[21,22], Fereydoun Hormozdiari[23,24], Dillon Lee[25], Aaron M. Wenger[26], Alex R. Hastie[27], Danny Antaki[28], Thomas Anantharaman[27], Peter A. Audano[1], Harrison Brand[5], Stuart Cantsilieris[1], Han Cao[27], Eliza Cerveira[6], Chong Chen[15], Xintong Chen[9], Chen-Shan Chin[26], Zechen Chong[15], Nelson T. Chuang[9], Christine C. Lambert[26], Deanna M. Church[29], Laura Clarke[20], Andrew Farrell[25], Joey Flores[30], Timur Galeey[21,22], David U. Gorkin[31,32], Madhusudan Gujral[28], Victor Guryev[7], William Haynes Heaton[29], Jonas Korlach[26], Sushant Kumar[21,22], Jee Young Kwon[6,33], Ernest T. Lam[27], Jong Eun Lee[34], Joyce Lee[27], Wan-Ping Lee[6], Sau Peng Lee[35], Shantao Li[21,22], Patrick Marks[29], Karine Viaud-Martinez[30], Sascha Meiers[3], Katherine M. Munson[1], Fabio C.P. Navarro[21,22], Bradley J. Nelson[1], Conor Nodzak[16], Amina Noor[28], Sofia Kyriazopoulou-Panagiotopoulou[29], Andy W.C. Pang[27], Yunjiang Qiu[32,36], Gabriel Rosanio[28], Mallory Ryan[6], Adrian Stütz[3], Diana C.J. Spierings[7], Alistair Ward[25], AnneMarie E. Welch[1], Ming Xiao[37], Wei Xu[29], Chengsheng Zhang[6], Qihui Zhu[6], Xiangqun Zheng-Bradley[20], Ernesto Lowy[20], Sergei Yakneen[3], Steven McCarroll[17,18,38], Goo Jun[39], Li Ding[40], Chong Lek Koh[41], Bing Ren[31,32], Paul Flicek[20], Ken Chen[15], Mark B. Gerstein[21,22,42,43], Pui-Yan Kwok[44], Peter M. Lansdorp[7,45,46], Gabor T. Marth[25], Jonathan Sebat[28,31,47], Xinghua Shi[16], Ali Bashir[10], Kai Ye[12,13,48], Scott E. Devine[9], Michael E. Talkowski[5,19,49], Ryan E. Mills[4,50], Tobias Marschall[8], Jan O. Korbel[3,20], Evan E. Eichler[1,51] & Charles Lee[6,33]

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA. [2]Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA. [3]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany. [4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. [5]Center for Genomic Medicine, Massachusetts General Hospital, Department of Neurology, Harvard Medical School, Boston, MA 02114, USA. [6]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA. [7]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, Groningen, AV NL-9713, The Netherlands. [8]Center for Bioinformatics, Saarland University and the Max Planck Institute for Informatics, 66123 Saarbrücken, Germany. [9]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA. [10]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [11]The School of Life Science and Technology of Xi'an Jiaotong University, 710049 Xi'an, China. [12]MOE Key Lab for Intelligent Networks & Networks Security, School of Electronics and Information Engineering, Xi'an Jiaotong University, 710049 Xi'an, China. [13]Ye-Lab For Omics and Omics Informatics, Xi'an Jiaotong University, 710049 Xi'an, China. [14]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA 02115, USA. [15]Department of Bioinformatics

and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [16]Department of Bioinformatics and Genomics, College of Computing and Informatics, The University of North Carolina at Charlotte, Charlotte, NC 28223, USA. [17]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [18]The Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [19]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [20]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom. [21]Yale University Medical School, Computational Biology and Bioinformatics Program, New Haven, CT 06520, USA. [22]Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA. [23]Biochemistry and Molecular Medicine, University of California Davis, Davis, CA 95616, USA. [24]UC Davis Genome Center, University of California, Davis, Davis, CA 95616, USA. [25]USTAR Center for Genetic Discovery and Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA. [26]Pacific Biosciences, Menlo Park, CA 94025, USA. [27]Bionano Genomics, San Diego, CA 92121, USA. [28]Beyster Center for Genomics of Psychiatric Diseases, Department of Psychiatry University of California San Diego, La Jolla, CA 92093, USA. [29]10X Genomics, Pleasanton, CA 94566, USA. [30]Illumina Clinical Services Laboratory, Illumina, Inc., 5200 Illumina Way, San Diego, CA 92122, USA. [31]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA 92093, USA. [32]Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA. [33]Department of Graduate Studies – Life Sciences, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun- gu, Seoul 03760, South Korea. [34]DNA Link, Seodaemun-gu, Seoul, South Korea. [35]TreeCode Sdn Bhd, Bandar Botanic, 41200 Klang, Malaysia. [36]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA. [37]School of Biomedical Engineering, Drexel University, Philadelphia, PA 19104, USA. [38]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [39]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77225, USA. [40]Department of Medicine, McDonnell Genome Institute, Siteman Cancer Center, Washington University School of Medicine, St. Louis, MI 63108, USA. [41]High Impact Research, University of Malaya, 50603 Kuala Lumpur, Malaysia. [42]Department of Computer Science, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA. [43]Department of Statistics and Data Science, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA. [44]Institute for Human Genetics, University of California–San Francisco, San Francisco, CA 94143, USA. [45]Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada. [46]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [47]Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA. [48]The First Affiliated Hospital of Xi’an Jiaotong University, 710061 Xi’an, China. [49]Center for Mendelian Genomics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [50]Department of Human

Genetics, University of Michigan, Ann Arbor, MI 48109, USA. [51]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.