Creation of distal enhancers defines human-specific features of interphase chromatin architecture in embryonic stem cells

Gennadi V. Glinsky[1]

[1] Institute of Engineering in Medicine

University of California, San Diego

9500 Gilman Dr. MC 0435

La Jolla, CA 92093-0435, USA

Correspondence: gglinskii@ucsd.edu

Web: http://iem.ucsd.edu/people/profiles/guennadi-v-glinskii.html

Running title: Evolution of human-specific regulatory networks

Key words: topologically associating domains; super-enhancers; super-enhancer domains; human-specific genomic regulatory sequences; chromatin loops; human ESC; pluripotent state regulators; NANOG; POU5F1 (OCT4); CTCF; methyl-cytosine deamination; recombination; Alu elements; LTR7 RNAs; L1 retrotransposition; LINE; LTR; LTR7/HERVH; LTR5_HS/HERVK; evolution of Modern Humans;

List of abbreviations

5hmC, 5-Hydromethylcytosine

CTCF, CCCTC-binding factor

DHS, DNase hypersensitivity sites

FHSRR, fixed human-specific regulatory regions

GRNs, genomic regulatory networks

HAR, human accelerated regions

hCONDEL, human-specific conserved deletions

hESC, human embryonic stem cells

HSGRL, human-specific genomic regulatory loci

HSNBS, human-specific NANOG-binding sites

HSTFBS, human-specific transcription factor-binding sites

LAD, lamina-associated domain

LINE, long interspersed nuclear element

lncRNA, long non-coding RNA

LTR, long terminal repeat

MADE, methylation-associated DNA editing

mC, methylcytosine

mESC, mouse embryonic stem cells

NANOG, Nanog homeobox

nt, nucleotide

POU5F1, POU class 5 homeobox 1

TAD, topologically associating domains

TE, transposable elements

TF, transcription factor

SE, super-enhancers

SED, super-enhancer domains

Abstract

Molecular and genetic definitions of human-specific changes to genomic regulatory networks (GRNs) contributing to development of unique to human phenotypes remain a highly significant challenge. Genome-wide proximity placement analysis of diverse families of human-specific genomic regulatory loci (HSGRL) identified topologically-associating domains (TADs) that are significantly enriched for HSGRL and designated rapidly-evolving in humans TADs (Genome Biol Evol. 2016 8; 2774-88). In this contribution, the genome-wide proximity placement analysis of human-specific genomic regulatory loci (HSGRL), hESC-enriched enhancers, super-enhancers (SEs), topologically-associating domains (TADS), and specific sub-TAD structures termed super-enhancer domains (SEDs) has been performed. In the hESC genome, 331 of 504 (66%) of SED-harboring TADs contain HSGRL and 68% of SEDs co-localize with HSGRL, suggesting that emergence of HSGRL may have rewired SED-associated GRNs within specific TADs by inserting novel and/or erasing existing non-coding regulatory sequences. Consequently, markedly distinct features of the principal regulatory structures of interphase chromatin evolved in the hESC genome compared to mouse: the SED quantity is 3-fold higher and the median SED size is significantly larger. Concomitantly, the overall TAD quantity is increased by 42% while the median TAD size is significantly decreased ($p = 9.11E-37$) in the hESC genome. Present analyses illustrate a putative global role for HSGRL in shaping the human-specific features of the interphase chromatin organization and functions, which are facilitated by accelerated creation rates of new enhancers associated with targeted placement of HSGRL at defined genomic coordinates. A trend toward the convergence of TAD and SED architectures of interphase chromatin in the hESC genome may reflect changes of 3D-folding patterns of linear chromatin fibers designed to enhance both regulatory complexity and functional precision of GRNs by creating predominantly a single gene (or a set of functionally-linked genes) per regulatory domain structures.

Introduction

Despite the remarkable progress illustrated by the near-comprehensive catalogue of changes within protein-coding genes that occurred during human evolution, the genetic and molecular basis defining unique to human phenotypic features remains largely unknown. These studies suggested that changes of protein-coding genes alone cannot explain most of the unique to human traits (1-5). Therefore, the hypothesis that unique to human phenotypes might result from human-specific changes to genomic regulatory sequences continues to remain highly relevant (6).

In recent years, the extensive search for human-specific genomic regulatory loci (HSGRL) has resulted in identification of thousands candidate HSGRL, a vast majority of which is located within non-protein coding regions of human genome (7-12). The rapidly-evolving DNA sequences in humans have been identified that are highly conserved across mammals and have acquired many sequence changes in humans since divergence from chimpanzees (13-16). Subsequent experimental analyses of these sequences, which were designated human accelerated regions (HARs), revealed that some HARs function as non-coding RNA genes expressed during the neocortex development (13) Pollard et al. 2006) and human-specific developmental enhancers (15). Recent computational and bioinformatics analyses and transgenic mouse experiments demonstrated that many HARs represent developmental enhancers (10).

Genome-wide analysis of transcription factor binding sites (TFBS) in human embryonic stem cells (hESC) demonstrated that transposable element-derived sequences, most notably LTR7/HERVH, LTR5_Hs/HERVK, and L1HS, harbor thousands of candidate human-specific TFBS (12). A majority of these candidate HSGRL appears to function as TFBS for NANOG, POU5F1 (OCT4), and CTCF proteins exclusively in hESC, suggesting their critical regulatory role during the early-stage embryogenesis (12). Interestingly, hESC-specific NANOG-binding sites appear enriched near the protein-coding genes regulating brain size, pluripotency long non-coding RNAs, hESC enhancers, and 5-hydroxymethylcytosine-harboring sequences immediately adjacent to TFBS. Consistent with the hypothesis that they contribute to development of human-specific phenotypes, candidate human-specific TFBS are placed near the coding genes associated with physiological development and functions of nervous and cardiovascular systems, embryonic development, behavior, as well as development of a diverse spectrum of pathological conditions such as cancer, diseases of

cardiovascular and reproductive systems, metabolic diseases, multiple neurological and psychological disorders (12; 17). Taken together, these studies lend credence to the hypothesis that HSGRL may function in human cells as regulatory DNA sequences.

Significantly, a constantly expanding catalog of candidate HSGRL includes nearly twenty thousand of genomic regulatory elements comprising the following multiple distinct families of HSGRL:

i) regions of human-specific loss of conserved regulatory DNA termed hCONDEL (7);

ii) human-specific epigenetic regulatory marks consisting of H3K4me3 histone methylation signatures at transcription start sites in prefrontal neurons (8);

iii) human-specific transcriptional genetic networks in the frontal lobe (9);

iv) conserved in humans novel regulatory DNA sequences designated human accelerated regions, HARs (10);

v) fixed human-specific regulatory regions, FHSRR (11);

vi) human-specific transcription factor-binding sites, HSTFBS (12);

vii) DNase I hypersensitive sites (DHSs) that are conserved in non-human primates but accelerated in the human lineage, haDHS (18);

viii) DNase I hypersensitive sites (DHSs) that are under accelerated evolution, ace-DHSs (19).

In the Drosophila genome and mammalian nucleus, beads on a string linear strands of interphase chromatin fibers are folded into continuous megabase-sized structures termed topologically associating domains (TADs) that are readily detectable by the high-throughput analysis of interactions of chemically cross-linked chromatin (20-23). It has been suggested that TADs represent evolutionary-conserved spatially-segregated neighborhoods of high local frequency of intrachromosomal contacts reflecting the likelihood of individual physical interactions between long-range enhancers and promoters of target genes in live cells (20; 24). Formal definition of TADs implies that neighboring TADs are separated by the sharp boundaries, across which the intrachromosomal contacts are relatively infrequent (20; 24). A genome-wide proximity placement analysis of 18,364 DNA sequences representing multiple diverse families of HSGRL within the context of the principal regulatory structures of the interphase chromatin identified a sub-set of TADs that are significantly enriched for HSGRL and termed rapidly-evolving in humans TADs (17). There are significant enrichment within

the boundaries of rapidly-evolving in humans TADs (revTAD) of hESC enhancers, primate-specific CTCF-binding sites, human-specific RNAPII-binding sites, hCONDELs, and H3K4me3 peaks with human-specific enrichment at transcription start sites (TSS) in prefrontal cortex neurons.

The phenotypic and transcriptional identities of embryonic stem cells (ESC) critically depend on continuing actions of key master transcription factors (25; 26), which govern the maintenance of the ESC's pluripotency state by forming constitutively active super-enhancers (27-29). Recent experiments show that regulation of pluripotency in ESC occurs within selected TADs by establishing specific sub-TAD structures designated super-enhancer domains (29). Super-enhancer domains (SEDs) are formed by the looping interactions between two CTCF-binding sites co-occupied by cohesin. In ESC, TADs' and SEDs' structures are designed to isolate super-enhancers (SEs) and target genes within insulated genomic neighborhoods to facilitate the functional precision of regulatory interactions between key genetic elements of pluripotency regulatory networks: SEs, SEs-driven cell identity genes and repressed genes encoding lineage-specifying developmental regulators (29). Precise nature of structural-functional relationships between the HSGRL and principal structural elements of the interphase chromatin defining a spatial architecture of the pluripotency regulatory network in hESC, namely SEs and SEDs, remains unexplored.

Results of the present analyses revealed novel insights into structural-functional features of HSGRL of potential mechanistic significance that may have contributed to the marked changes of the interphase chromatin structure in the human genome, which are exemplified by the increased number and reduced size of TADs. Importantly, the concomitant HSGRL-associated increase of both the quantity and size of SEDs was documented in the hESC genome. Collectively, these observations identify the increasing of both regulatory complexity and functional precision of genomic regulatory networks due to convergence of TAD and SED architectures as one of the main trend of the interphase chromatin structural changes driven by the HSGRL during the evolution of *H. sapiens*.

Materials and Methods

Data Sources and Analytical Protocols

Solely publicly available datasets and resources were used for this analysis as well as previously reported methodological approaches and computational pipelines validated for discovery of primate-specific gene and human-specific regulatory loci (10-12; 17; 30-32). The analysis is based on the University of California Santa Cruz (UCSC) LiftOver conversion of the coordinates of human blocks to corresponding non-human genomes using chain files of pre-computed whole-genome BLASTZ alignments with a minMatch of 0.95 (unless otherwise specified) and other search parameters in default setting (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Extraction of BLASTZ alignments by the LiftOver algorithm for a human query generates a LiftOver output "Deleted in new", which indicates that a human sequence does not intersect with any chains in a given non-human genome. This indicates the absence of the query sequence in the subject genome and was used to infer the presence or absence of the human sequence in the non-human reference genome. Human-specific regulatory sequences were manually curated to validate their identities and genomic features using a BLAST algorithm and the latest releases of the corresponding reference genome databases for time periods between April, 2013 and June, 2016.

Genomic coordinates of 3,127 topologically-associating domains (TADs) in hESC; 6,823 hESC-enriched enhancers; 6,322 conventional and 684 super-enhancers (SEs) in hESC; 231 SEs and 197 SEDs in mESC were reported in the previously published contributions (20; 27-29; 33). The primary inclusion criterion for selection of the human-specific genomic regulatory loci (HSGRL) analyzed in this contribution was the fact that they were identified in human cells lines and primary human tissues whose karyotype were defined as "normal". A total of 18,364 DNA sequences representing multiple diverse HSGRL families analyzed in this study were reported previously (17). The number of HSGRL placed within a given TAD was computed for every TAD in the hESC genome and the HSGRL placement enrichment was calculated as the ratio of observed values to expected values estimated from a random distribution model at the various cut-off thresholds. Datasets of NANOG-, POU5F1-, and CTCF-binding sites and human-specific TFBS in hESCs were reported previously (12; 34) and are publicly available. RNA-Seq datasets were retrieved from the UCSC data

repository site (http://genome.ucsc.edu/; (35) for visualization and analysis of cell type-specific transcriptional activity of defined genomic regions. A genome-wide map of the human methylome at single-base resolution was reported previously (36; 37) and is publicly available (http://neomorph.salk.edu/human_methylome). The histone modification and transcription factor chromatin immunoprecipitation sequence (ChIP-Seq) datasets for visualization and analysis were obtained from the UCSC data repository site (http://genome.ucsc.edu/; (38). Genomic coordinates of the RNA polymerase II (PII)-binding sites, determined by the chromatin integration analysis with paired end-tag sequencing (ChIA-PET) method, were obtained from the saturated libraries constructed for the MCF7 and K562 human cell lines (39). Genome-wide maps of interactions with nuclear lamina, defining genomic coordinates of human and mouse lamin-associated domains (LADs), were obtained from previously published and publicly available sources (40; 41). The density of TF-binding to a given segment of chromosomes was estimated by quantifying the number of protein-specific binding events per 1-Mb and 1-kb consecutive segments of selected human chromosomes and plotting the resulting binding site density distributions for visualization. Visualization of multiple sequence alignments was performed using the WebLogo algorithm (http://weblogo.berkeley.edu/logo.cgi). Consensus TF-binding site motif logos were previously reported (34; 42; 43).

The quantitative limits of proximity during the proximity placement analyses were defined based on several metrics. One of the metrics was defined using the genomic coordinates placing HSGRL closer to putative target protein-coding or lncRNA genes than experimentally defined distances to the nearest targets of 50% of the regulatory proteins analyzed in hESCs (44). For each gene of interest, specific HSGRL were identified and tabulated with a genomic distance between HSGRL and a putative target gene that is smaller than the mean value of distances to the nearest target genes regulated by the protein-coding TFs in hESCs. The corresponding mean values for protein-coding and lncRNA target genes were calculated based on distances to the nearest target genes for TFs in hESC reported by Guttman et al. (44). In addition, the proximity placement metrics were defined based on co-localization within the boundaries of the same TADs and the placement enrichment pattern of HSNBS located near the 251 neocortex/prefrontal cortex-associated genes, which identified the most significant enrichment of HSNBS placement at the genomic distances less than 1.5 Mb with a sharp peak of the enrichment p value at the distance of 1.5 Mb (Glinsky, GV. 2017. Human-

specific features of pluripotency regulatory networks in embryonic stem cells link fetal and adult brain development. Submitted).

The assessment of conservation of HSGRL in individual genomes of 3 Neanderthals, 12 Modern Humans, and the 41,000-year old Denisovan genome (45; 46) was carried-out by direct comparisons of corresponding sequences retrieved from individual genomes and the human genome reference database (http://genome.ucsc.edu/Neandertal/). Direct access to the specific Genome Browser tracks utilized for analyses and visualization:
http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg18&position=chr10%3A69713986-69714099&hgsid=393865029_yg7UixUE4a4awjjTahns4KTPkIl1 .

Recombination rates were downloaded from the HapMap Project (47) and the numbers of DNA segments with the recombination rates of 10 cM/Mb or greater were counted. This threshold exceeds ~10-fold the mean intensity of recombination rates in telomeric regions, which were identified as the regions with the higher recombination rates in the human genome. It is well known that over large genomic scales, recombination rates tend to be higher in telomeric as compared to centromeric chromosomal regions. In telomeric regions, the mean detected hotspot spacing is 90 kb and the mean intensity (total rate across the hotspot) per hotspot is 0.115 cM, whereas for centromeric regions the mean spacing is 123 kb and the mean intensity is 0.070 cM (47).

Statistical Analyses of the Publicly Available Datasets

All statistical analyses of the publicly available genomic datasets, including error rate estimates, background and technical noise measurements and filtering, feature peak calling, feature selection, assignments of genomic coordinates to the corresponding builds of the reference human genome, and data visualization, were performed exactly as reported in the original publications and associated references linked to the corresponding data visualization tracks (http://genome.ucsc.edu/). Any modifications or new elements of statistical analyses are described in the corresponding sections of the Results. Statistical significance of the Pearson correlation coefficients was determined using GraphPad Prism version 6.00 software. The significance of the differences in the numbers of events between the groups was calculated using two-sided

Fisher's exact and Chi-square test, and the significance of the overlap between the events was determined using the hypergeometric distribution test (48).

Results

Markedly distinct features of super-enhancers in the hESC genome compared with mESC

The sustained expression of key cell identity genes and repression of genes encoding lineage-specifying developmental regulators is essential for maintaining ESC identity and pluripotency state. These processes are governed by the master TFs OCT4 (POU5F1), SOX2, and NANOG (OSN), that function by establishing SEs regulating cell identity genes, including master TFs themselves (27; 28). The spatial relationships between SEs and HSGRL in the hESC genome within the context of TADs have not been investigated. To this end, the placement enrichment analysis was carried out to identify all TADs in hESC genome that harbor SEs and examine the association of SEs and HSGRL. There are 504 TADs (16%) harboring 642 SEs (94%) in the hESC genome. Remarkably, significant placement co-enrichments were observed between SEs, HSTFBS, and HARs residing within the boundaries of 279 TADs in the hESC genome (**Table 1**). In total, 279 TADs (8.9%) harbor 369 SEs (57.5%) that are co-localized with 300 HSTFBS and 564 HARs. When other HSGRL families (17) were considered in co-localization analyses, 331 TADs (10.6%) harboring 436 SEs (67.9%) were found to co-localize with HSGRL.

In human genome, there are approximately 7,000 high-confidence hESC-enriched enhancers (27; 33), less than 10% of which are defined as SEs (27). Placement of high-confidence hESC enhancers is significantly enriched within the revTADs (17), implying possible genome-wide associations of HSGRL and hESC-enriched enhancers. Indeed, such associations became apparent when genomic co-localization analyses were performed to assess numbers of HSGRL and hESC-enriched enhancers residing within the boundaries of TADs. These analyses revealed that there is a significant direct correlation between the numbers of hESC-enriched enhancers and HARs that are located within the same revTADs (**Fig. 1**). Genome-wide placement and/or retention of HARs appears enriched within hESC enhancer-harboring TADs and seems to favor TADs containing larger numbers of hESC-enriched enhancers (**Fig. 1**).

Co-localization of a majority of SEs with HSGRL suggest that HSGRL may affect the structural-functional features of SEs in the hESC genome. Consistent with this hypothesis, genomes of human and mouse ESC manifest markedly distinct features associated with SE structures and functions. Despite strikingly similar genome sizes and numbers of protein-coding genes, hESC genome contains 3-fold more SEs compared to mouse: there are 684 SEs in the genome of hESC (27) and 231 SEs in the genome of mESC (28). In the genome of hESC, only twenty-five of mESC SEs (11%) are represented as conserved orthologous sequences having genomic architecture of SEs, suggesting that 89% of mESC SEs lost the SE features in humans. Furthermore, 96% of SEs in the genome of hESC acquired structural-functional features of SEs during evolution after Euarchonta and Glires split 88 million years ago. The median size of SEs in hESC appears significantly larger compared to the median SE size in mESC ((9,589 bp versus 8,667 bp, respectively; $p = 0.017$). Detailed size distribution analyses demonstrated that accumulation of SEs in hESC genome is associated with increased number of large SEs and decreased number of small SEs compared to mouse genome: there are 7-fold increase of very large SEs having size more than 30 Kb, consistent ~3-fold increase of SEs having size range from 2 – 30 Kb, and a marked 38-fold depletion of small SEs having size less than 2 Kb (**Table 2; Fig. 2**).

Collectively, these data indicate that structural-functional features of SEs are markedly distinct in the hESC genome compared to the mouse, which appear associated with the enrichment of HSGRL within SE-harboring TADs. It will be of interest to determine whether the HSGRL placement within SE-harboring TADs exerts biologically-meaningful effects on SE functions. Targeted placements and/or retention of HSGRL may increase density of enhancer elements within selected TADs, which would result in the merger of conventional enhancer units into super-enhancer structures containing the exceptionally high level of transcription-enabling apparatus to drive and continually maintain high expression of associated target genes. This idea is supported by the findings that nearly 60% (1571) of the HARs overlap at least one of the common markers of enhancer activity in human cells (10).

TAD structural features are markedly altered in the hESC genome compared with mESC

There are 3,127 TADs in the hESC genome, which is 42% more than 2,200 TADs in the genome of mESC (20). Correspondingly, there are 87,883 CTCF-binding sites in hESC (34), 29,018 (33%) of which represent primate-specific CTCF-binding sequences (12). The median size of TADs in hESC is significantly smaller compared to the median size of TADs in mESC (680 Kb versus 880 Kb; p = 9.11E-37). Detailed size distribution analyses of TADs in human and mouse ESC (**Fig. 2; Table 3**) revealed that there is a ~ 2-fold depletion of large TADs having size > 2,000 Kb and consistently increased numbers of medium-size and small TADs having size range from 100 - 1,000 Kb (**Table 3**). Structural-functional features of the revTADs in the hESC genome appear markedly distinct from the TADs of the orthologous sequences in the mouse ESC genome. Of the 60 revTAD (17), nineteen (32%) are placed within the primate-specific sequences that failed to align to the mouse reference genome database sequence. Remaining revTADs appear to evolve by erasing the existing and establishing new domain boundaries within orthologous DNA sequences via two distinct mechanisms: 1) domain & boundary crossing (hESC TADs appear to cross boundaries of 2 to 4 orthologous TADs in mESC genome); and 2) domain shrinking & boundary creation (smaller hESC TADs are placed within the boundaries of larger orthologous TAD sequences in mESC genome).

The above considerations prompted additional analyses of relationships between the SEs and TADs in the genomes of human and mouse ESC (**Fig. 3**). These analyses revealed that in hESC there are significant direct correlations between the size of TADs and SE's span defined as a number of bp between the two most distant SEs located within a given TAD (**Fig. 3A**). Similar trends were observed in the mESC genome, however, the correlation coefficient values were not statistically significant. Correlation patterns observed for the 60 revTADs for placements of hESC-enriched enhancers, HARs, HSTFBS, and size of TADs (17) were validated on a larger set of 147 revTADs (**Fig. 3B**). Genome-wide, the highly significant direct correlation was discovered between the size of TADs and the number of hESC-enriched enhancers located within TADs (**Fig. 3C**). These observations are conceptually coherent because TAD boundaries were inferred from the relative prevalence and directionality of interchromosomal interactions along the chromosome length (20), which are predominantly mediated by the enhancers' activities and detected as enhancer-promoter and enhancer-enhancer interactions in the Hi-C analyses. It will be of interest to determine experimentally whether the size

and boundaries of the adjacent and neighboring TADs can be altered by increasing the density, structure, and activity of the resident enhancers.

Potential mechanisms of HSGRL-mediated effects on principal regulatory structures of interphase chromatin

Present analyses provide a conceptual framework for understanding genome-scale regulatory changes during evolution within the context of the principal regulatory structures of the interphase chromatin to reflect an apparent trend toward increasing complexity of genomic regulatory networks (**Fig. 4**). Experimental observations at the foundation of building blocks of the genome's evolution model are described in the previous sections and additional considerations are focused on the analyses of potential contributions of hESC-enriched enhancers, SEs, and SEDs to these processes.

According to the model, one of the key elements of the evolution of genomic regulatory networks is the creation of new enhancer elements (**Fig. 4**). Conventional enhancers comprise discrete DNA segments occupying a few hundred base pairs of the linear DNA sequence and harboring multiple TFBS. SEs consist of clusters of conventional enhancers that are densely occupied by the master transcription factors and Mediator (Whyte et al., 2013). Therefore, it is logical to expect that creation of new TFBS and increasing density of TFBS would increase the probability of the emergence of new enhancer elements. It follows that creation of new enhancers and increasing their density would facilitate the emergence of new SE structures. This sequence of events would imply that evolutionary time periods required for creation of TFBS, enhancers, and SEs are shortest for TFBS, intermediate for enhancers, and longest for SEs. To test this assumption, estimates of creation time periods for enhancers and SEs in the hESC genome were calculated (**Table 4**) and compared to the previously reported estimates of creation time periods of TFBS (12). Consistent with the model expectations, the estimated creation time is markedly longer for SEs compared to enhancers for conserved (8-fold), primate-specific (17-fold), and human-specific (63-fold) regulatory sequences (**Table 4**). Furthermore, the estimated creation time periods for enhancers appear several fold longer compared to the creation time estimates for TFBS sequences in both chimpanzee and humans (12). One of the intriguing results of this analysis is that 87% of new SEs and 75% of new enhancers in the hESC genome were created within

conserved sequences (**Table 4**). Similarly, a majority (67%) of new binding sites for NANOG (59,221 of 88,351 TFBS) and CTCF (58,865 of 87,883 TFBS) proteins in the hESC genome compared with mESC are located within conserved sequences (**Supplemental Fig. S3**).

Notably, the estimated creation time appears accelerated in humans compared to chimpanzee for both SEs (7-fold) and enhancers (27-fold), suggesting that human genomes were acquiring new regulatory elements and increasing the regulatory complexity and precision of genomic regulatory networks at the markedly accelerated pace.

CTCF-binding sites play a crucial role in defining the TAD boundaries (20; 49) and in establishing the SED architecture (29). Knocking down the expression of CTCF and cohesin genes to reduce their protein concentrations results in a decrease in the number of intra-TAD chromatin loops (50-52), increase in the number of inter-TAD chromatin interactions crossing the TAD boundaries (52), and chromatin compaction (53). Several lines of evidence are in agreement with the hypothesis that creation of new CTCF-binding sites may have contributed to the rewiring of interphase chromatin architecture and genomic regulatory networks during primate evolution:

i) In the hESC genome, 29,018 of 87,883 (33%) CTCF-binding sites represent primate-specific sequences (12; 34);
ii) Of the 23,709 constitutive CTCF-binding sites implicated in defining TAD boundaries in the human genome (Li et al., 2013), 6,787 sites (28.6%) failed alignment to the mouse genome, suggesting that they represent primate-specific sequences;
iii) Two hundred eighty seven HARs (10.5%) are located within 10 Kb of the 336 constitutive CTCF-binding sites;
iv) Five hundred seventy one HARs (20.8%) are located within 1 Kb of 953 CTCF-binding sites;
v) Activation of retrotransposons and lineage-specific repeat-driven dispersion of CTCF-binding sites has produced species-specific expansions of CTCF binding in mammalian genomes and these new CTCF-binding sites function as chromatin domain insulators and transcriptional regulators (54).

Detailed proximity placement and HARs/TFBS co-localization analyses identified 123 HARs located within 1 Kb from 127 high-confidence CTCF-binding sites, which were validated in 23 different human cells

lines. Notably, 123 of 127 (97%) high-confidence CTCF-binding sites represent overlapping CTCF/RAD21-binding sites, which is consistent with their putative regulatory role in establishing TAD boundaries and/or SED architecture. A significant majority of 123 HARs located near high-confidence overlapping CTCF/RAD21-binding sites harbor at least one TFBS in human cells (76 HARs; 62%).

Significant increase of TAD and SED numbers in human genomes compared with mouse indicates that chromatin folding profiles are dramatically altered, reflecting more frequent intrachromosomal segmentation of linear DNA fibers and enhanced likelihoods of formation of isolated intra-segmental looping structures. These changes of chromatin folding profiles and DNA fibers' packaging patterns would impose more stringent requirement on DNA elasticity, particularly, near the TAD and SED boundaries due to bending of DNA double helix. Consequently, changes of DNA sequence and/or chromatin structure may be necessary to accommodate these new structural requirements. It has been reported that SINE repeats are enriched at the TAD boundaries (20), suggesting that insertion of repetitive elements may contribute to changes of DNA elasticity near putative bending sites. Consistent with this hypothesis, a survey of SED sequences in hESC revealed an apparent systematic placement of Alu elements within ~5 Kb windows near SED boundaries (**Fig. 5**). Clusters of closely-spaced sequences of at least three Alu elements belonging to most ancient AluJ (~65 million years old), second oldest AluS (~30 million years old), and currently active modern AluY sub-families (55) were observed frequently, suggesting that placement and/or retention of Alu elements at these sites were occurring for millions of years and continues at the present time.

Taken together with the recent report that nearly 60% of the HARs overlap at least one of the common markers of enhancers in human cells (10), these observations strongly argue that creation of new enhancer elements is one of the key events defining evolution of genomic regulatory networks. Marked acceleration of the TFBS and enhancers' creation processes in humans may have contributed to genome-scale rearrangements of principal regulatory structures of the interphase chromatin, leading to the increased complexity and enhanced precision of genomic regulatory networks and emergence of human-specific phenotypes. Frequent locations of Alu sequences near the putative DNA bending sites suggest their potential role in regulation of the elasticity of chromatin fibers by reducing nucleosome placements due to the low affinity of Alu elements to nucleosome binding (56) and/or facilitating strand invasion reactions between neighboring

DNA strands leading to Alu/Alu recombination events (57; 58). This model is in agreement with the observations that Alu elements appear preferentially retained in GC-rich and gene-rich regions of the human genome (59).

Conservation patterns of HSGRL in individual human genomes

All analyses conducted so far were performed using the reference genome databases and not the individual human genomes. To address this limitation, the assessment of conservation of HSGRL in individual genomes of 3 Neanderthals, 12 Modern Humans, and the 41,000-year old Denisovan genome (45; 46) was carried-out by direct comparisons of corresponding sequences retrieved from individual genomes and the human genome reference database (http://genome.ucsc.edu/Neandertal/). Full-length sequence alignments with no gaps of the individual genome sequences to the corresponding sequences in the human genome reference databases of both hg18 and hg19 releases were accepted as the evidence of sequence conservation in the individual human genome.

Sequences of all analyzed to date HSGRL appear conserved in individual human genomes (**Fig. 6** and **Supplemental Fig. S4**), albeit a significant inter-individual variability in degree of conservations of specific HSGRL is apparent. Conservation of all HSGRL are consistently at the lowest level in the Neanderthals' genome, suggesting that creation and/or retention rates of HSGRL are enhanced in Modern Humans. The results of these analyses indicate that sequences of HSGRL with assigned biochemical functions, e.g., specific TFBS or Lamin B1 (LMNB1)-binding sites, which are residing within HARs, exhibit markedly higher conservation levels compared to sequences of HARs harboring the corresponding HSGRL. This conclusion remains valid for HSGRL sequences with assigned specific biochemical or biological functions that were associated with HARs by proximity placement analyses (**Figs. 6A, B**). HSGRL sequences manifesting the relatively high conservation levels in the Neanderthals' genome appear most conserved in the individual genomes of Modern Humans as well, including the 41,000-year old Denisovan genome (**Fig. 6B**).

Consistent patterns of significant direct correlations between conservation profiles of distinct seemingly unrelated HSGRL sequences in individual human genomes were observed (**Fig. 6C** and **Supplemental Fig. S4**). One possible interpretation of these observations is that HSGRL conservation patterns reflect intrinsic

features of the individual human genome, integration of which can gauge the overall capacity of a genome to create and retain HSGRL. This idea was tested by calculating for each individual human genome the genomic fitness scores integrating into a single numerical value sequence conservation data of 909 HSGRL (**Fig. 6C;** bottom panel). Notably, individual human genomes appear markedly distinct based on the results of this analysis with the lowest genomic fitness score of 0.9 in Neanderthals, followed by the scores of 1.89; 2.22; and 2.27 for individual genomes of Native American, Denisova cave, and Mongolian subjects, respectively. On the other end of the spectrum, the genomic fitness scores of 4.51; 4.72; and 4.73 were obtained for genomes of Yoruba (West Africa), French (Western Europe), and San (Southern Africa) individuals, respectively.

These observations are highly congruent with the recent definition of structural variations-free (fixed) human-specific regulatory regions (FHSRR) based on the stringent formal analysis of the HSGRL intra-species patterns of variations within human population using exome and full genome sequencing database of 1,092 individuals from the 1000 Genomes Project Consortium (11; 60). Collectively, the consistent evidence of HSGRL conservation in individual human genomes are in accord with the hypothesis of their putative functional role in defining human-specific phenotypes.

Discussion

Placing several meters of linear DNA fibers into micron-sized nucleus while maintaining the genome integrity and ensuring the functional precision of genomic regulatory networks represent a set of formidable challenges successfully addressed during the evolution of eukaryotes. The elegant evolutionary solutions to these problems are illustrated by the recent pioneering work on the 3D structures of interphase chromosomes. These studies revealed specific folding patterns of the linear chromosome fibers into spatially-segregated domain-like segments experimentally-defined as the continuous megabase-sized topologically associating domains (TADs). TADs are readily and reproducibly detectable by the high-throughput analysis of inter-chromosomal interactions of chemically cross-linked chromatin (20-23). It has been suggested that one of the principal creative events in the continuing chromatin domain architecture remodeling process during human evolution is the HSGRL-enabled emergence of new enhancer elements (17), increasing density of which would enhance the probability of structural transition from conventional enhancers to novel SE structures and the subsequent

formation of new SEDs (**Fig. 4**). Significantly, primate-specific and human-specific sequences appear to contribute to creation of interacting pairs of conserved and/or newly created overlapping CTCF/cohesin-binding sites flanking novel SEs, which represents the essential structural requirement for the formation of functional SEDs. On the evolutionary scale, these processes involve continuing removal of old and creation of new regulatory sequences through multiple trial-and-error events enabled by retrotransposition, cytosine MADE, and recombination (12; 17). Collectively, the ensemble of these HSGRL-associated changes of the nuclear regulatory architecture may facilitate the enhanced precision of regulatory interactions between enhancers and target genes within specifically reconstructed insulated SED neighborhoods and rewired genome-wide TAD networks. Taken together, these observations imply that the convergence of TAD and SED architectures, which is exemplified in the human genome by concomitant processes of increasing both quantity and size of SEDs and the increasing number and size reduction of TADs, might represent one of the main directions of the interphase chromatin structural changes during the evolution of genomic regulatory networks.

Recent studies documented several human-specific features of the interphase chromatin architecture, evolution of which appear associated with the targeted placement and/or retention of HSGRL. Genome-wide proximity placement analyses of 18,364 DNA sequences representing distinct families of HSGRL identified a small fraction of TADs in the human genome, which manifest a statistically significant accumulation of HSGRL compared to the expected values estimated based on a random distribution model (17). This set of TADs, termed rapidly-evolving in humans TADs (revTADs), acquired the maximum enrichment levels of 16-fold for HARs; 17-fold for hESC-FHSRRs; 50-fold for HSTFBS; and 88-fold for DHS-FHSRRs ($p < 0.0001$ in all instances). Follow-up analyses of sixty revTADs, which were defined based on the enrichment of HARs and HSTFBS, revealed that all reported to date HSGRL and human-specific epigenetic signatures associated with embryonic development appear significantly enriched within the revTAD boundaries (17). One of the features of revTADs was revealed by the correlation analyses of HARs and HSTFBS: placements of different types of HSTFBS manifest significant positive correlations, whereas placements of HARs and HSTFBS exhibit inverse correlation profiles (17). Consistent with the hypothesis of the increased precision of genomic regulatory networks during evolution, a majority of revTADs in human genome tend to harbor a small number of coding genes: 65% of revTADs contain five or less protein-coding genes, among which eight revTADs harbor just one

protein-coding gene and two revTADs contain only non-coding RNA genes. Several revTADs contain large clusters of functionally-related protein-coding genes: three revTADs on chr11 harbor 61 genes encoding olfactory receptors; three consecutively-spaced revTADs on chr19 harbor 31 genes encoding zinc finger proteins; and one revTAD on chr20 contains eight beta-defensins' genes. Overview of the RNAseq data documented two other features of revTADs: i) clearly discernable tissue-specific patterns of non-coding RNA expression; and ii) a prominent presence of non-coding RNA transcripts in human brain regions.

Correlation screens of proximity placement patterns of different HSGRL residing within the revTADs revealed markedly distinct correlation profiles of individual members of HSGRL families and recombination rates within the host revTADs. These analyses readily distinguish placement patterns of HARs and HSTFBS: revTADs having high recombination scores tend to accumulate large numbers of HARs while revTADs with low recombination scores harbor high numbers of HSTFBS (17). Consistent with the requirements of DNA strand interactions and strand invasion for the recombination process, significant direct correlations were observed between recombination scores and intra-chromosomal contacts within revTADs (17). Therefore, placements of HARs and HSTFBS within revTADs appear associated with distinct molecular processes. Placement of HARs within revTADs exhibits significant positive correlation with high recombination rates, which seems to connect the biogenesis of HARs with recombination mechanisms. In contrast, significant inverse correlation between HSTFBS placement and recombination rates as well as location of 99% of HSTFBS within TE – derived DNA sequences (12) strongly implicate TE activity in the biogenesis of HSTFBS. Since HARs by definition are located within highly evolutionary conserved DNA sequences, results of these analyses suggest that molecular mechanisms driving the emergence of regulatory loci evolved by exaptation of ancestral DNA (61; 62) may be associated with meiotic recombination as well.

One of the notable features of potential mechanistic significance is the association of hESC-enriched enhancers and SEDs with HSGRL (**Table 1**). Taking into account that there are ~4,000 HSTFBS in the hESC genome (12) and nearly 60% of HARs overlap at least one of the common markers of enhancers in human cells (10), it seems logical to propose that creation of new enhancer elements leading to increasing density of conventional enhancers in selected genomic regions is one of the key events defining the increasing genomic complexity as the main direction of evolution of GRNs. Consistent with this idea, there are significant direct

correlations between the placements within TADs of HARs and hESC-enriched enhancers (**Fig. 1**), indicating that there is an apparent trend of the placement preference of HARs within TADs harboring hESC-enriched enhances. It follows, that higher density of conventional enhancers would increase the probability of local structural transitions of chromatin architecture to SEs and SEDs, provided other local structural requirements are in place or co-evolved (see below). In agreement with this model, there are 3-fold more SEs in the hESC genome compared with the mESC and the median size of hESC SEs is significantly larger (**Fig. 2**).

Concomitantly, marked changes of TAD structural features in the hESC genome compared with mESC were observed, which are particularly striking for revTADs. Genome-wide, there are 42% more TADs in the hESC genome and the median size of hESC TADs is significantly smaller (**Fig. 2**). These changes of TAD structural features remain consistent when the corresponding comparisons were made between the similar in size individual human and mouse chromosomes (**Fig. 2**), indicating that hESC genome contains the increased number of the predominantly smaller size TADs. Consistent with these observations, comparisons of the high-resolution 3D maps of human and mouse genomes, which were recently obtained by Rao et al. (63) using in situ Hi-C at a resolution range of 1 – 5 Kb, revealed that there are ~3-fold more contact domains formed by the concomitantly increased numbers of long-range chromatin loops in the human GM12878 B-cell lymphoblasts compared with the mouse CH12 B-cell lymphoblasts (63).

In the hESC genome, structural changes of the revTADs seem particularly evident: 32% of revTADs are located within primate-specific genomic regions and the remaining revTADs appear evolved via mechanisms of boundary crossing, domain mergers, and creation of new boundaries within larger TADs of orthologous mouse sequences. Collectively, this dramatic changes of the regulatory infrastructure of interphase chromatin can be explained by the model of convergence of TAD and SED architecture (**Fig. 4**). According to this model, the high density of conventional enhancers increases the likelihood of transition to larger in size SE structures, formation of new SEDs, and increasing segmentation of genomic regions into insulated regulatory neighborhoods of large SED/small TAD structures. Consistent with this model, significant correlations were observed between the size of TADs and genomic span of hESC SEs residing within TADs (**Fig. 3A**). There are significant direct correlations between the revTAD sizes and the numbers of hESC-enriched enhancers residing within revTADs (**Fig. 3B**). Genome-wide, the highly significant direct correlation

was observed between the size of TADs and the number of hESC-enriched enhancers located within TADs (**Fig. 3C**).

In addition to enabling the creation of new enhancer elements, increasing density of conventional enhancers, and facilitating transition to SED structures, potential mechanisms of HSGRL-mediated effects on principal regulatory structures of interphase chromatin are likely involve emergence of overlapping CTCF/cohesin sites and LMNB1-binding sites as well as continuing insertion of Alu clusters near the putative DNA bending sites (**Figs. 4 & 5**). Collectively, the ensemble of these structural changes facilitated by the targeted placements and retention of HSGRL at specific genomic locations would enable the emergence of new SED structures and remodeling of existing TADs to drive evolution of GRNs (**Fig. 4**). Presented models of dynamic transitions of the interphase chromatin principal regulatory structures are in accord with the results of recent high-resolution in situ Hi-C experiments demonstrating that human genomes are partitioned into contact domains consisting of ~10,000 loops accommodating functional links between enhancers and promoters (63). Consistent with the idea that chromatin loops frequently demarcate the boundaries of contact domains, anchors at the loop bases typically occur at the contact domain boundaries and involve binding of two CTCF/cohesin sites in a convergent orientation with the asymmetric binding motifs of interacting sites aligned to face each other (63).

Compelling evidence of conservation in individual human genomes of different families of HSGRL (11; 12; 17; **Fig. 6; Supplemental Fig. S4**) are in accord with the hypothesis of their putative functional role in defining human-specific phenotypes. Observations of highly consistent patterns of significant direct correlations between conservation profiles of distinct seemingly unrelated HSGRL sequences in individual human genomes were observed (**Fig. 6C** and **Supplemental Fig. S4**) suggest that HSGRL conservation patterns reflect intrinsic features of the individual human genome, integration of which can help to assess the overall capacity of a genome to create and retain HSGRL. In light of emerging evidence highlighting the role of HSGRL in development of therapy-resistant malignancies (12; 17), it will be of interest to explore experimentally the potential practical utility of the outlined herein concepts by building and validating the working models of defined super-enhancers' domains and associated TAD structures in the hESC genome (Glinsky, GV. 2017.

Human-specific features of pluripotency regulatory networks in embryonic stem cells link fetal and adult brain development. Submitted).

Concluding remarks

In general terms, chromosomes can be viewed as the physical conduits of genetic information enabling its secure storage, maintenance, transfer, and efficient translation into a diverse spectrum of specific phenotypes. To a large degree all these processes are facilitated by the linear code of DNA sequences and enabled by the chromatin structures defining a 3D architecture of interphase chromosomes, which can be considered collectively as the chromatin folding code. Present analyses suggest that increasing regulatory complexity in human genomes associated with targeted placements of many thousands of HSGRL has a major effect on the principal regulatory structures of interphase chromatin, namely TADs, SEs, and SEDs, perhaps, contributing to the creation of human-specific chromatin folding code. Recently reported half-lives and mean-lifetimes of enhancers were estimated at 296 and 427 hundred million years, respectively (62). Comparisons of these estimates with the estimates of time periods required for creation of enhancers and SEs (**Table 4**) seem to indicate that new enhancer elements are created at a markedly faster pace compared with their decay time span. This marked evolutionary dichotomy of time period requirements for creation of new enhancers compared to the enhancers' loss provides an underlying mechanism explaining why the increasing genomic complexity represents a major trend during the evolution of genomic regulatory networks.

Exaptation of ancestral DNA was identified as a main mechanism of creation of human-specific enhancers active in embryonic limb (61) and as a prevalent creation mechanism of recently evolved enhancers during the mammalian genome evolution (62). Consistently, a vast majority of distinct classes of regulatory elements in the hESC genome appears created on conserved DNA sequences (**Table 4**; **Supplemental Fig. S3**), suggesting that exaptation of ancestral DNA constitutes a main mechanism of creation of new regulatory sequences in the human genome. Significantly, exaptation of ancestral DNA appears to generate overall many more recently evolved regulatory sequences than can be attributed to the repeat-driven expansion mechanisms. Therefore, these observations provide further support to the hypothesis that meiotic recombination is the predominant mechanism responsible for creation of new regulatory loci during evolution.

The concept of evolution of evolvability posits that a genotypic feature evolves not due to its functional effect, but due to its effect on the ability of DNA to evolve more quickly (64). Considering the evolution of enhancers as an example of evolution of evolvability (65; 66) and taking into account the potential role of HSGRL in creation of human-specific networks of enhancers, SEs, and SEDs, the apparent mechanistic links of these processes to several key enzymatic systems should be regarded as highly promising novel molecular targets of potential therapeutic significance. Of particular relevance in this context is the idea that the regulatory balance of HAT and HDAC activities at specific genomic loci as well as the performance of enzymatic systems regulating cytosine recovery/methyl-cytosine deamination cycles and DNA recombination processes may play a major role during the evolution of regulatory DNA sequences and emergence of genomic regulatory networks controlling unique to human physiological and pathological phenotypes. Emerging evidence implicating HSGRL derived from stem cell-associated retroviral sequences in the pathogenesis of clinically lethal human malignancies (67-69) seems particularly intriguing in this regard.

Supplemental Information

Supplemental information includes Supplemental Figures S1–S4 and can be found with this article online.

Author Contributions

This is a single author contribution. All elements of this work, including the conception of ideas, formulation, and development of concepts, execution of experiments, analysis of data, and writing of the paper, were performed by the author.

Acknowledgements

This work was made possible by the open public access policies of major grant funding agencies and international genomic databases and the willingness of many investigators worldwide to share their primary research data. I would like to thank my colleagues for their valuable critical contributions during the informal review and formal peer review process of this work.

References

1. Chimpanzee Sequencing and Analysis Consortium. 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87.
2. Green RE et al. 2010. A draft sequence of the Neanderthal genome. Science; 328, 710–722.
3. Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, Knight JR, Mullikin JC, Meader SJ, Ponting CP, Lunter G, Higashino S, Hobolth A, Dutheil J, Karakoç E, Alkan C, Sajjadian S, Catacchio CR, Ventura M, Marques-Bonet T, Eichler EE, André C, Atencia R, Mugisha L, Junhold J, Patterson N, Siebauer M, Good JM, Fischer A, Ptak SE, Lachmann M, Symer DE, Mailund T, Schierup MH, Andrés AM, Kelso J, Pääbo S. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486: 527-31.
4. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PL, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-9.
5. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prüfer K, de Filippo C, Meyer M, Zwyns N, Salazar-García DC, Kuzmin YV, Keates SG, Kosintsev PA, Razhev DI, Richards MP, Peristov NV, Lachmann M, Douka K, Higham TF, Slatkin M, Hublin JJ, Reich D, Kelso J, Viola TB, Pääbo S. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014; 514: 445-9.
6. King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188: 107-16.
7. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, Wenger AM, Bejerano G, Kingsley DM. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature **471**: 216-9.

8. Shulha HP, Crisci JL, Reshetov D, Tushir JS, Cheung I, Bharadwaj R, Chou HJ, Houston IB, Peter CJ, Mitchell AC, Yao WD, Myers RH, Chen JF, Preuss TM, Rogaev EI, Jensen JD, Weng Z, Akbarian S. 2012. Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. PLoS Biol **10**: e1001427.
9. Konopka G, Friedrich T, Davis-Turak J, Winden K, Oldham MC, Gao F, Chen L, Wang GZ, Luo R, Preuss TM, Geschwind DH. 2012. Human-specific transcriptional networks in the brain. Neuron **75**: 601-17.
10. Capra, J.A., Erwin, G.D., McKinsey, G., Rubenstein, J.L., Pollard, K.S. 2013. Many human accelerated regions are developmental enhancers. Philos Trans R Soc Lond B Biol Sci. 368 (1632): 20130025.
11. Marnetto D, Molineris I, Grassi E, Provero P. 2014. Genome-wide identification and characterization of fixed human-specific regulatory regions. Am J Hum Genet **95**: 39-48.
12. Glinsky, GV. 2015. Transposable elements and DNA methylation create in embryonic stem cells human-specific regulatory sequences associated with distal enhancers and non-coding RNAs. Genome Biol Evol. 7: 1432-1454.
13. Pollard KS, Salama SR, Lambert N, et al. (16 co-authors). 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443: 167–172.
14. Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. Science 314 (5800): 786.
15. Prabhakar S, Visel A, Akiyama JA, et al. (13 co-authors). 2008. Human specific gain of function in a developmental enhancer. Science 321 (5894): 1346–1350.
16. Bird C, Stranger B, Liu M, Thomas D, Ingle C, Beazley C, Miller W, Hurles M, Dermitzakis E. 2007. Fast-evolving noncoding sequences in the human genome. Genome Biol. 8(6): R118.
17. 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: 2774-88.
18. Gittelman RM, et al. 2015. Comprehensive identification and analysis of human accelerated regulatory DNA. Genome Res. 25: 1245–1255.

19. Dong X, Wang X, Zhang F, Tian W. 2016. Genome-wide identification of regulatory sequences undergoing accelerated evolution in the human genome. Mol Biol Evol. 33: 2565-75.

20. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376–380.

21. Hou, C., Li, L., Qin, Z.S., and Corces, V.G. 2012. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol. Cell 48, 471–484.

22. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485: 381–385.

23. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148: 458–472.

24. Gorkin DU, Leung D, Ren B. 2014. The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell **14**: 762-75.

25. Young, RA. 2011. Control of the embryonic stem cell state. Cell 144: 940–954.

26. Ng, HH and Surani, MA. 2011. The transcriptional and signalling networks of pluripotency. Nat. Cell Biol. 13, 490–496.

27. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre´, V., Sigova, A.A., Hoke, H.A., and Young, R.A. 2013. Super-enhancers in the control of cell identity and disease. Cell 155, 934–947.

28. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319.

29. Dowen J.M., Fan Z.P., Hnisz D., Ren G., Abraham B.J., Zhang L.N., Weintraub A.S., Schuijers J., Lee T.I., Zhao K., Young R.A. 2014. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell 159: 374–387.

30. Kent, W.J. 2002. BLAT - the BLAST-like alignment tool. Genome Res. 12, 656-664.

31. Tay, S.K., Blythe, J., and Lipovich, L. 2009. Global discovery of primate-specific genes in the human genome. Proc. Natl. Acad. Sci. USA 106, 12019-12024.

32. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human-mouse alignments with BLASTZ. Genome Res. 13, 103–107.

33. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, Yang H, Wang T, Lee AY, Swanson SA, Zhang J, Zhu Y, Kim A, Nery JR, Urich MA, Kuan S, Yen CA, Klugman S, Yu P, Suknuntha K, Propson NE, Chen H, Edsall LE, Wagner U, Li Y, Ye Z, Kulkarni A, Xuan Z, Chung WY, Chi NC, Antosiewicz-Bourget JE, Slukvin I, Stewart R, Zhang MQ, Wang W, Thomson JA, Ecker JR, Ren B.. 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 153: 1134-48.

34. Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., and Bourque, G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 42, 631-634.

35. Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., Raney, B.J., Pohl, A., Malladi, V.S., Li, C.H., Lee, B.T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B.M., Fujita, P.A., Dreszer, T.R., Diekhans, M., Cline, M.S., Clawson, H., Barber, G.P., Haussler, D., and Kent, W.J. 2013. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 41, D64-69.

36. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., and Ecker, J.R. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.

37. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, Yu M, Tonti-Filippini J, Heyn H, Hu S, Wu JC, Rao A, Esteller M, He C, Haghighi FG, Sejnowski TJ, Behrens MM, Ecker JR. 2013. Global epigenomic reconfiguration during mammalian brain development. Science.341, 1237905.

38. Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., Lee, B.T., Barber, G.P., Harte, R.A., Diekhans, M., Long, J.C., Wilder, S.P., Zweig, A.S., Karolchik, D., Kuhn, R.M., Haussler, D., and Kent, W.J. 2013. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 41, D56-63.

39. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., Sim, H.S., Peh, S.Q., Mulawadi, F.H., Ong, C.T., Orlov, Y.L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K.I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M.J., Cheung, E., Liu, E., Sung, W.K., Snyder, M., and Ruan, Y. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84-98.

40. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948-951.

41. Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., Reinders, M., Wessels, L., and van Steensel, B. 2010. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol. Cell. 38, 603-613.

42. Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., Rando, O.J., Birney, E., Myers, R.M., Noble, W.S., Snyder, M., and Weng, Z. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 22, 1798-1812.

43. Ernst, J., and Kellis, M. 2013. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. Genome Res. 23, 1142-1154.

44. Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E., and Lander, E.S. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477:295-300.

45. Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.J., Kelso, J., Slatkin, M., Pääbo, S. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 468, 1053-1060.
46. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andrés, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J., Pääbo, S. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science. 338, 222-226.
47. The International Hapmap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
48. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, GM. 1999. Systematic determination of genetic network architecture. Nat. Genet. 22, 281-285.
49. Li, Y., Huang, W., Niu, L., Umbach, D.M., Covo, S., Li, L. 2013. Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. BMC Genomics. 14: 553.
50. Sofueva S, Yaffe E, Chan WC, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, Pollard SM, Schroth GP, Tanay A, Hadjur S.. 2013. Cohesin-mediated interactions organize chromosomal domain architecture. The EMBO J. 32: 3119–3129.
51. Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, Ing-Simmons E, Lenhard B, Giorgetti L, Heard E, Fisher AG, Flicek P, Dekker J, Merkenschlager M. 2013. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. 23: 2066–2077.
52. Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RW, van de Corput MP, van de Werken HJ, Knoch TA, van IJcken WF, Grosveld FG, Ren B, Wendt KS. 2014. Cohesin and CTCF

differentially affect chromatin architecture and gene expression in human cells. Proc Natl Acad Sci USA 111: 996-1001.

53. Tark-Dame M, Jerabek H, Manders EM, Heermann DW, van Driel R. 2014. Depletion of the chromatin looping proteins CTCF and cohesin causes chromatin compaction: insight into chromatin folding by polymer modelling. PLoS Comput Biol. 10: e1003877.
54. Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148:335–348.
55. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE. 2008. Active Alu retrotransposons in the human genome. Genome Res. 18:1875-83.
56. Huda A, Mariño-Ramírez L, Landsman D, Jordan IK. 2009. Repetitive DNA elements, nucleosome binding and human gene expression. Gene 436:12-22.
57. Lee W, Mun S, Kang K, Hennighausen L, Han K. 2015. Genome-wide target site triplication of Alu elements in the human genome. Gene 561:283-91.
58. Morales ME, White TB, Streva VA, DeFreece CB, Hedges DJ, Deininger PL. 2015. The contribution of Alu elements to mutagenic DNA double-strand break repair. PLoS Genet. 11:e1005016.
59. Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921.
60. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.
61. Cotney, J., Leng, J., Yin, J., Reilly, S.K., DeMare, L.E., Emera, D., Ayoub, A.E., Rakic, P., and Noonan, J.P. 2013. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell 154: 185–196.
62. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, Turner JM, Bertelsen MF, Murchison EP, Flicek P, Odom DT. 2015. Enhancer evolution across 20 mammalian species. Cell 160: 554-66.

63. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159: 1665-80.
64. Wagner GP, Altenberg, L. 1996. Perspective: Complex adaptations and the evolution of evolvability. Evolution. 50 (3): 967-976.
65. He X, Duque TS, Sinha S. 2012. Evolutionary origins of transcription factor binding site clusters. Mol Biol Evol. 29:1059-70.
66. Duque T, Samee MA, Kazemian M, Pham HN, Brodsky MH, Sinha S. 2014. Simulations of enhancer evolution provide mechanistic insights into gene regulation. Mol Biol Evol. 31:184-200.
67. Glinsky GV. 2015. Viruses, stemness, embryogenesis, and cancer: a miracle leap toward molecular definition of novel oncotargets for therapy resistant malignant tumors? Oncoscience 2: 751–754.
68. Glinsky GV. 2016. Activation of endogenous human stem cell-associated retroviruses (SCARs) and therapy-resistant phenotypes of malignant tumors. Cancer Lett. 376: 347–359.
69. Glinsky GV. 2016. Single cell genomics reveals activation signatures of endogenous SCAR's networks in aneuploid human embryos and clinically intractable malignant tumors. Cancer Lett. 381: 176-93.

Figure Legends

Figure 1. Genome-wide associations between the number of HARs and the quantity of hESC-enriched enhancers located within TADs.

A. Direct correlation between the numbers of HARs and hESC-enriched enhancers located within the individual revTADs.

B-D. In the hESC genome, TADs segregated into sub-groups with increasing numbers of HARs contain concomitantly higher proportion of hESC-enriched enhancer-harboring TADs compared with TADs without hESC enhancers. The fractions of TADs with and without hESC enhancers are shown for a total of 3,062 TADs (hg19 release of human reference genome database) in the hESC genome (B), for 1,135 TADs harboring 2,609 HARs (C), for 2,075 TADs containing 6,703 hESC-enriched enhancers (D). In the figure (D) the average numbers of hESC-enriched enhancers are shown for sub-groups of TADs harboring 0; 1; 2; and at least 5 HARs. P values designate the statistical significance of the differences of the hESC enhancer numbers between the neighboring sub-groups of TADs. Calculations of the HAR numbers located within TADs were performed after converting the genomic coordinates of 3,127 TADs (Dixon et al., 2012) from the hg18 to hg19 release of the human reference genome database using the LiftOver algorithm.

Figure 2. Markedly distinct structural features of SEs and TADs in genomes of human and mouse ESCs.

A. Size distribution analyses revealed significantly increased size of SEs and decreased size of TADs in the hESC genome compared with mouse. Top two panels show the median sizes of SEs (top left panel) and TADs (top right panel) sorted in a descending order based on individual TAD sizes and segregated into ten sub-groups at 10% increments. Bottom two panels show the median (bottom left panel) and average (bottom right panel) sizes of TADs for individual chromosomes of human and mouse ESC genomes. Note that median sizes of TADs on 17 of 20 (85%) chromosomes in the hESC genome (except chr7, chr13, and chr14) are smaller compared to median TAD sizes on chromosomes in the mESC genome.

B. Numbers of both SEs (top left panel) and TADs (top right panel) are increased in the hESC genome compared with mouse. Bottom two panels illustrate that numbers of TADs are decreased in differentiated cells

compared to ESC both in humans (bottom left panel) and mice (bottom right panel).

Figure 3. Placements of hESC enhancers and HARs within TADs are directly correlated with the size of TADs in the hESC genome.

A. Significant direct correlations between the size of TADs and SE's span defined as a number of bp between the two most distant SEs located within a given TAD in the hESC genome (top two panels). Top left panel shows a correlation profile between the size of 504 TADs and genomic span of 642 SEs located within TADs. Top right panel shows a correlation profile between the size of 103 TADs harboring at least two SEs and genomic span of 241 SEs residing within TADs. Percentiles within shaded areas indicate the percent of TADs containing HSGRL within a designated set, which increases concomitantly with the increasing quantity of SEs located within TADs and larger TAD size. Note that similar trends were observed in the mESC genome, however, the correlation coefficient values were not statistically significant (bottom two panels).

B. Patterns of significant direct correlations between the size of 147 revTADs and numbers of hESC-enriched enhancers (top left panel) and numbers of HARs (top right panel) located within the revTADs. The bottom left panel illustrates the profile of significant correlation between the numbers of hESC-enriched enhancers and HARs residing within the revTADs (direct correlation; bottom left panel). In contrast, the bottom right panel shows the previously reported inverse correlation between the numbers of HSTFBS and HARs residing within the revTADs (Glinsky, 2016).

C. Genome-wide, there is a highly significant direct correlation between the numbers of hESC-enriched enhancers located within TADs and the average size of corresponding TADs harboring hESC-enriched enhancers. The numbers of hESC-enriched enhancers located within each individual TAD were quantified and TADs harboring 0 to 9 enhancers were segregated into subgroups harboring the same numbers of enhancers. The numbers of TADs in each subgroup are indicated. Eighty-one TADs harboring ten or more hESC-enriched enhancers (range 10-30) were segregated into one subgroup with the average enhancers' content of 12.5 per TAD. The average TAD sizes were computed for each subgroup of TADs and the results were plotted to assess the correlation pattern.

Figure 4. Mechanisms of HSGRL-mediated effects on principal regulatory structures of the interphase chromatin contributing to the evolution of genomic regulatory networks.

Collectively, the ensemble of these structural changes facilitated by the targeted placements and retention of distinct families of HSGRL at specific genomic locations would enable the emergence of new SED structures and remodeling of existing TADs to increase the complexity and enhance the precision of genomic regulatory networks. See text for details and relevant references.

Figure 5. Clusters of Alu elements in the vicinity of putative DNA bending sites near the borders of SEDs and TADs.

A. Clusters of Alu elements near the borders of the ID3 SED on chr1:23,878,033-23,894,300 (16,268 bp). Dotted lines depict the genomic positions of the overlapping CTCF/cohesin-binding sites, interactions of which form the anchor base of the ID3 SED.

B. Clusters of Alu elements near the NANOG SED left border on chr12:7,864,594-7,869,500 (4,907 bp).

C. Clusters of Alu elements near the NANOG SED right border on chr12:8,012,400-8,017,400 (5,001 bp).

Arrows in the figures (B) and)C) point the overlapping CTCF/cohesin-binding sites, interactions between which form the anchor base of the NANOG SED.

Note that clusters of closely-spaced sequences of at least three Alu elements belonging to most ancient AluJ (~65 million years old), second oldest AluS (~30 million years old), and currently active modern AluY sub-families are observed, suggesting that placement and/or retention of Alu elements at these sites were occurring for millions of years and continues at the present time (see text for details).

Figure 6. Conservation patterns of HSGRL in individual human genomes.

Conservation of HSGRL in individual genomes of 3 Neanderthals, 12 Modern Humans, and the 41,000-year old Denisovan genome was carried-out by direct comparisons of corresponding sequences retrieved from individual genomes and the human genome reference database (http://genome.ucsc.edu/Neandertal/). Full-length sequence alignments with no gaps of the individual human genome sequences to the corresponding sequences in the human genome reference databases were accepted as the evidence of sequence

conservation in the individual human genome.

A. Top two panels show conservation patterns of HSGRL located within the revTADs for 57 HARs and associated 75 CTCF-binding sites (top left panel); 22 HARs harboring 23 LMNB1-binding sites and 23 LMNB1-binding sites residing within 22 HARs (top right panel). Bottom two panels show conservation patterns of 90 HARs harboring 93 LMB1-binding sites (bottom left panel) and 55 HARs harboring 55 CTCF-binding sites and 55 CTCF-binding sites located within 55 HARs (bottom right panel).

B. Top two panels show conservation patterns of 123 HARs associated with 127 high-confidence overlapping CTCF/RAD21-binding sites (top left pane) and 127 high-confidence overlapping CTCF/RAD21-binding sites associated with 123 HARs (top right panel). Bottom two panels show conservation patterns of 69 CTCF/RAD21-binding sites conserved in all Neandertals' genome (bottom left panel) and 152 human-specific NANOG-binding sites highly-conserved in individual human genomes (bottom right panel). Note that conservation of all HSGRL are consistently at the lowest level in the Neanderthals' genome, suggesting that creation and/or retention rates of HSGRL are enhanced in Modern Humans. However, HSGRL sequences manifesting the relatively high conservation levels in the Neanderthals' genome appear most conserved in the individual genomes of Modern Humans as well, including the 41,000-year old Denisovan genome (bottom left panel). Sequences of HSGRL with assigned biochemical functions, e.g., specific TFBS or Lamin B1 (LMNB1)-binding sites, which are residing within HARs exhibit markedly higher conservation levels compared to sequences of HARs harboring the corresponding HSGRL. This conclusion remains valid for HSGRL sequences with assigned specific biochemical or biological functions that were associated with HARs by proximity placement analyses (figures A, B).

C. Direct correlations of conservation profiles of DNA sequences of distinct HSGRL in individual human genomes (top two panels) and markedly different values of genomic fitness scores (GFS) integrating into a single numerical value sequence conservation data of 909 HSGRL in individual human genomes (bottom panel). Note that values of GFS are inversely correlated with the variation coefficients of GFS values in individual human genomes, consistent with the hypothesis that GFS reflects the intrinsic property of an individual genome to create and/or retain the HSGRL.

Table 1. Association of hESC super-enhancers with human-specific regulatory loci

Genomic Features*	TADs	H1 hESC-SE	hsTFBS	HARs
Genome	3,062	642	3803	2745
HARs + SE	168	212	0	406
Enrichment Factor	1.0	6.0	NA	2.8
HSTFBS + SE	56	72	146	0
Enrichment Factor	1.0	6.1	2.1	NA
HARs + HSTFBS + SE	55	85	154	158
Enrichment Factor	1.0	7.4	2.3	3.3
SE + HSGRL	279	369	300	564

HARs, Human Accelerated Regions; HSTFBS, human-specific Transcription Factor-Binding Sites; SE, Super Enhancers; HSGRL, human-specific genomic regulatory loci; H1 hESC, H1 human Embryonic Stem Cells; NA, not applicable; * - genomic coordinates of the hg19 human genome reference database;

Table 2. Size distribution analysis of super-enhancers in human and mouse ESC

Category	hESC super-enhancers		mESC super-enhancers		Fold enrichment in hESC
Size, bp	Number	Percent	Number	Percent	
> 30,000	29	4.2	4	1.7	7.25
>20,000	102	14.9	28	12.1	3.64
>10,000	331	48.4	108	46.8	3.06
>5,000	554	81.0	154	66.7	3.60
>2,000	683	99.9	193	83.5	3.54
1,000-2,000	0	0.0	21	9.1	0.00
<1,000	1	0.1	17	7.4	0.06
Total	684	100.0	231	100.0	2.96

hESC, human embryonic stem cells; mESC, mouse embryonic stem cells

Table 3. Size distribution analysis of topologically-associating domains in genomes of human and mouse ESC

Category	hESC TADs		mESC TADs		Enrichment/Depletion in hESC
Size, Kbp	Number	Percent	Number	Percent	
> 3,000	23	0.7	55	2.5	0.42
>2,000	160	5.1	255	11.6	0.63
>1,000	955	30.5	982	44.6	0.97
>500	2154	68.9	1745	79.3	1.23
>200	3054	97.7	2166	98.5	1.41
>100	3125	99.9	2200	100.0	1.42
<100	2	0.1	0	0.0	NA
Total	3127	100.0	2200	100.0	1.42

TADs, topologically-associating domains

Table 4. Estimates of creation time periods during evolution of SEs and enhancers in the hESC genome

1.1. Estimates of super-enhancers (SE) creation time in the human ESC genome

Number of SE in hESC genome	Shared with mESC	All new SE in hESC*	New SE on conserved sequences*	Primate-specific SE sequences*	Human-specific SE sequences**
684	25 (3.7%)	659	573	83	3
Number of new super-enhancers per 100,000 years		0.88	0.76	0.11	0.81
Number of years for creation of one super-enhancer		113,636	131,579	909,091	123,457

1.2. Estimates of enhancers' creation time in the human ESC genome

Number of enhancers in hESC genome	Shared with mESC	All new enhancers*	New enhancers on conserved sequences*	Primate-specific enhancers sequences*	Human-specific enhancers sequences**
7006	682 (9.7%)	6,324	4,714	1,420	190
Number of new enhancers per 100,000 years		8.4	6.3	1.9	51
Number of years for creation of one enhancer		11,905	15,873	52,632	1,961

* Calculated based on estimates of Humans & Chimpanzees split 13 million years ago and 88 million years from Euarchonta & Glires (Gliriformes) split resulting in the estimated evolutionary timeline of ~75 million years;
** Calculated based on estimates of Modern Humans & Neanderthals split 370,000 years ago;

Estimates for human-specific SE and enhancers were calculated based on the assumption that all human-specific sequences emerged after Modern Humans & Neanderthals split 370,000 years ago;

Estimates of creation time of primate-specific SE and enhancers were based on the assumption that all primate-specific sequences emerged before Modern Humans/Chimpanzees split 13 million years ago.

Figure 1.

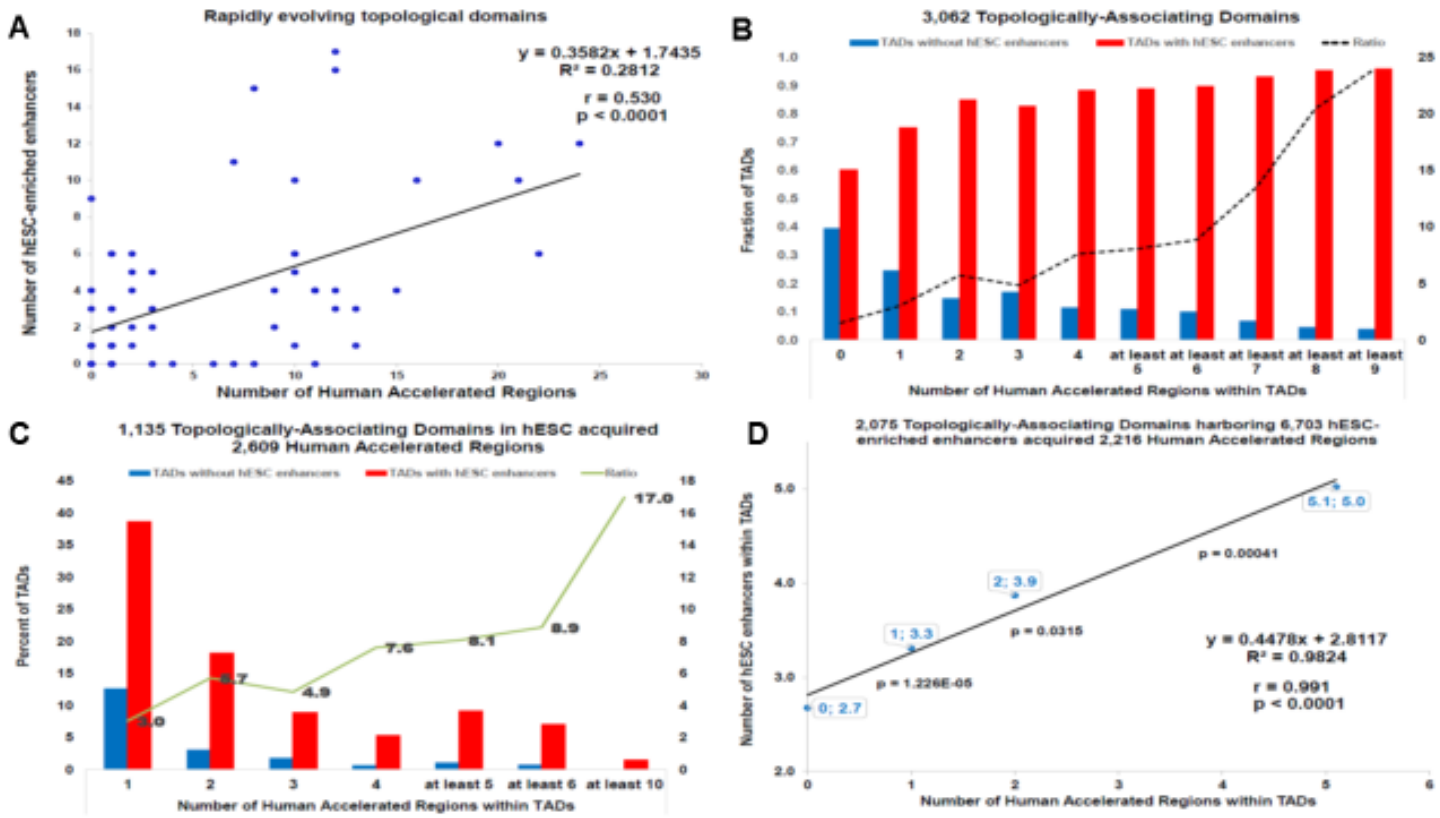

A
Rapidly evolving topological domains
y = 0.3582x + 1.7435
R² = 0.2812
r = 0.530
p < 0.0001
Number of hESC-enriched enhancers
Number of Human Accelerated Regions
B
3,062 Topologically-Associating Domains
TADs without hESC enhancers
TADs with hESC enhancers
Ratio
Fraction of TADs
0
1
2
3
4
at least 5
at least 6
at least 7
at least 8
at least 9
Number of Human Accelerated Regions within TADs
C
1,135 Topologically-Associating Domains in hESC acquired 2,609 Human Accelerated Regions
TADs without hESC enhancers
TADs with hESC enhancers
Ratio
Percent of TADs
1.0
3.7
4.9
7.6
8.1
8.9
17.0
1
2
3
4
at least 5
at least 6
at least 10
Number of Human Accelerated Regions within TADs
D
2,075 Topologically-Associating Domains harboring 6,703 hESC-enriched enhancers acquired 2,216 Human Accelerated Regions
Number of hESC enhancers within TADs
0; 2.7
1; 3.3
2; 3.9
5.1; 5.0
p = 1.226E-05
p = 0.0315
p = 0.00041
y = 0.4478x + 2.8117
R² = 0.9824
r = 0.991
p < 0.0001
Number of Human Accelerated Regions within TADs

Figure 2.

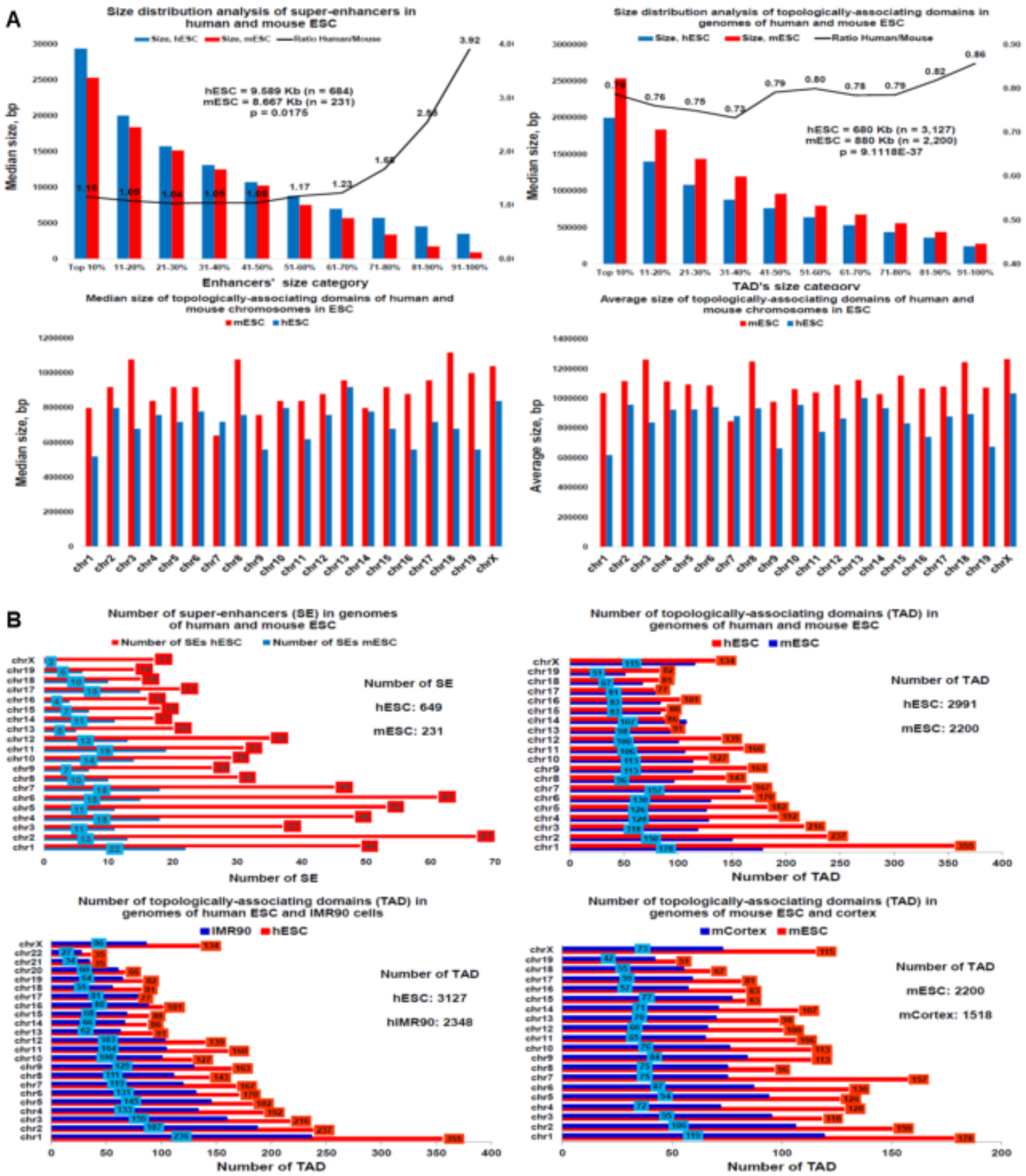
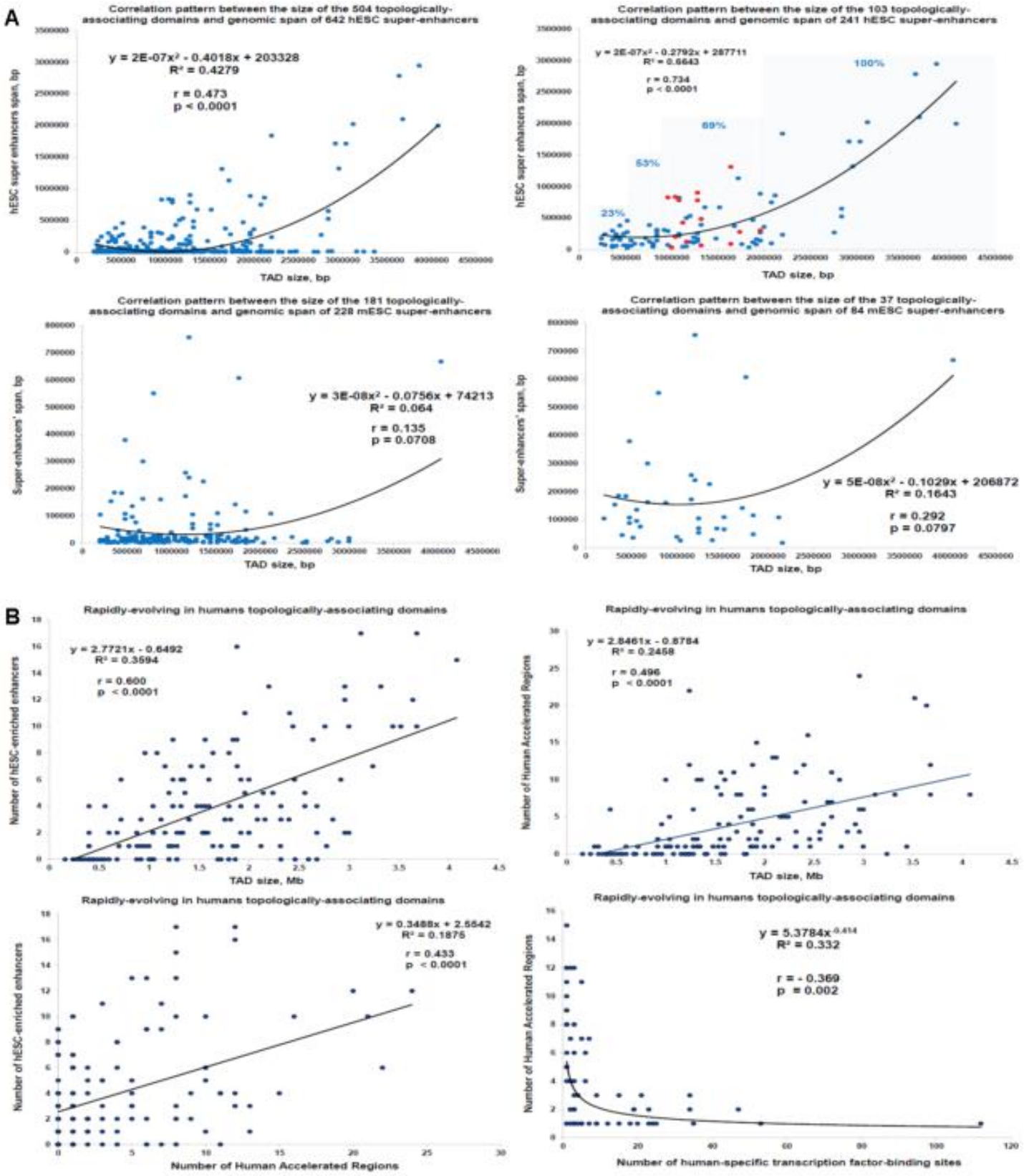

A
Size distribution analysis of super-enhancers in human and mouse ESC
Size, hESC
Size, mESC
Ratio Human/Mouse
hESC = 9.589 Kb (n = 684)
mESC = 8.667 Kb (n = 231)
p = 0.0175
Median size, bp
Enhancers' size category
Size distribution analysis of topologically-associating domains in genomes of human and mouse ESC
hESC = 680 Kb (n = 3,127)
mESC = 880 Kb (n = 2,200)
p = 9.1118E-37
TAD's size category
Median size of topologically-associating domains of human and mouse chromosomes in ESC
mESC
hESC
Average size of topologically-associating domains of human and mouse chromosomes in ESC
Average size, bp
B
Number of super-enhancers (SE) in genomes of human and mouse ESC
Number of SEs hESC
Number of SEs mESC
Number of SE
hESC: 649
mESC: 231
Number of topologically-associating domains (TAD) in genomes of human and mouse ESC
Number of TAD
hESC: 2991
mESC: 2200
Number of topologically-associating domains (TAD) in genomes of human ESC and IMR90 cells
IMR90
hESC
hIMR90: 2348
Number of topologically-associating domains (TAD) in genomes of mouse ESC and cortex
mCortex
mESC
mCortex: 1518

Figure 3.


A
Correlation pattern between the size of the 504 topologically-associating domains and genomic span of 642 hESC super-enhancers
y = 2E-07x² - 0.4018x + 203328
R² = 0.4279
r = 0.473
p < 0.0001
hESC super enhancers span, bp
TAD size, bp
Correlation pattern between the size of the 103 topologically-associating domains and genomic span of 241 hESC super-enhancers
y = 2E-07x² - 0.2792x + 287711
R² = 0.6643
r = 0.734
p < 0.0001
23%
53%
69%
100%
hESC super enhancers span, bp
TAD size, bp
Correlation pattern between the size of the 181 topologically-associating domains and genomic span of 228 mESC super-enhancers
y = 3E-08x² - 0.0756x + 74213
R² = 0.064
r = 0.135
p = 0.0708
Super-enhancers' span, bp
TAD size, bp
Correlation pattern between the size of the 37 topologically-associating domains and genomic span of 84 mESC super-enhancers
y = 5E-08x² - 0.1029x + 206872
R² = 0.1643
r = 0.292
p = 0.0797
Super-enhancers' span, bp
TAD size, bp
B
Rapidly-evolving in humans topologically-associating domains
y = 2.7721x - 0.6492
R² = 0.3594
r = 0.600
p < 0.0001
Number of hESC-enriched enhancers
TAD size, Mb
Rapidly-evolving in humans topologically-associating domains
y = 2.8461x - 0.8784
R² = 0.2458
r = 0.496
p < 0.0001
Number of Human Accelerated Regions
TAD size, Mb
Rapidly-evolving in humans topologically-associating domains
y = 0.3488x + 2.5642
R² = 0.1876
r = 0.433
p < 0.0001
Number of hESC-enriched enhancers
Number of Human Accelerated Regions
Rapidly-evolving in humans topologically-associating domains
y = 5.3784x^-0.414
R² = 0.332
r = - 0.369
p = 0.002
Number of Human Accelerated Regions
Number of human-specific transcription factor-binding sites

C

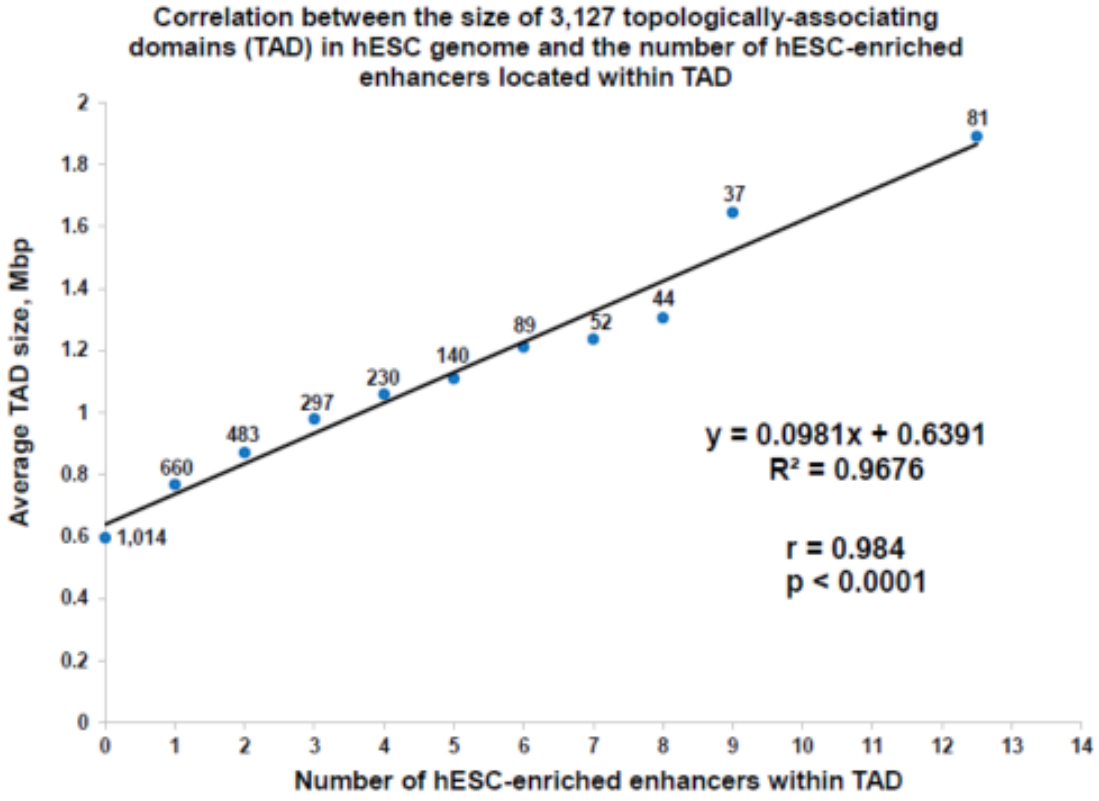

Correlation between the size of 3,127 topologically-associating domains (TAD) in hESC genome and the number of hESC-enriched enhancers located within TAD
Average TAD size, Mbp
Number of hESC-enriched enhancers within TAD
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1,014
660
483
297
230
140
89
52
44
37
81
y = 0.0981x + 0.6391
R² = 0.9676
r = 0.984
p < 0.0001

Figure 4.

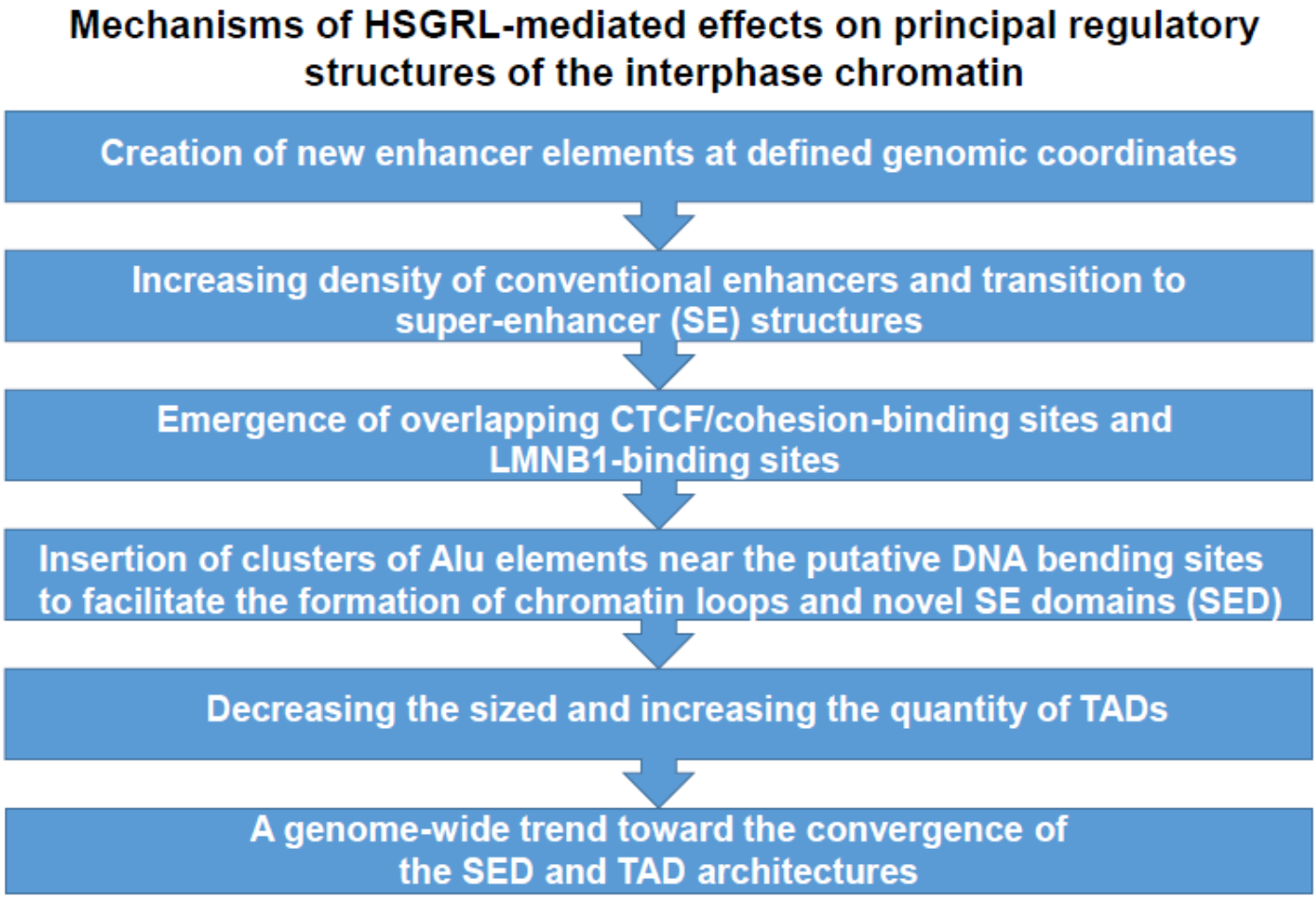
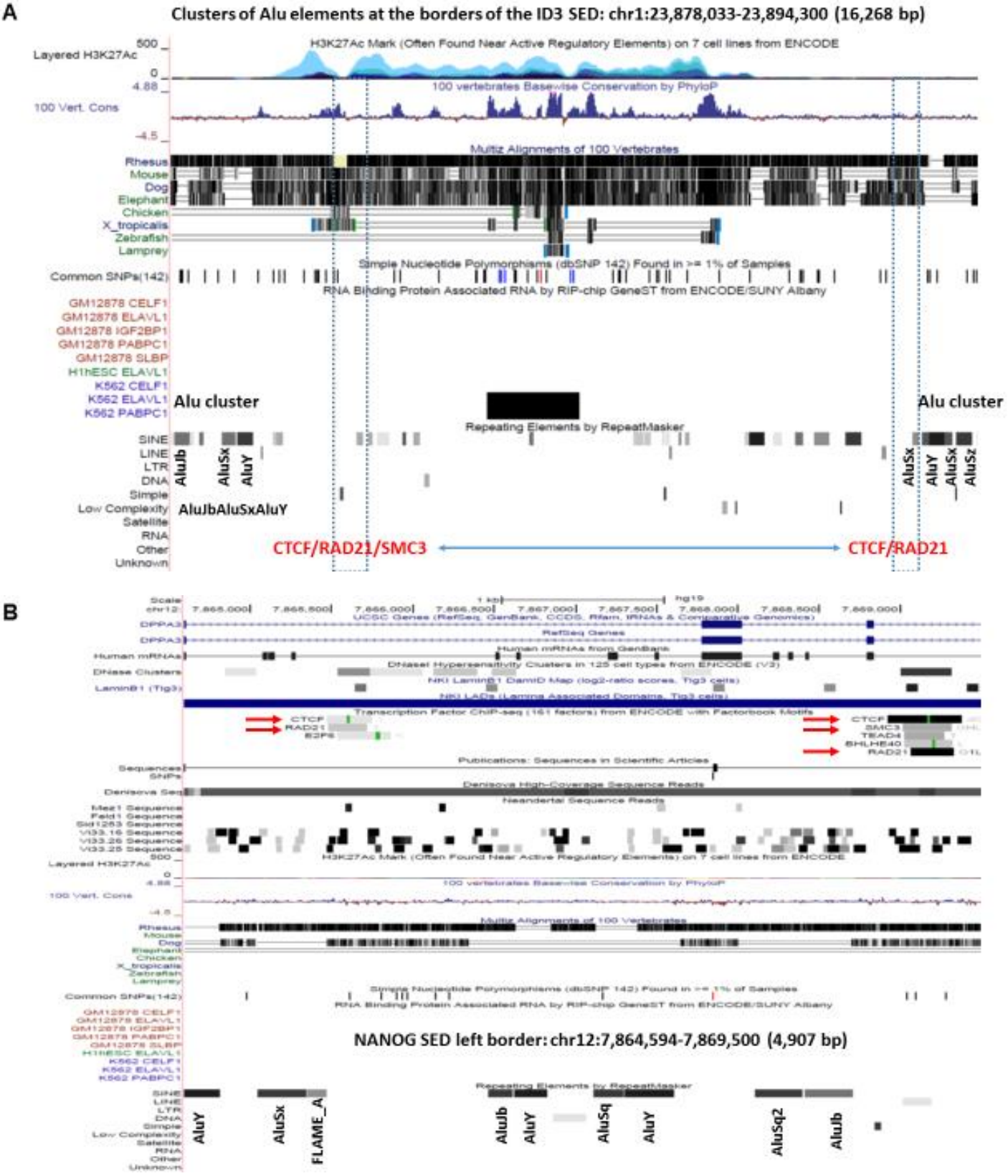

Mechanisms of HSGRL-mediated effects on principal regulatory structures of the interphase chromatin
Creation of new enhancer elements at defined genomic coordinates
Increasing density of conventional enhancers and transition to super-enhancer (SE) structures
Emergence of overlapping CTCF/cohesion-binding sites and LMNB1-binding sites
Insertion of clusters of Alu elements near the putative DNA bending sites to facilitate the formation of chromatin loops and novel SE domains (SED)
Decreasing the sized and increasing the quantity of TADs
A genome-wide trend toward the convergence of the SED and TAD architectures

Figure 5.


A
Clusters of Alu elements at the borders of the ID3 SED: chr1:23,878,033-23,894,300 (16,268 bp)
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
100 Vert. Cons
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(142)
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
RNA Binding Protein Associated RNA by RIP-chip GeneST from ENCODE/SUNY Albany
GM12878 CELF1
GM12878 ELAVL1
GM12878 IGF2BP1
GM12878 PABPC1
GM12878 SLBP
H1hESC ELAVL1
K562 CELF1
K562 ELAVL1
K562 PABPC1
Alu cluster
Alu cluster
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
AluJb
AluSx
AluY
AluJbAluSxAluY
AluSx
AluY
AluSx
AluSz
CTCF/RAD21/SMC3
CTCF/RAD21
B
NANOG SED left border: chr12:7,864,594-7,869,500 (4,907 bp)
AluY
AluSx
FLAME_A
AluJb
AluY
AluSq
AluY
AluSq2
AluJb

C

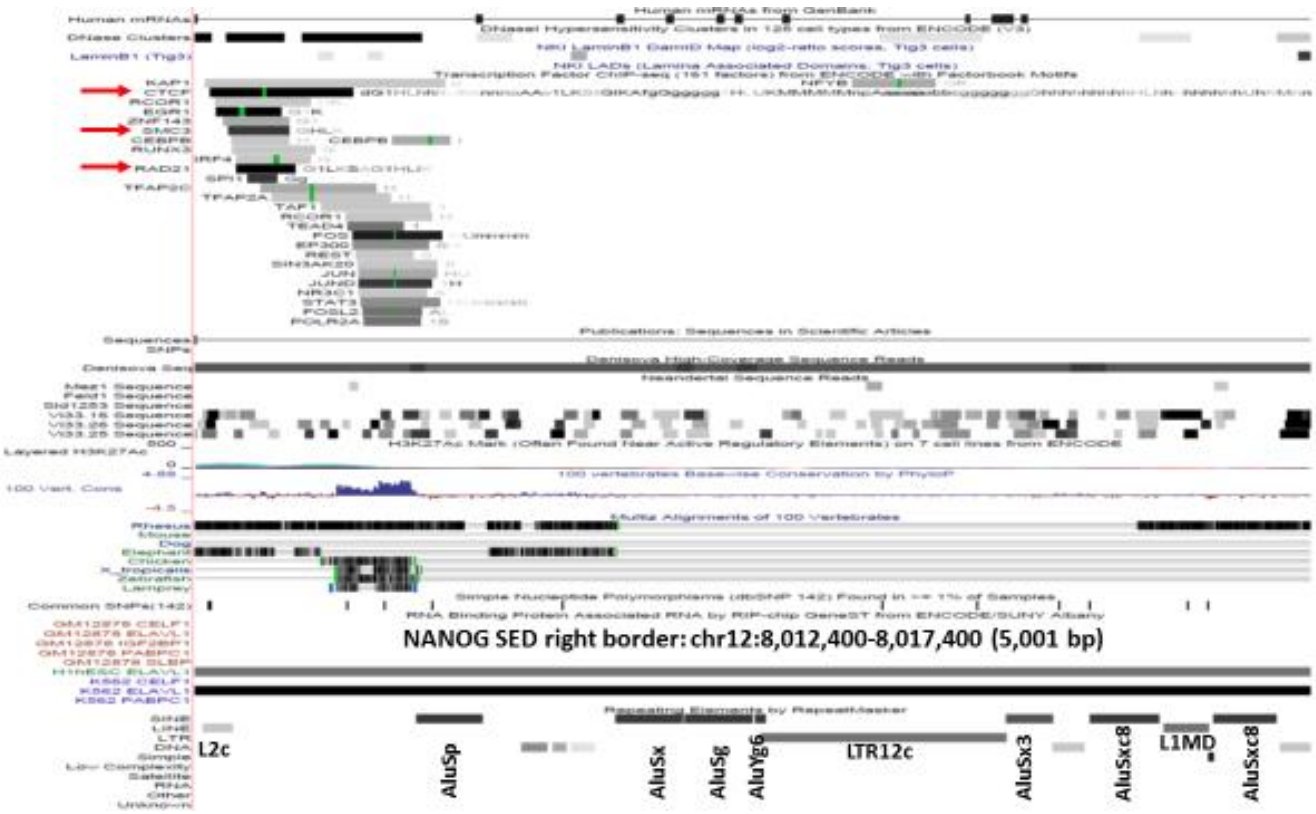
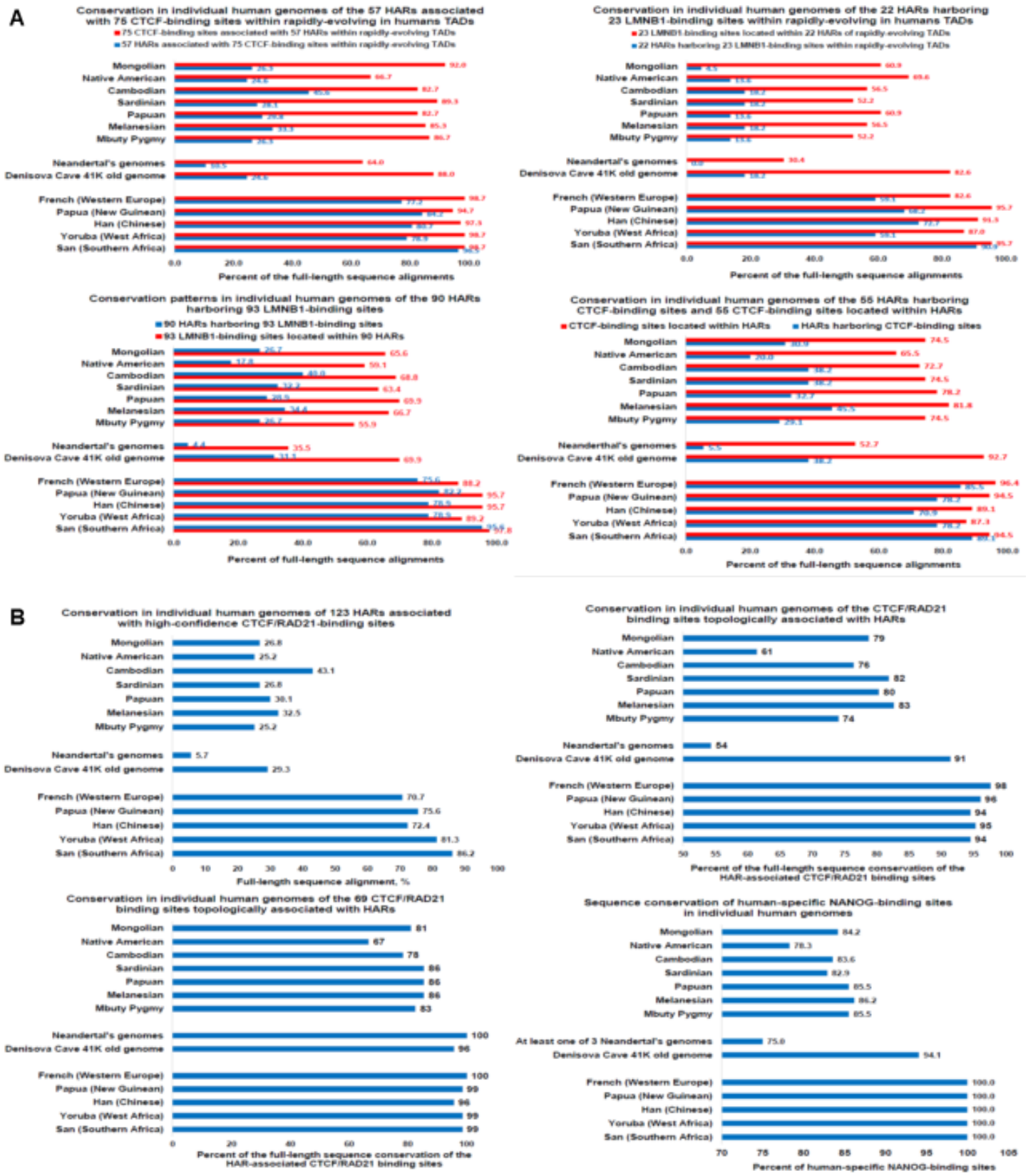
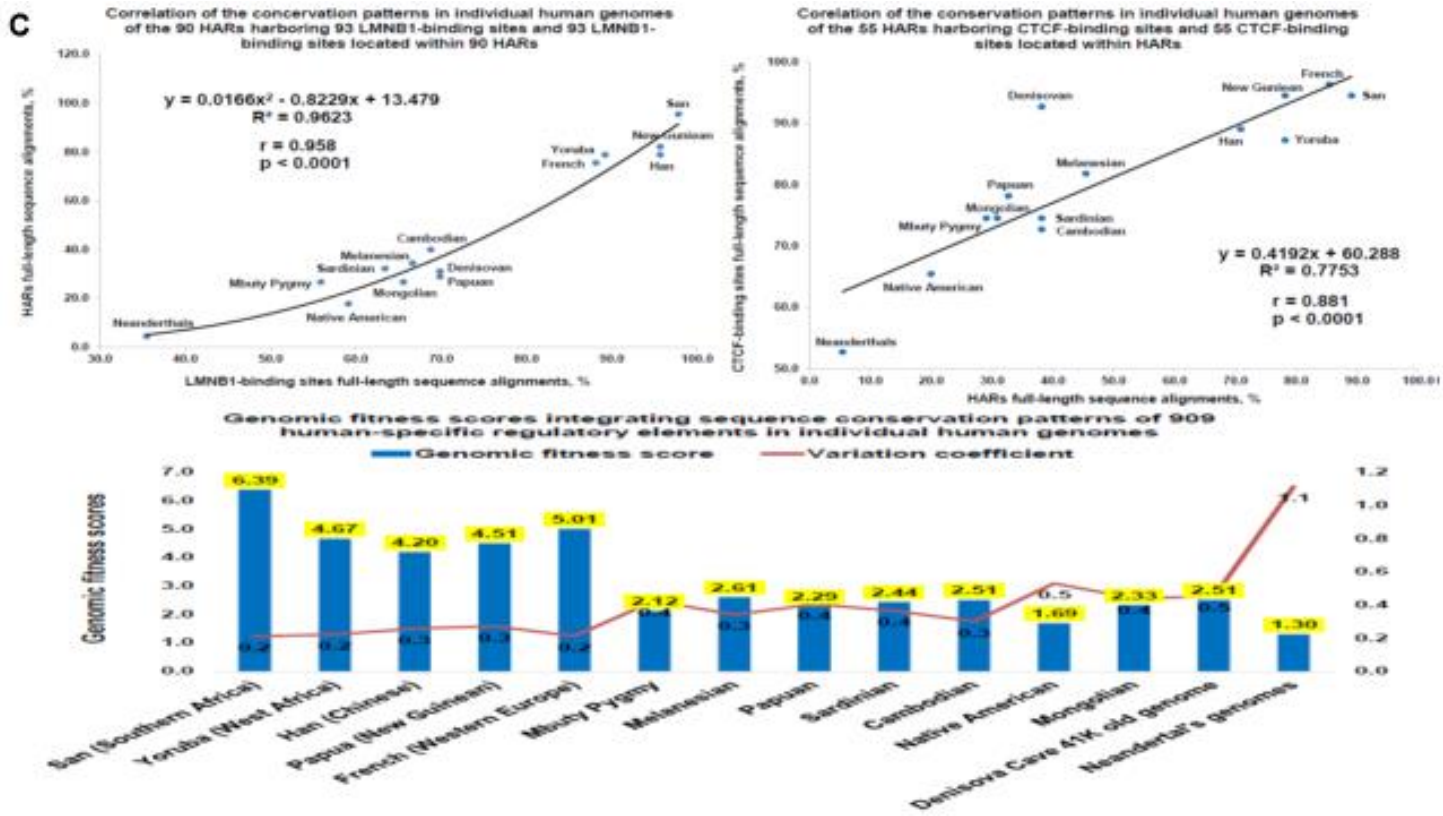

NANOG SED right border: chr12:8,012,400-8,017,400 (5,001 bp)
L2c
AluSp
AluSx
AluSg
AluYg6
LTR12c
AluSx3
AluSxc8
L1MD
AluSxc8

Figure 6.


A
Conservation in individual human genomes of the 57 HARs associated with 75 CTCF-binding sites within rapidly-evolving in humans TADs
75 CTCF-binding sites associated with 57 HARs within rapidly-evolving TADs
57 HARs associated with 75 CTCF-binding sites within rapidly-evolving TADs
Mongolian
Native American
Cambodian
Sardinian
Papuan
Melanesian
Mbuty Pygmy
Neandertal's genomes
Denisova Cave 41K old genome
French (Western Europe)
Papua (New Guinean)
Han (Chinese)
Yoruba (West Africa)
San (Southern Africa)
Percent of the full-length sequence alignments
Conservation in individual human genomes of the 22 HARs harboring 23 LMNB1-binding sites within rapidly-evolving in humans TADs
23 LMNB1-binding sites located within 22 HARs of rapidly-evolving TADs
22 HARs harboring 23 LMNB1-binding sites within rapidly-evolving TADs
Percent of the full-length sequence alignments
Conservation patterns in individual human genomes of the 90 HARs harboring 93 LMNB1-binding sites
90 HARs harboring 93 LMNB1-binding sites
93 LMNB1-binding sites located within 90 HARs
Percent of full-length sequence alignments
Conservation in individual human genomes of the 55 HARs harboring CTCF-binding sites and 55 CTCF-binding sites located within HARs
CTCF-binding sites located within HARs
HARs harboring CTCF-binding sites
Percent of the full-length sequence alignments
B
Conservation in individual human genomes of 123 HARs associated with high-confidence CTCF/RAD21-binding sites
Full-length sequence alignment, %
Conservation in individual human genomes of the CTCF/RAD21 binding sites topologically associated with HARs
Percent of the full-length sequence conservation of the HAR-associated CTCF/RAD21 binding sites
Conservation in individual human genomes of the 69 CTCF/RAD21 binding sites topologically associated with HARs
Percent of the full-length sequence conservation of the HAR-associated CTCF/RAD21 binding sites
Sequence conservation of human-specific NANOG-binding sites in individual human genomes
At least one of 3 Neandertal's genomes
Percent of human-specific NANOG-binding sites


C
Correlation of the concervation patterns in individual human genomes of the 90 HARs harboring 93 LMNB1-binding sites and 93 LMNB1-binding sites located within 90 HARs
y = 0.0166x² - 0.8229x + 13.479
R² = 0.9623
r = 0.958
p < 0.0001
HARs full-length sequence alignments, %
LMNB1-binding sites full-length sequence alignments, %
San
New Guinean
Yoruba
French
Han
Cambodian
Melanesian
Sardinian
Denisovan
Papuan
Mixty Pygmy
Mongolian
Native American
Neanderthals
Corelation of the conservation patterns in individual human genomes of the 55 HARs harboring CTCF-binding sites and 55 CTCF-binding sites located within HARs
y = 0.4192x + 60.288
R² = 0.7753
r = 0.881
p < 0.0001
CTCF-binding sites full-length sequence alignments, %
HARs full-length sequence alignments, %
Genomic fitness scores integrating sequence conservation patterns of 909 human-specific regulatory elements in individual human genomes
Genomic fitness score
Variation coefficient
Genomic fitness scores
6.39
4.67
4.20
4.51
5.01
2.12
2.61
2.29
2.44
2.51
1.69
2.33
2.51
1.30
San (Southern Africa)
Yoruba (West Africa)
Han (Chinese)
Papua (New Guinean)
French (Western Europe)
Mixty Pygmy
Melanesian
Papuan
Sardinian
Cambodian
Native American
Mongolian
Denisova Cave 41K old genome
Neanderthal's genomes