

Adam, Eve, and genes: on humankind's explosive origin

***Human-specific DNA regions and their
significance***

Richard v. Sternberg
Biologic Institute

We all know the story...



Lucy's DNA

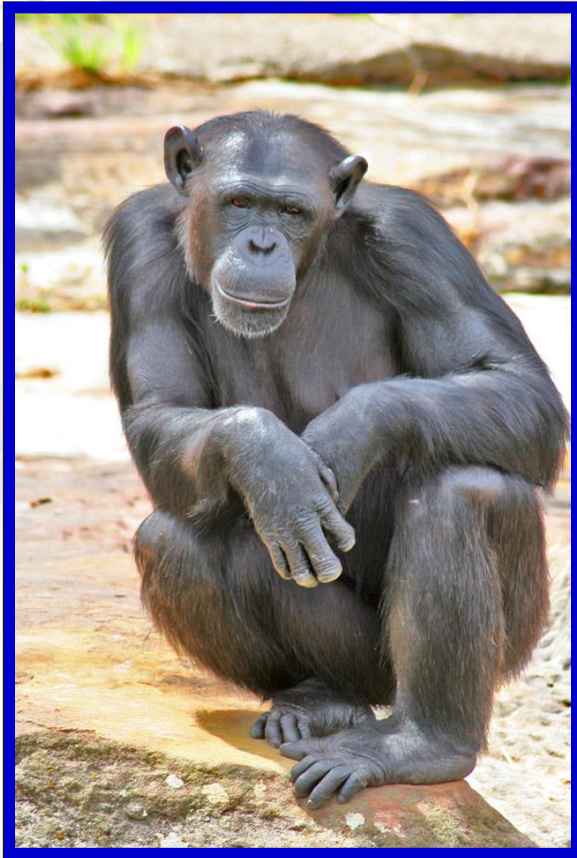
Mutations



Sophia's DNA



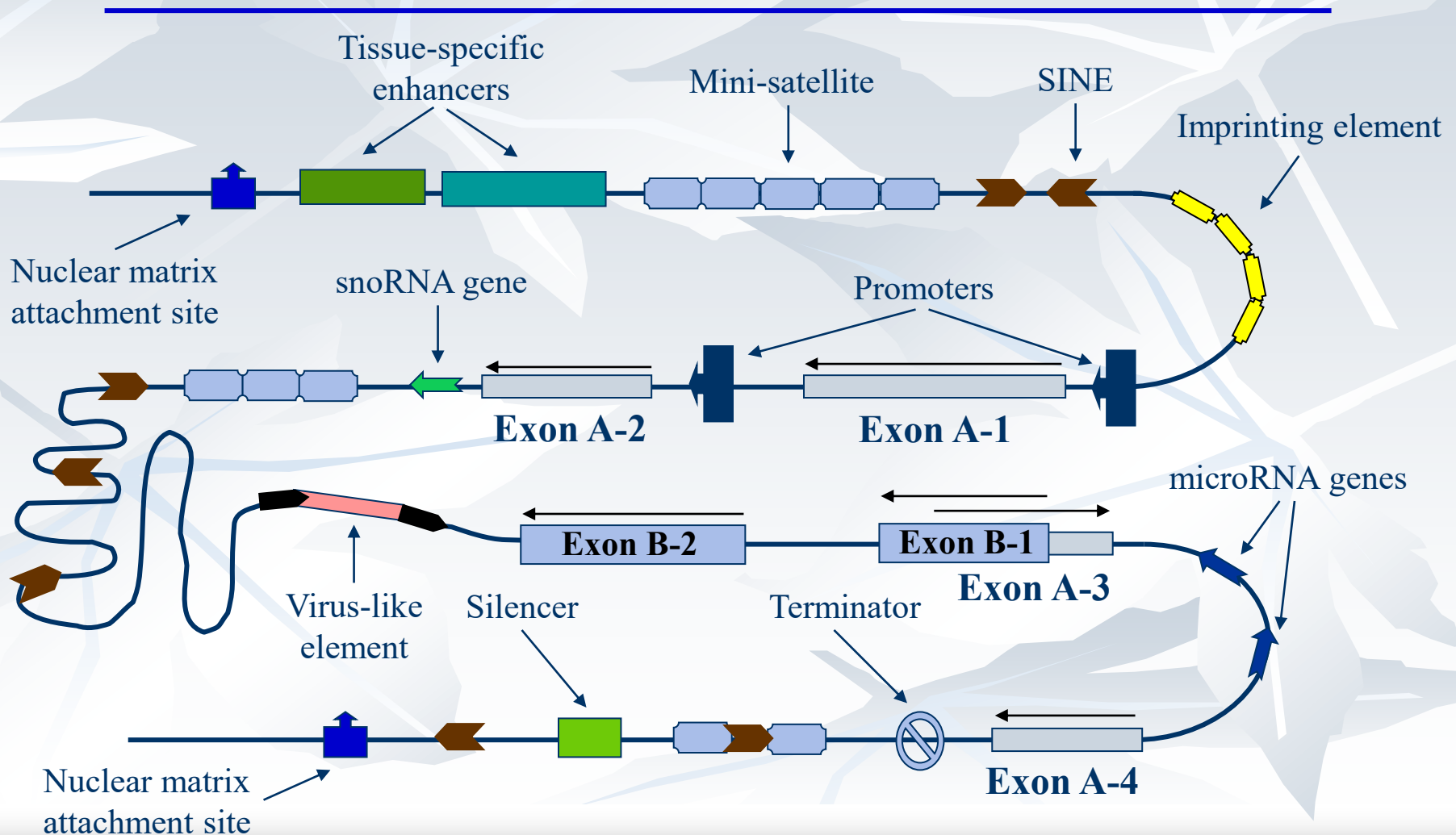
As a result of recent common ancestry, a mere 1% DNA difference is said to exist between any chimp and a human.



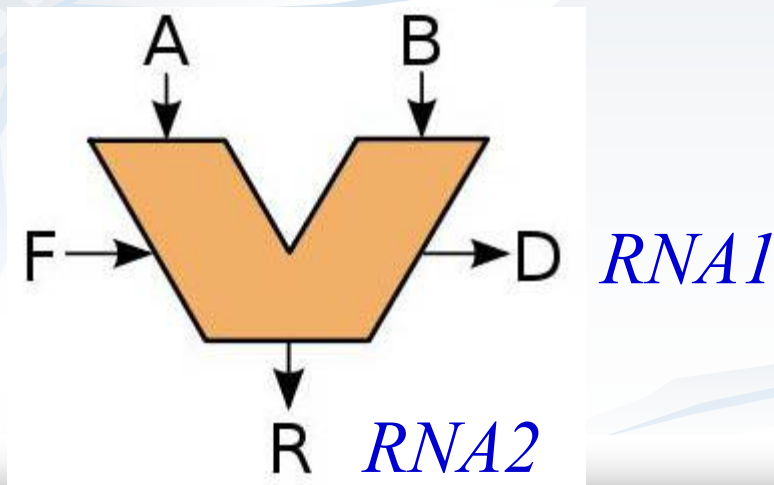
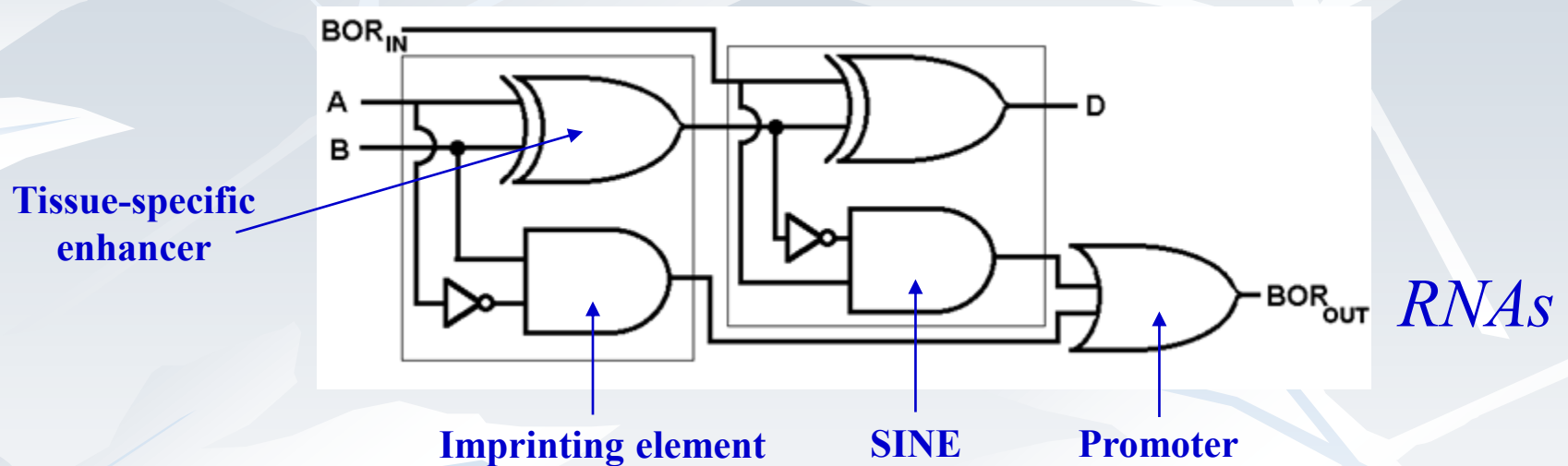


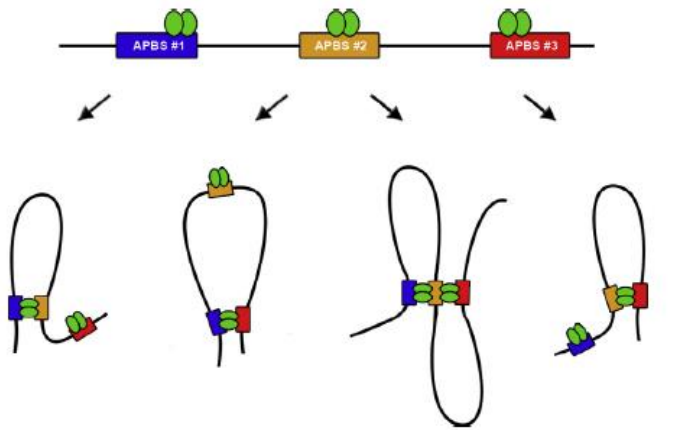
**But is this true? Let's begin with
this...**

Principle 1: A typical (animal) ‘gene’ consists of interleaved, interspersed, multilevel, and overlapping “data files.”



Principle 2: This order permits a 'gene' to be formed into circuits differentially.



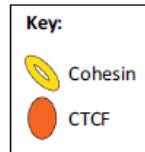
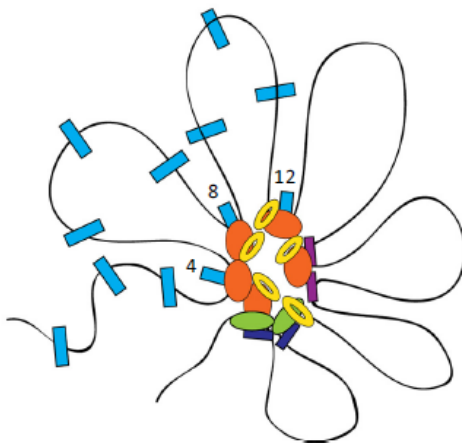
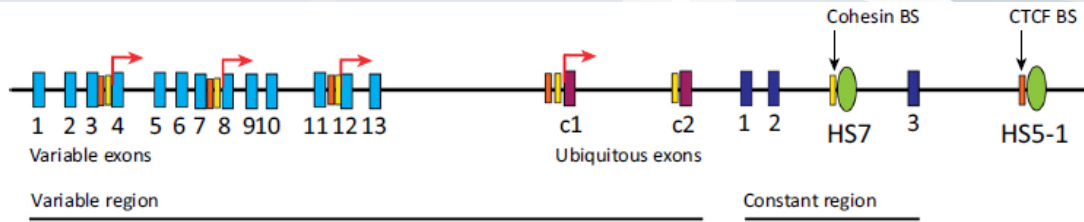


Architectural proteins, transcription, and the three-dimensional organization of the genome

<http://dx.doi.org/10.1016/j.febslet.2015.05.025>

Caelin Cubeñas-Potts, Victor G. Corces*

Chromatin folding indeed allows different circuits to be formed.

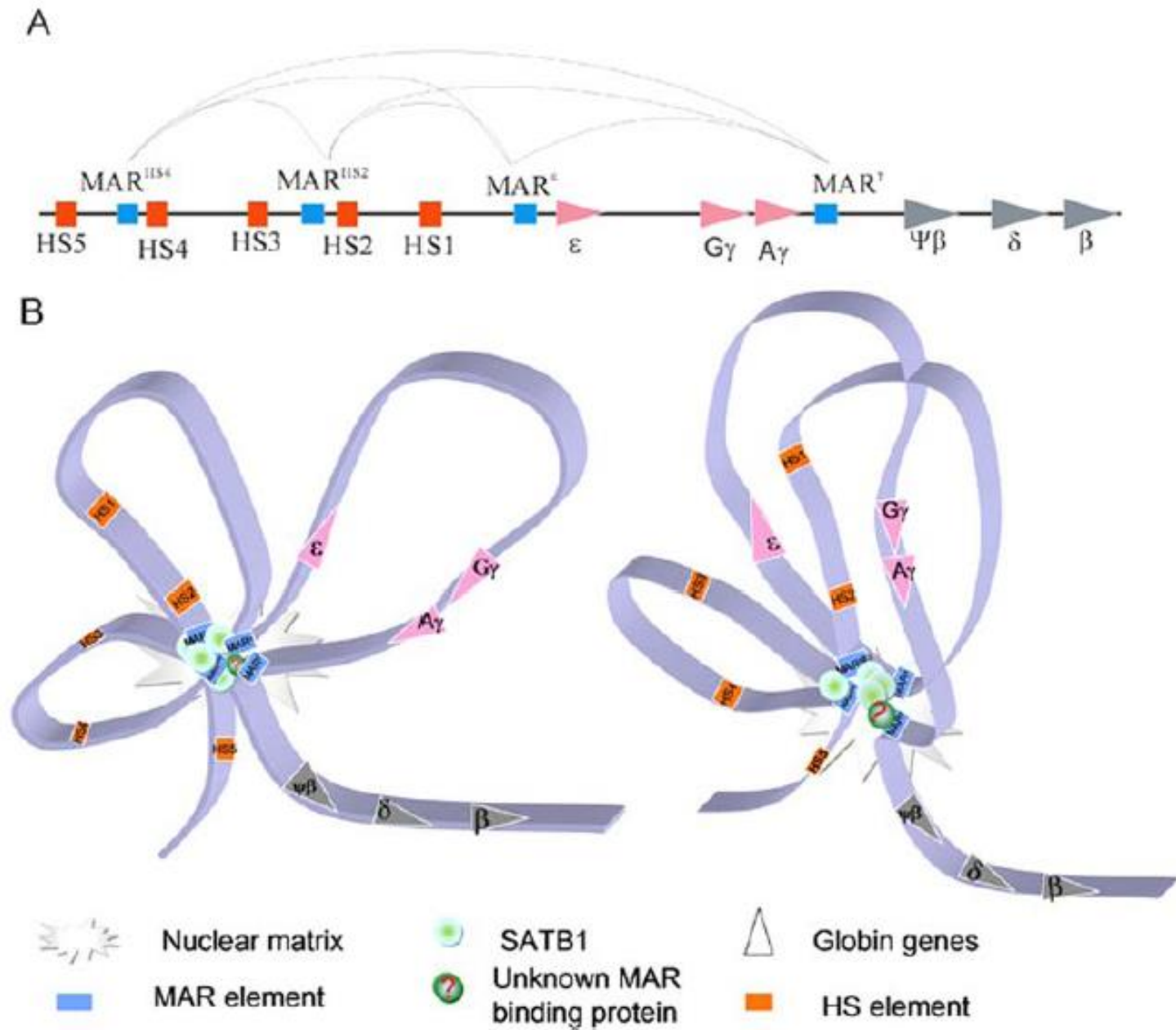


The human protocadherin A (PCDH α) gene cluster

Architectural proteins: regulators of 3D genome organization in cell fate

Elena Gómez-Díaz and Victor G. Corces

Trends in Cell Biology, November 2014, Vol. 24, No. 11



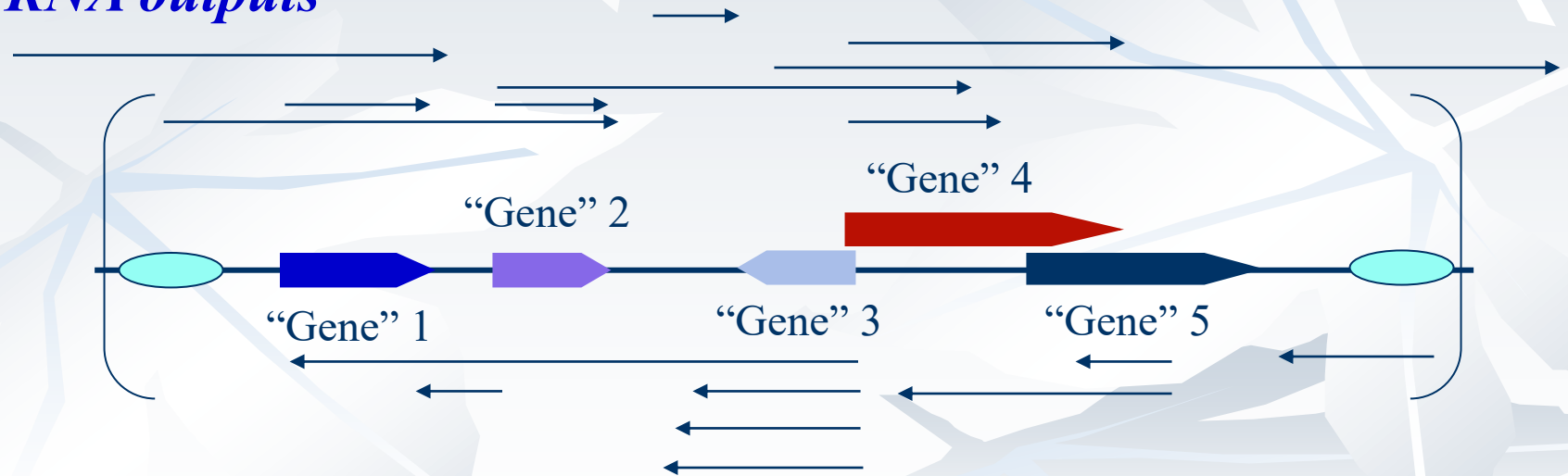
Inter-MAR Association Contributes to Transcriptionally Active Looping Events in Human β -globin Gene Cluster

Li Wang¹, Li-Jun Di¹, Xiang Lv, Wei Zheng, Zheng Xue, Zhi-Chen Guo, De-Pei Liu^{1*}, Chi-Chuan Liang

February 2009 | Volume 4 | Issue 2 | e4629

Principle 3: Gene data files are clustered into higher-order “folders” along a chromosome. This arrangement enables different types of RNAs to be encoded on both strands.

RNA outputs



Principle 4: Gene “folders” are in turn arranged into “superfolders.”

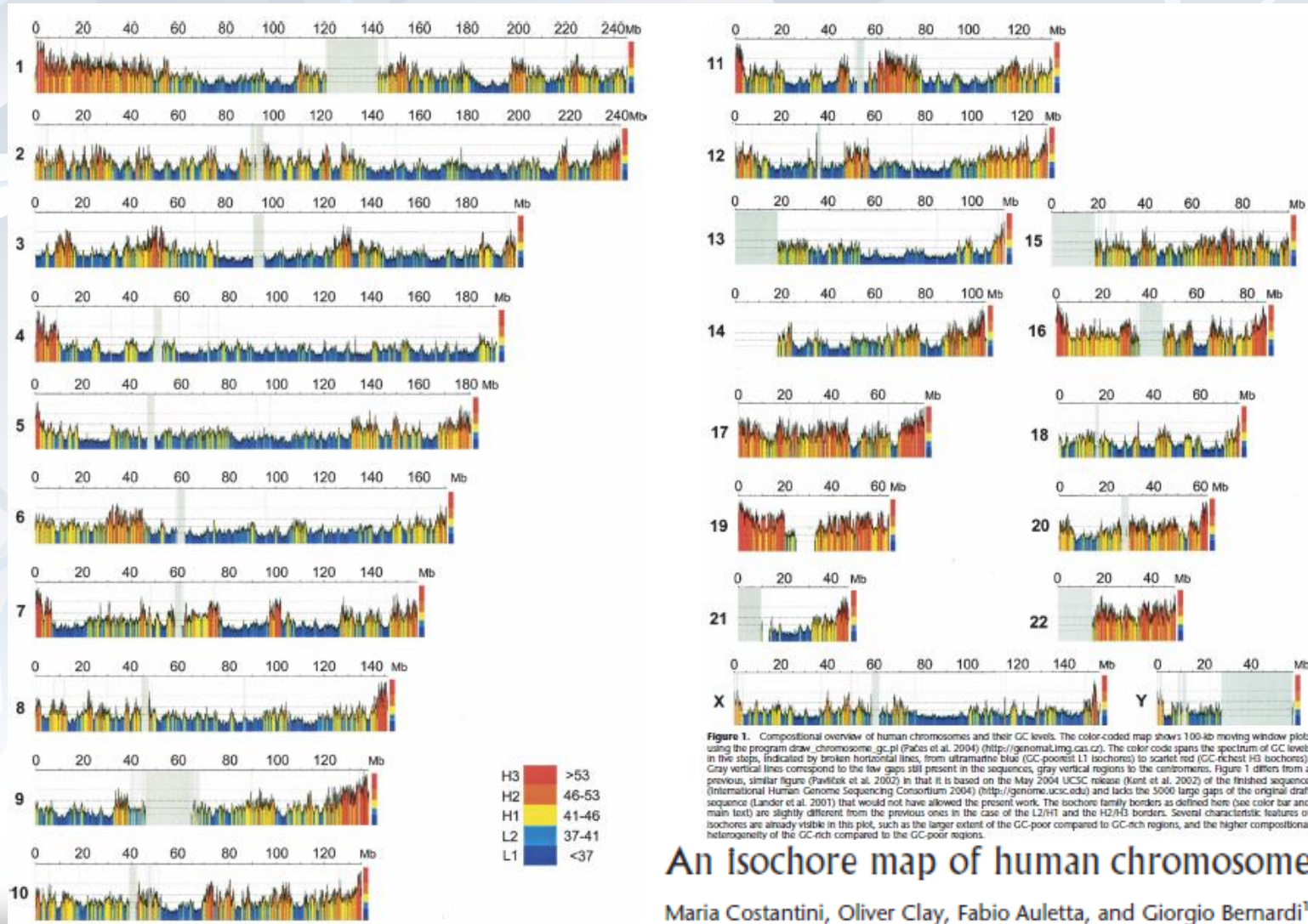
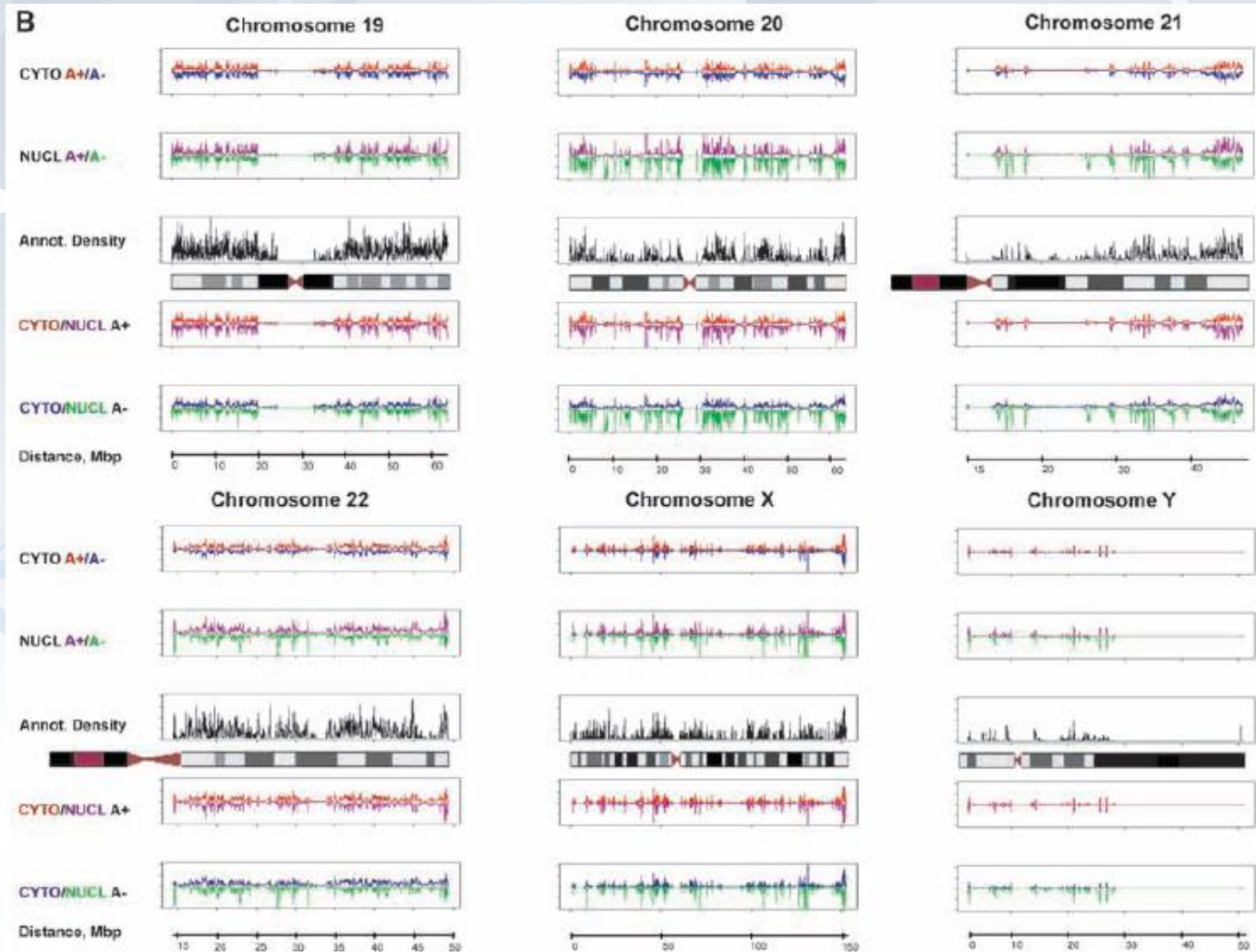


Figure 1. Compositional overview of human chromosomes and their GC levels. The color-coded map shows 100-kb moving window plots using the program `draw_chromosome_gc.pl` (Pačes et al. 2004) (<http://genoma.limg.cas.cz>). The color code spans the spectrum of GC levels in five steps, indicated by broken horizontal lines, from ultramarine blue (GC-poorest L1 isochores) to scarlet red (GC-richest H3 isochores). Gray vertical lines correspond to the low gaps still present in the sequences, gray vertical regions to the centromeres. Figure 1 differs from a previous, similar figure (Pavlidis et al. 2002) in that it is based on the May 2004 UCSF release (Kent et al. 2002) of the finished sequence (International Human Genome Sequencing Consortium 2004) (<http://genome.ucsc.edu>) and lacks the 5000 large gaps of the original draft sequence (Lander et al. 2001) that would not have allowed the present work. The isochore family borders as defined here (see color bar and main text) are slightly different from the previous ones in the case of the L2/H1 and the H2/H3 borders. Several characteristic features of isochores are already visible in this plot, such as the larger extent of the GC-poor compared to GC-rich regions, and the higher compositional heterogeneity of the GC-rich compared to the GC-poor regions.

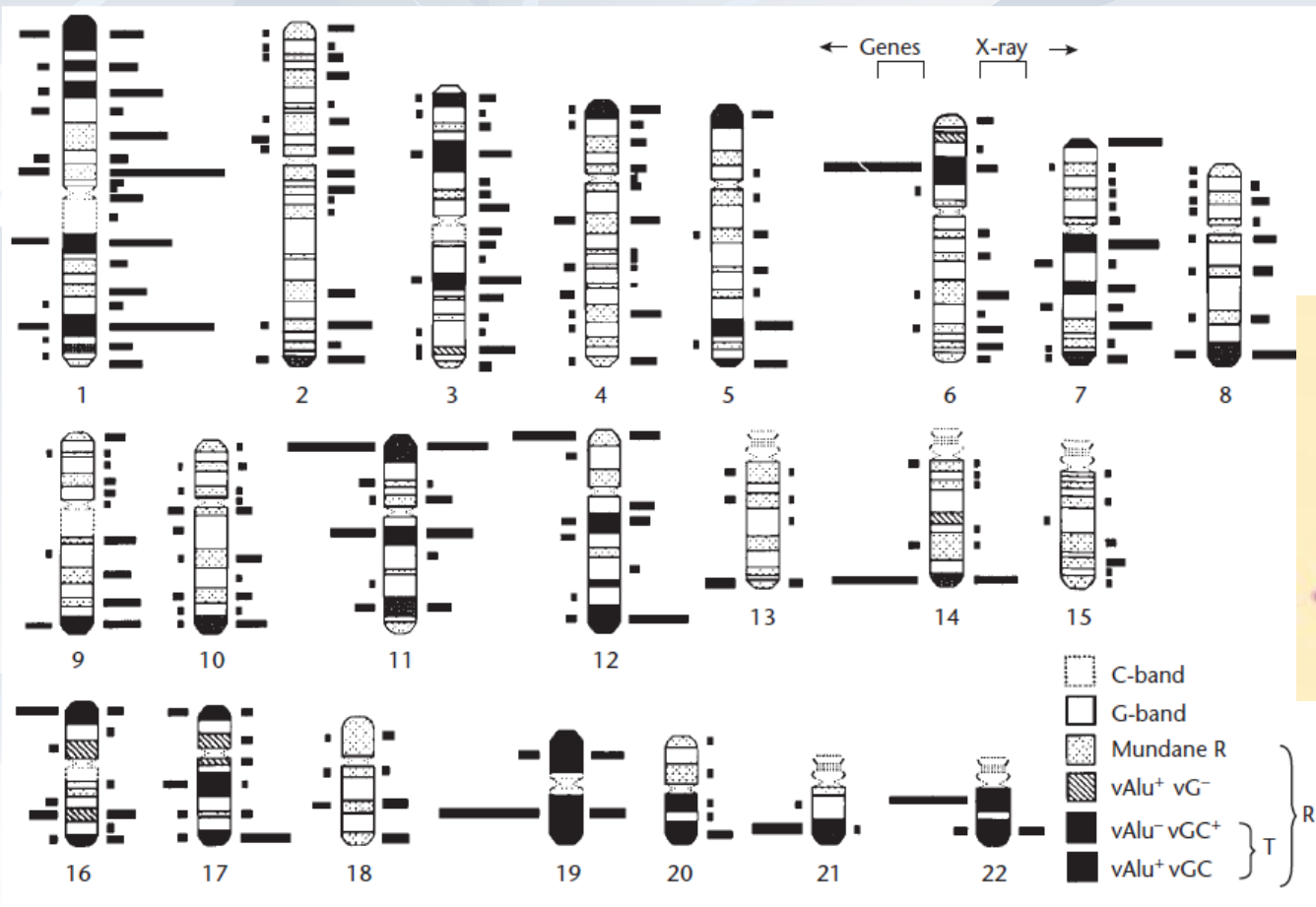
An Isochore map of human chromosomes

Maria Costantini, Oliver Clay, Fabio Auletta, and Giorgio Bernardi¹

Different “superfolders” encode different classes of RNA outputs.



And chromosome “superfolders” are in turn ordered into banding patterns...



Chromosomal Bands and Sequence Features

ENCYCLOPEDIA OF LIFE SCIENCES © 2005,

...such as those of CpG islands.

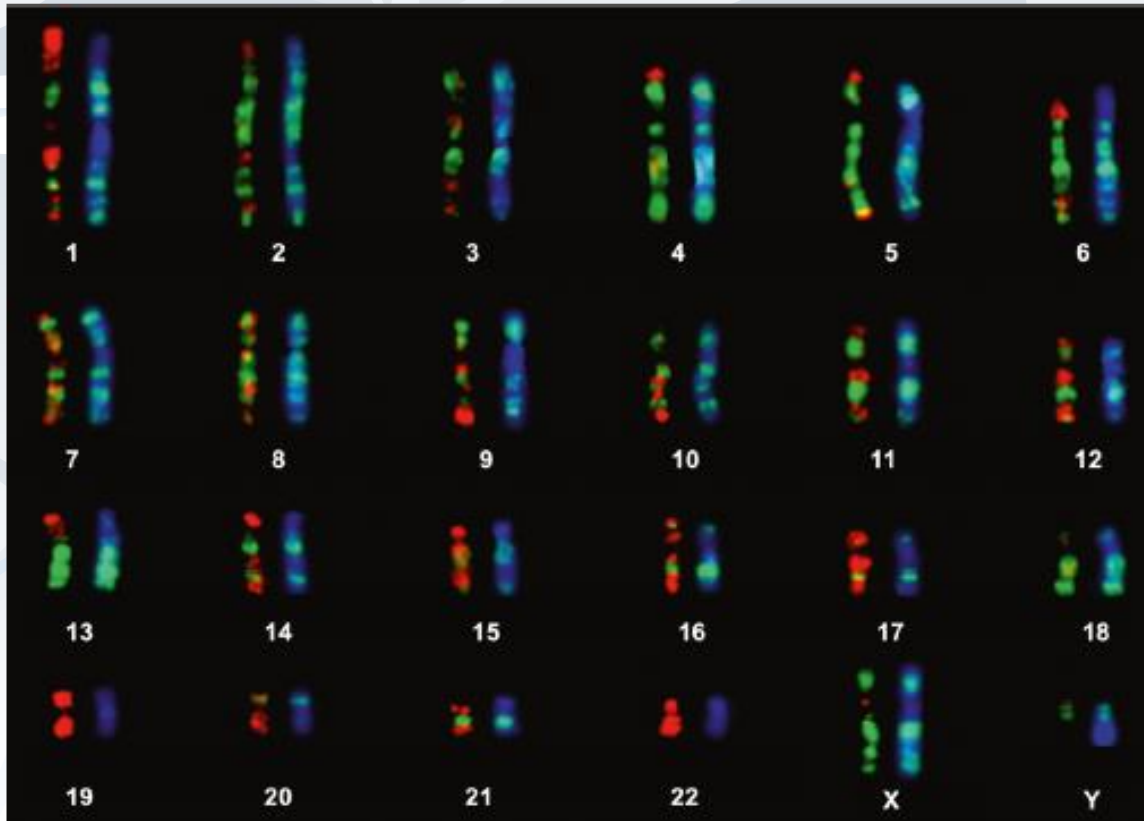



Fig. 2 Fluorescence in situ hybridisation (FISH) reveals the distribution of CpG islands across the human genome. For each metaphase chromosome, the hybridisation signal from CpG islands (red) is shown on the left of each pair. 4,6-Diamidino-2-phenyl indole (DAPI)-stained chromosomes are on the left. Late replicating G-bands are shown in green. Modified from Craig and Bickmore (1994)

Patterns in the genome

Wendy A. Bickmore ¹

This brings us to “junk DNA”...

Such are known to be replete with experimentally demonstrated functions:

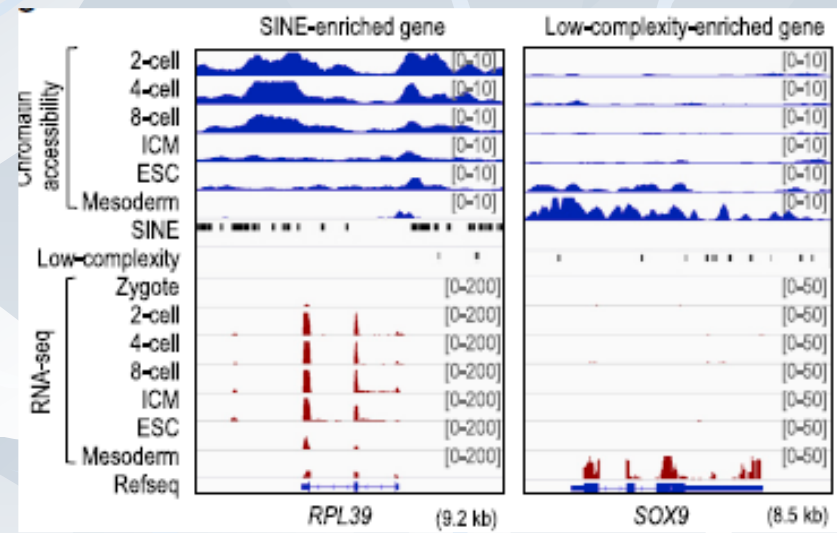
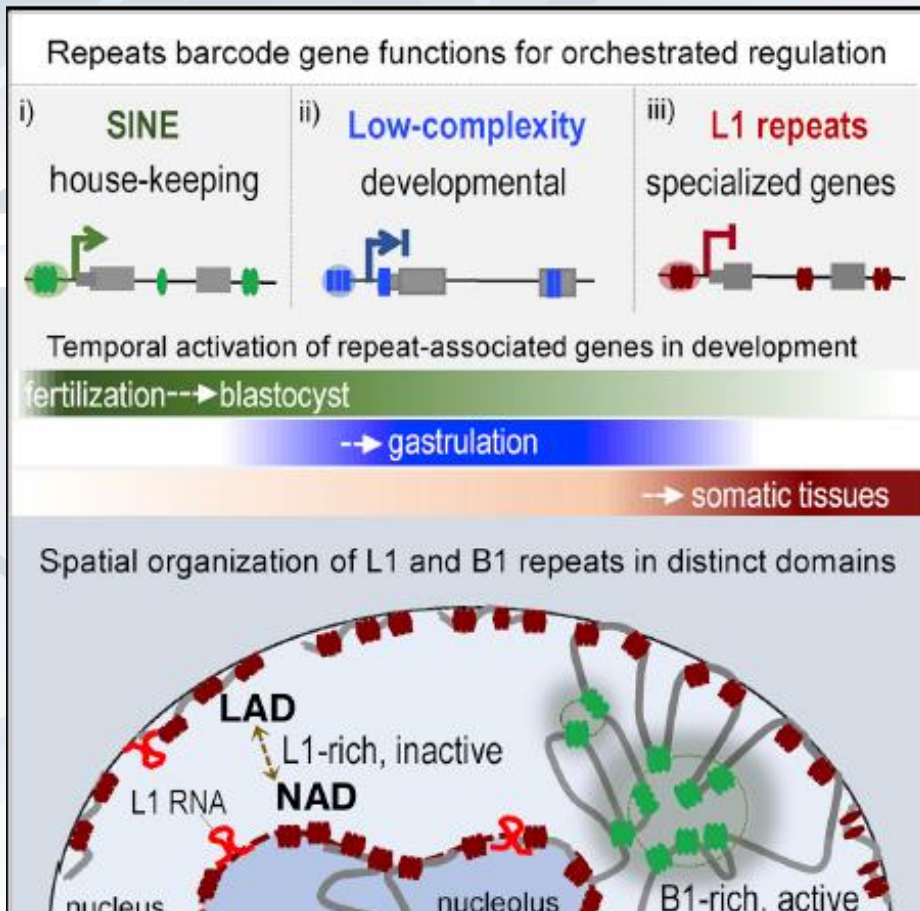
Highlights

- SINE, L1, and low-complexity repeats barcode genes with distinct functions
- Genomic repeats dictate the time and level of gene expression during development
- L1-enriched genes are sequestered in the inactive NAD/LAD domains for silencing
- L1 RNA promotes the nuclear localization and repression of L1-enriched genes

Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation

J. Yuyang Lu, Wen Shao, Lei Chang, ...,
Miguel Ramalho-Santos, Yujie Sun,
Xiaohua Shen

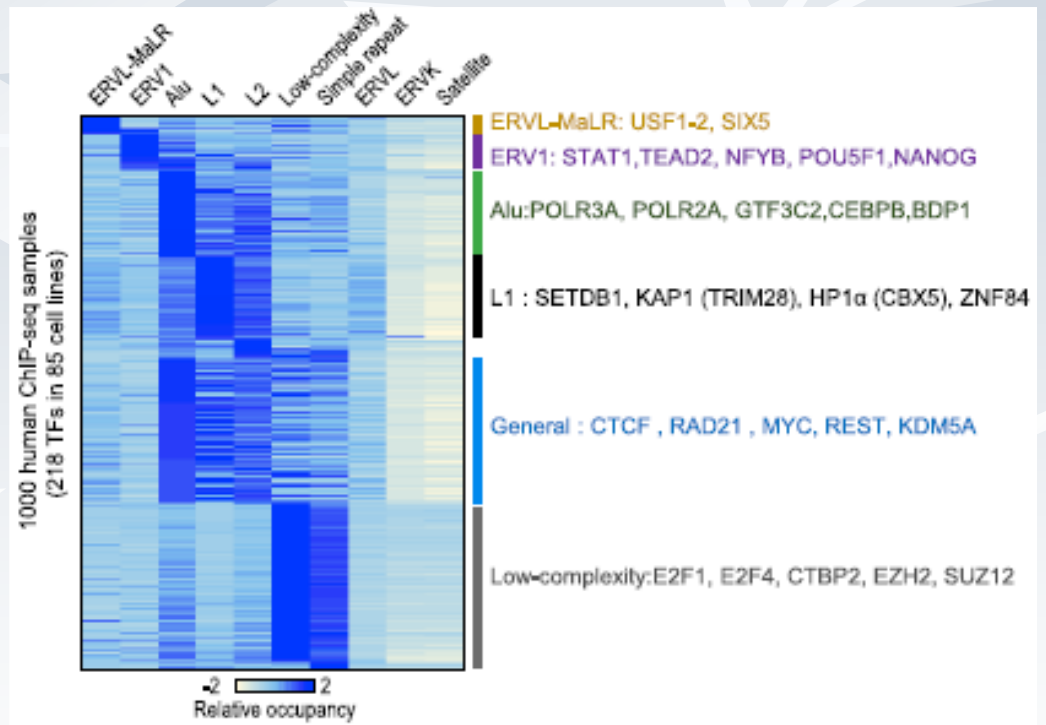
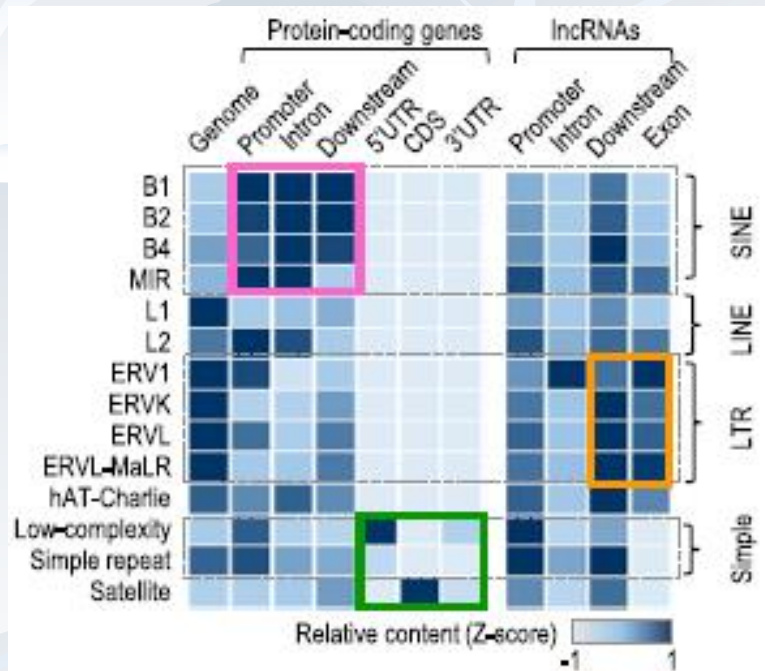
Lu et al., 2020, Cell Reports 30, 3296–3311



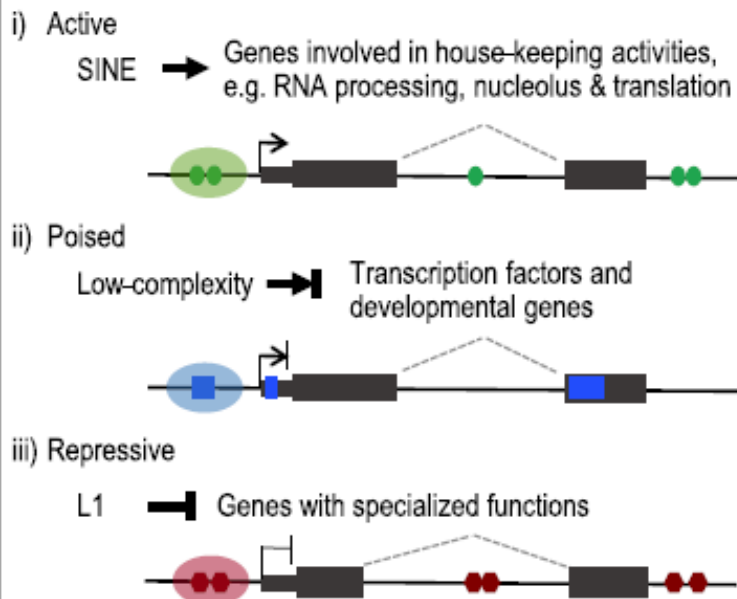
Chromatin accessibility of genic repeats

	2-cell					Pluripotent		Lineage-specific		
	Early	Early	Late	4-cell	8-cell	ICM	ESC	Meso	Endo	NPC
SINE	0.8	1.6	3.0	2.4	2.4	1.3	1.2	0.8	0.8	0.6
B1	1.1	2.5	4.8	3.7	3.4	1.7	1.6	0.9	1.0	0.8
B2	0.7	1.3	2.5	1.9	2.0	1.1	0.9	0.5	0.5	0.4
B4	0.6	1.0	1.8	1.5	1.7	1.0	0.9	0.6	0.6	0.5
ERVL	0.5	1.0	1.9	1.6	1.5	0.7	0.6	0.3	0.4	0.3
ERV1	0.4	0.7	1.1	0.9	1.0	0.8	0.7	0.4	0.5	0.3
ERVK	0.2	0.4	0.6	0.6	0.7	0.4	0.2	0.1	0.2	0.2
ERVL-MaLR	0.2	0.4	0.7	0.7	0.7	0.3	0.3	0.2	0.2	0.1
L1	0.1	0.1	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1
L2	0.6	0.8	1.1	1.0	1.3	1.0	1.0	0.9	0.8	0.7
Satellite	0.8	0.6	0.8	0.8	1.3	0.8	1.1	0.5	0.6	0.4
Low-complexity	1.2	1.0	1.2	1.2	1.0	1.3	1.7	2.0	2.2	2.4

0 2 (observed / random)

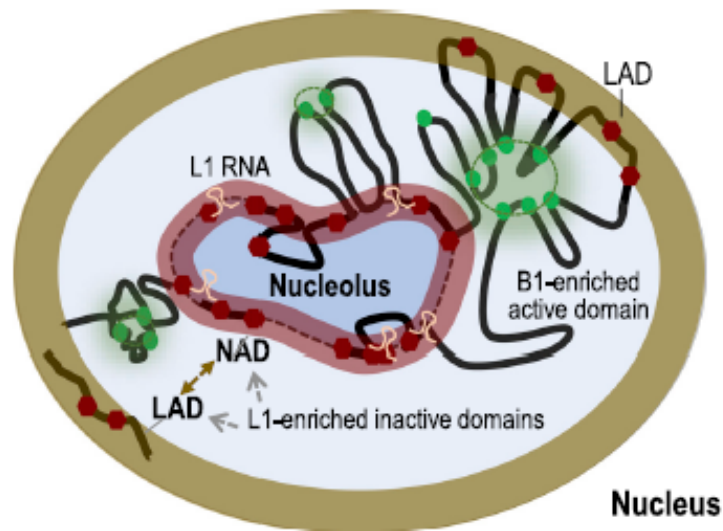


Gene function and transcription activity in ESCs

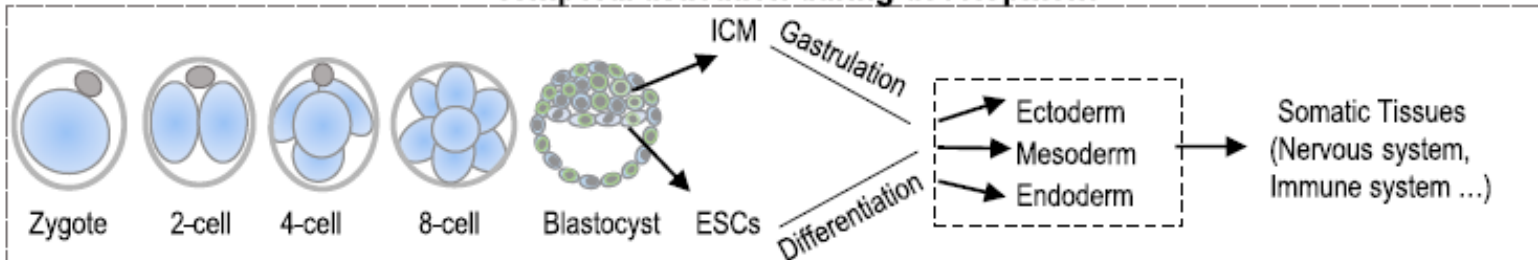


Nuclear organization

- Active domains: B1-enriched nuclear interior
- Inactive domains: L1-enriched NADs & LADs



Temporal activation during development

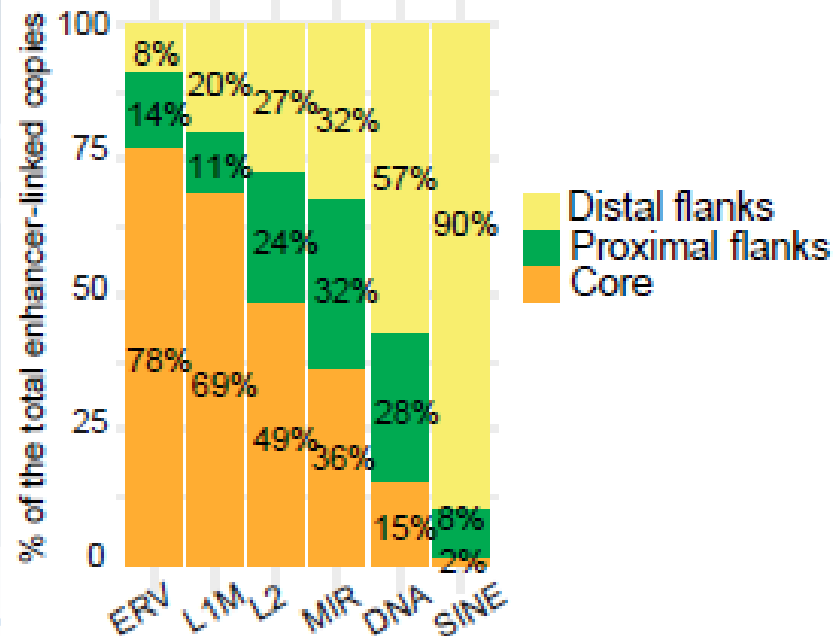


SINE


Low-complexity

L1

Proportional localization of enriched TEs in enhancer domains



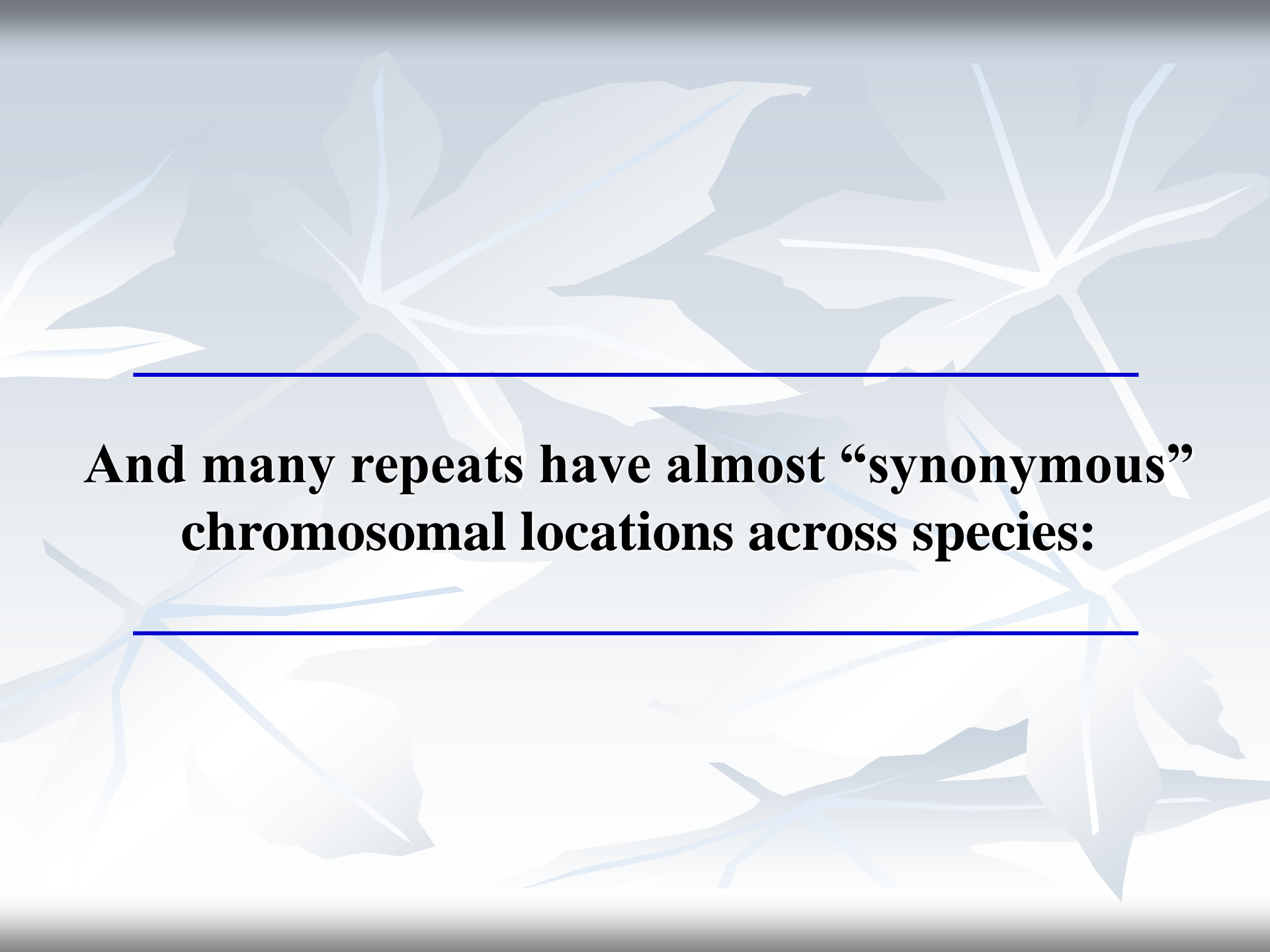
Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers

Mengliang Ye^a, Christel Goudot^a, Thomas Hoyle^a, Benjamin Lemoine^b, Sebastian Amigorena^{a,1}, and Elina Zueva^{a,1} 

www.pnas.org/cgi/doi/10.1073/pnas.1912008117

Predicted TF or TF family	logo	Enhancer		Gene desert		similar sequence in the consensus
		e-value	%	e-value	%	
ORR1						
ETS (Etv-Ets-Gabpa)		2.6e-244	50%	5.1e-30	22%	CAGGAAGT(T/G)
		2.5e-92	49%	5.1e-77	27%	CAGGAAGT(T/G)
		4.5e-28	46%	-	-	TTCCTCT
RUNX (1,2,3)		3.9e-64	16%	-	-	TGTGGTTT (AAACCACA)
Lin54		4.8e-31	27%	4.7e-22	21%	TTTGAATG (CATTCAAA)
Max_Myc		6.3e-18	30%	5.5e-11	14%	ACACTTGGT
MTD						
E2f/Erf_Fli1		2.8e-58	60%	5.1e-30	27%	TTCCTGC (GCAGGAA)
		1.5e-48	53%	-	-	TTCCTGC (GCAGGAA)
		5.8e-41	60%	2.1e-30	57%	TTCCTGC
Runx1		2e-61	43%	1.5e-15	29%	CTGTGGG (CCCACAG)
RMER						
Sp1, Klf, E2f2		1.3e-13	44%	-	-	TCCCTTCCCC
		4.8e-08	44%	-	-	CCCCTCCCC
Rel_RelA_Bcl6		9.8e-22	55%	-	-	TCCCTTCCCC
Tcf7_Lef1		5.8e-07	20%	-	-	AGACCAAC, TTTGGTCT
Tead3		6.8e-09	35%	-	-	ACCATACC
MTE						
ETS (Etv, Gabpa, Elk)		3.7e-44	41%	2e-17	36%	AGGAGAAA, AGGAGACA
Sp1, Klf		2.5e-18	30%	3.3e-13	26%	ACCCACCC
Zbtb26_Smad4		4.7e-07	15%	-	-	ATCTAGAAT
RLTR						
Klf1, RUNX		1.3e-10	53%	-	-	TGTGGTT
Prdm1_RelA		2.8e-08	32%	-	-	GAAAGTC
Zfp523_Zfp143		9.5e-07	34%	-	-	ACTAAAACA
MLT						
Rbpj		0.015	15%	-	-	TCCCCCA
Sp1/2_, Klf		0.025	12%	-	-	CCCTCCC
Hic1		0.054	11%	-	-	GCCACC
Forkhead, Znf384		0.009	22%	-	-	AAATAAAT

MIR						
Zfp787		9.5e-27	24%	5.8e-13	13%	GGGCCTCAGTTTC (GGA AACTGAG)
Zfp788		4.5e-20	18%	-	-	
Tbp		1.7e-15	16%	-	-	GTAAAATG (CATTTTAC)
Nfat		4.5e-20	16%	6.3e-14	23%	GTAAAATGG
Nr4f2_Essra		4.70e-16	19%	-	-	GTGACCT (AGGTCAC)
Gata		2.6e-11	11%	-	-	AGATGA
L2						
Sry/Zfp422_384/Forkhead		3e-19	20%	6.3e-21	13%	AATAAAA
		4.7e-48	18%	1.9e-24	10%	AAAAAACAAAAA
Sp1, Klf_Znf263		6.3e-51	15%	-	-	CCCCTCCCC
Fli1		3.1e-11	21%	-	-	AGGAG
		3.5e-46	12%	-	-	CACACA
L1						
Sry/Zfp422_384/Forkhead		1e-14	42%	-	-	TATTTTA (ATAAAT)
		1.1e-34	37%	-	-	AAAAACAAA
Setbp1_Ahctf1		1.5e-81	42%	3e-22	12%	TATTTTAA
Sp1/Klf, Znf148		3.4e-64	40%	ns		CCCCCCT(CT)CCCC
		7.4e-95	26%	9.8e-10	16%	CACACCCA
DNAhat						
Sreb1		7.40e-38	54%	-	-	GGGTCACCACAA (TTGTGGTGACCC)
		5.50e-48	51%	3.5e-31	40%	TAAAGGGTC
Rxra_Rxb_Zfp852		1.3e-142	46%	-	-	GACCCCT
B2						
Zfp384/Forkhead		1e-79	30%	1e-56	23%	ATAAAAATAAAA
Fos_Jun		7.1e-205	47%	-	-	GATGGCTCA
B1						
		1e-300	10%	1e-300	2%	AAAAAACAAAA
		3.5e-138	18%	2.3e-63	9%	AAAAAACAAAA
B4						
		1e-300	13%	9.8e-300	11%	CACACACACACA



**And many repeats have almost “synonymous”
chromosomal locations across species:**

Alu and B1 Repeats Have Been Selectively Retained in the Upstream and Intronic Regions of Genes of Specific Functional Classes

Aristotelis Tsirigos*, Isidore Rigoutsos*

PLoS Computational Biology

December 2009 | Volume 5 | Issue 12 | e1000610

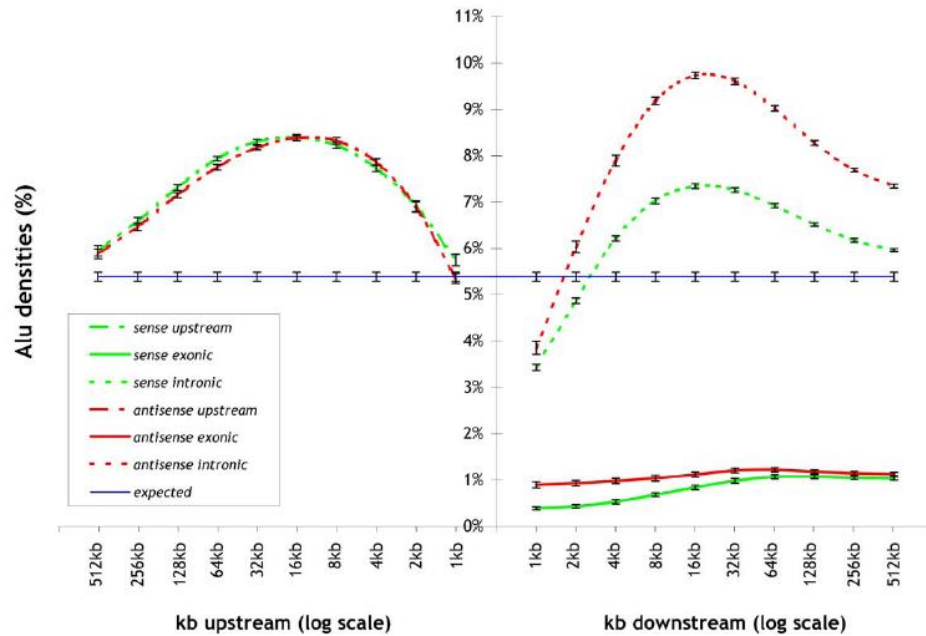


Figure 1. Alu densities upstream and downstream of known genes as a function of distance from the gene transcript start position.

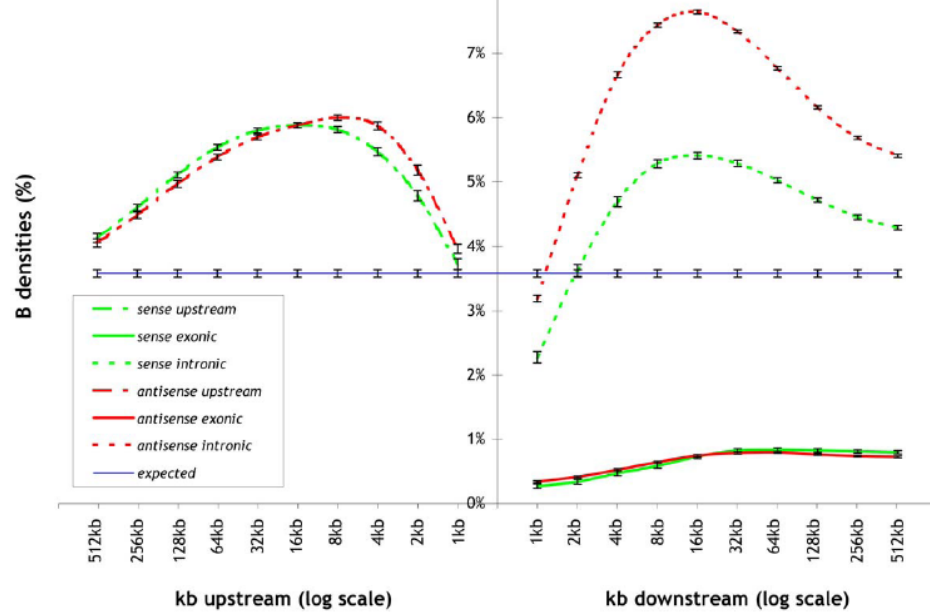
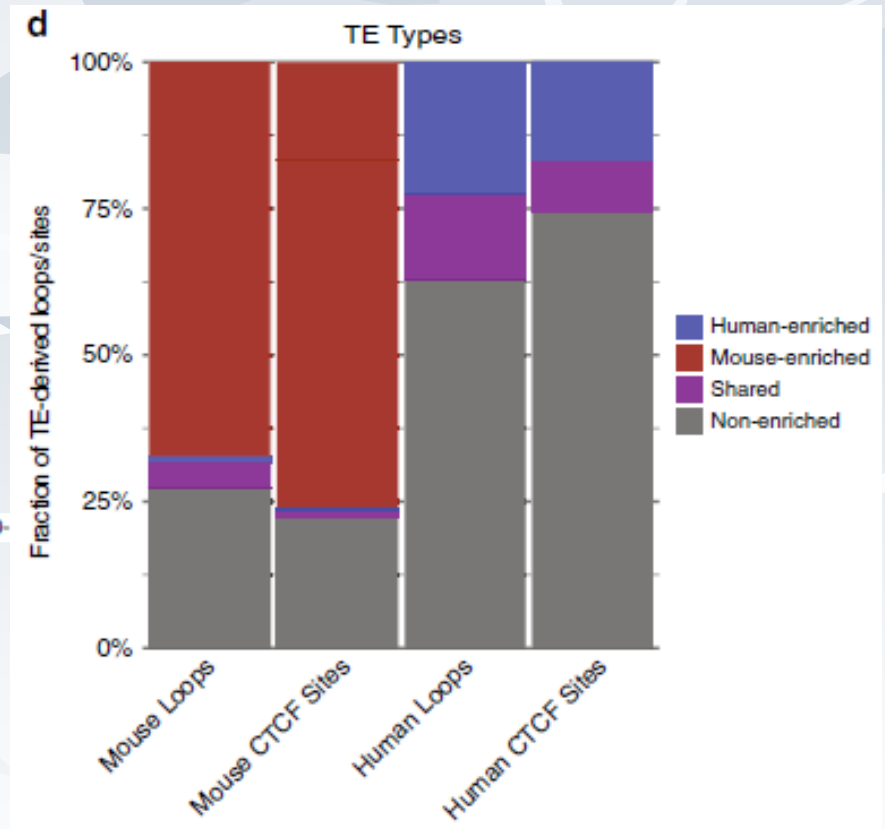


Figure 2. B element (B1, B2, B4) densities upstream and downstream of known genes as a function of distance from the gene transcript start position. Green and red curves correspond to B element instances in the sense and antisense orientation respectively.

Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes

Adam G. Diehl¹, Ningxin Ouyang¹ & Alan P. Boyle^{1,2}

NATURE COMMUNICATIONS | (2020)11:1796 | <https://doi.org/10.1038/s41467-020-15520->



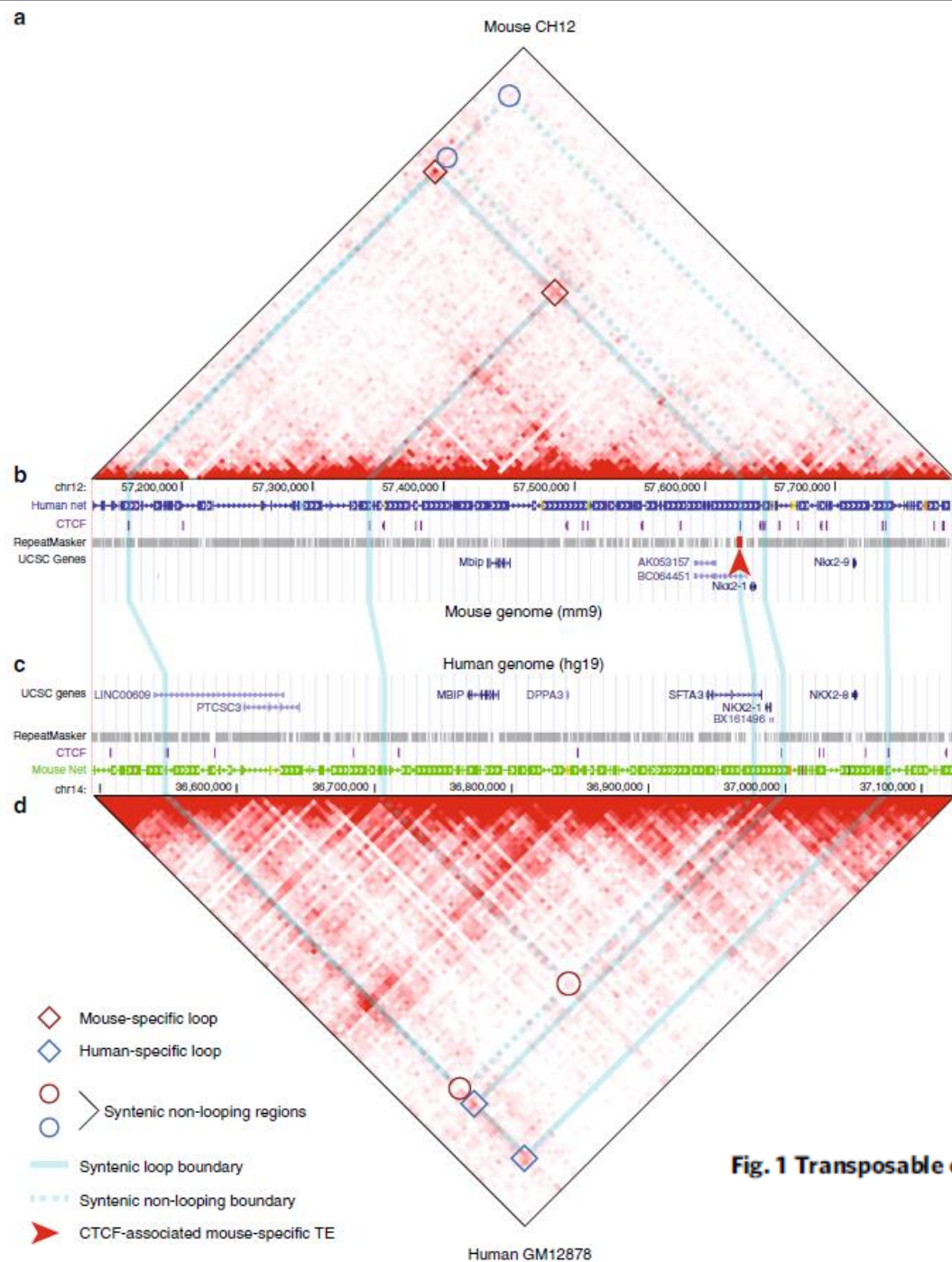
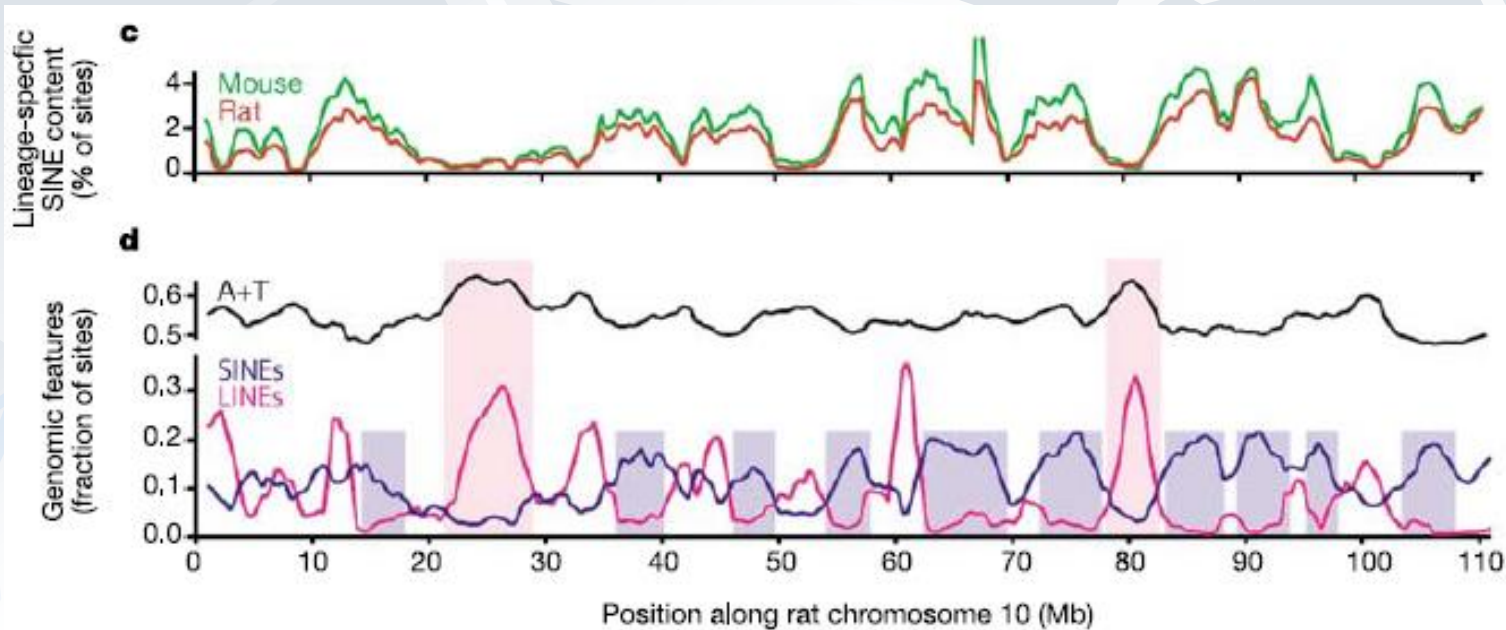


Fig. 1 Transposable element insertions create novel species-specific loop contacts.



The overall “data” pattern along a megafolder is the same *but* the species-specific details of the logic gates are different.



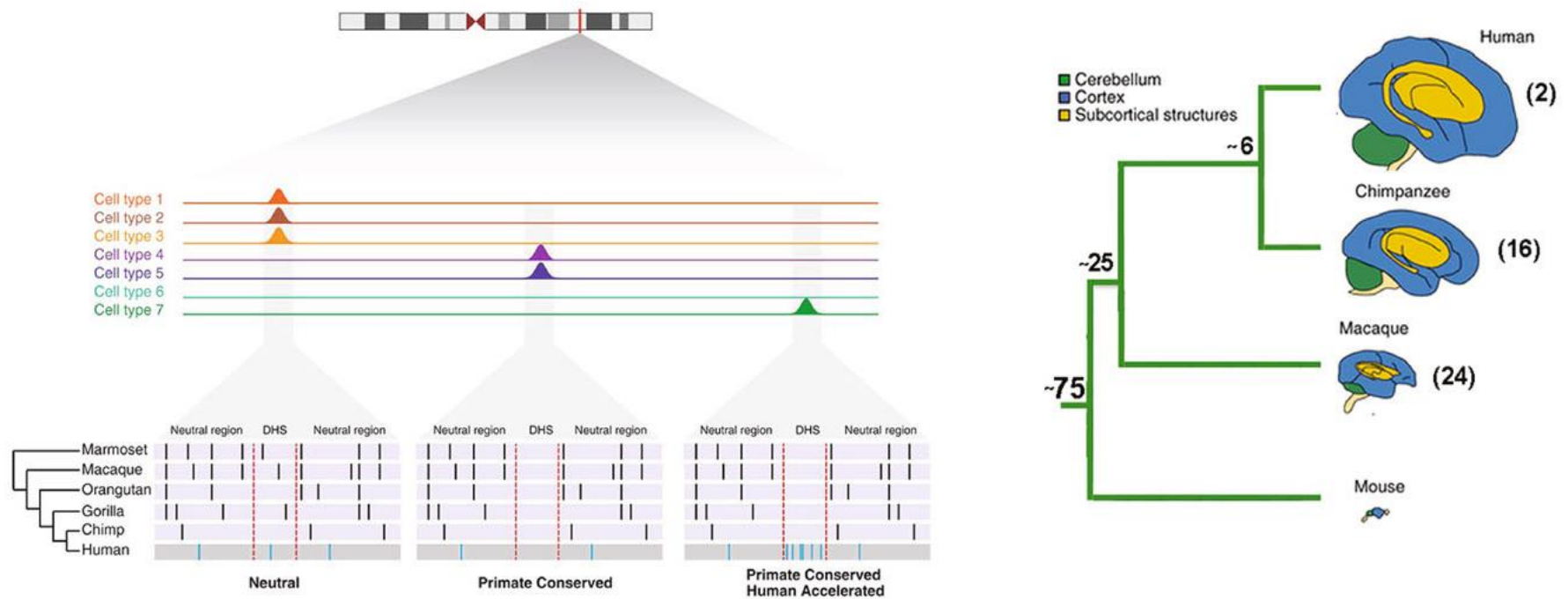
Genome sequence of the Brown Norway rat yields insights into mammalian evolution

NATURE | VOL 428 | 1 APRIL 2004



There's much more...

But in 2006 a number of so-called “Human Accelerated Regions” (HARs) were discovered. These have a divergence pattern that exceeds the rate of mutation.

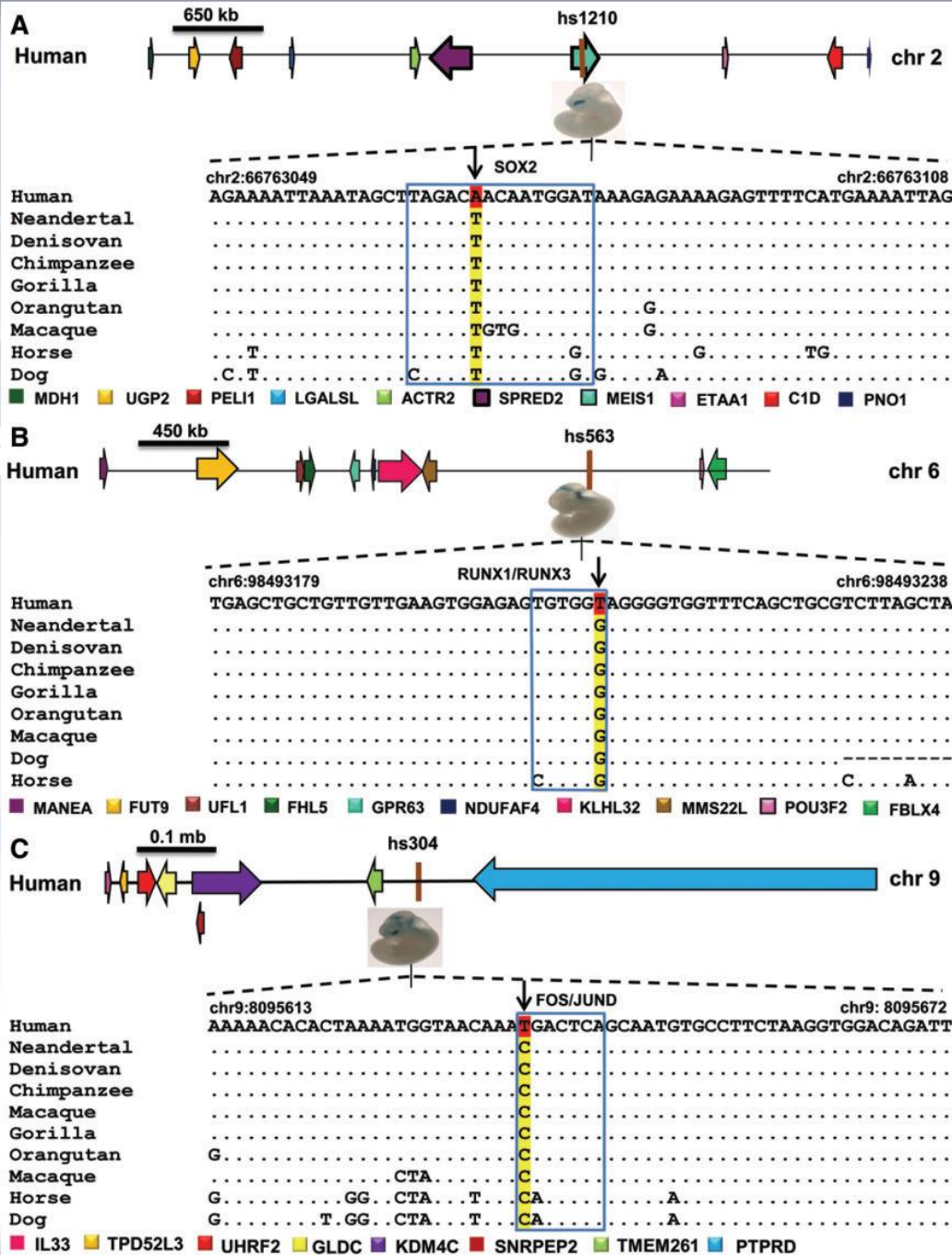


Franchini LF, Pollard KS. 2017. Human evolution: the non-coding revolution. *BMC Biol.* 15(1): 89.

Most of these:

- *have DNA-letter changes that are significantly greater than chromosome-wide, neutral substitutions;*
 - *are about 260 letters in length;*
 - *are non-randomly distributed (most are near the ends of chromosomes);*
 - *97% occur in segments long thought to be “junk”; and...*
 - *are disproportionately found near brain-specific regulatory sequences.*
-

**And of such most have been shown
to enhance gene expression (they
are “enhancers”)**



Zehra R, Abbasi AA. 2018. *Homo sapiens*-specific binding site variants within brain exclusive enhancers are subject to accelerated divergence across human population. *Genome Biol Evol.* 10(3): 956-966.

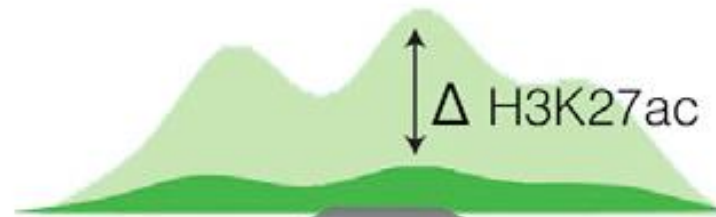
Human-unique protein-binding sites in 15 brain-specific enhancers...

SN	ID	GRCh37/hg19	Brain Domain	TF	TFBS
1	hs37	chr16: 54650598–54651882	Forebrain	PEA3	ACWTCK
2	hs1210	chr2: 66762515–66765088	Forebrain	SOX2 ^a	NNNANAACAAWGRNN
3	hs526	chr4: 1613479–1614106	Forebrain	NF1B	CTGGCASGV
4	— hs563	— chr6: 98491829–98493238	— Hindbrain	POU3F2 RUNX1/3 ^a	NWAAYA TGTGGT
5	hs1366	chr6: 38358690–38360084	Midbrain	TCFAP2B	CCCCAGGC
6	hs1632	chr11: 116521882–116522627	Midbrain	ZIC1	VGGGGAGS
7	hs1726	chr18: 49279374–49281480	Hindbrain	—	—
8	hs1526	chr2: 104353933–104357342	Forebrain	SOX9	RNACAAAGGVN
9	— hs847	— chr4: 42150091–42151064	— Forebrain	PBX1 LEF1	NYAYMCATCAAWNWN NNTCAAAGN
10	hs540	chr13: 71358093–71359507	Forebrain	MEF2A	TATTTWANM
11	hs1019	chr7: 20838843–20840395	Forebrain	—	—
12	hs192	chr3: 180773639–180775802	Forebrain	—	—
13	hs1301	chr11: 16423269–16426037	Forebrain	—	—
14	hs430	chr19: 30840299–30843536	Midbrain	—	—
15	hs304	chr9: 8095553–8096166	Mid/Fore	FOS/JUND ^a	TGACTCA/TGACTCAN
	—	—	—	NR2F1	TGACCTY
	—	—	—	NURR1	YRRCCTT

In addition, a number of these also modulate cranio-facial differences between chimps and humans:



Sequence analysis
of causative mutations



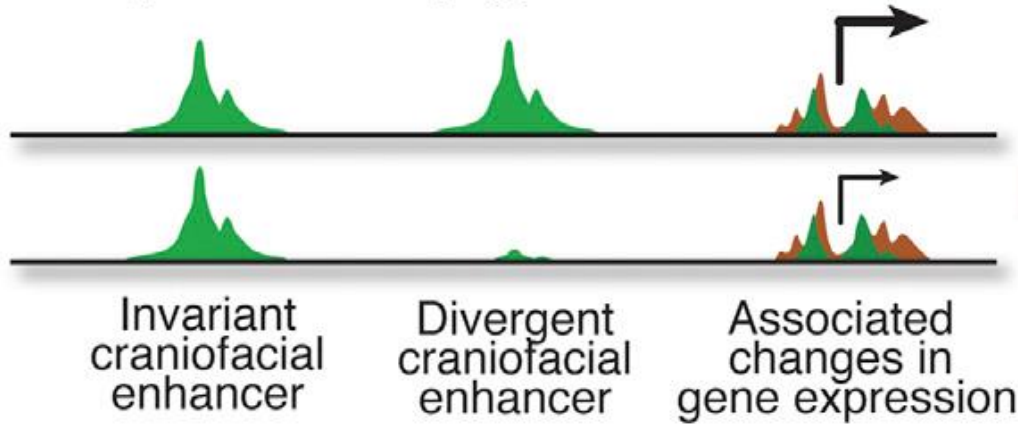
Coordinator motif



human AAATGAAAAACACATGT
chimp AAATGAAAAATACATGT

Comparative Epigenomics

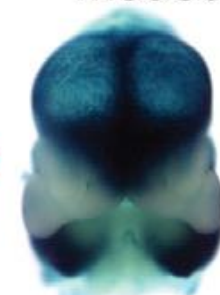
Validation mouse E11.5 face



Invariant
craniofacial
enhancer

Divergent
craniofacial
enhancer

Associated
changes in
gene expression



human
enhancer



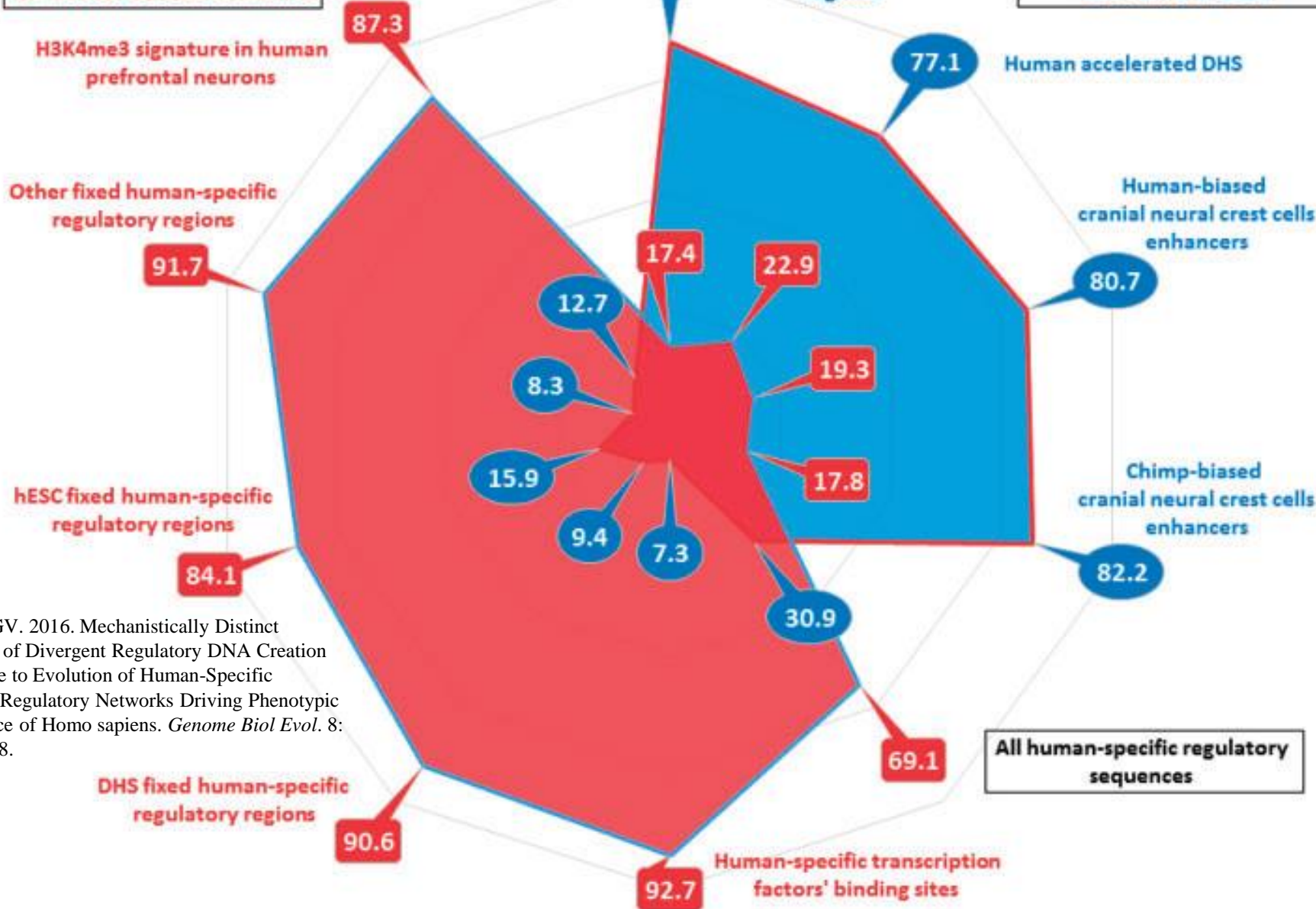
chimp
enhancer



There are also thousands (>18,000) of “human-specific regulatory sequences” (HSRSs) that are derived from retroviral-like elements.

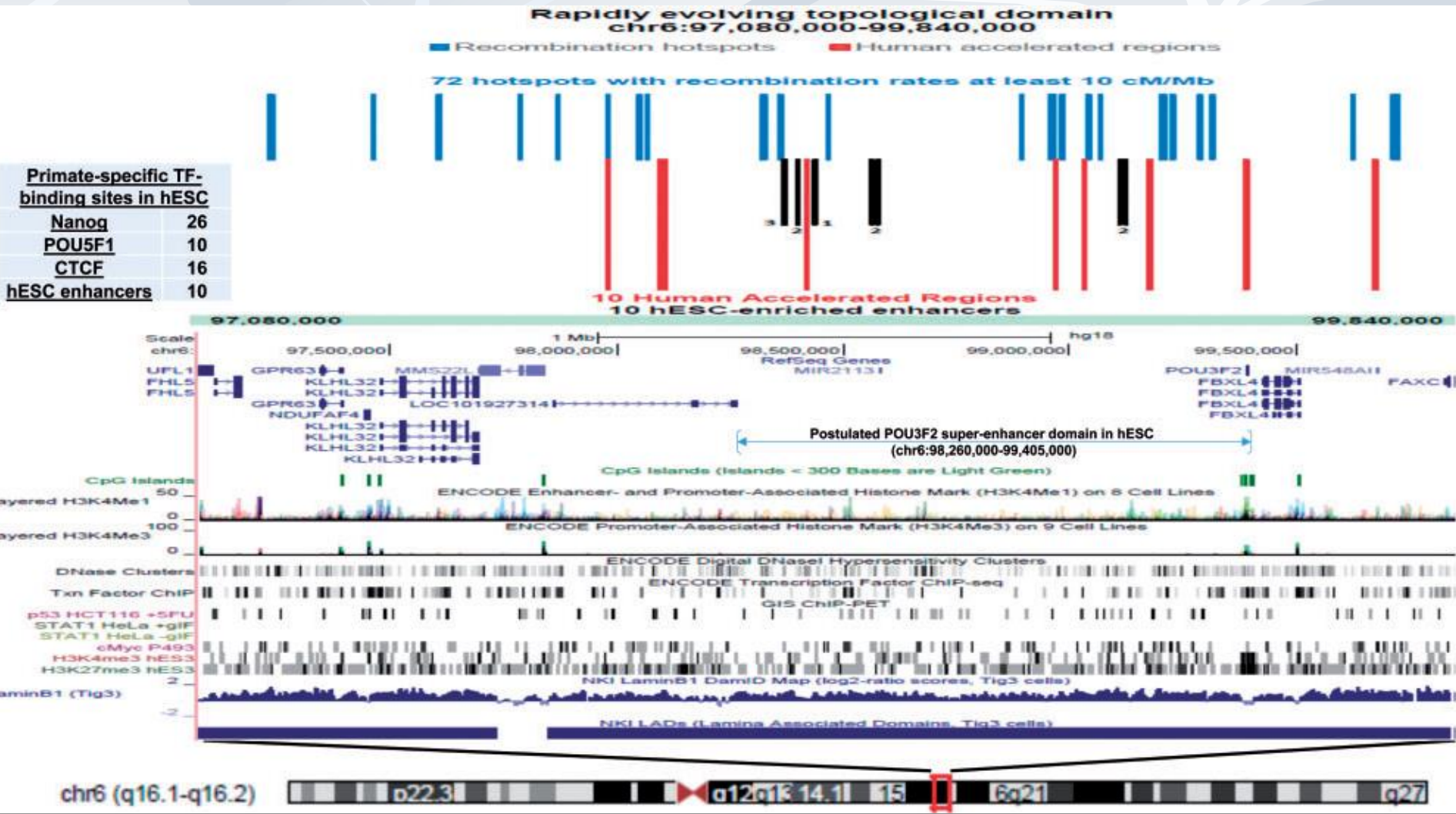
HUMAN-SPECIFIC EXPANSION OF TRANSPOSABLE ELEMENTS

EXAPTATION OF ANCESTRAL REGULATORY DNA



Glinsky GV. 2016. Mechanistically Distinct Pathways of Divergent Regulatory DNA Creation Contribute to Evolution of Human-Specific Genomic Regulatory Networks Driving Phenotypic Divergence of Homo sapiens. *Genome Biol Evol.* 8: 2774-2788.

Many of these reside in dense, high-complexity regions that are differentially folded into topologically-associated domains:

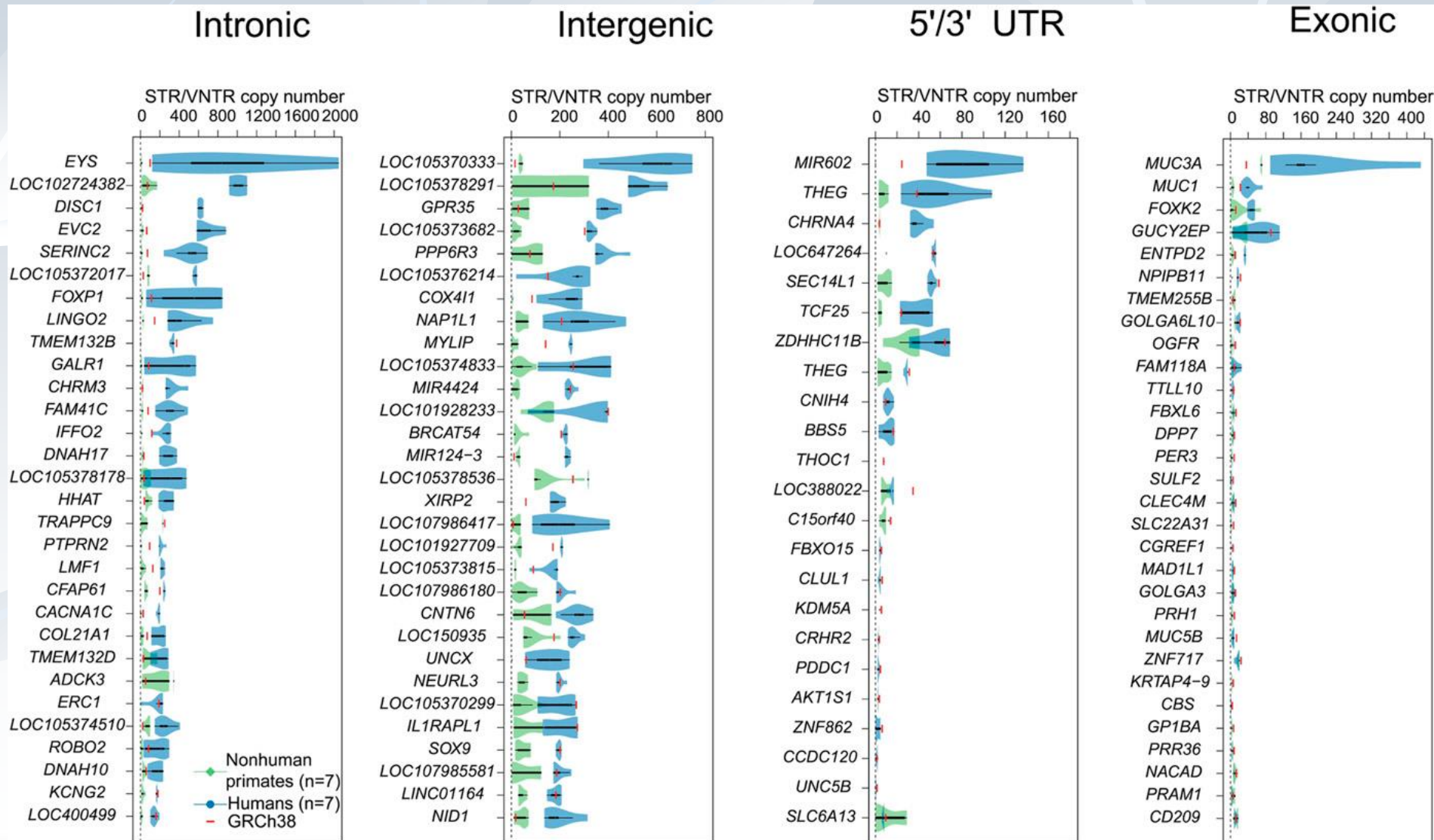


Genomic features of 60 HAR-based topologically-associating domains:

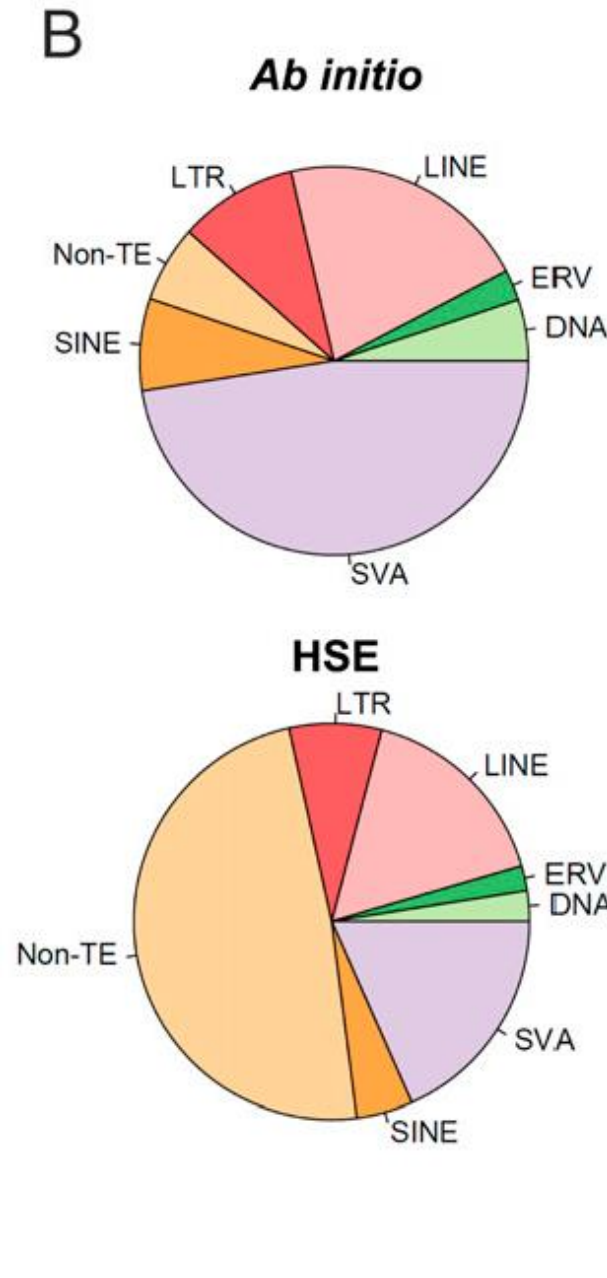
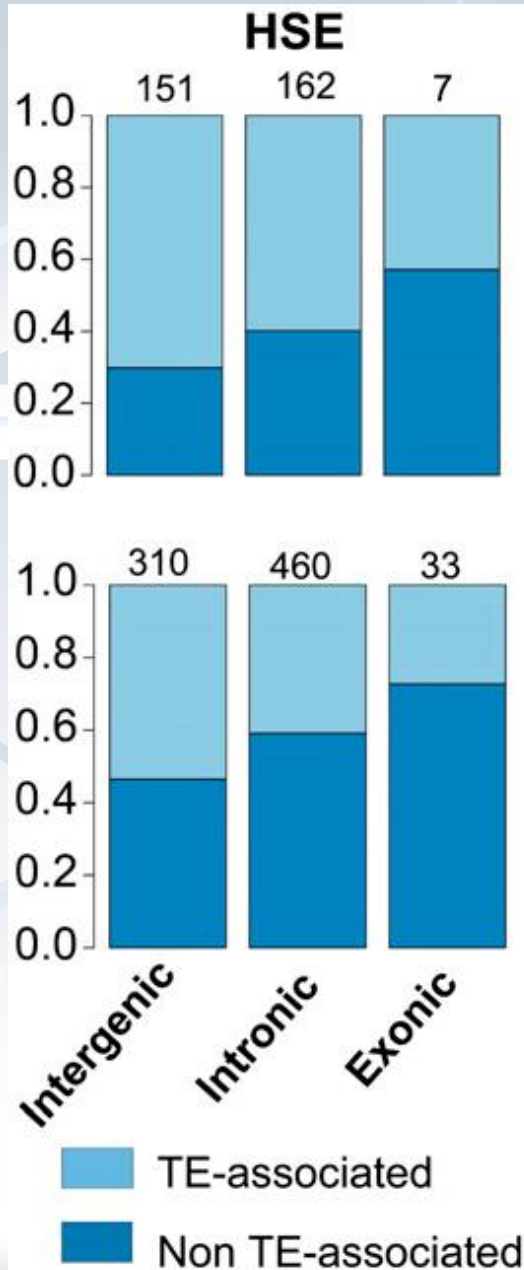
Genomic features	Genome	revTADs	Expected	Enrichment	P-value
Human Accelerated Regions (HARs)	2,745	378	53	7.4	<0.0001
Human-specific TFBS	3,803	1,370	73	18.8	<0.0001
Lamina-associated domains (LADs)	1,344	54	26	2.1	0.0019
Human-specific CTCF-binding sites	591	312	11	28.4	<0.0001
Human-specific NANOG-binding sites	826	192	16	12	<0.0001
Human-specific RNAPII-binding sites	290	181	6	30.2	<0.0001
Human-specific regulatory regions identified in H1-hESC	1,932	109	37	2.9	<0.0001
Human-specific regulatory regions identified in multiple cells	4,249	417	82	5.1	<0.0001
DHS-defined human-specific regulatory regions	2,118	558	41	13.6	<0.0001
Human-specific conservative deletions (CONDELs)	583	29	11	2.6	<0.0001
Human ESC enhancers	6,823	240	131	1.8	<0.0001
Human-specific transcriptional network in the brain	6,622	147	127	1.2	0.3856
Primate-specific CTCF-binding sites	29,081	1,269	558	2.3	<0.0001
H3K27ac peaks with human-specific enrichment in embryonic limb at E33 stage	780	31	15	2.1	0.0238
H3K4me3 peaks with human-specific enrichment in prefrontal cortex (PFC) neurons	410	29	8	3.6	<0.0001

hESC, human embryonic stem cells; TFBS, transcription factor-binding site; HARs, human accelerated region; LAD, lamina-associated domain; TAD, topologically- associating domain; RNAPII, RNA polymerase II; PFC, prefrontal cortex; DHS, DNase hypersensitive sites; CONDELs, conservative deletions; E33, embryonic day 33...

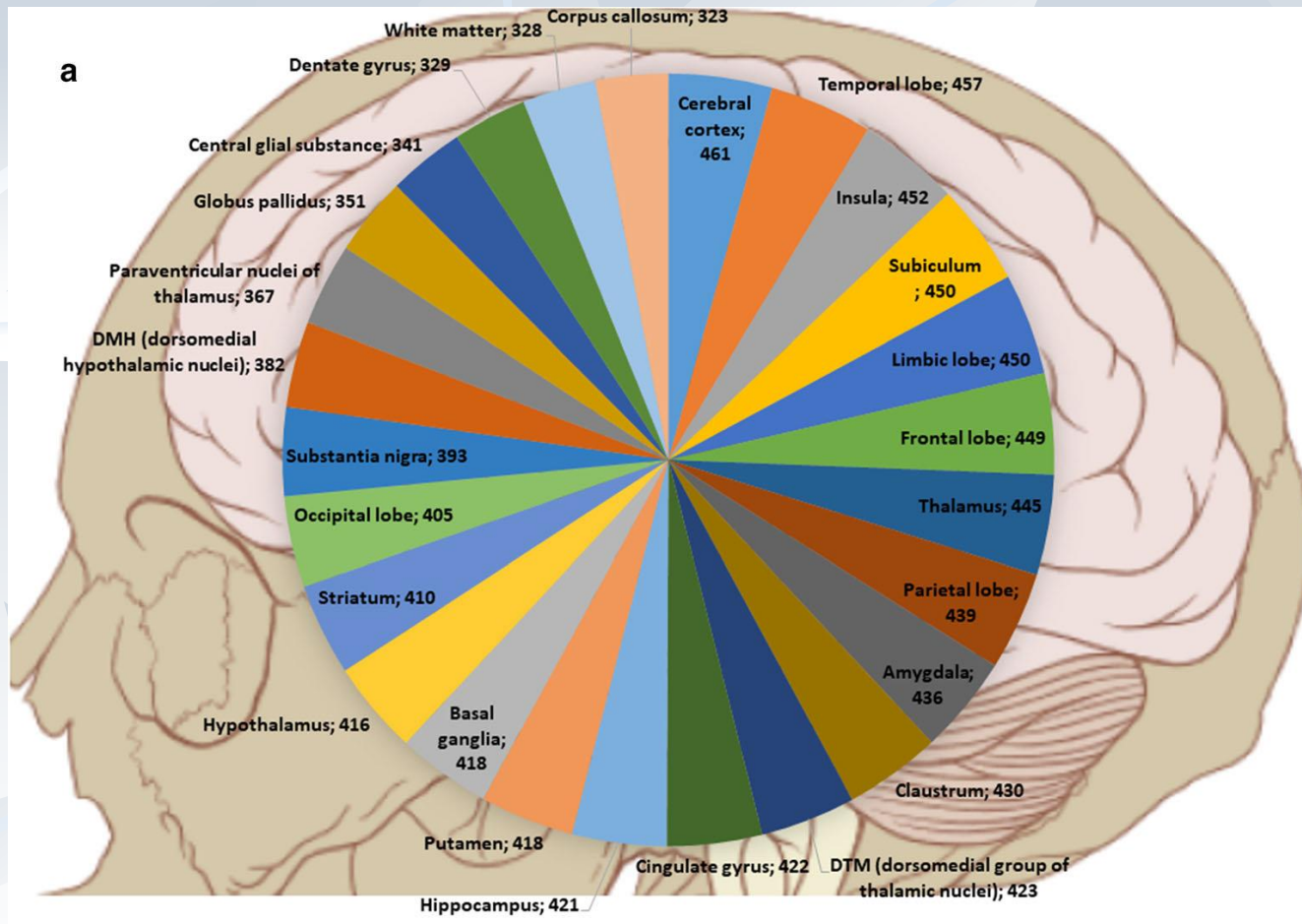
Moreover, at least 1,584 “short tandem repeats” are unique to our DNA:



Sulovari A, Li R, Audano PA, et al. 2019. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc Natl Acad Sci U S A*. 116(46): 23243-23253.

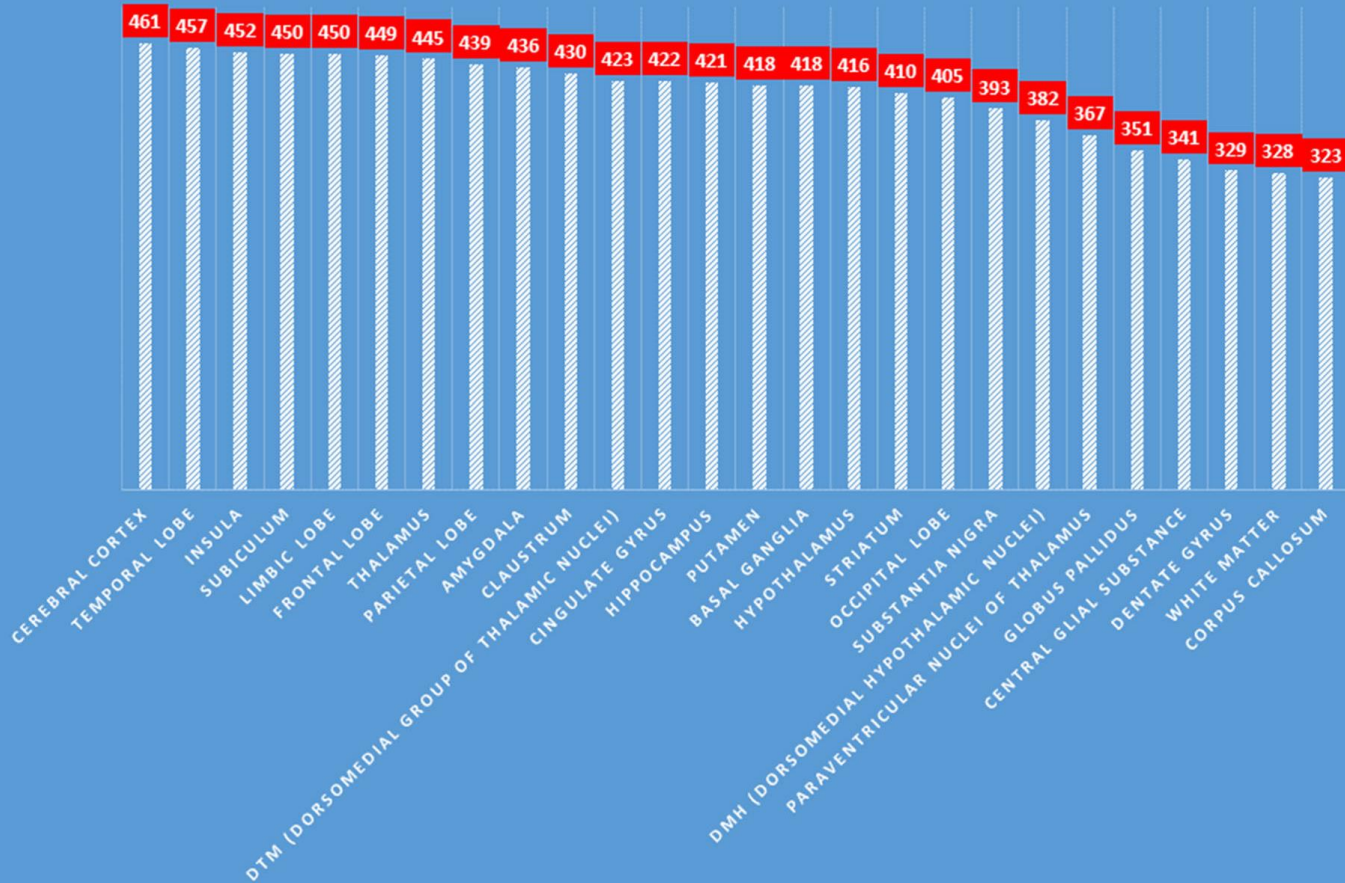


These in turn are associated with so-called “jumping genes” or “transposable elements”.

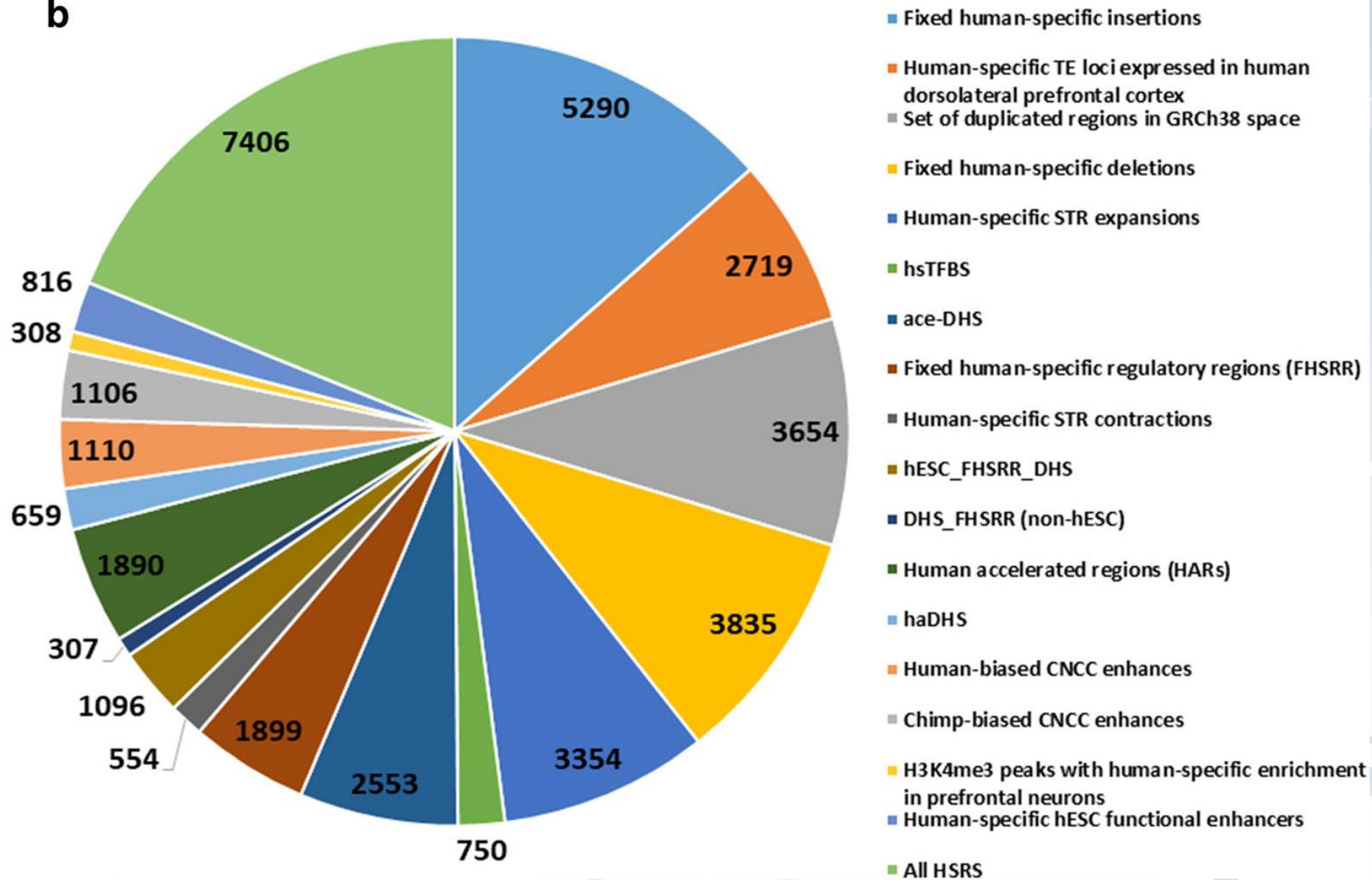


**35,074
human-
specific DNA
letter changes
are linked to
6640 of 8405
brain-
expressed
genes.**

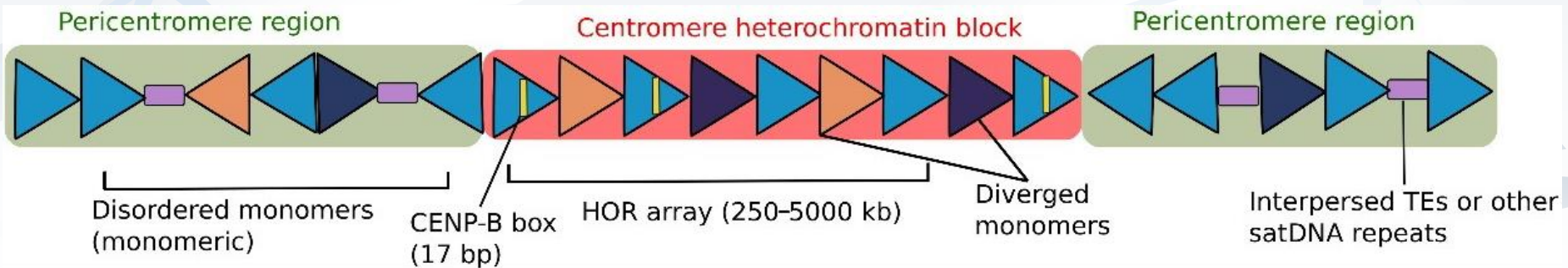
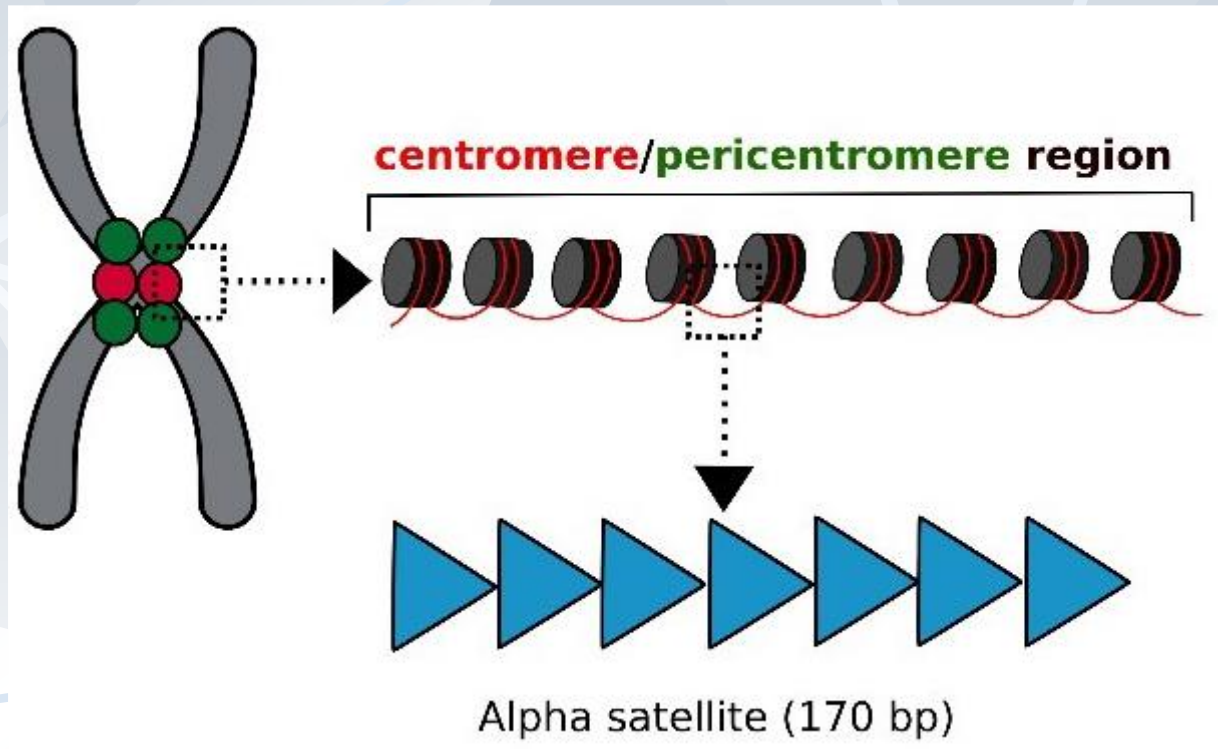
NUMBER OF HUMAN BRAIN REGIONS' MARKER GENES LINKED TO 35,074 HUMAN-SPECIFIC SINGLE NUCLEOTIDE CHANGES



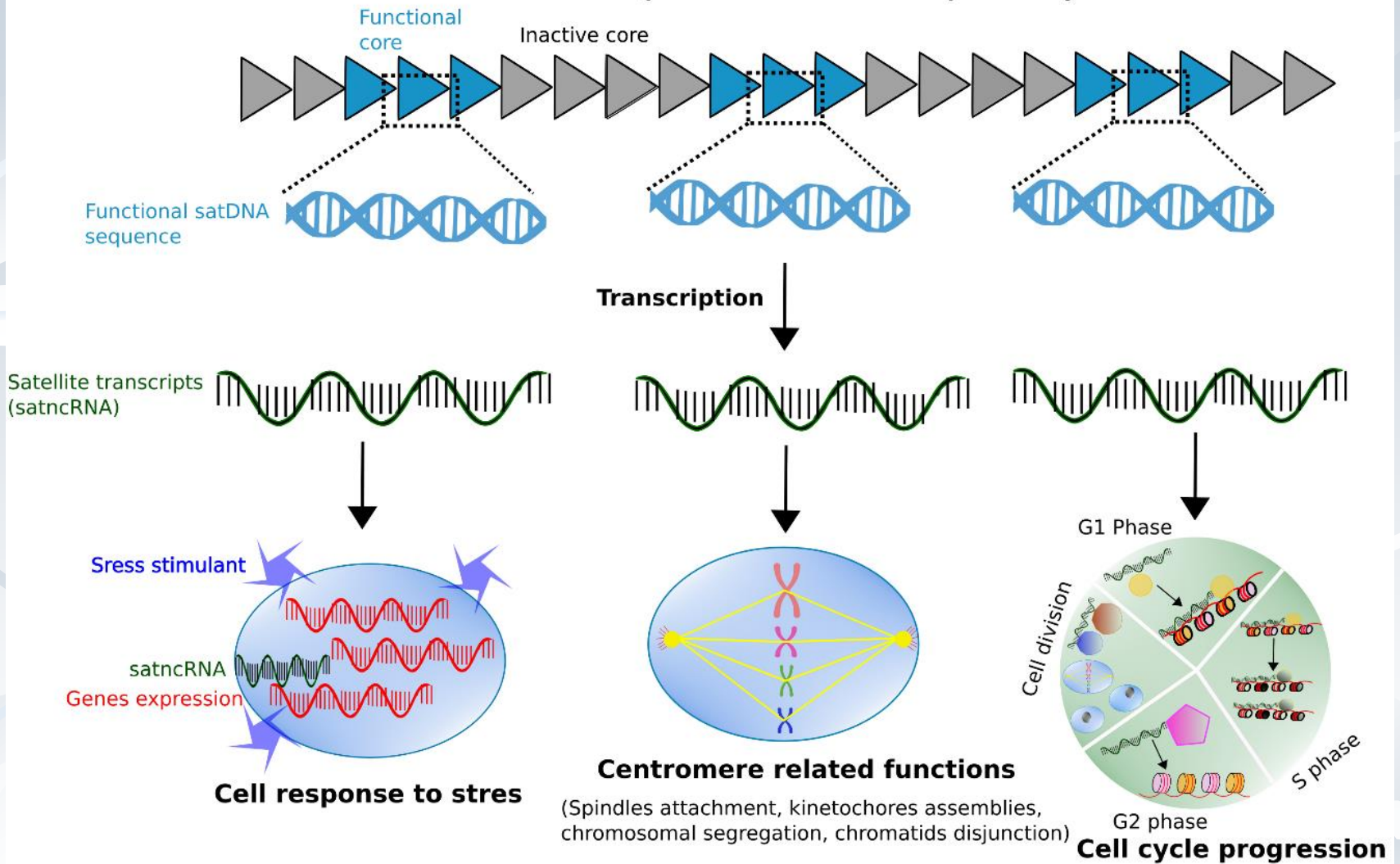
Distinct families of regulatory DNA sequences make up 59,089 human-specific regulatory sequences (HSRS) in or near 8405 ‘genes’; these are neuro-regulatory and linked to retroviral-like elements.

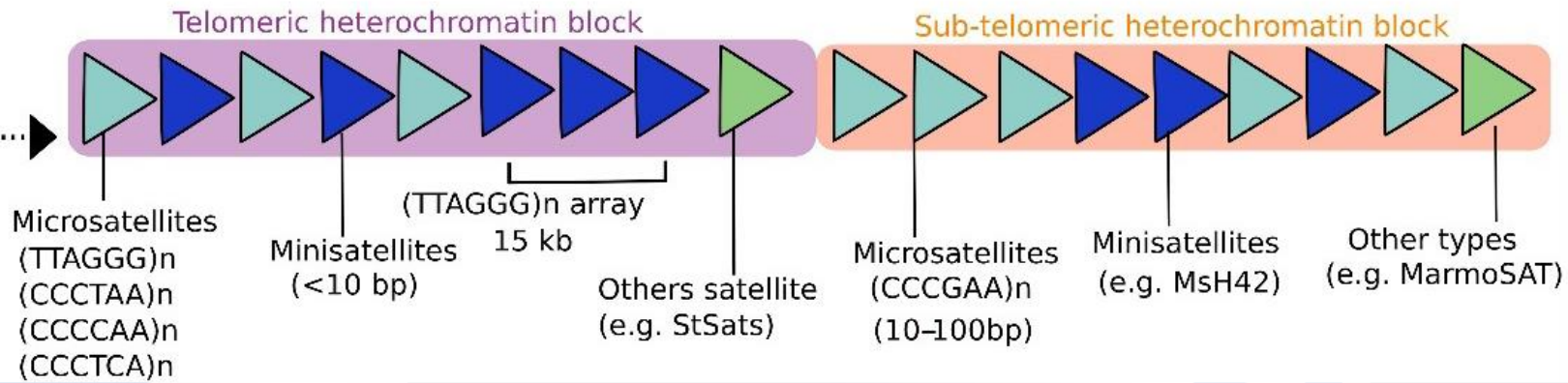
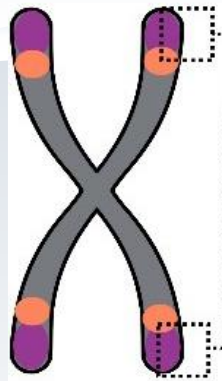
b

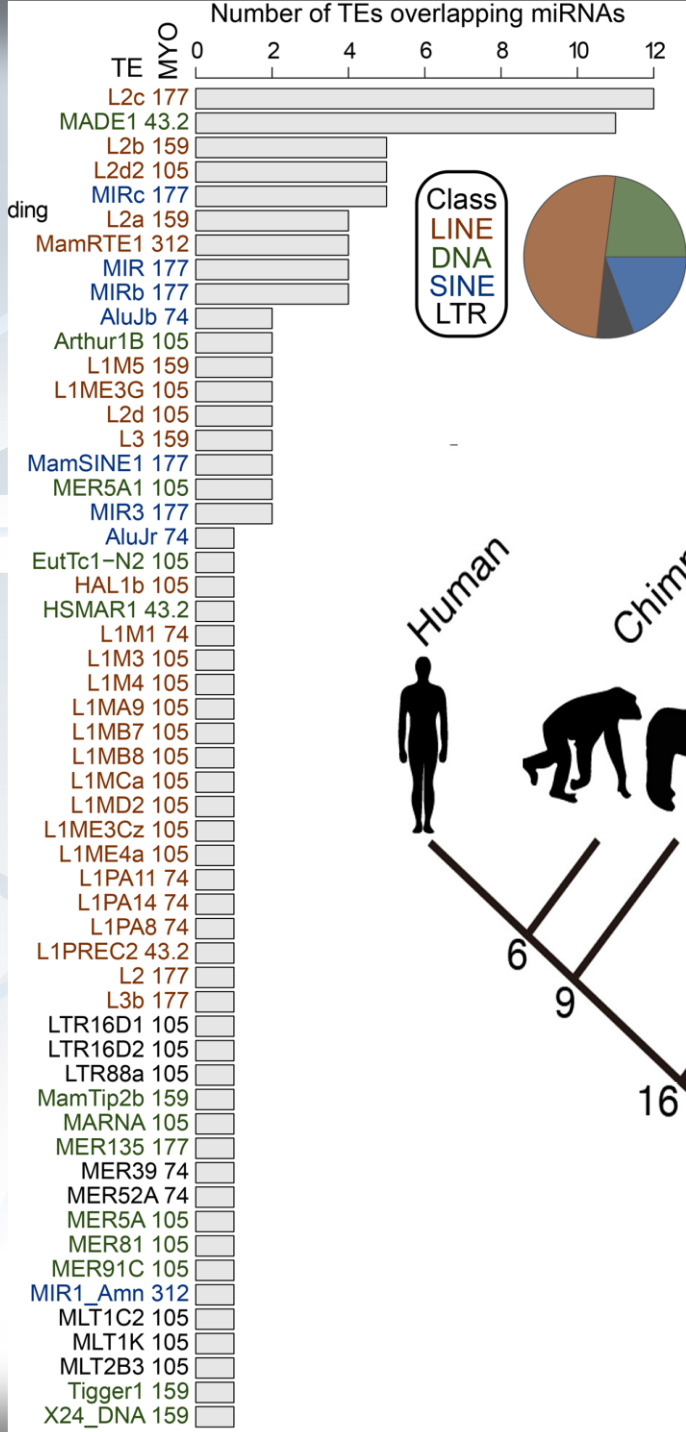
The “dark matter” of our genome also has distinct features.



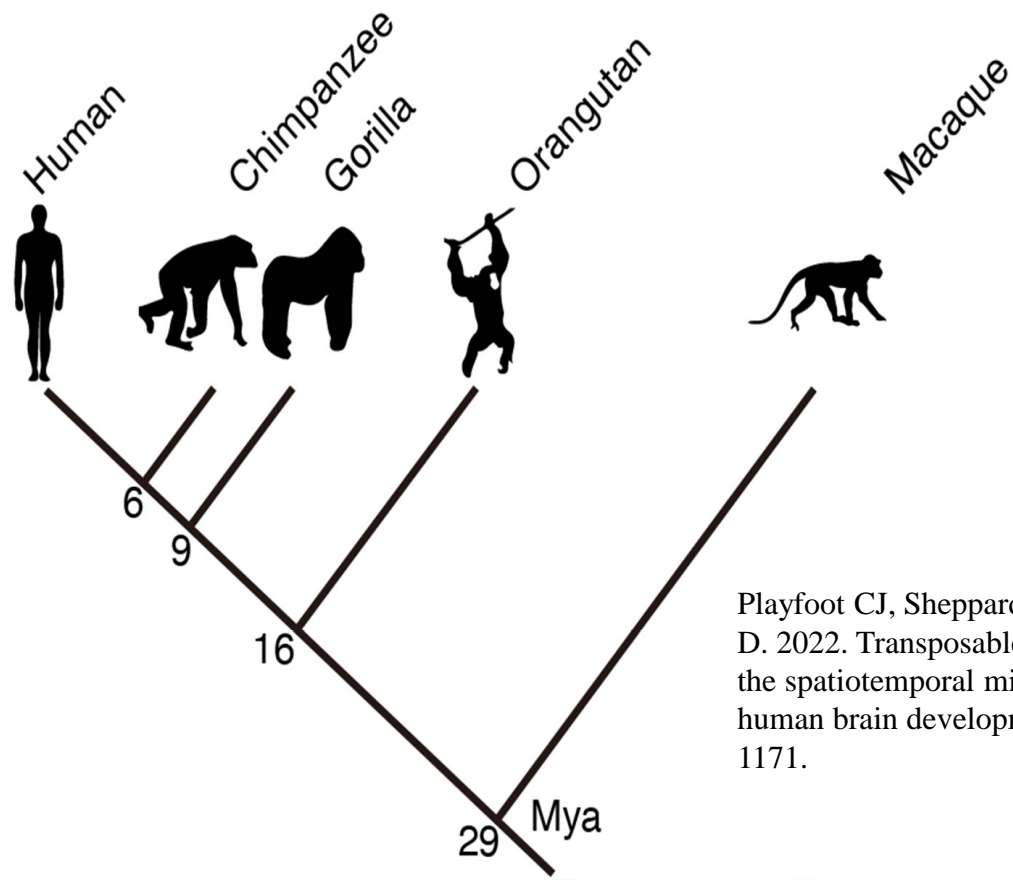
Centromeric/pericentromeric satDNA repeats array



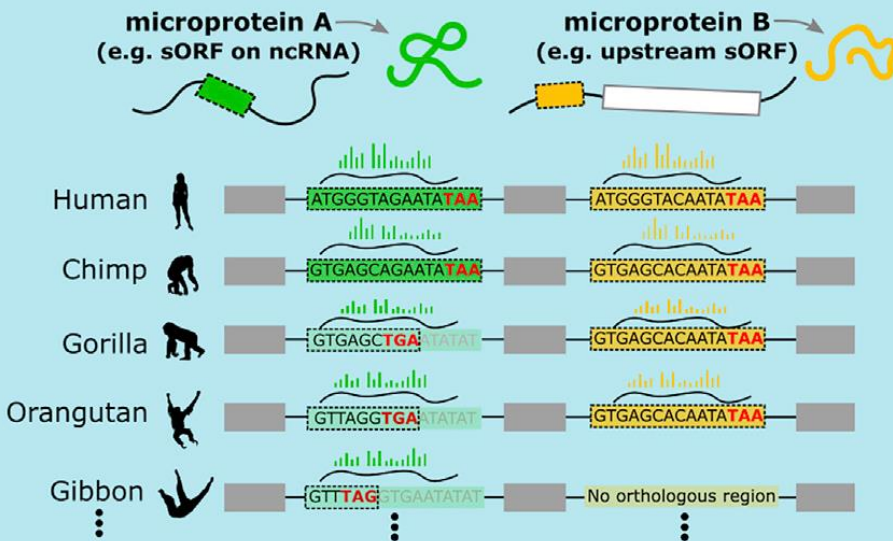




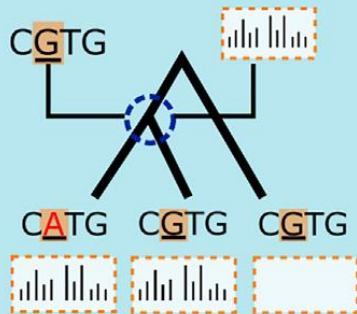
A number of human “jumping genes” also encode various brain-expressed “micro-RNAs”.



Playfoot CJ, Sheppard S, Planet E, and Trono D. 2022. Transposable elements contribute to the spatiotemporal microRNA landscape in human brain development. *RNA* 28:1157-1171.

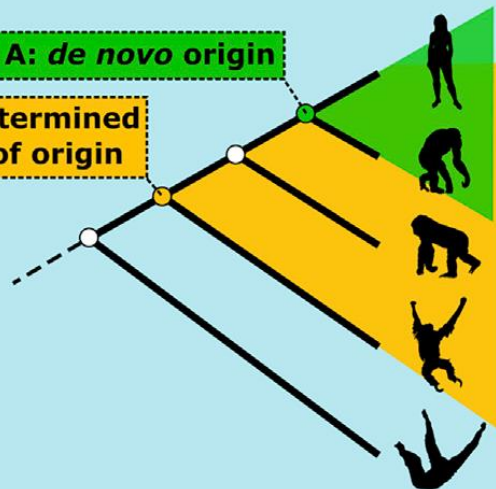


Ancestral Sequence Reconstruction + Inference of ancestral transcription

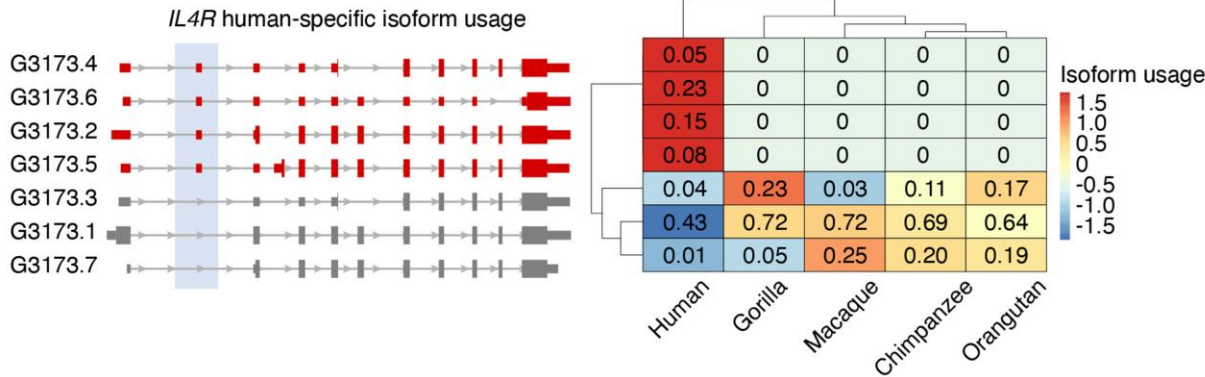
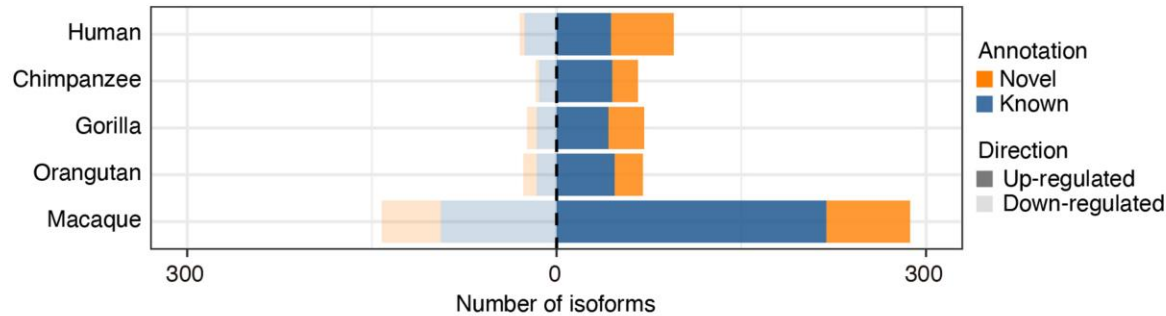
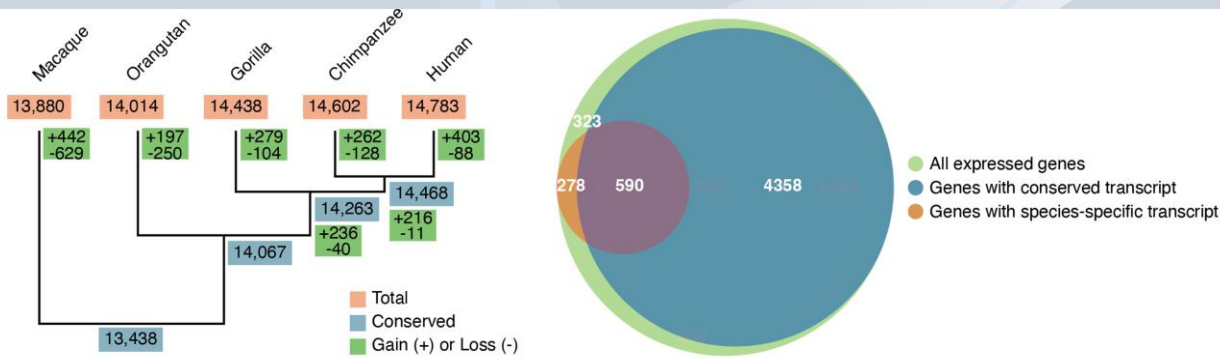


A: de novo origin

B: undetermined mode of origin



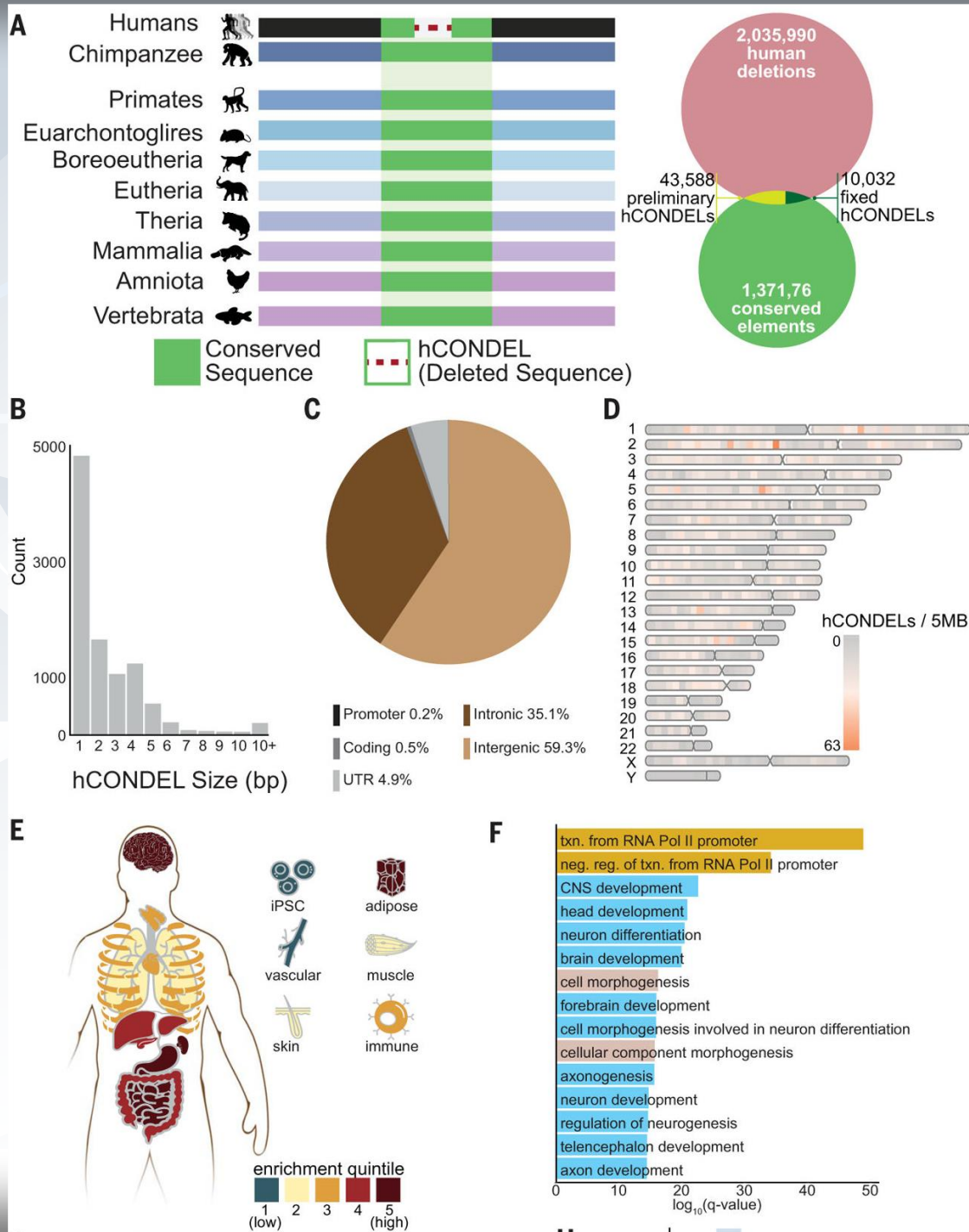
Thousands of “micro-proteins” are also encoded by human-specific DNAs, which have no counterparts in chimps and other primates.



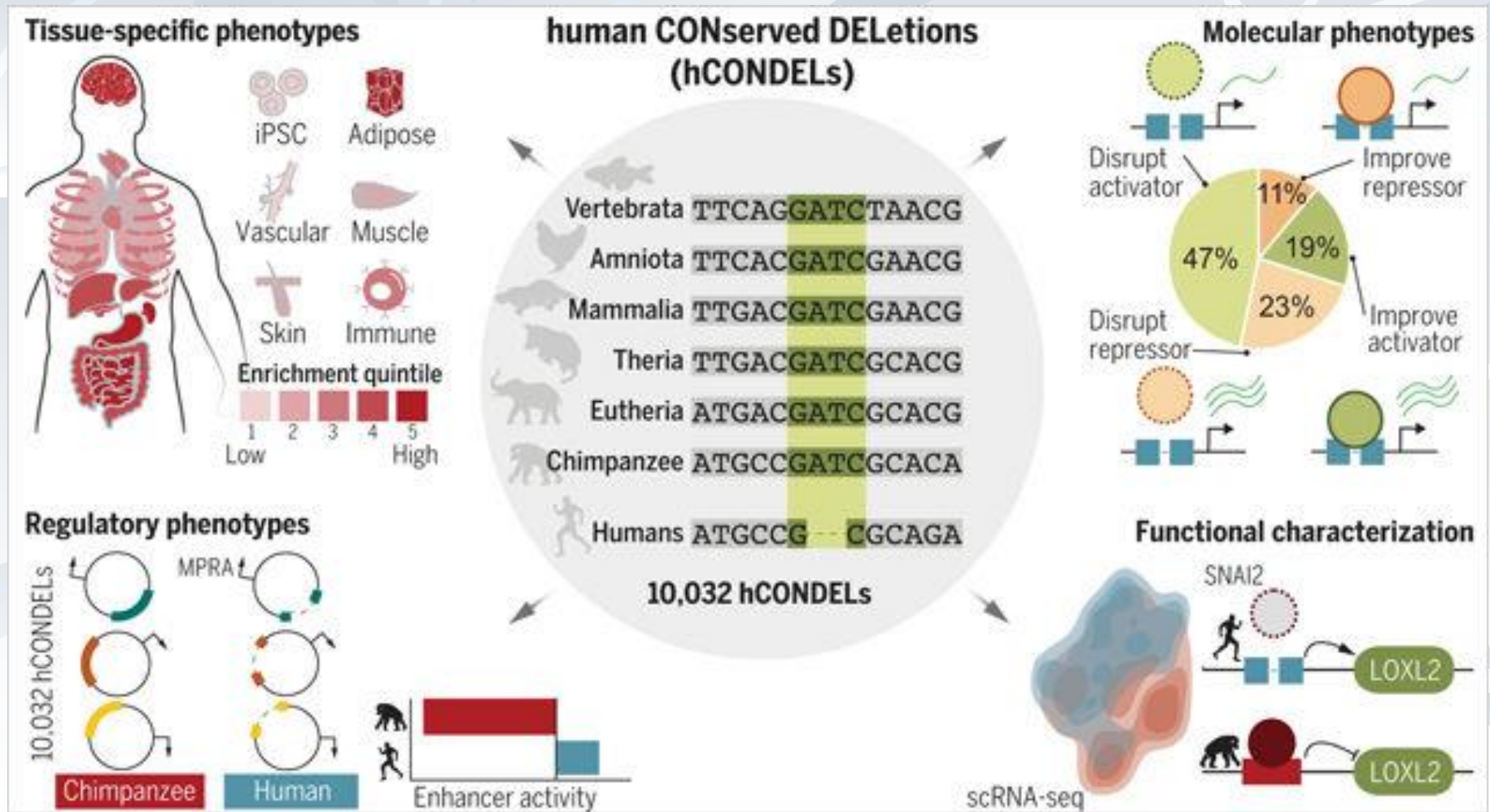
How genes are “transcribed” or used by our cells is human-specific too in hundreds of instances.

On the other hand, there are *millions* of human-specific absences all throughout our chromosomes, which are distinctly non-random.

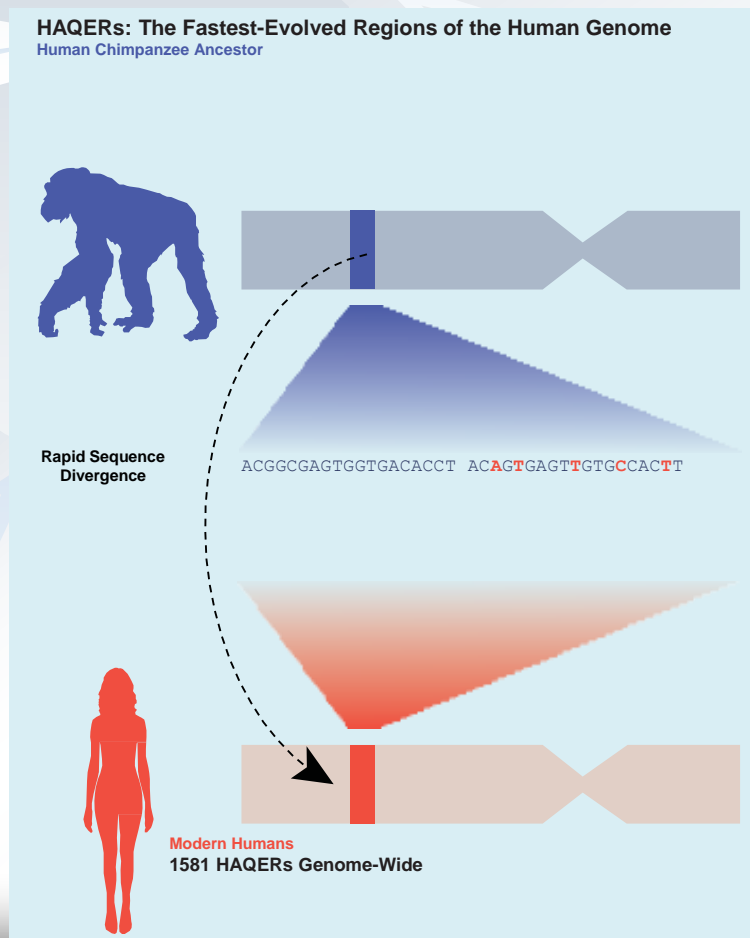
Xue JR, Mackay-Smith A, Mouri K, et al. 2023. The functional and evolutionary impacts of human-specific deletions in conserved elements. *Science* 380(6643): eabn2253.



Human-specific absences often occur in regions that are highly conserved in other animals.



Then again there are human DNA enhancers that are highly divergent from chimps and other primates:

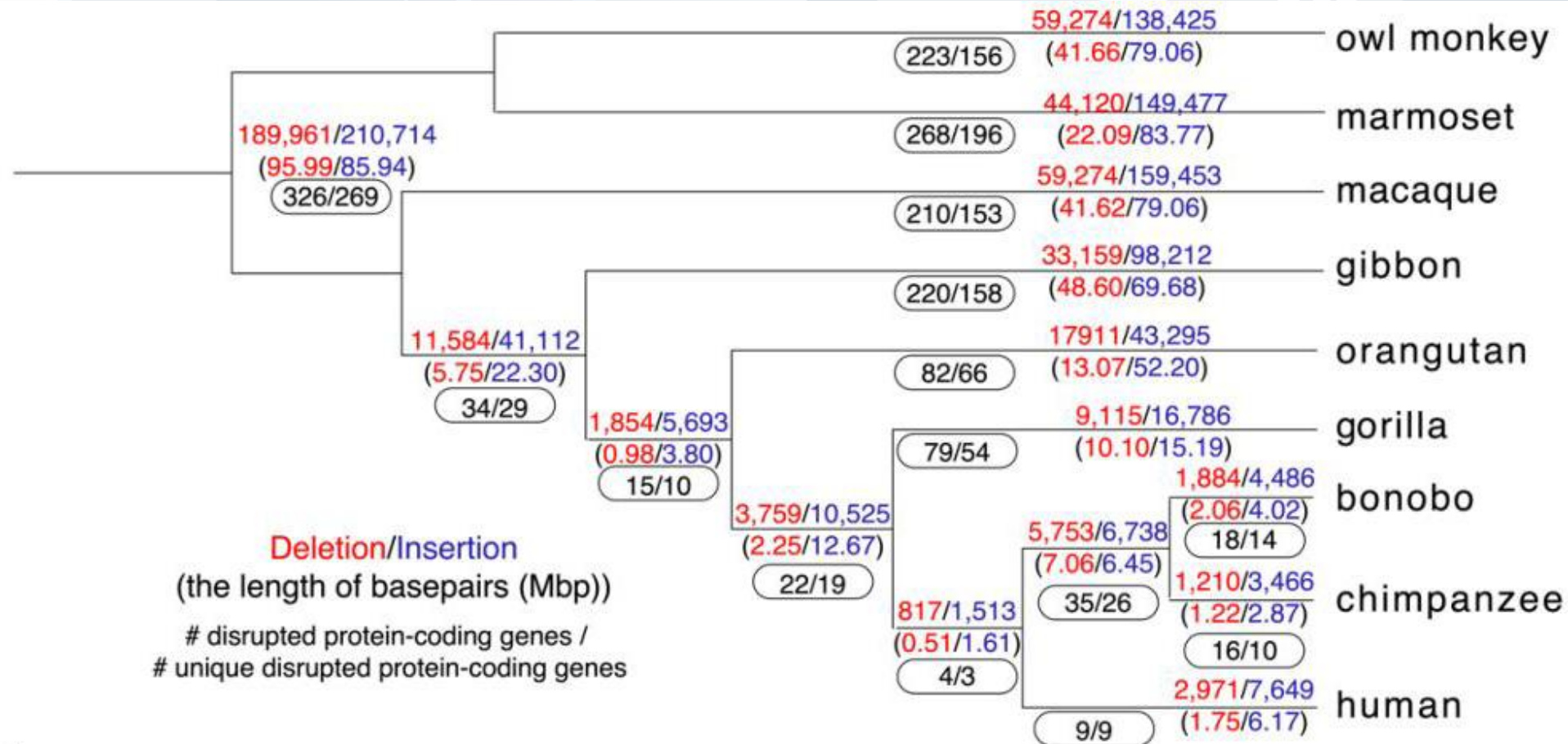


Once more, these are located in or near genes that are integral to our neurological systems.

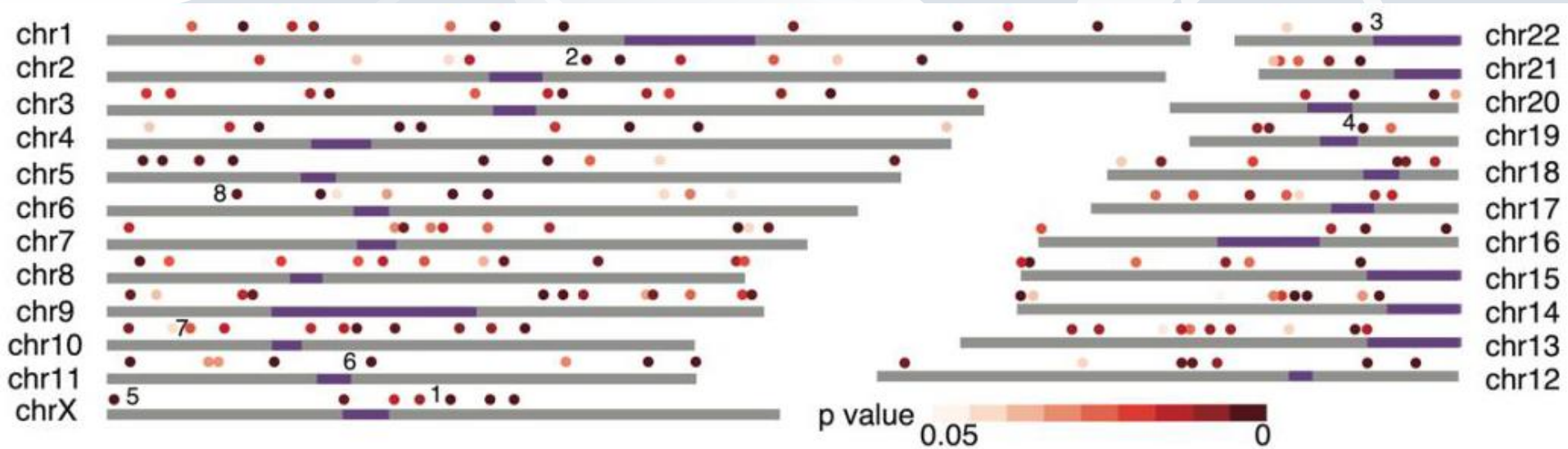
Mangan RJ, Alsina FC, Mosti F, et al. 2022. Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* 185(24): 4587-4603.

**The upshot of all this is that
hundreds, thousands, and millions
of our DNA code-letters had
explosive origins — *and* became
newly functional in a short period
of time!**

To close: some facts to ponder...



Hotspots in primate chromosomes where hundreds (754) of rearrangements are non-random in their occurrence and position.



Chimpanzee DNA

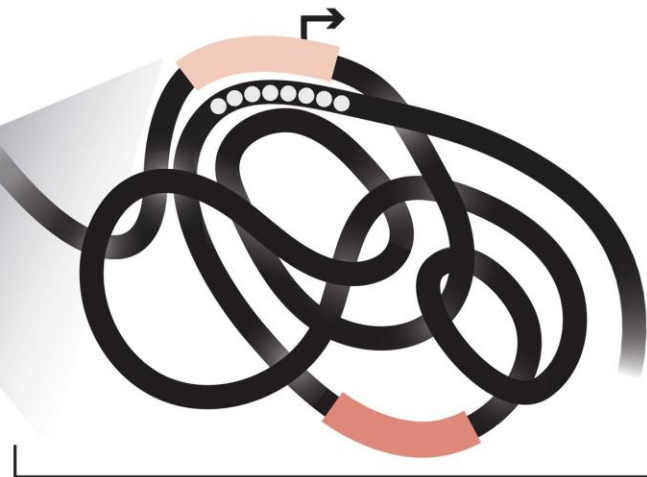
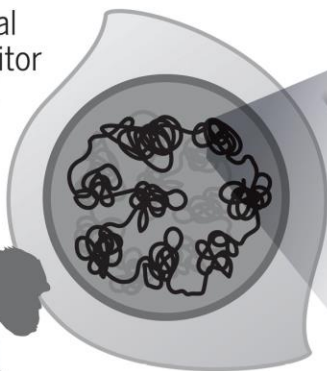
Gene A

Gene B

Human accelerated region (HAR)



Neural progenitor cell



Topologically associating domain (TAD)

Human DNA

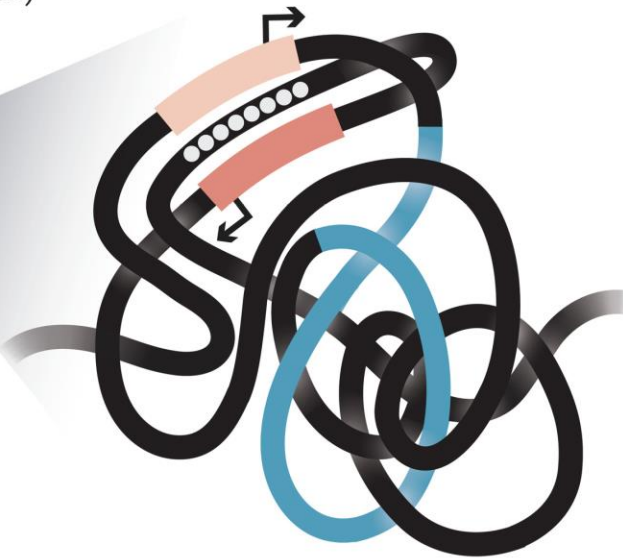
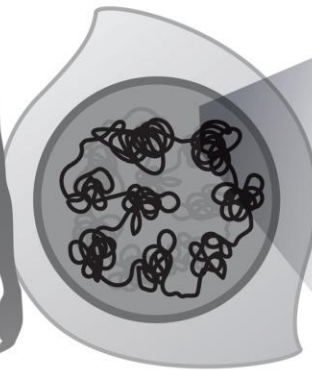
Gene A

Human-specific structural variant (hsSV)

(Insertion)

Gene B

HAR



**Next:
our
genetics
in 3D
and 4D!**