Supplemental Methods

Peak calling and cross-species comparison

We mapped all other FASTQ files to hg38 or panTro5 genomes using BWA with default options (Li 2013). Following the ENCODE recommendation, we used MACS2 with default options for peak calling (Landt et al. 2012). We applied the broad peaks option to H3K9me3 ChIP-seq data and the narrow peak option to the other ChIP-seq datasets. For narrow peaks with multiple replicates, we further applied IDR to all possible pairs of replicates (Li et al. 2011). We only included peak regions confirmed by at least one IDR run in the final peak output. For H3K9me3 broad peaks, we only included peaks with a Q-value < 0.05 that appeared at least three times in 9 human samples or at least two times in 7 chimpanzee samples.

Enrichment analysis

We calculated the number of indel-RRR overlapping using BEDTools intersect function (`bedtools intersect -a indel.bed -b RRR.bed`). To calculate the enrichment of overlap between two sets of intervals, we used the BEDTools fisher function to perform Fisher's exact test within the entire hg38 genome (`bedtools fisher -a a.bed -b b.bed -g hg38.chromSize`) (Quinlan and Hall 2010). We used the hypergeometric distribution function phyper with log.p=TRUE in R to calculate the extremely small P-values. We also directly performed Fisher's exact test using a contingency table if the number of intersections was available (R code: fisher.test(table2×2, alternative="two.sided")). To perform the permutation test, we also used BEDTools to shuffle one set of coordinates 1000 times with the --noOverlapping option, followed by an intersection with the other set of coordinates. We used two-tailed tests for all P-value-related statistics.

Data visualization

The deepTools (Ramirez et al. 2016) options used in different figures: Figure 5A,C: scale-regions –metagene; Figure 6: scale-regions for TE insertion plot and –referencePoint center for pre-insertion sites; Figure 7A,B: --referencePoint TES --sortUsing region_length; Supplemental_Fig_S6: --reference Point center –sortUsing region_length –binSize binsize. In Supplemental_Fig_S7, we calculated profile using deepTools plotProfile and exported data using –outFileNameData option. Linear smoothing was performed using polynomial function.

Reference

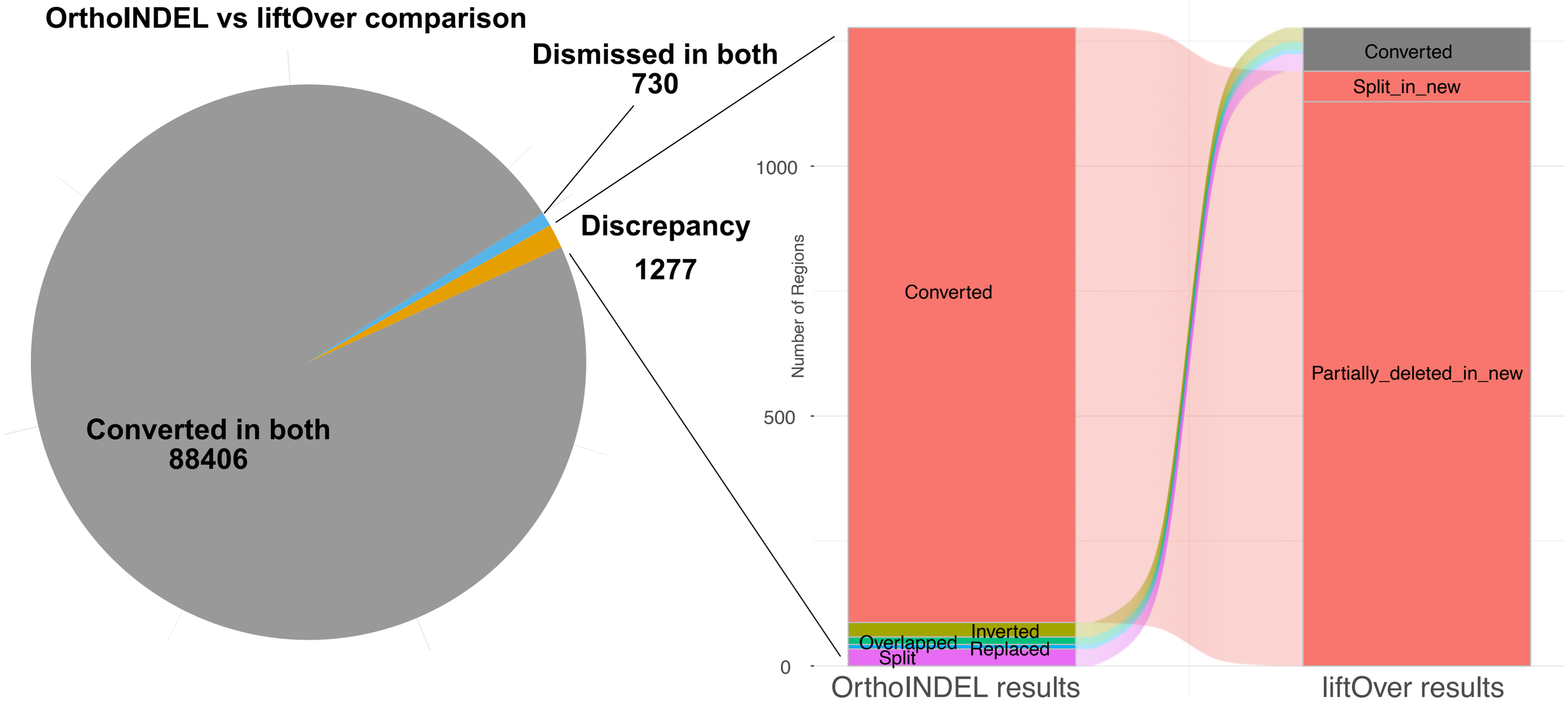
Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22**: 1813–1831.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arxivorg*.

Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**: 1752–1779.

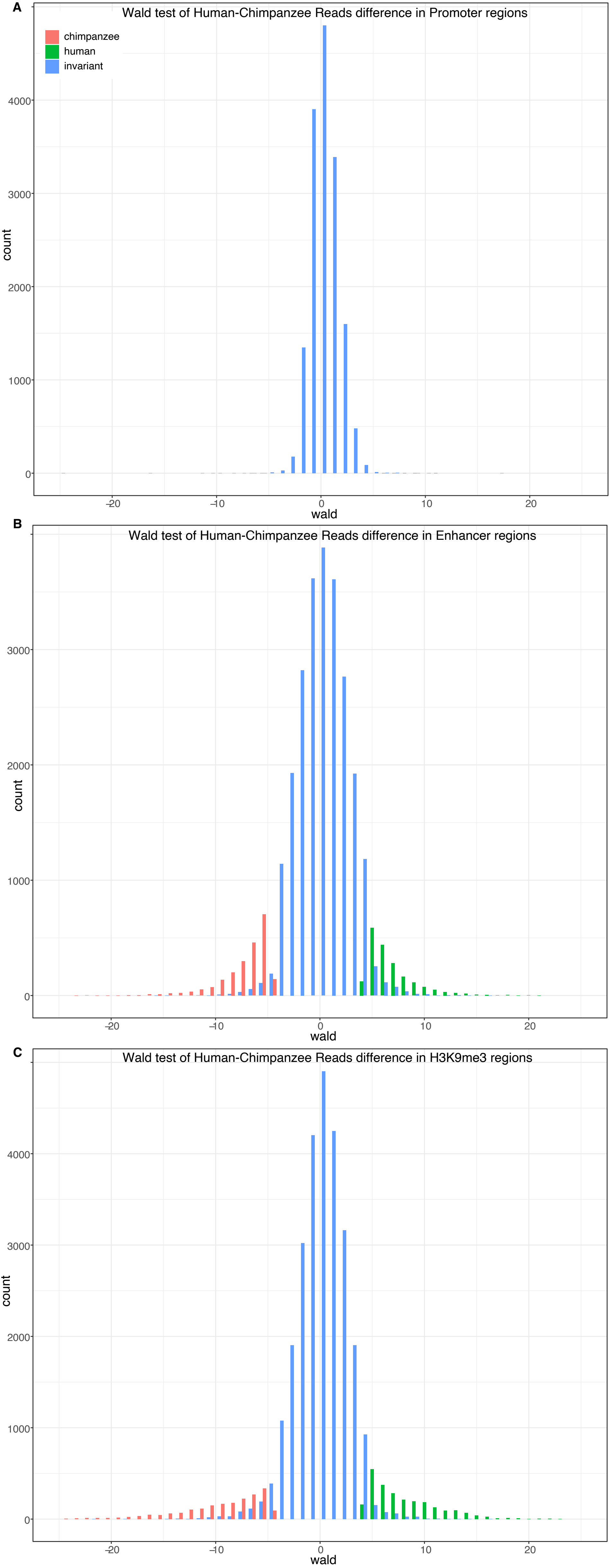
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Ramirez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**: W160–5.



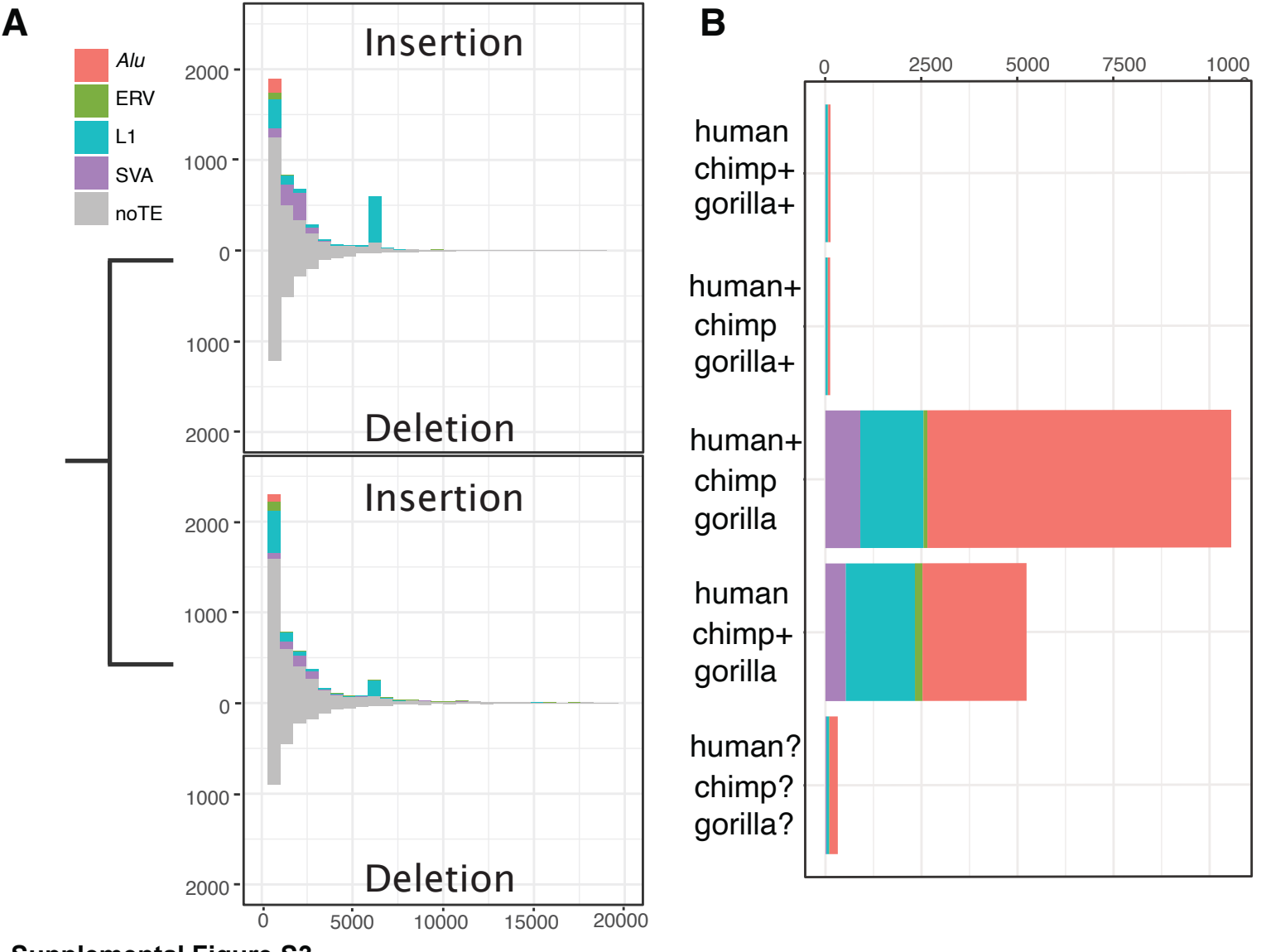
Supplemental Figure S1

Comparison between UCSC liftOver and OrthoINDEL. On the left is a pie chart showing number of regions can be converted from hg38 to panTro5 by both liftOver and OrthoINDEL, dismissed by both pipelines, or can only be converted by one of them. On the right is a bar chart showing the discrepancy between the two pipelines. On the left is a bar showing number of regions in OrthoINDEL results. Those failed to convert were classified as Inverted, Overlapped, Replaced or Split. On the right is a bar showing the same number in liftOver results. Those dismissed by it were categorized as Split_in_new or Partially_deleted_in_new. The bar plot comparison shows the majority of discrepancy are those regions converted by OrthoINDEL but dismissed by liftOver because of partially deletion.



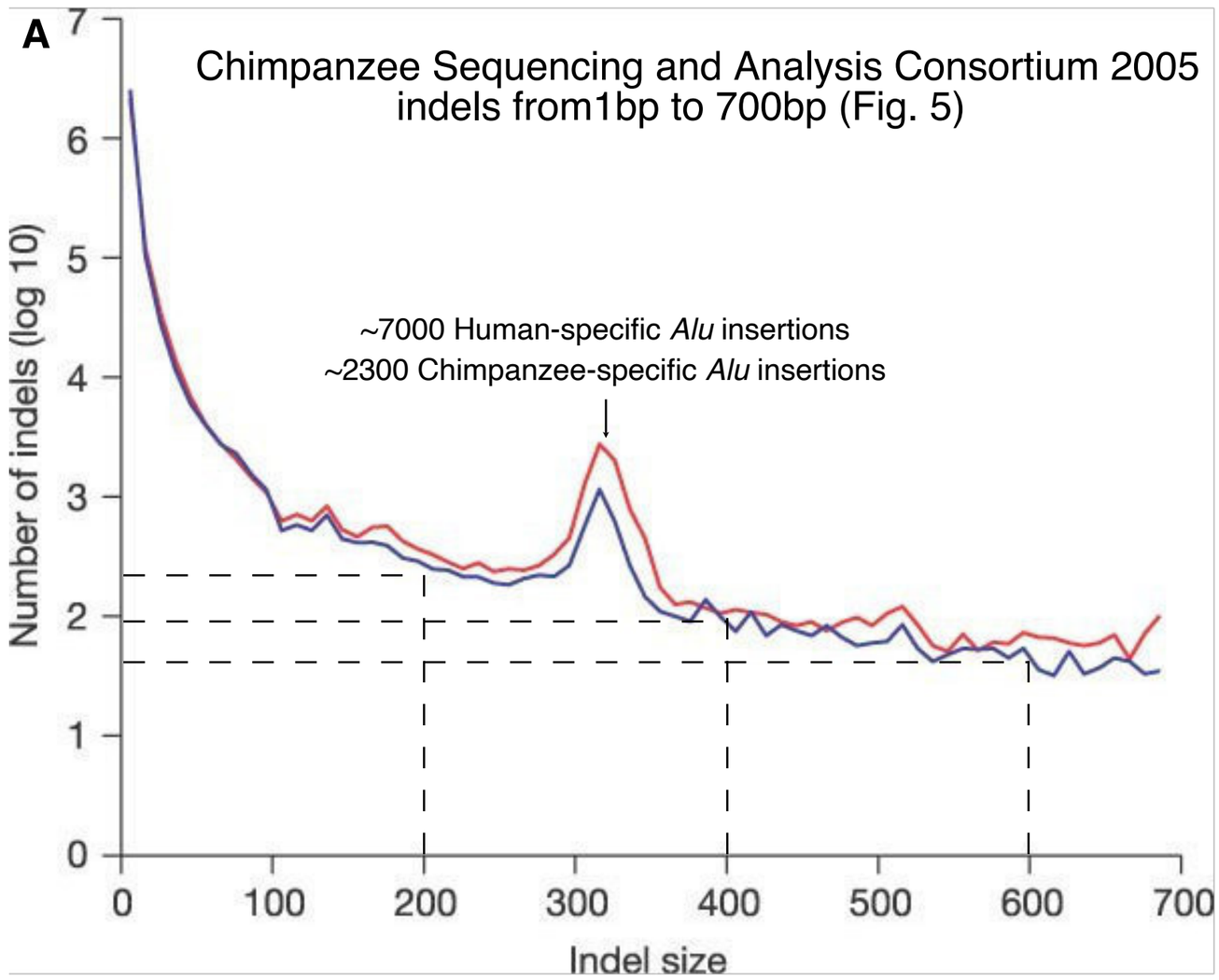
Supplemental Figure S2

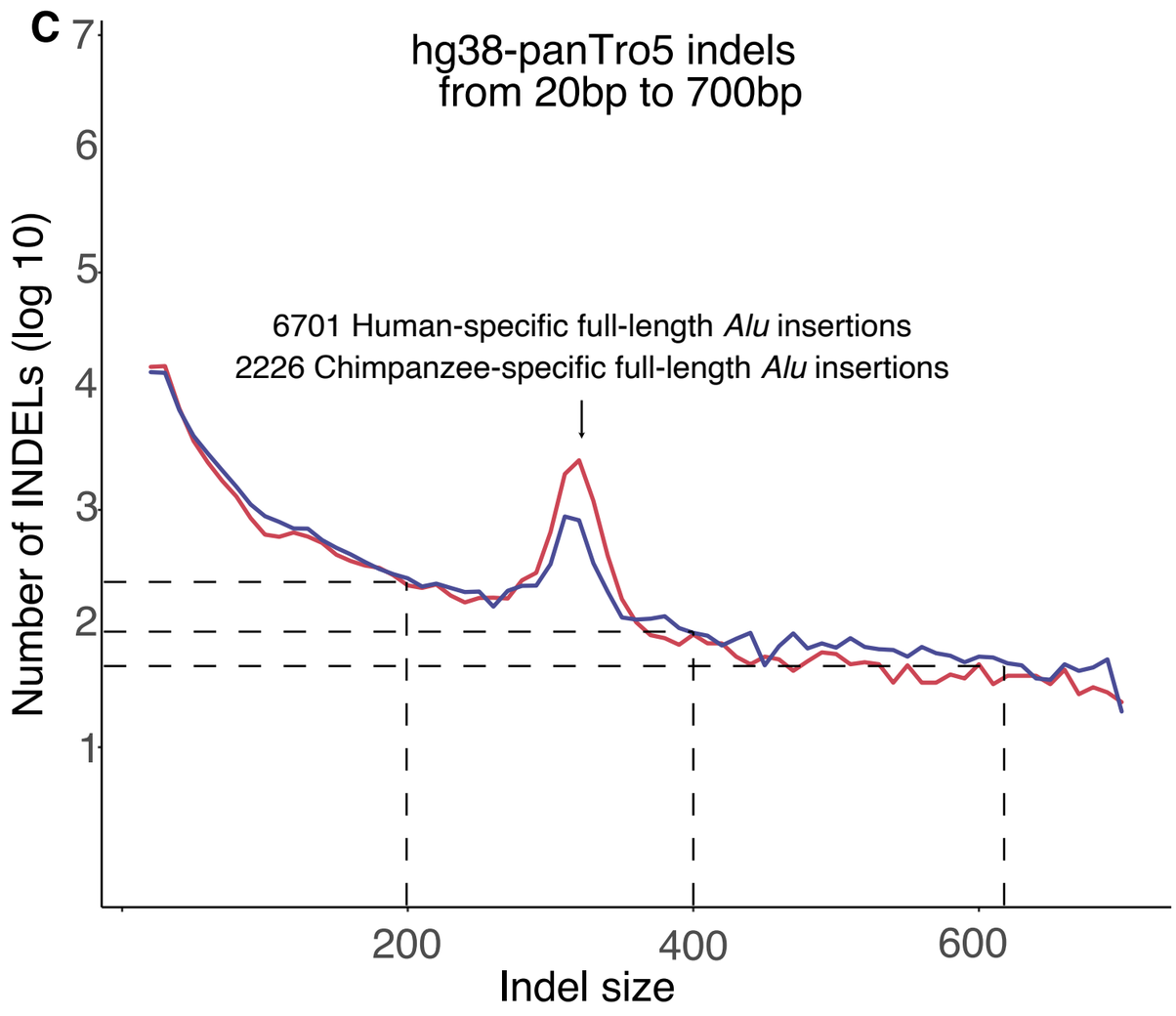
Wald static distribution histogram of invariant or lineage-biased promoter, enhancer and H3K9me3 regions. X-axis shows the Wald statistic of human read counts minus chimpanzee read counts. Human-biased, chimpanzee-biased and invariant regions were labeled with green, red and blue. A: Wald statistic distribution of promoter regions; B: Wald statistic distribution in enhancer regions; C: Wald statistic distribution in H3K9me3 regions.

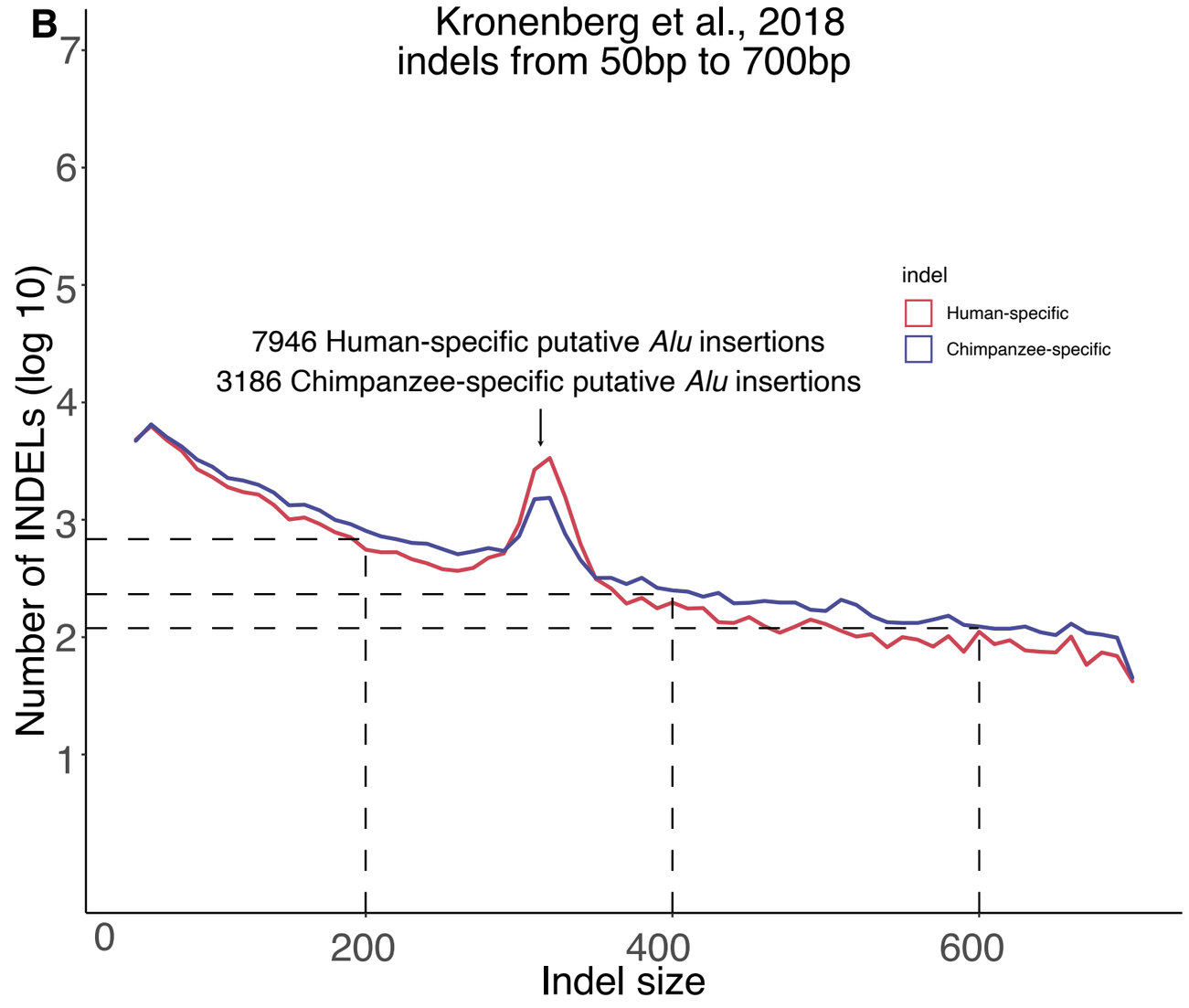


Supplemental Figure S3

A. All indels landing in regions with 200bp flanking sequences mappability > 0.7 between human and chimpanzee. The number of insertions and deletions in each lineage is plotted in a back-to-back histogram with indel length on the x-axis and the number of indels of different lengths on the y-axis. Colors distinguish indels based on TE classification ("noTE", not derived from TE). B. The majority of human/chimpanzee-specific TE insertions are absent in the gorilla genome. The same color scheme as in Figure 1a is used to illustrate *Alu*, ERV, L1, and SVA insertions. From top to bottom, bars represent the number of TE insertions only present in chimpanzee, only present in human, shared by human and gorilla but absent in chimpanzee, and shared by chimpanzee and gorilla but absent in human. The last bar provides the number of human/chimpanzee-specific TE insertions without clearly defined orthologous loci in the gorilla genome.

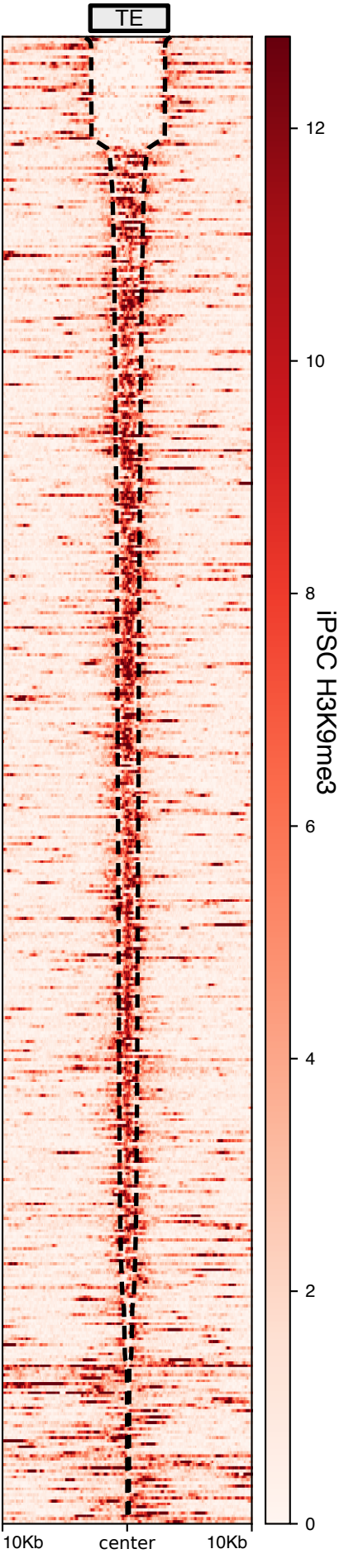






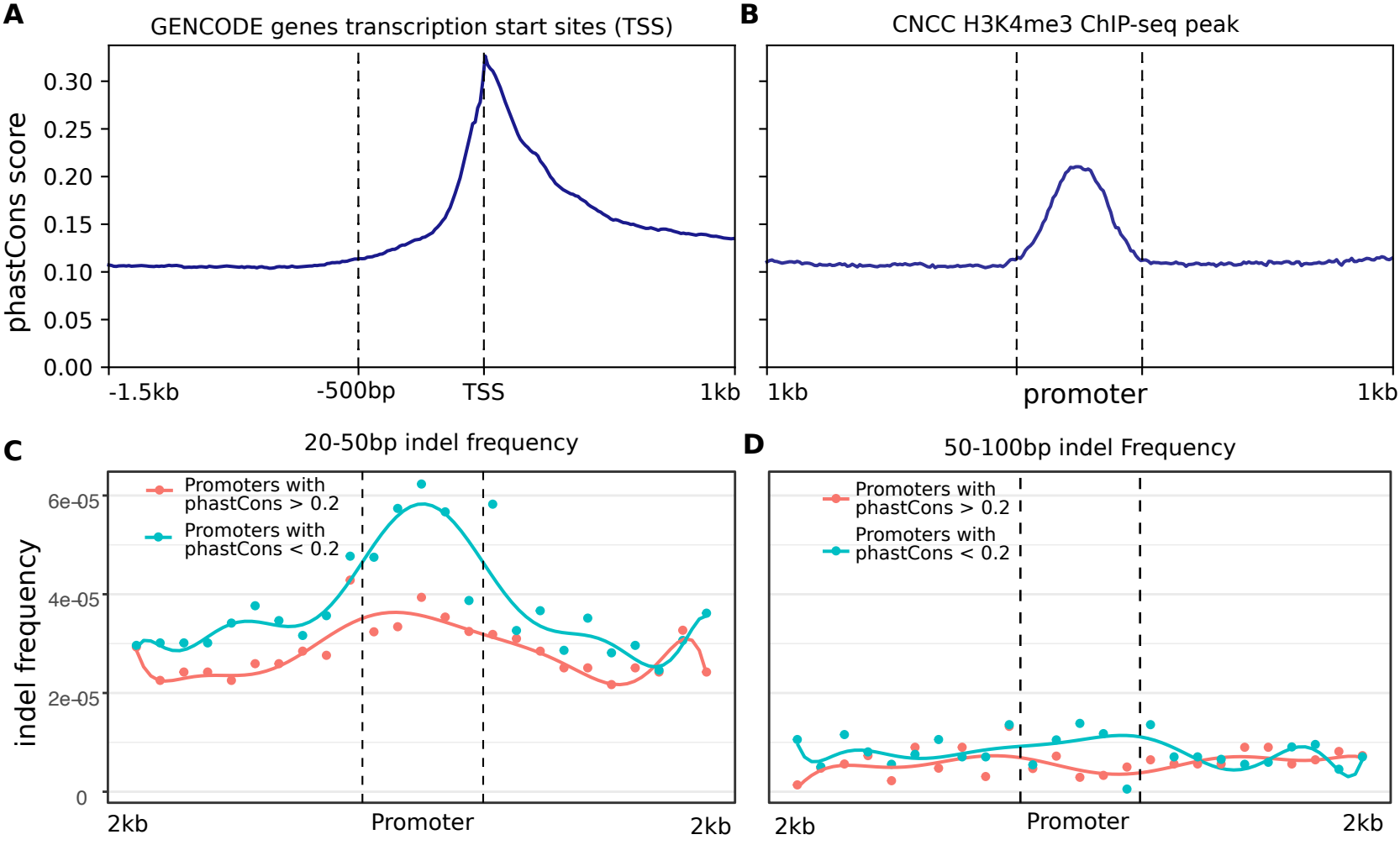
Supplemental Figure S4

Indel call set comparison between Chimpanzee Sequencing and Analysis Consortium 2005, Kronenberg et al 2018 and hg38-panTro5 indel called in this study. A. Species-specific indel length distribution (adapted from Chimpanzee Sequencing and Analysis Consortium 2005 Fig.5). B. Indel length distribution from 50bp to 700bp of Kronenberg et al 2018. C. Indel length distribution from 20bp to 700bp of indel identified in this study. D Venn Diagram overlap between hg38-panTro5 indels (≥50bp) identified by DASVC in this study with indels identified by Kronenberg et al 2018.



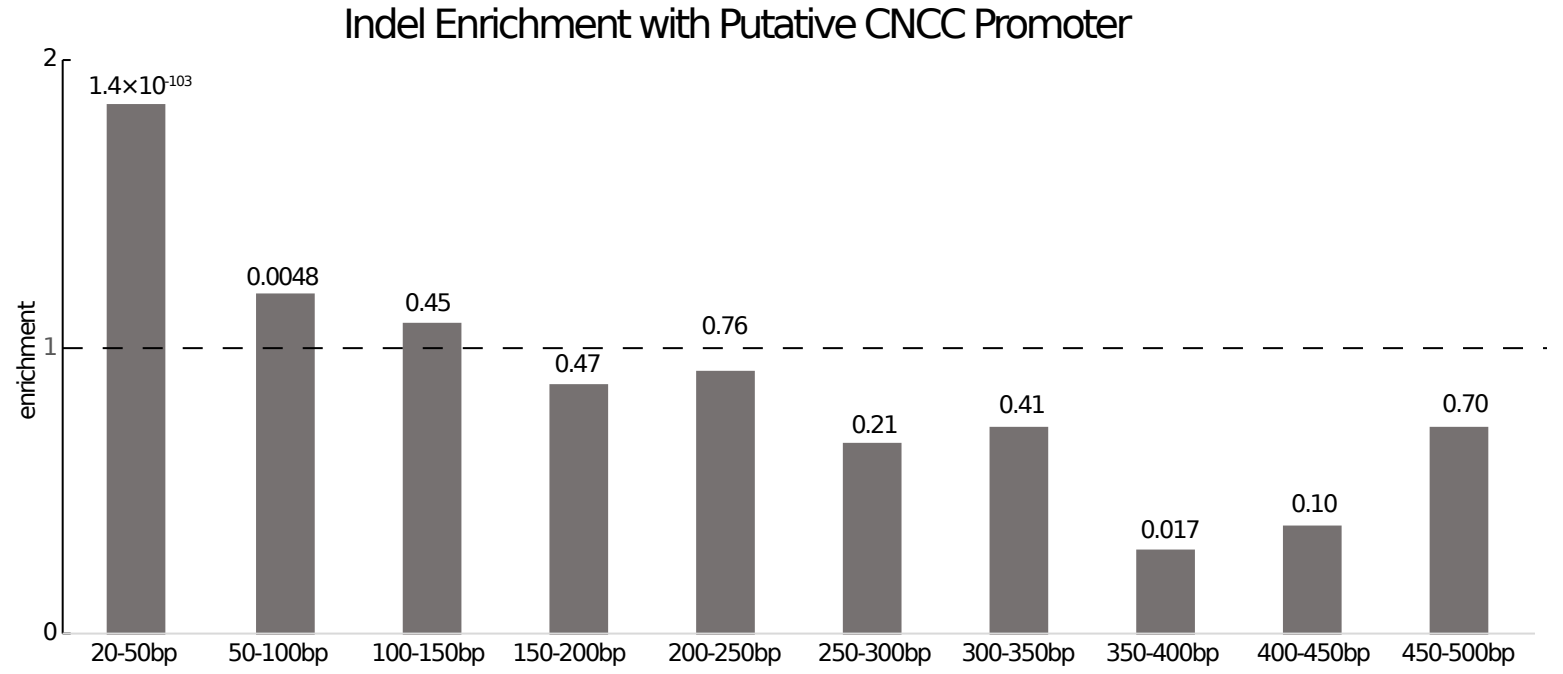
Supplemental Figure S5

New human-biased heterochromatin regions associated with human-specific TE insertions did not spread beyond the TE insertion. The TE insertion boundary is illustrated, and iPSC H3K9me3 signals are shown as a heatmap.



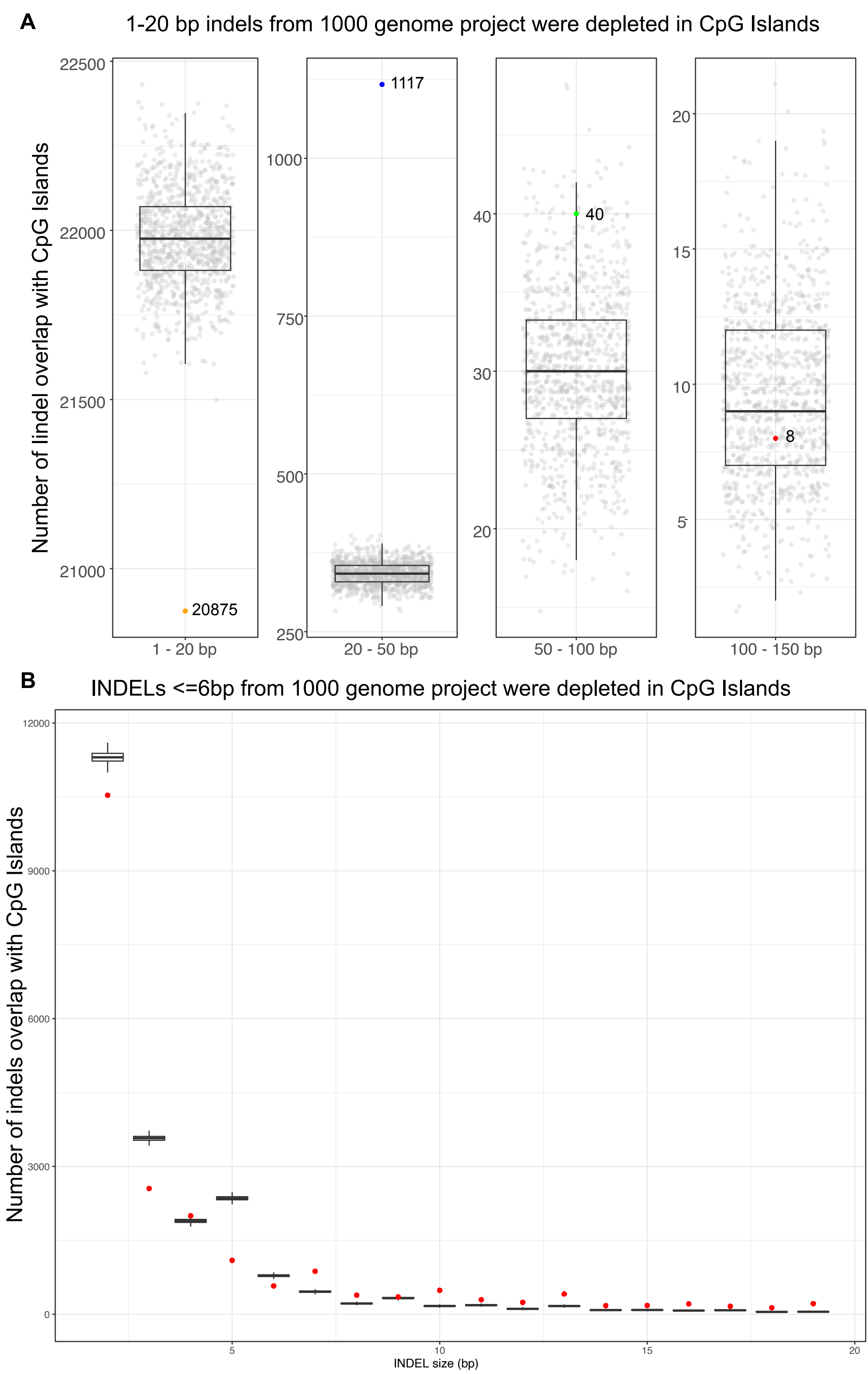
Supplemental Figure S6

Conservation of promoters and its relationship to indel rate. A. Aggregated phastCons score of annotated GENCODE promoters. The annotated GENCODE promoter is defined as 500bp region directly upstream of transcription start sites. B. Aggregated phastCons score of putative CNCC promoters. The putative CNCC promoter is defined by the H3K4me3 ChIP-seq peak. C. 20-50bp indel frequency of more conserved putative CNCC promoters (phastCons score > 0.2) and less conserved putative CNCC promoters (phastCons score < 0.2). D. 50-100bp indel frequency of more conserved putative CNCC promoters (phastCons score > 0.2) and less conserved putative CNCC promoters (phastCons score < 0.2).



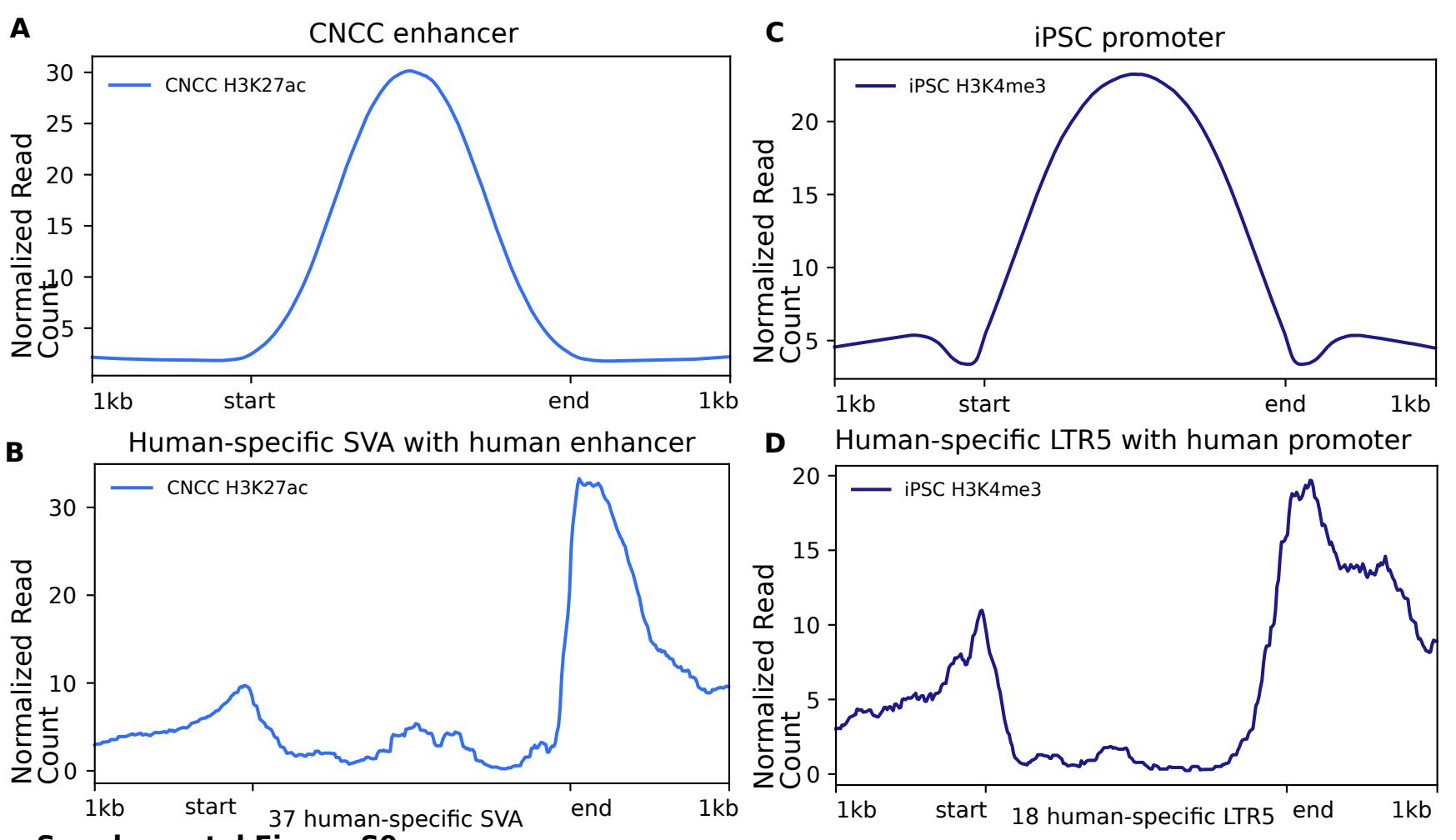
Supplemental Figure S7

Enrichment of indels of different sizes within putative CNCC promoters. Fisher's exact test enrichment values are represented on the y-axis, and the p-value is labeled at the top of each bar. A dashed horizontal line with enrichment =1 is also plotted for comparison.



Supplemental Figure S8

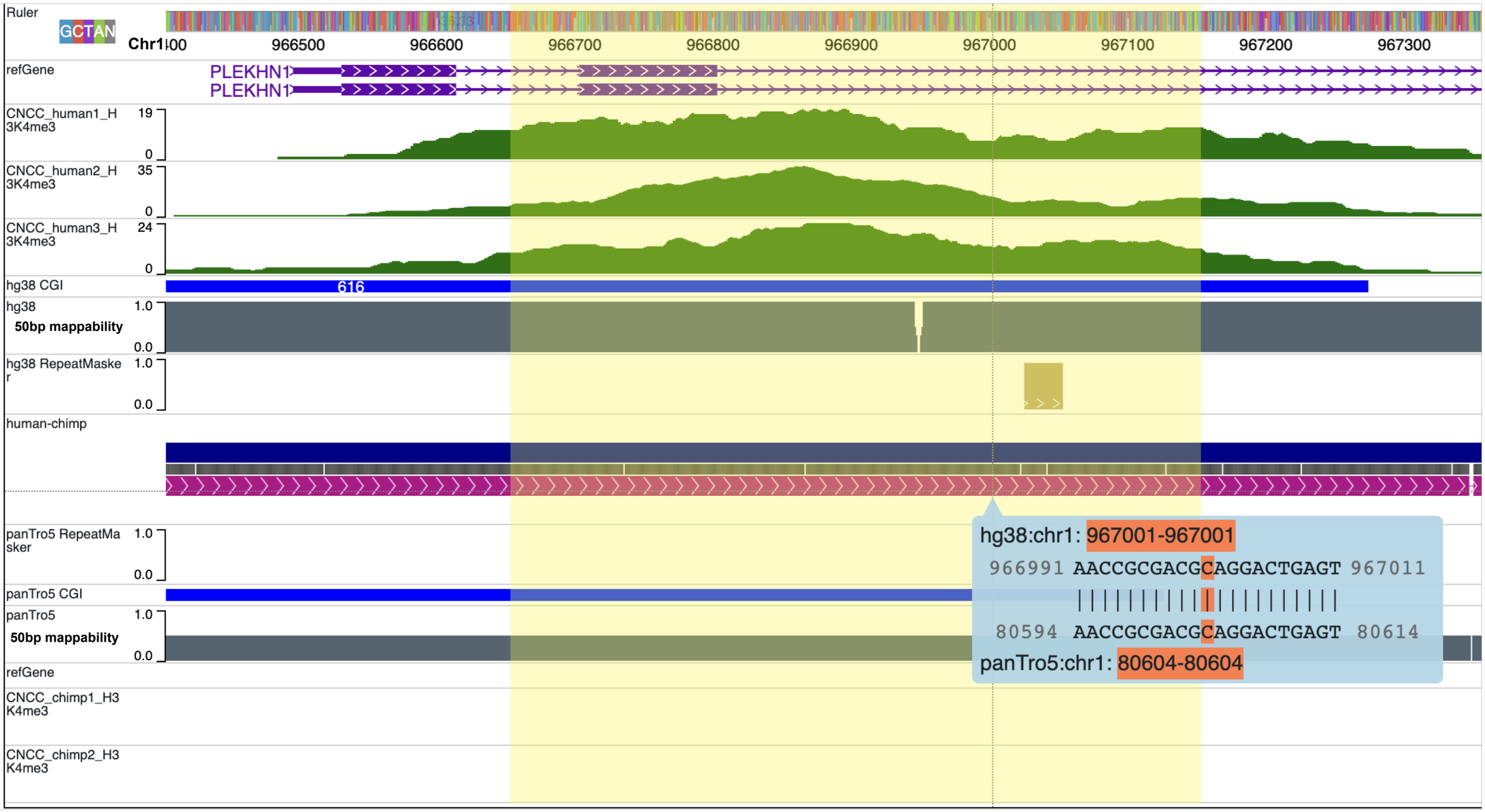
Comparing the intersection between CpG Islands and 1000 randomly shuffled indels called by 1000 Genomes Project with the observed intersection number. The 1000 times shuffled numbers were displayed as grey dots and box plot. The observed number is show as a red dot. A. The top panel shows the comparison separating indels to 1-20bp, 20-50bp, 50-100bp and 100-150bp. B. The lower panel shows the comparison of all indels from 1-20bp separated by 1bp window.



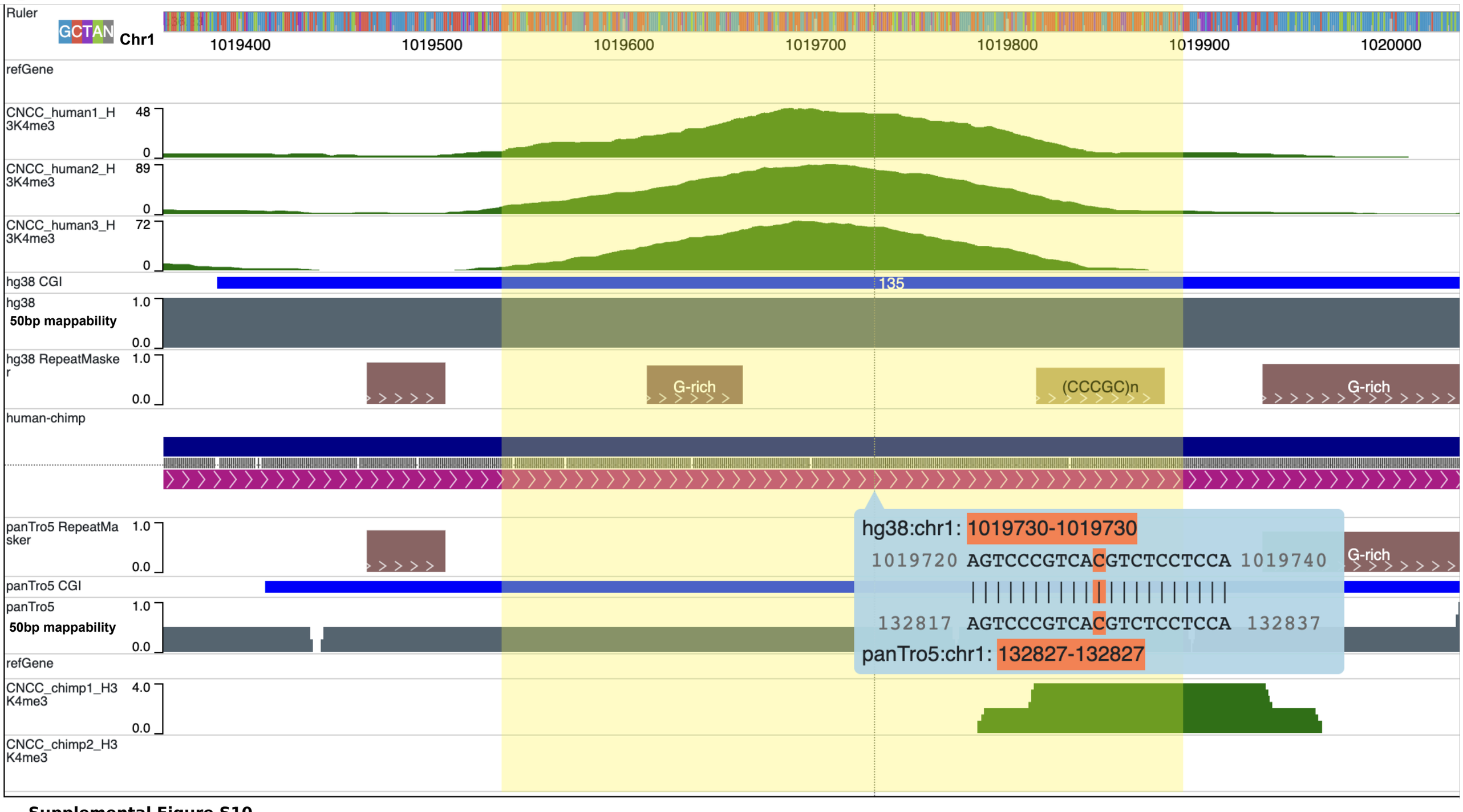
Supplemental Figure S9

Aggregated Chip-seq profile of different putative RRR. A. Aggregated CNCC H3K27ac Chip-seq profile of all putative CNCC enhancers. B. Aggregated CNCC H3K27ac Chip-seq profile of 37 human-specific SVA insertions that associated putative CNCC enhancers. C. Aggregated iPSC H3K4me3 Chip-seq profile of GENCODE annotated promoters. D. Aggregated iPSC H3K4me3 Chip-seq profile of 18 human-specific LTR5 insertions associated with iPSC H3K4me3 Chip-seq peaks.

A



B



Supplemental Figure S10
Mappability at syntenic regions can differ between human and chimpanzee reference genomes. Two example regions show H3K4me3 peaks only in the human genome. 50bp mappability in the two regions is high in the human genome while mappability is reduced in the chimpanzee genome. We filtered these types of regions out to avoid false positive lineage-biased regions.

Supplemental Table S1

Comparison between OrthoINDEL and liftOver. Columns from left to right: hg38 chromosome, hg38 start, hg38 end, OrthoINDEL results, liftOver results. Each line represents a region in hg38 genome, column 4 and 5 represents whether that region can be converted to panTro5 genome.

Supplemental Table S2

All shared and lineage-biased RRRs not associated with indels found in this study. Columns from left to right: hg38 start, hg38 end, hg38 strain, MACS2 score in hg38, IDR integer score in hg38, panTro5 chromosome, panTro5 start, panTro5 end, panTro5 strain, MACS2 score in panTro5, IDR integer score in panTro5, lineage, type.

Supplemental Table S3

All INDELs identified between the hg38 and panTro5 genomes in this study. Each column from left to right represents: hg38 chromosome, hg38 start, hg38 end, indel type defined by DASC, INDEL size, hg38-pantro5 alignment strain, panTro5 chromosome, panTro5 start, panTro5 end, INDEL lineage, TE subfamily, TE class, TE strain.

Supplemental Table S4

INDELs overlapping RRRs. Each column from left to right represents: hg38 chromosome, indel hg38 start, indel hg38 end, indel type defined by DASVC, indel size, hg38-pantro5 alignment strain, indel panTro5 chromosome, indel panTro5 start, indel panTro5 end, indel lineage, TE

subfamily, TE class, TE strain, CRE hg38 start, RRR hg38 end, RRR hg38 strain, RRR MACS2 score in hg38, RRR IDR integer score in hg38, RRR panTro5 chromosome, RRR panTro5 start, RRR panTro5 end, RRR panTro5 strain, RRR MACS2 score in panTro5, RRR IDR integer score in panTro5, CRE lineage, RRR type, RRR-indel overlapping size in hg38.

Supplemental Table S5

Number of indels overlapping regulatory/repressive regions. For each type of indels (separated to four different columns), how many of them can be found or cannot be found in different regions (12 rows including all putative invariant/lineage-biased promoter, enhancer, H3K9me3 regions. Indels outside of putative CNCC promoter/enhancer are defined as "not promoter/enhancer in CNCC", outside of iPSC H3K9me3 regions are defined as "not H3K9me3 region in iPSC", indels outside of as any regulatory/repressive region are called "Neither promoter/enhancer in CNCC or H3K9me3 region in iPSC").

No. of indels intersect with putative regulatory/repressive regions:	**human insertion**	**chimp insertion**	**human deletion**	**chimp deletion**
Do not intersect with promoter/enhancer in CNCC	51768	46630	12411	11289
invariant promoter CNCC	794	895	296	257
Human-biased promoter CNCC	0	0	0	0
Chimp-biased promoter CNCC	0	0	0	0
invariant enhancer CNCC	1059	934	266	231
Human-biased enhancer CNCC	116	57	19	29
Chimp-biased enhancer CNCC	100	123	49	27

No. of indels intersect with putative regulatory/repressive regions:	human insertion	chimp insertion	human deletion	chimp deletion
Do not intersect with H3K9me3 region in iPSC	49622	44968	11999	10879
invariant H3K9me3 region iPSC	3358	3041	926	814
Human-biased H3K9me3 region iPSC	755	164	46	114
Chimp-biased H3K9me3 region iPSC	103	466	70	29
Do not intersect with either promoter/enhancer in CNCC or H3K9me3 region in iPSC	47658	43032	11392	10351