Epigenomic differences in the human and chimpanzee genomes are associated with structural variation

Xiaoyu Zhuo, Alan Y Du, Erica C Pehrsson, et al.

Genome Res. published online December 10, 2020
Access the most recent version at doi:10.1101/gr.263491.120

Epigenomic differences in the human and chimpanzee genomes are associated with structural variation

Xiaoyu Zhuo[1,2], Alan Y. Du[1,2], Erica C. Pehrsson[1,2], Daofeng Li[1,2] and Ting Wang[1,2,3]

1 Washington University School of Medicine in St. Louis, Department of Genetics, Saint Louis, Missouri, USA

2 The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, MO, USA

3 McDonell Genome Institute, Washington University School of Medicine, St Louis, MO, USA

ABSTRACT

Structural variation (SV), including insertions and deletions (indels), is a primary mechanism of genome evolution. However, the mechanism by which SV contributes to epigenome evolution is poorly understood. In this study, we characterized the association between lineage-specific indels and epigenome differences between human and chimpanzee to investigate how SVs might have shaped the epigenetic landscape. By intersecting medium-to-large human-chimpanzee indels (20bp-50kb) with putative promoters and enhancers in cranial neural crest cells (CNCC) and repressed regions in induced pluripotent cells (iPSC), we found that 12% indels overlap putative regulatory and repressed regions (RRRs), and 15% of these indels are associated with lineage-biased RRRs. Indel-associated putative enhancer and repressive regions are ~1.3 and ~3 times as likely to be lineage-biased, respectively, as those not associated with indels. We found a 2-fold enrichment of medium-sized indels (20bp to 50bp) in CpG island (CGI)-containing promoters than expected by chance. Lastly, from human-specific transposable element insertions, we identified putative regulatory elements, including NR2F1-bound putative CNCC enhancers derived from SVAs and putative iPSC promoters derived from LTR5s. Our results demonstrate that different types of indels are associated with specific epigenomic diversity between human and chimpanzee.

INTRODUCTION

The question of what makes us uniquely human has long been of interest (Darwin 1871). Comparative genomics has sought the genetic basis of human-specific traits (King and Wilson 1975; The Chimpanzee Sequencing and Analysis Consortium 2005; Kronenberg et al. 2018; Rogers and Gibbs 2014; Wall 2013), including human-specific gene gain/loss or regions under accelerated evolution (Pollard et al. 2006; Zhu et al. 2007; Enard et al. 2002; Atkinson et al. 2018; Florio et al. 2018; Fiddes et al. 2018; Suzuki et al. 2018; Franchini and Pollard 2017). In addition, epigenetic and transcriptomic differences also contribute to human-specific phenotypes (Prescott et al. 2015; Gallego Romero et al. 2015; Trizzino et al. 2017; Ward et al. 2018; Danko et al. 2018; Pai et al. 2011; Hernando-Herraez et al. 2013; Eres et al. 2019). However, how structural variations (SVs) affect human-specific functions is just beginning to be explored (Gordon et al. 2016; Fudenberg and Pollard 2019).

SVs, which include deletions, duplications, inversions, insertions, and translocations, are responsible for the majority of genetic differences within populations and between species. The 1000 Genomes Project estimated that an individual carries a median of 8.9 Mb of SVs versus 3.6 Mb of single nucleotide variants (SNVs) (Sudmant et al. 2015). Long-read sequencing of two haploid human genomes revealed that the majority of SVs were novel, suggesting that the impact of SVs is underestimated (Huddleston et al. 2017). SVs also contribute to inter-species divergence. In 2005, the Chimpanzee Sequencing and Analysis Consortium reported ~90 Mb of insertions or deletions (indels) between human and chimpanzee; in contrast, SNVs constituted only ~35 Mb (The Chimpanzee Sequencing and Analysis Consortium 2005).

Non-coding cis-regulatory elements (CREs) play a critical role in gene regulation (The ENCODE Project Consortium 2004). One powerful method to identify putative functional elements is epigenomic profiling (Ernst and Kellis 2012; Roadmap Epigenomics Consortium et al. 2015). For example, H3K4me3 is usually associated with promoters, and H3K27ac is associated with both active promoters and active enhancers. In contrast, H3K9me3 is associated with heterochromatin, a repressed state characterized by densely packed DNA and low gene expression (Becker et al. 2015). By applying epigenomic profiling to related organisms ("comparative epigenomics"), we can compare epigenetic signature across species between syntenic regions and investigate the birth and death of regulatory elements during evolution (Xiao et al. 2012; Lowdon et al. 2016). Although studies have investigated enhancer evolution between human and chimpanzee (Trizzino et al. 2017; Gallego Romero et al. 2015; Ward et al. 2018; Prescott et al. 2015), the impact of SV on these elements has not been studied.

Here, we developed a novel computational strategy to define syntenic regions that contain indels. Using publicly available epigenomic datasets from human and chimpanzee, we defined the association between indels and epigenetic differences between the two species. We explored both epigenomic conservation and innovation in association with medium to large indels (20bp to 50kb), and how lineage-specific transposable element (TE) insertions contribute to new putative functional elements. Our findings indicate that SVs and epigenomic changes between human and chimpanzee are significantly interrelated.

RESULTS

Development of a novel method to find orthologous regions overlapping large indels

Conventional comparative genomic/epigenomic methods rely on tools such as UCSC liftOver to retrieve syntenic regions between species based on alignments between genome assemblies (Kuhn et al. 2013). However, these tools usually fail to return syntenic regions when the synteny is disrupted by medium to large SVs. To overcome this obstacle, we developed a new pipeline that combines CrossMap (Zhao et al. 2014), an alternative to liftOver, with our newly developed tool called OrthoINDEL (Methods). Instead of filtering syntenic regions using the minimum percentage of bases that can be converted to the new assembly, CrossMap outputs the syntenic region as multiple blocks split by alignment gaps. OrthoINDEL then concatenates the fragmented orthologs output by CrossMap if they are continuous or separated only by indels (Fig. 1). This way, our pipeline stringently converts regions from the source genome to their syntenic coordinates in the target genome even if they overlap large indels. In contrast, without sacrificing specificity, the UCSC liftOver dismisses regions with a large fraction absent in the target genome. To illustrate this improvement, we performed the same genomic coordinates conversion from human to chimpanzee using either OrthoINDEL or liftOver. We found OrthoINDEL successfully converted ~1000 more regions that were dismissed as "partially deleted" or "split in new" by liftOver (Supplemental_Fig_S1, Supplemental_Table_S1). Thus, OrthoINDEL is better suited to retrieve syntenic regions overlapping indels.

Indel-associated enhancers and H3K9me3 regions are more likely to be lineage-biased

By analyzing two human-chimpanzee comparison ChIP-seq datasets (Prescott et al. 2015, Ward et al. 2018), we identified ~15,000 putative promoters (H3K4me3 ChIP-seq peaks), ~27,000 putative enhancers (H3K27ac peaks outside of H3K4me3 peaks) and ~31,000 H3K9me3 regions (H3K9me3 broad peaks) in each species that constitute ~40Mb, 67Mb and 300Mb, respectively (Fig. 2A). Together, we define them as putative regulatory and repressed regions (RRRs). We found support for ~88% of putative CNCC promoters by overlapping them with annotated FANTOM5 or GENCODE promoters (~84% and ~85% in GENCODE and FANTOM5, respectively) (Frankish et al. 2019; The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014).

Next, we applied the OrthoINDEL pipeline to establish the syntenic relationship between human and chimpanzee RRRs. We further classified regions as human-biased, chimpanzee-biased, or invariant based on both peak-calling and ChIP-seq reads difference defined by DESeq2 (Methods; Supplemental_Table_S2) (Landt et al. 2012; Love et al. 2014). A datahub displaying the processed data over human and chimpanzee syntenic regions is available at the WashU Epigenome Browser (Li et al. 2019) (https://epigenomegateway.wustl.edu/browser/?sessionFile=https://wangftp.wustl.edu/~xzhuo/CNCC/publicationSession.json).

We annotated all indels larger than 20bp between the human (hg38) and chimpanzee (panTro5) reference genomes using the DASVC pipeline (Methods) (Gordon et al. 2016). In total, we defined 193,180 medium-to-large indels (20bp to 50kbp) encompassing 95.8Mb (~3%

of the human haploid genome). We selected ~127,000 of them (42.2Mb) with a defined ancestral state (using gorilla genome as an outgroup) and located within nonrepetitive regions for epigenomic analysis (Methods) (Supplemental_Fig_S3). We also annotated TE-derived insertions within these indels (Methods) (Fig. 2B, Supplemental_Table_S3). The overall number and length of our indels were in excellent agreement with previously published results (The Chimpanzee Sequencing and Analysis Consortium 2005; Kronenberg et al. 2018) (Supplemental_Fig_S4).

Next, we characterized the association between indels and putative RRRs. We identified ~15,000 indel-overlapping putative RRRs by intersecting their coordinates using BEDTools (Quinlan and Hall 2010) (Supplemental_Table_S4). Indels are slightly depleted in putative RRRs instead of being uniformly distributed in the genome (Fisher's exact test enrichment ratio 0.94, P-value 2.4×10^{-9}). We found that ~88% (112,433/127,350) of indels do not overlap any putative RRRs in the study (Supplemental_Table_S5), 10% of indels overlap with invariant elements, while the remaining 1.8% (2267 indels) are associated with lineage-biased elements (Fig. 2C). An association between a human-specific indel and a human-biased putative RRR (135 with enhancer and 801 with H3K9me3 heterochromatin) could suggest that the indel may have created the RRR in the human lineage. On the other hand, association between a human-specific indel and a chimpanzee-biased element (149 with enhancer and 173 with H3K9me3 heterochromatin) could suggest that the indel may have disrupted an ancestral RRR in the human lineage. The same logic applies to chimpanzee lineage indels (Fig. 3, Supplemental_Table_S5).

In accordance with previously reported findings, we found that all except five putative promoters are invariant between the two species (Fig. 2C). In contrast, ~85% of putative enhancer and repressed regions are invariant (Prescott et al. 2015; Ward et al. 2018) (Fig. 2C). Compared with those not associated with indels, putative enhancers associated with indels are ~30% more likely to be lineage-biased (Fisher's exact test P-value 4.7×10^{-6}), and H3K9me3 regions associated with indels are about three times as likely to be lineage-biased (Fisher's exact test P-value 10^{-760}) (Fig. 2C). This result suggests that indels have a moderate association with putative enhancers and a strong association with H3K9me3 regions.

The enrichment of different indels with putative RRRs

We sought to understand if different size categories of indels had different association with putative RRRs. We divided non-TE-derived indels to four groups (20-50bp, 50-500bp, 500-5kb, 5k-50kb) and separated TE-derived insertions by TE class. We defined indels from 20 to 50bp as medium-sized indels, and indels >=50bp as large indels following conventions (The 1000 Genomes Project Consortium 2015; Kronenberg et al. 2018). We calculated the enrichment of the intersection between these indel categories with putative RRRs over the genomic background using Fisher's exact test (Fig. 4). We plotted the enrichment ratio and p-value of all pairs with at least one intersection (Fig. 4).

We noticed three main trends. First, medium-sized indels (20-50bp) are enriched in invariant putative CNCC promoters. In contrast, indels larger than 500bp are depleted in invariant putative promoters. Second, we found indels longer than 5kb are depleted in invariant putative CNCC enhancers. Third, in H3K9me3 repressed regions, lineage-specific sequences longer than

500bp (insertions > 500bp in that lineage and deletions > 500bp in the other lineage) are enriched for H3K9me3 regions from the same lineage (Fig. 4). As an example, we found that human lineage insertions and chimpanzee lineage deletions > 500bp are enriched in human-biased H3K9me3 regions. None of the lineage-biased putative CNCC promoters overlap indel (Fig. 4), and the enrichment/depletion of indels with putative invariant promoters is almost identical to their enrichment/depletion with all putative CNCC promoters.

We separated TE-derived insertions by TE class and performed the same enrichment analysis (Fig. 4). Lineage-specific SVA insertions are significantly enriched in both putative lineage-biased CNCC enhancers and iPSC H3K9me3 repressed regions for both species, which implies that newly inserted SVA elements may have provided CNCC-specific enhancers and been targeted by the repressive marks in both species (Fig. 4). ERV insertions are only enriched as lineage-biased H3K9me3 repressed regions in the chimpanzee lineage, contributed mainly by the H3K9me3-modified chimpanzee-specific PTERV regions (Supplemental_Table_S4) (Yohn et al. 2005). Previous studies in fly and yeast suggest that repressive marks can spread beyond the heterochromatin boundary and affect nearby genes (Elgin and Reuter 2013; Obersriebnig et al. 2016; Greenstein et al. 2018). However, we found here that H3K9me3 marks rarely expand beyond 2kb outside newly inserted TE boundaries (Supplemental_Fig_S5). In general, out of the SVs associated with biased chromatin marks, insertions tend to be associated with creation, rather than destruction, of putative RRRs.

We further explore the enrichment of indels from 20-50bp with putative CNCC promoters and the enrichment of SVA insertions with putative CNCC enhancers in the following sections.

Enrichment of medium-sized indels within CpG islands

Promoters are considered conserved elements due to their low nucleotide substitution rate (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014). We also found both putative CNCC promoters and GENCODE-annotated promoters are conserved using phastCons score (Supplemental_Fig_S6A, B). In contrast, using Fisher's exact test, we found that medium-sized (20-50bp) indels are highly enriched within putative CNCC promoters, whereas indels >500bp are depleted within putative CNCC promoters (Fig. 4). To characterize the relationship between indel size and their enrichment in promoters at a finer resolution, we separated indels <500bp by size at 50bp intervals and tested their enrichment with putative CNCC promoters using Fisher's exact test. We found that 50-100bp indels are barely enriched, and all indel bins >100bp are not statistically enriched (Supplemental_Fig_S7). To validate our observation, we further calculated indel frequency around annotated genes for 20-50bp and 50-100bp indels. Again, we found that 20-50bp indels, but not 50-100bp indels, have elevated frequency immediately upstream of transcription start sites (Fig. 5A). We also observed slightly lower indel frequency in more conserved promoters (phastCons >0.2) than in less conserved promoters (phastCons <0.2), indicating that conserved promoters have fewer indels (Supplemental_Fig_S6C, D). To test whether the enrichment of 20-50bp indels in promoters also exists in the human population, we extracted variants from the 1000 Genomes Project Phase 3 release (The 1000 Genomes Project Consortium 2015; Sudmant et al. 2015) and repeated the analysis. Consistent with our inter-species observation, only medium-sized indels (20-50 bp) within the human population are enriched in the promoters (Fig. 5B).

Seventy percent of human gene promoters contain CpG islands (CGIs) (Saxonov et al. 2006). Many indels in promoters are located within UCSC annotated CGIs, suggesting that indels within CGIs could be driving the observed high indel rate in promoters (Kuhn et al. 2013). We separated promoters into CGI and non-CGI promoters, and found that 20-50bp indels are enriched only in CGI promoters (enrichment ratio 2.60, P-value 10^{-762}) but not in non-CGI promoters (enrichment ratio 1.06, P-value 0.02).

Since indels > 150bp are not enriched within putative CNCC promoters (Supplemental_Fig_S7), we analyzed the enrichment of 20-50bp, 50-100bp, and 100-150bp indels with promoters and CGIs. To verify the Fisher's exact test results (Fig. 4 and Supplemental_Fig_S7), we directly compared observed number of intersections with the expected distribution based on random sampling (Fig. 5C) (Methods). CGIs outside promoters are enriched for all three sizes of indels ($P < 0.001$, permutation test), and the enrichment of indels in CGIs within promoters decreases as a function of indel size from 20bp to 150bp. In contrast, promoters without CGIs are barely enriched for any indel category (Fig. 5C). Thus, consistent with previous findings, our data suggests that CGIs are hot spots for indels in evolution (Tian et al. 2011; Kiktev et al. 2018).

Indels from 100bp to 150bp are less enriched in CGIs within promoters than the smaller 20-50bp and 50-100bp indels, potentially because they may drastically influence regulatory activity and have thus been selected against during evolution. Medium-sized indels (20-50bp), on the other hand, seem to have been tolerated, as promoters with CGIs are highly epigenetically conserved between human and chimpanzee (Fig. 4).

Lastly, we characterized the enrichment of indels in CGIs within the human population using the variants from 1000 Genomes Project Phase 3 release and found the same pattern (Supplemental_Fig_S8) (The 1000 Genomes Project Consortium 2015). We also characterized the CGI enrichment of small indels (< 20bp). Small indels (1-20bp), especially indels with length <= 6bp, are depleted in CGIs. Since 98% of 1000 Genome Project indels are smaller than 20bp, our conclusion is consistent with the previous observation that indels are depleted within CGIs (Neininger et al. 2019), but highlights the novel finding of enrichment of medium sized indels in CGIs.

Lineage-specific TEs give rise to putative promoters and enhancers and gradually become repressed during evolution

Lineage-specific SVA insertions are highly enriched in both putative enhancers and repressed regions (Fig. 4). For the 37 SVA elements that overlap with human-biased putative enhancer regions, the aggregated H3K27ac ChIP-seq profile shows areas of elevated signal similar to that in putative CNCC enhancers (Supplemental_Fig_S9A, B). To better illustrate their enrichment pattern, we plotted aggregated ChIP-seq profiles around all human-specific SVAs and compared them to the profiles of their orthologous pre-insertion sites in chimpanzee (Fig. 6A). In addition to the previously described ChIP-seq data, we also included iPSC H3K27ac from both species and human iPSC H3K4me3 (The ENCODE Project Consortium 2004; Gallego Romero et al. 2015). We also analyzed human-specific LTR5 insertions in a similar fashion (Supplemental_Fig_S9C, D).

We used the mappability score as a measurement of repetitiveness and the propensity of a genomic region to produce uniquely mappable reads (Derrien et al. 2012). Lineage-specific SVA and LTR5 have low mappability scores (Fig. 6A). Therefore, their corresponding ChIP-seq signals were likely underestimated. Indeed, with the exception of H3K9me3 ChIP-seq, which was sequenced using 100bp paired-end reads, all other ChIP-seq datasets generated using single-end reads have close to zero signal over the low mappability regions. Nevertheless, we found a strong H3K27ac signature suggesting putative enhancer activity on the 3' flanking region of SVA in CNCC and a strong promoter signature (H3K4me3 and H3K27ac) on both flanking regions of LTR5 in iPSCs in human but not in chimpanzee. Conversely, we found H3K4me3 ChIP-seq signal 3' of LTR5 insertions in CNCC, but these were found in both human and chimpanzee, suggesting that in this case the epigenomic mark may be independent of the TE insertion (Fig. 6A).

To understand how epigenetic profiles might evolve over time, we extended our analysis to related TE insertions that are shared by both species. SVA proliferated and diverged in the human genome in a similar fashion as the amplification of L1 (Khan et al. 2006, Hancks and Kazazian 2016). SVAs in the human genome were classified into six subfamilies (SVA-A to SVA-F). The expansion of subfamilies from SVA-A to SVA-D predates the human-chimpanzee split, and SVA-E and SVA-F expanded after human-chimpanzee divergence (Wang et al. 2005). We defined all human-specific SVAs as SVA-human and classified human-chimpanzee shared SVAs based on their subfamilies (from SVA-A to SVA-D). We plotted H3K27ac, H3K9me3 and mappability profiles of different SVA subfamilies in Fig. 6B. We found that the 3' boundary H3K27ac signal in CNCC decreases as SVA subfamily ages, with the greatest signal in human-

specific subfamilies. In contrast, H3K9me3 signal intensifies with increasing SVA age in iPSC (Fig. 6B).

Next, we similarly classified LTR5s based on lineage-specificity and subfamilies (from oldest to youngest: LTR5A-shared, LTR5B-shared, and LTR5Hs-shared to LTR5-human) and performed the same analysis using H3K4me3 data in iPSC. Similar to SVAs, the strength of the LTR5 promoter signature in iPSC is negatively correlated with age. However, LTR5s are not marked by H3K9me3 in iPSC (Fig. 6C).

NR2F1 binding is correlated with enhancer signature on the 3' end of SVA in CNCC

Our data thus far predicted a putative enhancer within SVAs close to their 3' end. The H3K27ac signal is likely “hidden” due to low mappability, but part of it extends into mappable regions, as we observed in Fig. 6. However, we cannot observe the boundary ChIP-seq signal at the 5' end of a full-length (1.6kb) SVA, which might be too long for the enhancer signal to extend beyond. To determine the precise location of the putative enhancer, we extracted all SVAs with complete 3' ends in the human genome, sorted them based on size, anchored them at the 3' end, and annotated them with CNCC H3K27ac signal (Fig. 7A). Again, we observed the elevated H3K27ac signal on the 3' end of SVAs (Fig. 7A). However, once the SVA length was reduced to 300-500 bp, a similar boundary H3K27ac signal on the 5' end emerged. This result is consistent with the hypothesis that the enhancer signature originating within SVAs can extend beyond the low mappability region. The boundary H3K27ac signal disappeared on both ends of shorter SVA copies, suggesting that further truncation resulted in a loss of the internal enhancer (Fig. 7A, B).

Zooming in onto the 5' end of these shorter SVA fragments revealed a strong NR2F1 binding motif (Fig. 7B, C). Importantly, the disappearance of the boundary H3K27ac signal correlated with the truncation of the NR2F1 motif (Fisher's exact test P-value 7×10^{-5}) (Fig. 7C). NR2F1 is a critical regulator in CNCC (Rada-Iglesias et al. 2012). Indeed, the NR2F1 and H3K27ac ChIP-seq signals co-occur in SVAs in CNCC (Fig. 7B) (Prescott et al. 2015). However, not every putative human-biased CNCC SVA enhancer has a related NR2F1 peak (Fig. 7D, E), suggesting the possible involvement of other transcription factors in these SVA-derived putative enhancers.

DISCUSSION

In this study, we systematically characterized how medium to large indels are correlated with differences in the epigenome in the human and chimpanzee lineages. We found that indels are enriched in putative lineage-biased enhancers and H3K9me3 repressed regions. We should note that all genomic enrichments were estimated using a whole genome random distribution as the background. However, genomic features are not randomly distributed and the assumption may not always be appropriate. Yokoyama et al. described a substitution-based framework to model birth-death of lineage-specific functional elements (Yokoyama et al. 2014). We demonstrated here that in addition to substitutions, indels can also contribute to the birth-death of regulatory and repressed elements. Our strategy is readily applicable to other comparative epigenomic datasets, and we have made our processed data available in our comparative browser (Li et al. 2019).

Mutation rate varies depending on genomic region. It has been reported that the substitution rate in closed chromatin regions is higher than the rate in open chromatin regions (Makova and Hardison 2015; Fortin and Hansen 2015). Lunter et al. found the highest indel rates in regions with extremely high and extremely low GC content (Lunter et al. 2006). Using macaque as an outgroup, Kvikstad et al. reported a curvilinear relationship between human lineage indel rate and GC content, and they also found weak anti-correlation between insertion rate and number of CGIs at 1Mb genomic windows (Kvikstad et al. 2007). Specific to CGIs, the substitution rate at CpG sites is higher than the mutation rate at other sites because of the high deamination rate of 5-methyl cytosine (Coulondre et al. 1978). Cohen et al. found that the lack of methylation of CGIs can explain their maintenance without implying purifying selection in primates (Cohen et al. 2011). However, CGIs are often associated with genome instability (Deaton and Bird 2011; Du et al. 2014).

We found that CGIs are hotspots for medium/large indels in hominids. Kiktev et al. demonstrated that the high-GC region in yeast has a high deletion/duplication rate resulting from DNA polymerase slippage (Kiktev et al. 2018). Thus, analogous to well-characterized SVs in human coding exons (Montgomery et al. 2013; Challis et al. 2015), high-GC regions are prone to forming a single stranded DNA loop due to polymerase slippage during DNA replication (Tian et al. 2011), which likely results in the increased indel rate. However, GC rich regions are prone to sequencing error and we cannot completely rule out the possibility that some indels we called were caused by the difficulty of sequencing and calling variants in these GC rich regions.

By comparing sequence conservation with epi-conservation, Xiao et al. reported elevated epi-conservation of H3K27ac, H3K27me3, and methylated CpGs (but not H3K4me3) in rapidly evolving sequences and proposed that epi-conservation could buffer some deleterious mutations (Xiao et al. 2012). Here, we report the conservation of H3K4me3 marks between human and chimpanzee despite an elevated rate of medium-sized indels, further supporting the concept of epi-conservation and its potential role in buffering the impact of mutations.

Britten and Davidson proposed the gene battery model in the 1970s to explain the evolution of regulatory networks (Britten and Davidson 1971). They proposed that a single "activator gene" can control a "battery of genes" by interacting with diffused repetitive sequences throughout the genome. Since then, TEs have been repeatedly demonstrated to contribute novel regulatory elements (Feschotte 2008; Wang et al. 2007; Lynch et al. 2011; Chuong et al. 2017). However, in primates, especially in the human lineage, there have been conflicting reports about the regulatory role of TEs. Trizzino et al. found specific TE subfamilies enriched in liver enhancers (Trizzino et al. 2017). On the other hand, Ward et al. could not find significant contribution of TEs to gene regulation in pluripotent stem cells (Ward et al. 2018). We report clear signals of TE-derived, tissue-specific putative enhancers and promoters unique to human or chimpanzee. We identified LTR5 as putative promoters in iPSC, while Fuentes et al. found them to have enhancer activity in human embryonal carcinoma NCCIT cells (Fuentes et al. 2018). We could not find a large impact on nearby gene expression associated with these TE-derived enhancers with our limited dataset. It is possible that these new enhancers do not regulate the closest genes. Alternatively, they may provide functional redundancy instead of inventing new regulation (Osterwalder et al. 2018; Choudhary et al. 2020). We also report that

346 TE-associated heterochromatin displays limited spreading (Supplemental_Fig_S5). We found a
347 rapid conversion of TE epigenetic modification from active to repressive states as a function of
348 age in young TEs. This discovery echoes a previous report of the transition of repressive marks
349 from cytosine methylation to H3K9me3 as ERVs age in the human genome (Ohtani et al. 2018).
350 These data suggest that although many TEs carry regulatory elements, the host rapidly and
351 continuously silences such activity during evolution.

352 Most genomic analyses rely on second-generation sequencing, which produces short reads,
353 restricting our ability to detect signals from repetitive, low mappability regions. To overcome
354 this limitation, we investigated not only the epigenetic signal from within TEs but also from
355 flanking regions. By comparing TE insertions with orthologous pre-insertion sites in another
356 species, we can infer that the epigenetic signal originates from these highly repetitive regions.
357 Our approach expands the application of second-generation sequencing and reveals that there
358 are more potentially functional elements hidden in unmapped territories. However, the
359 sensitivity of our method is limited by the distance of the element to the boundary, read length,
360 DNA fragment size and other factors.

361 Only four different TEs have been actively transposing in the human lineage. Of the four, we
362 found that two are associated with putative enhancers and promoters using data from only two
363 cell types. Although most TEs are neutrally evolving in the genome, our discovery suggested
364 that many CREs carried by TEs were active upon insertion. Our finding begs more thorough
365 investigation of CREs derived from recently inserted TEs in more cell types and between
366 different species.

METHODS

Indel identification

We applied the DASVC tool to annotate indels between human and chimpanzee (Gordon et al. 2016). The DASVC tool was downloaded from (https://github.com/zeeev/DASVC) and was used with default parameters to identify 20bp to 50kb indels between the hg38 and panTro5 genomes. We further processed DASVC output using a Python script (refine_calledSV.py) to remove segmental replacements and extract the exact coordinates in both species. To identify indels that occurred in mappable regions, we calculated the average 75bp-mappability score of the flanking 200bp of all indels in both species, and selected those indels with a mappability score > 0.7 in both flanking regions in both species (Derrien et al. 2012).

To differentiate deletion in one lineage from insertion in the other, we identified orthologous coordinates for all indels in the gorilla reference genome (gorGor5). We considered an indel to be a deletion if the indel region is present in the gorilla genome, and an insertion if the region is absent in the gorilla.

A new chimpanzee reference genome panTro6 was published shortly after we started this project (Kronenberg et al. 2018). PanTro6 closed 52% of remaining gaps, but it does not refute the high-quality panTro5 reference genome in assembled regions. Therefore, the conclusions we reach here using panTro5 should remain valid.

TE-derived insertion annotation

To annotate TE-derived insertions within the identified indels, we intersected the indel list with RepeatMasker annotations for both the hg38 and panTro5 genomes using BEDTools (Smit et al.; Quinlan and Hall 2010). To avoid calling fragmented TEs as separate TE insertion events, we defined an indel as a TE-derived insertion if it was derived from a single TE insertion event. Indels containing only an *Alu* element, a full-length endogenous retrovirus (ERV), or a solo long terminal repeat (LTR) were counted as *Alu*/ERV insertions. Solitary LTRs were included because they are derived from an ERV insertion followed by non-homologous recombination (Mager and Stoye 2015). Due to the prevalence of 5' end truncation during target-primed reverse transcription (TPRT) (Luan et al. 1993), we also tolerated incomplete 5' ends in defining L1 and SVA insertions. One limitation to our rigorous approach is that we did not annotate lineage-specific solo-LTRs, where part of the solo-LTR aligned with the 5' end LTR of the full-length ERV and part aligned with the 3'end, as TE-derived insertions. In addition to the known actively transposing TE subfamilies described above, we also identified a few lineage-specific LTR12C elements (Supplemental_Table_S3).

Since TE insertions are homoplasy-free and unidirectional (Bashir et al. 2005; Ray et al. 2006), the orthologous regions corresponding to most human-/chimpanzee-specific TE insertions should be found as pre-insertion sites in the gorilla genome. As expected, the orthologous locations of 98% (15,803 of 16,068) of lineage-specific TE insertions are pre-insertion sites in the gorilla genome (Supplemental_Fig_S4B), whereas the remaining 2% were present as TE

insertions. These cases could be explained by incomplete lineage sorting, as demonstrated before (Kronenberg et al. 2018; Ray et al. 2006).

Peak calling and cross-species comparison

We downloaded both CNCC and iPSC raw read FASTQ files from NCBI GEO repository, accessions GSE70751, GSE61343 and GSE96712 (Prescott et al. 2015; Gallego Romero et al. 2015; Ward et al. 2018). Human H3K27ac iPSC ChIP-seq data and the associated input BAM files mapped to hg38 were downloaded from the ENCODE portal experiment ENCSR729ENO. We called ChIP-seq peaks using ENCODE recommendations with MACS2 and IDR thresholding (Supplemental Methods) (Li et al. 2011; Landt et al. 2012; Li 2013).

We applied CrossMap to identify orthologous segments for each peak in the other species (Zhao et al. 2014). For each peak region, CrossMap outputs all fragmented orthologous loci separated by any indel >1 bp. We developed a new tool, OrthoINDEL, that processes fragmented syntenic regions from the CrossMap output file. OrthoINDEL uses two parameters to filter fragmented regions. A maximum distance of 50kb was used to filter out fragments with too large a separation. We used a minimum distance of 50bp to define continuous fragments. To be defined as an indel, the split fragments we required to be continuous in one species. Since we focus on indels, our pipeline removes other SVs including inversions. Lastly, we filtered out regions with average 50bp-mappability < 0.7 in either species (compared with indel identification, we used more stringent mappability criteria here to eliminate false positive lineage-biased RRRs) (Derrien et al. 2012). The last filtering step is critical to filter out false positive lineage-biased regions (Supplemental_Fig_S10). With this pipeline, we defined

stringent 1:1 syntenic regions between human and chimpanzee tolerating indels up to 50kb that can be converted reciprocally using OrthoINDEL.

Peak calling is sensitive to sequencing coverage and background signal. To better differentiate invariant peaks from lineage-biased peaks, we counted the number of reads mapped to each peak in both species and applied DESeq2 to quantify peak intensity difference (Love et al. 2014). We classified regions as "human-biased" if a peak was called only in the human genome by MACS2 and the number of ChIP-seq reads in the human peak is significantly higher than the number in its orthologous region in the chimpanzee genome (DESeq2, $q<0.0001$); "chimpanzee-biased" regions were defined in a similar fashion. Lastly, we classified regions as "invariant" if ChIP-seq peaks were called in both species or if a peak was called in only one species, but the difference of ChIP-seq reads number between syntenic regions is not significant.

Enrichment analysis

We calculated the number of indel-RRR overlaps using the BEDTools intersect function (Quinlan and Hall 2010). To perform Fisher's exact test with the hg38 genome as background, we used the BEDTools fisher function. The permutation test used BEDTools to shuffle indel coordinates and intersect with CGIs and promoters (Supplemental Methods). Briefly, we counted the number of intersections of human-chimpanzee indels ranging from 20-50bp, 50-100bp and 100-150bp with CGIs within promoters, CGIs outside of promoters and promoters without CGIs, respectively. We then randomly shuffled indel coordinates 1000 times and repeated the intersection.

Identification of transcription factor binding motifs

We used FIMO from the MEME suite to find potential transcription factor binding sites within SVA elements (Bailey et al. 2009; Grant et al. 2011).

Data visualization

We generated bigWig files using methylQA and displayed them on the WashU Epigenome Browser (Li et al. 2015). All data are visualized on the WashU Epigenome Browser (Li et al. 2019). All ChIP-seq data were normalized to reads per genomic content (RPGC) using deepTools bamcoverage (Ramirez et al. 2016). Binding profiles and heatmaps were generated using deepTools2 (Supplemental Methods) (Ramirez et al. 2016).

DATA ACCESS

All processed data are accessible on the WashU Comparative Epigenome browser: (https://epigenomegateway.wustl.edu/browser/?sessionFile=https://wangftp.wustl.edu/~xzhuo/CNCC/publicationSession.json). Our pipeline and scripts generated in this study are available on GitHub (https://github.com/xzhuo/indel_epi_landscape) and as Supplemental Code.

ACKNOWLEDGMENTS

We thank all members of the Wang lab for their helpful suggestions; Dr. Zev Kronenberg from PacBio for his help with the DASVC pipeline and sharing with us their apes indel variant-calling result; and Silas Hsu from UIUC for building the updated WashU Epigenome Browser. Finally,

we want to thank three reviewers for their constructive criticisms. This manuscript is dedicated to the memory of Yimin Yang (1960--2018).

X.Z. is supported in part by 5R25DA027995.

A.Y.D. is supported by a grant from NHGRI (no. T32 HG000045).

E.C.P. is supported by a Postdoctoral Fellowship, PF-17-201-01, from the American Cancer Society.

T.W. is supported by NIH grants R01HG007175, U24ES026699, U01CA200060, U01HG009391, and U41HG010972, and by the American Cancer Society Research Scholar grant RSG-14-049-01-DMC.

AUTHOR CONTRIBUTIONS

X.Z. and T.W. conceived and implemented the study. X.Z. performed the analysis and wrote the paper; A.Y.D. contributed to data analysis; D.L. contributed to data visualization and interpretation; E.C.P. and A.Y.D. edited the paper; and T.W. supervised the study. All authors read and approved the final manuscript.

DISCLOSURE DECLARATION

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1: Comparison of UCSC liftOver with OrthoINDEL. Briefly, CrossMap splits syntenic regions into fragments separated by any gap in the alignment. OrthoINDEL concatenates the fragments split by indels and returns syntenic regions containing these indels. UCSC liftOver does not convert the third example (yellow), where a large portion from the species1 region is absent in species2. OrthoINDEL enables us to retrieve syntenic regions with large indels and filter out other SVs such as inversions. Rectangles represent one-to-one alignments from species1 to species2. Diamonds represents insertions in species1. Triangles represents an inverted region between the two species.

Figure 2: All putative RRRs and indels between human and chimpanzee and the number of overlaps between them. A. Length distribution of all putative RRRs and their orthologs. Violin plots show the length of putative promoter, enhancer and H3K9me3 repressed regions. For regions called in only one species, the length of their syntenic regions are plotted in the other species. Size in the human genome is shown on the left; size in the chimpanzee on the right. The average length of each distribution is marked and labeled. B. Size distribution of indels 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 a TE insertion). C. The number of lineage-biased/invariant putative CNCC promoters, CNCC enhancers, and iPSC H3K9me3 heterochromatin regions with or without indel association. Regions were separated into those without an indel or with one of the four types of

indels. Colors distinguish putative RRR invariant between the two species or biased in either lineage. The percentage of each category is displayed in a stacked histogram with the number of occurrences labeled.

Figure 3: Examples of indels associated with different CREs on the WashU Epigenome Browser. Human-chimpanzee track represents pairwise alignment between the human (blue) and chimpanzee (pink) genomes. A. A human-specific insertion associated with a chimpanzee-biased enhancer. B. A chimpanzee-specific deletion associated with a human-biased enhancer. C. A chimpanzee-specific insertion associated with a human-biased H3K9me3 heterochromatin region. D. A human-specific deletion associated with a chimpanzee-biased H3K9me3 heterochromatin region.

Figure 4: Enrichment of indel categories with putative RRR categories. Each dot in the matrix represents the enrichment of one type of indel within a specific putative RRR. Indels were first separated into human insertions, human deletions, chimpanzee insertions, or chimpanzee deletions and then further subdivided by size or TE classification. In addition to lineage-specific HERVK(HML2), chimpanzee-specific ERV insertions also include PTERV insertions absent from human genome. Putative CNCC promoter, CNCC enhancer, and iPSC H3K9me3 heterochromatin regions are presented from left to right. Each putative RRR is further classified horizontally into human-chimpanzee invariant, human-biased, and chimpanzee-biased regions. Human-biased and chimpanzee-biased putative CNCC promoters are not shown in the figure because they do not intersect with any indel. The enrichment P-value (BEDTools Fisher's exact test) is displayed

as the background greyscale in the matrix, and the enrichment ratio is displayed using both the color and size of each dot. Combinations with no intersections are crossed out.

Figure 5: Enrichment of 20-50bp indels in promoters/CGI. A. Frequency of human-chimpanzee indels of sizes 20-50bp and 50-100bp in and around annotated human genes. Metaplots display genes with introns removed and 5 kb flanking regions surrounding the transcription start site and end site. B. Frequency of human population indels found in the 1000 Genomes Project around the same promoter regions described in Fig. 5A. C. Comparing the observed number of indels intersecting with CpG island within promoters, CpG island outside of promoters and promoters without CpG island with the same intersections between randomly shuffled indels and the three CpG island or promoter regions. All indels were shuffled 1000 times. The density distributions of the numbers of shuffled indel intersections are illustrated by black lines. The observed numbers are indicated with a red vertical line with the observation/expectation ratio (O/E) and P value at the top of each graph.

Figure 6: ChIP-seq read count, normalized to reads per genomic content, surrounding repetitive TE insertions reveals putative hidden CREs. A. ChIP-seq signal profiles and 50bp mappability over all human-specific SVA and LTR5 insertions in the human genome (top) and their orthologous pre-insertion sites in the chimpanzee genome (bottom), with 1kb flanking regions. B. Profile of CNCC H3K27ac ChIP-seq, iPSC H3K9me3 ChIP-seq, and 50bp mappability around SVA insertions in the human genome, with 1 kb flanking regions. SVA subfamilies are distinguished by color with copy numbers inside parenthesis. C. Profile of iPSC H3K4me3 ChIP-seq, iPSC H3K9me3 ChIP-seq, and 50bp mappability around LTR5 insertions in the human

genome, with 1kb flanking regions. LTR5 subfamilies are distinguished by color with copy numbers labeled.

Figure 7: NR2F1 binding profile in CNCCs coincides with the putative SVA enhancer profile. A. Heatmaps of H3K27ac (50bp single-end) and NR2F1 (202bp paired-end) ChIP-seq signal in CNCC over all SVA elements with complete 3' ends (outlined by dotted line) and 3kb flanking regions in the human genome sorted by length (top to bottom, longest to shortest). B. Zoomed-in view of the heatmaps displaying 53 5' truncated SVA elements from 587 bp to 300 bp boxed in Fig. 7a. High H3K27ac and NR2F1 signal are visible on both ends, and flanking H3K27ac ChIP-seq signal correlates with NR2F1 ChIP-seq signal. C. Nucleotide sequence alignment of the 10 truncated SVA elements boxed in Fig. 7B. Elements with strong CNCC H3K27ac signals are marked by checkmarks on the right; elements without CNCC H3K27ac signal are marked by a cross. The NR2F1 binding motif is provided at the bottom of the alignment. D. A human SVA insertion associated with a human-biased enhancer and NR2F1 binding. E. A human SVA insertion associated with a human-biased enhancer but no NR2F1 binding. Note the low mappability, indicating high repetitiveness, over the SVA insertions and the difference in NR2F1 ChIP-seq peaks between E and F.

REFERENCES

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

563 Atkinson EG, Audesse AJ, Palacios JA, Bobo DM, Webb AE, Ramachandran S, Henn BM. 2018.
564 No Evidence for Recent Selection at FOXP2 among Diverse Human Populations. *CELL* **174**:
565 1424–1435.e15.

566 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009.
567 MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**: W202–8.

568 Bashir A, Ye C, Price AL, Bafna V. 2005. Orthologous repeats and mammalian phylogenetic
569 inference. *Genome Research* **15**: 998–1006.

570 Becker JS, Nicetto D, Zaret KS. 2015. H3K9me3-Dependent Heterochromatin: Barrier to Cell
571 Fate Changes. *Trends in genetics□: TIG* **32**: 29–41.

572 Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation
573 on the origins of evolutionary novelty. *Q Rev Biol* **46**: 111–138.

574 Challis D, Antunes L, Garrison E, Banks E, Evani US, Muzny D, Poplin R, Gibbs RA, Marth G, Yu F.
575 2015. The distribution and mutagenesis of short coding INDELs from 1,128 whole exomes.
576 *BMC Genomics* **16**: 143.

577 The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the
578 chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

579 Choudhary MN, Friedman RZ, Wang JT, Jang HS, Zhuo X, Wang T. 2020. Co-opted transposons
580 help perpetuate conserved higher-order chromosomal structures. *Genome Biol* **21**: 16.

581 Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from
582 conflicts to benefits. *Nat Rev Genet* **18**: 71–86.

583 Cohen NM, Kenigsberg E, Tanay A. 2011. Primate CpG islands are maintained by heterogeneous
584 evolutionary regimes involving minimal selection. *CELL* **145**: 773–786.

585 Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution
586 hotspots in Escherichia coli. **274**: 775–780.

587 Danko CG, Choate LA, Marks BA, Rice EJ, Wang Z, Chu T, Martins AL, Dukler N, Coonrod SA, Tait
588 Wojno ED, et al. 2018. Dynamic evolution of regulatory element ensembles in primate
589 CD4+ T cells. *Nature Ecology & Evolution* **2**: 537–548.

590 Darwin C. 1871. *The Descent of Man, and Selection in Relation to Sex*. Createspace Independent
591 Publishing Platform.

592 Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010–
593 1022.

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. ed. C.A. Ouzounis. *PLoS ONE* **7**: e30377.

Du X, Gertz EM, Wojtowicz D, Zhabinskaya D, Levens D, Benham CJ, Schäffer AA, Przytycka TM. 2014. Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. *Nucleic Acids Research* **42**: 12367–12379.

Elgin SCR, Reuter G. 2013. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. *Cold Spring Harb Perspect Biol* **5**: a017780–a017780.

Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, Pääbo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**: 869–872.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636–640.

Eres IE, Luo K, Hsiao CJ, Blake LE, Gilad Y. 2019. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. ed. H.S. Malik. *PLoS Genet* **15**: e1008278.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462–470.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397–405.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, et al. 2018. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *CELL* **173**: 1356–1369.e22.

Florio M, Heide M, Pinson A, Brandl H, Albert M, Winkler S, Wimberger P, Huttner WB, Hiller M. 2018. Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife Sciences* **7**.

Fortin J-P, Hansen KD. 2015. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology* **16**: 180.

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

625 Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C,
626 Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and
627 mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

628 Fudenberg G, Pollard KS. 2019. Chromatin features constrain structural variation across
629 evolutionary timescales. *Proceedings of the National Academy of Sciences* **116**: 2175–2180.

630 Fuentes DR, Swigut T, Wysocka J. 2018. Systematic perturbation of retroviral LTRs reveals
631 widespread long-range effects on human gene regulation. *eLife Sciences* **7**.

632 Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Zhou X, Ward MC, Banovich NE, Kagan CL,
633 Burnett JE, Huang CH, Mitrano A, et al. 2015. A panel of induced pluripotent stem cells from
634 chimpanzees: a resource for comparative functional genomics. *eLife Sciences* **4**: e07103.

635 Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A,
636 Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science*
637 **352**: aae0344–aae0344.

638 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif.
639 *Bioinformatics* **27**: 1017–1018.

640 Greenstein RA, Jones SK, Spivey EC, Rybarski JR, Finkelstein IJ, Al-Sady B. 2018. Noncoding RNA-
641 nucleated heterochromatin spreading is intrinsically labile and requires accessory elements
642 for epigenetic stability. *eLife Sciences* **7**: e0159292.

643 Hancks DC, Kazazian HH. 2016. Roles for retrotransposon insertions in human disease. *Mobile*
644 *DNA* **7**: 9.

645 Hernando-Herraez I, Prado-Martinez J, Garg P, Fernandez-Callejo M, Heyn H, Hvilsom C,
646 Navarro A, Esteller M, Sharp AJ, Marques-Bonet T. 2013. Dynamics of DNA methylation in
647 recent human and great ape evolution. ed. Y. Gilad. *PLoS Genet* **9**: e1003763.

648 Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay
649 TA, Munson KM, Kronenberg ZN, Vives L, et al. 2017. Discovery and genotyping of structural
650 variation from long-read haploid genome sequence data. *Genome Research* **27**: 677–685.

651 Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human
652 LINE-1 retrotransposons since the origin of primates. **16**: 78–87.

653 Kiktev DA, Sheng Z, Lobachev KS, Petes TD. 2018. GC content elevates mutation and
654 recombination rates in the yeast Saccharomyces cerevisiae. *Proceedings of the National*
655 *Academy of Sciences* **115**: E7109–E7118.

656 King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**:
657 107–116.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* **360**: eaar6343.

Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* **14**: 144–161.

Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. 2007. A Macaque's-Eye View of Human Insertions and Deletions: Differences in Mechanisms. *PLOS Computational Biology* **3**: e176.

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 D, Hsu S, Purushotham D, Sears RL, Wang T. 2019. WashU Epigenome Browser update 2019. *Nucleic Acids Research* **47**: W158–W165.

Li D, Zhang B, Xing X, Wang T. 2015. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods* **72**: 29–40.

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.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lowdon RF, Jang HS, Wang T. 2016. Evolution of Epigenetic Regulation in Vertebrate Genomes. *Trends Genet*.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *CELL* **72**: 595–605.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Computational Biology* **2**: e5.

Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**: 1154–1159.

Mager DL, Stoye JP. 2015. Mammalian Endogenous Retroviruses. *Microbiol Spectr* **3**: MDNA3–0009–2014.

691 Makova KD, Hardison RC. 2015. The effects of chromatin organization on variation in mutation
692 rates in the genome. *Nature Reviews Genetics* **16**: 213–223.

693 Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B,
694 Karczewski KJ, Smith KS, et al. 2013. The origin, evolution, and functional impact of short
695 insertion-deletion variants identified in 179 human genomes. *Genome Research* **23**: 749–
696 761.

697 Neininger K, Marschall T, Helms V. 2019. SNP and indel frequencies at transcription start sites
698 and at canonical and alternative translation initiation sites in the human genome. *PLoS*
699 *ONE* **14**: e0214816.

700 Obersriebnig MJ, Pallesen EMH, Sneppen K, Trusina A, Thon G. 2016. Nucleation and spreading
701 of a heterochromatic domain in fission yeast. *Nat Commun* **7**: 11518.

702 Ohtani H, Liu M, Zhou W, Liang G, Jones PA. 2018. Switching roles for DNA and histone
703 methylation depend on evolutionary ages of human endogenous retroviruses. *Genome*
704 *Research* **28**: 1147–1157.

705 Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y,
706 Plajzer-Frick I, Pickle CS, et al. 2018. Enhancer redundancy provides phenotypic
707 robustness in mammalian development. *Nature* **554**: 239–243.

708 Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. 2011. A genome-wide study of DNA
709 methylation patterns and gene expression levels in multiple human and chimpanzee
710 tissues. ed. G. Gibson. *PLoS Genet* **7**: e1001316.

711 Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G,
712 Baertsch R, et al. 2006. Forces shaping the fastest evolving regions in the human genome.
713 *PLoS Genet* **2**: e168.

714 Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut T,
715 Wysocka J. 2015. Enhancer Divergence and cis-Regulatory Evolution in the Human and
716 Chimp Neural Crest. *CELL* **163**: 68–83.

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

719 Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, Wysocka J. 2012. Epigenomic
720 annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell*
721 *Stem Cell* **11**: 633–648.

722 Ramirez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T.
723 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic*
724 *Acids Research* **44**: W160–5.

Ray DA, Xing J, Salem A-H, Batzer MA. 2006. SINEs of a nearly perfect character. *Syst Biol* **55**: 928–935.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Rogers J, Gibbs RA. 2014. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat Rev Genet*.

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA* **103**: 1412–1417.

Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. *<http://wwwrepeatmaskerorg>*.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F, et al. 2018. Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *CELL* **173**: 1370–1384.e16.

Tian X, Strassmann JE, Queller DC. 2011. Genome nucleotide composition shapes variation in simple sequence repeats. *Molecular Biology and Evolution* **28**: 899–909.

Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Research* **27**: 1623–1633.

Wall JD. 2013. Great ape genomics. *ILAR J* **54**: 82–90.

Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. *Journal of Molecular Biology* **354**: 994–1007.

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci USA* **104**: 18613–18618.

Ward MC, Zhao S, Luo K, Pavlovic BJ, Karimi MM, Stephens M, Gilad Y. 2018. Silencing of transposable elements may not be a major driver of regulatory evolution in primate iPSCs. *eLife Sciences* **7**: 166.

Xiao S, Xie D, Cao X, Yu P, Xing X, Chen C-C, Musselman M, Xie M, West FD, Lewin HA, et al. 2012. Comparative epigenomic annotation of regulatory DNA. *CELL* **149**: 1381–1392.

Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, Eichler MY, McPherson JD, Zhao S, Pääbo S, et al. 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *Plos Biol* **3**: e110.

Yokoyama KD, Zhang Y, Ma J. 2014. Tracing the Evolution of Lineage-Specific Transcription Factor Binding Sites in a Birth-Death Framework. *PLOS Computational Biology* **10**: e1003771.

Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**: 1006–1007.

Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D. 2007. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol* **3**: e247.

Fig. 1

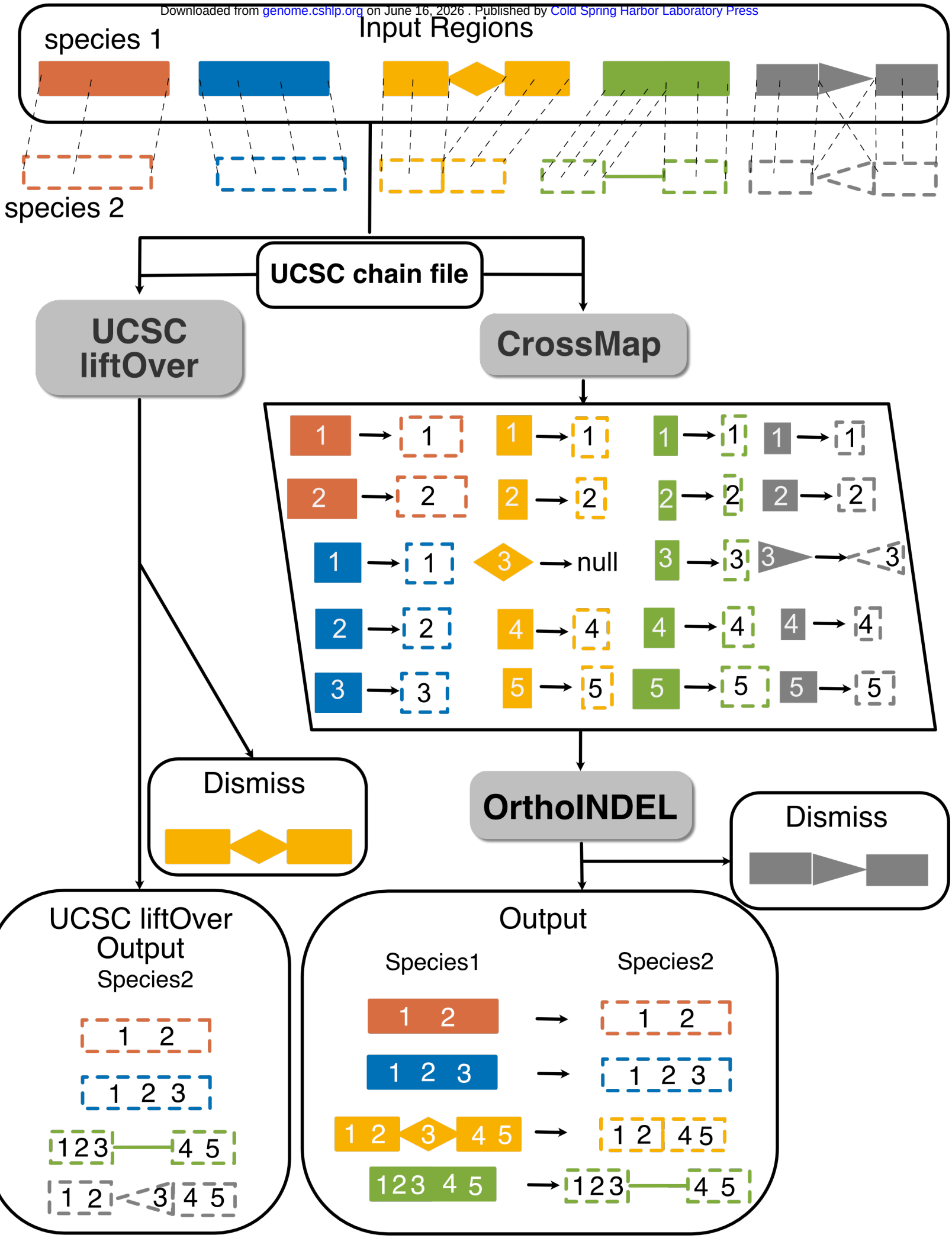


Fig.2

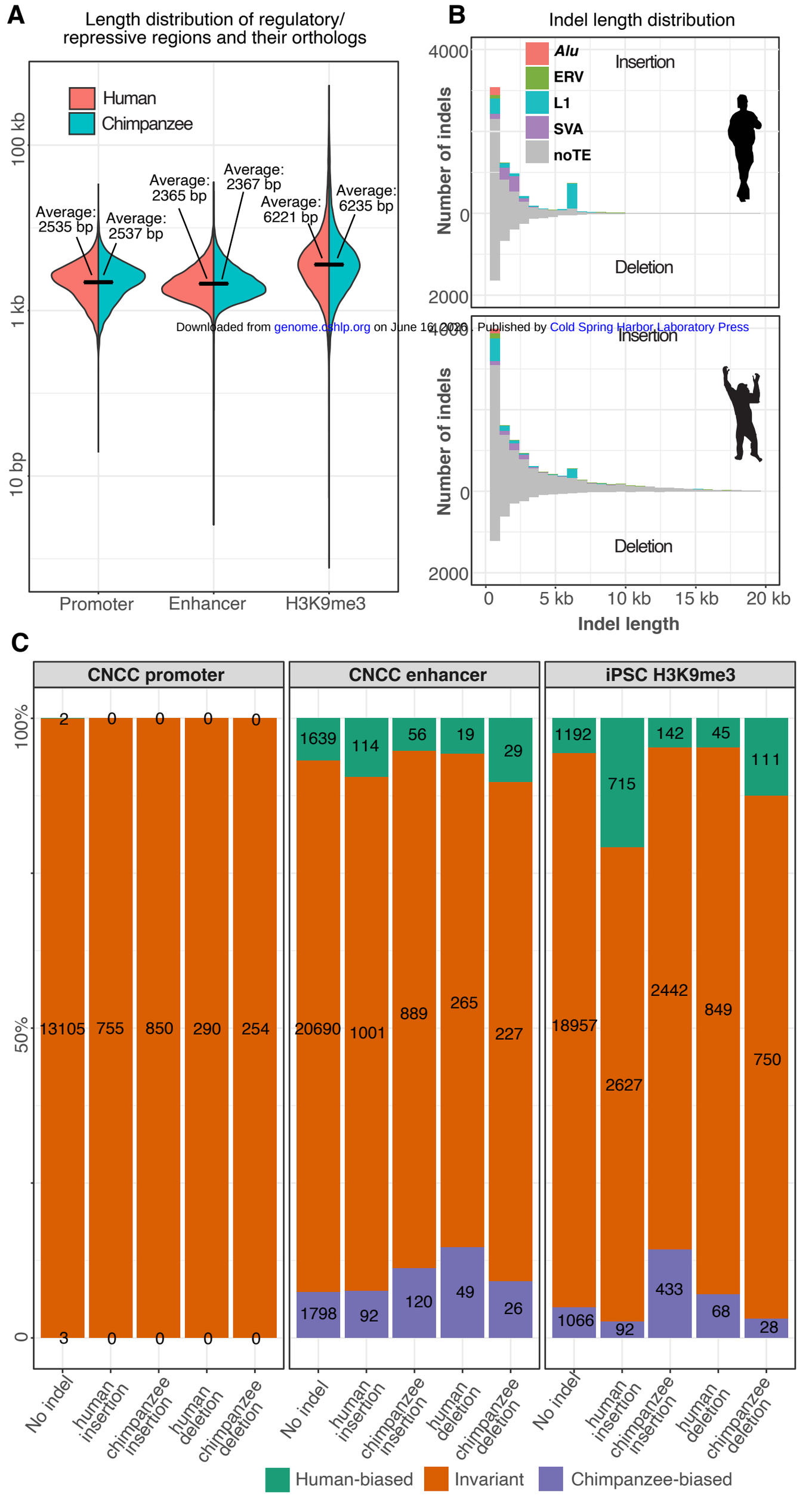


Fig. 3

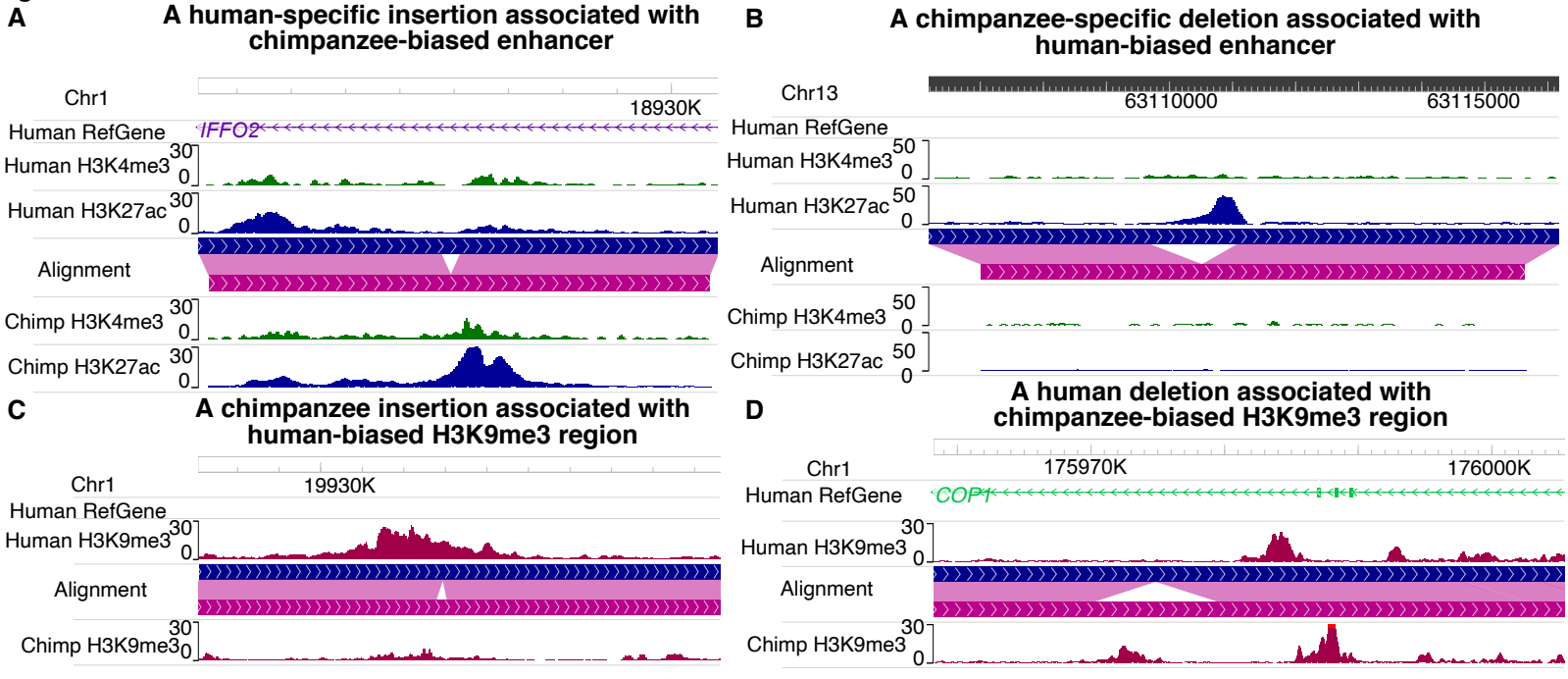

A
A human-specific insertion associated with chimpanzee-biased enhancer
Chr1
18930K
Human RefGene
IFFO2
Human H3K4me3
Human H3K27ac
Alignment
Chimp H3K4me3
Chimp H3K27ac
B
A chimpanzee-specific deletion associated with human-biased enhancer
Chr13
63110000
63115000
Human RefGene
Human H3K4me3
Human H3K27ac
Alignment
Chimp H3K4me3
Chimp H3K27ac
C
A chimpanzee insertion associated with human-biased H3K9me3 region
Chr1
19930K
Human RefGene
Human H3K9me3
Alignment
Chimp H3K9me3
D
A human deletion associated with chimpanzee-biased H3K9me3 region
Chr1
175970K
176000K
Human RefGene
COPA
Human H3K9me3
Alignment
Chimp H3K9me3

Fig. 4

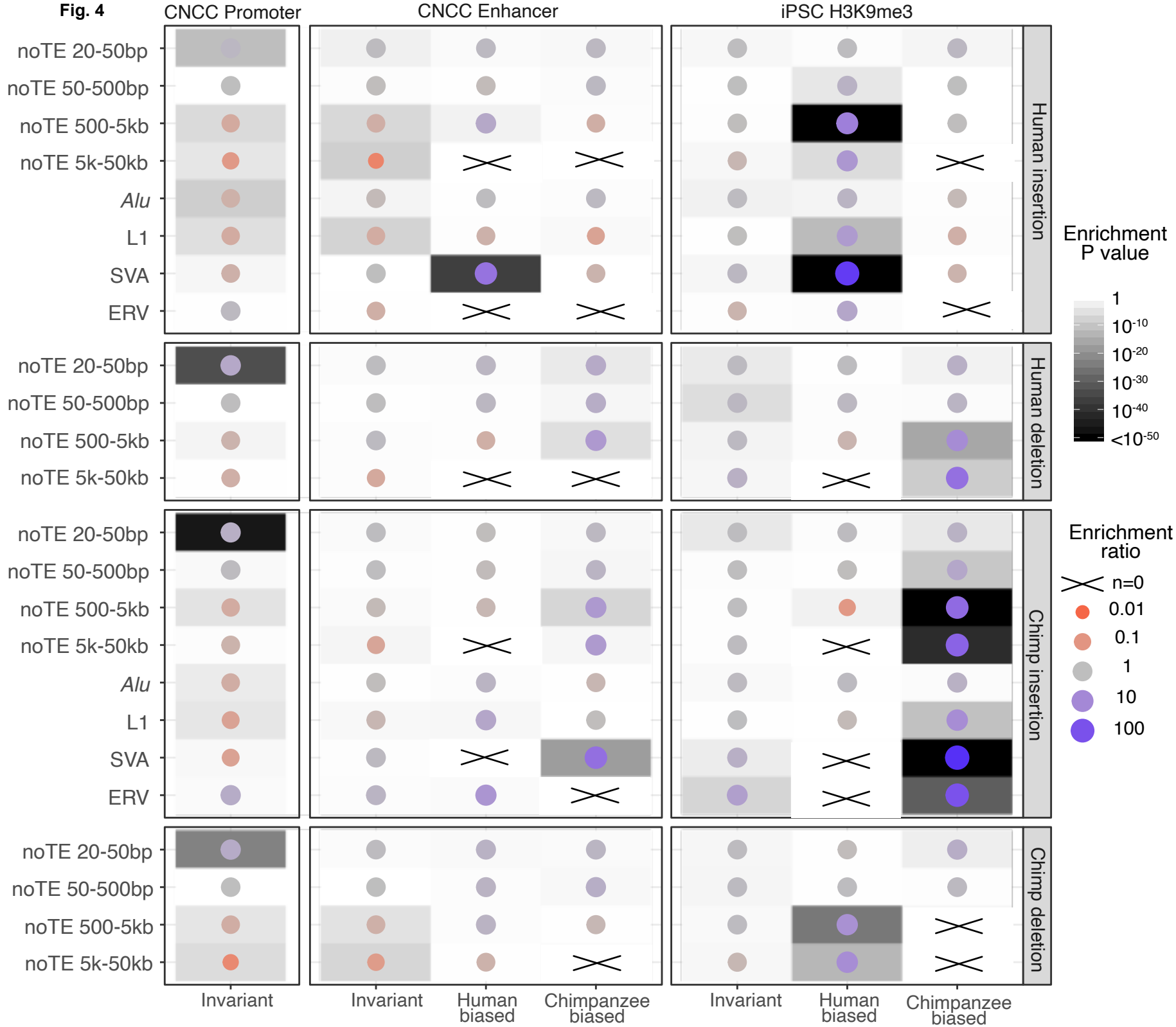

CNCC Promoter
CNCC Enhancer
iPSC H3K9me3
noTE 20-50bp
noTE 50-500bp
noTE 500-5kb
noTE 5k-50kb
Alu
L1
SVA
ERV
Human insertion
Human deletion
Chimp insertion
Chimp deletion
Invariant
Human biased
Chimpanzee biased
Enrichment P value
1
10-10
10-20
10-30
10-40
<10-50
Enrichment ratio
n=0
0.01
0.1
1
10
100

Fig. 5

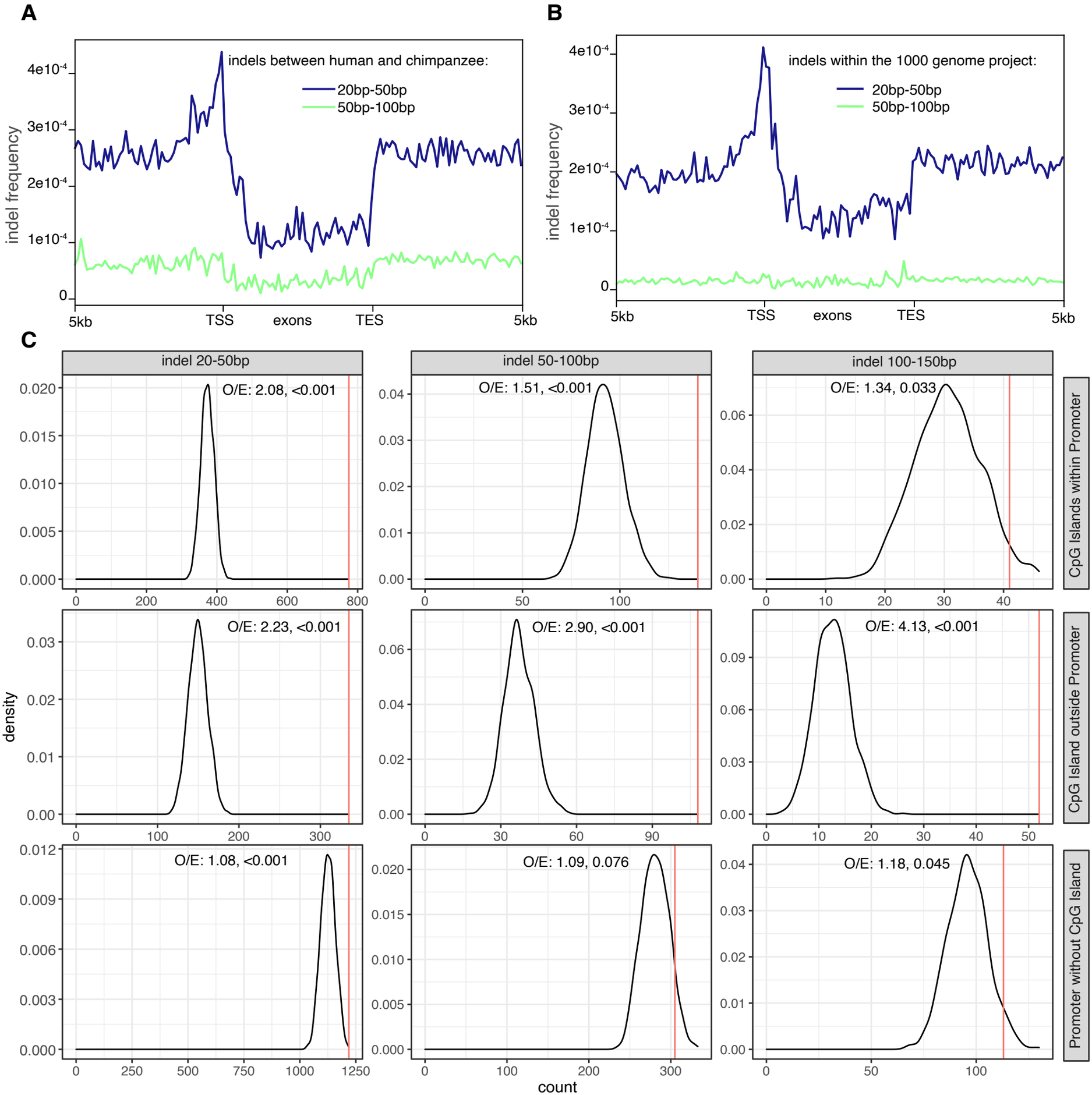

A
indels between human and chimpanzee:
20bp-50bp
50bp-100bp
indel frequency
4e10-4
3e10-4
2e10-4
1e10-4
0
5kb
TSS
exons
TES
5kb
B
indels within the 1000 genome project:
20bp-50bp
50bp-100bp
indel frequency
4e10-4
3e10-4
2e10-4
1e10-4
0
5kb
TSS
exons
TES
5kb
C
indel 20-50bp
indel 50-100bp
indel 100-150bp
CpG Islands within Promoter
CpG Island outside Promoter
Promoter without CpG Island
O/E: 2.08, <0.001
O/E: 1.51, <0.001
O/E: 1.34, 0.033
O/E: 2.23, <0.001
O/E: 2.90, <0.001
O/E: 4.13, <0.001
O/E: 1.08, <0.001
O/E: 1.09, 0.076
O/E: 1.18, 0.045
density
count

Fig. 6

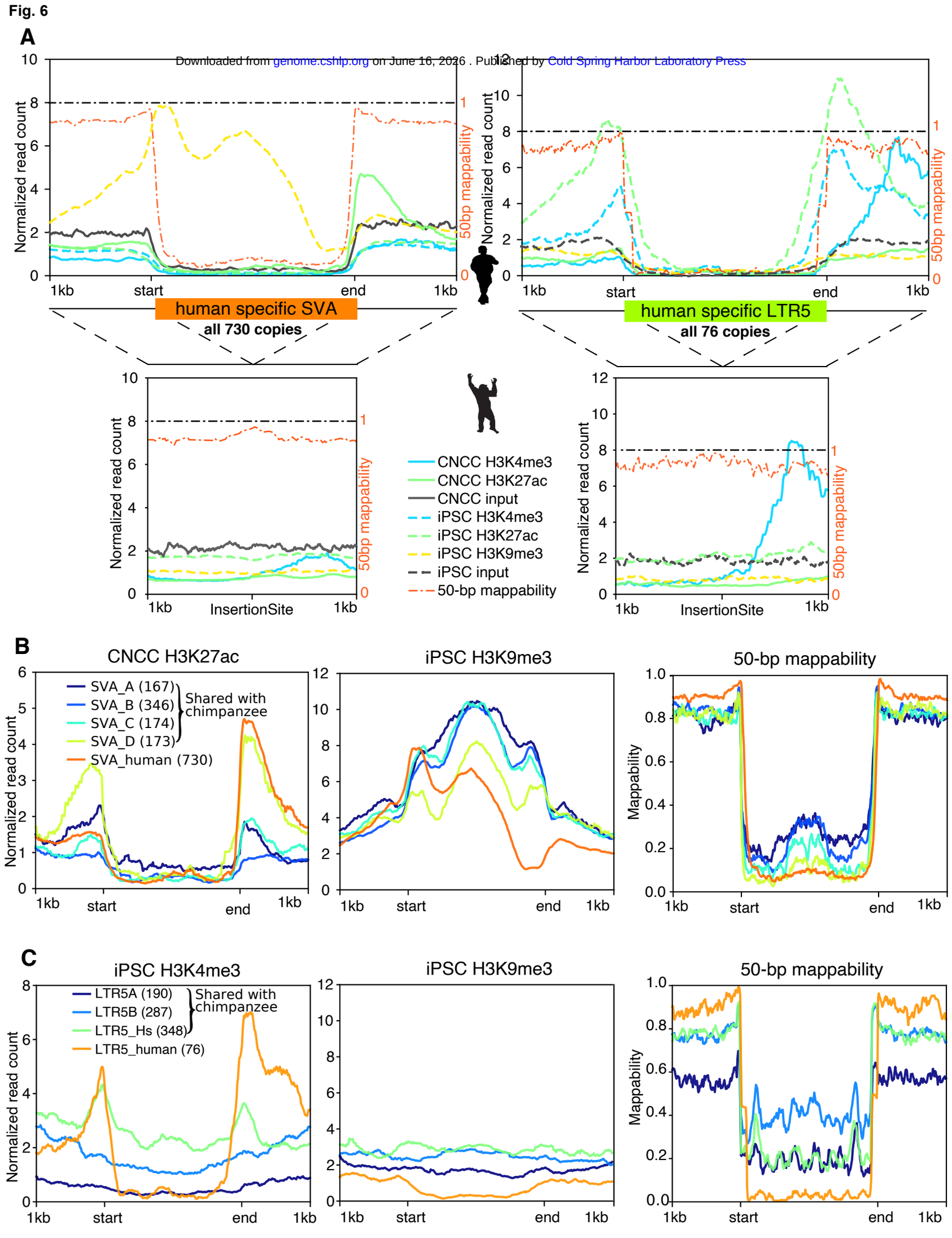
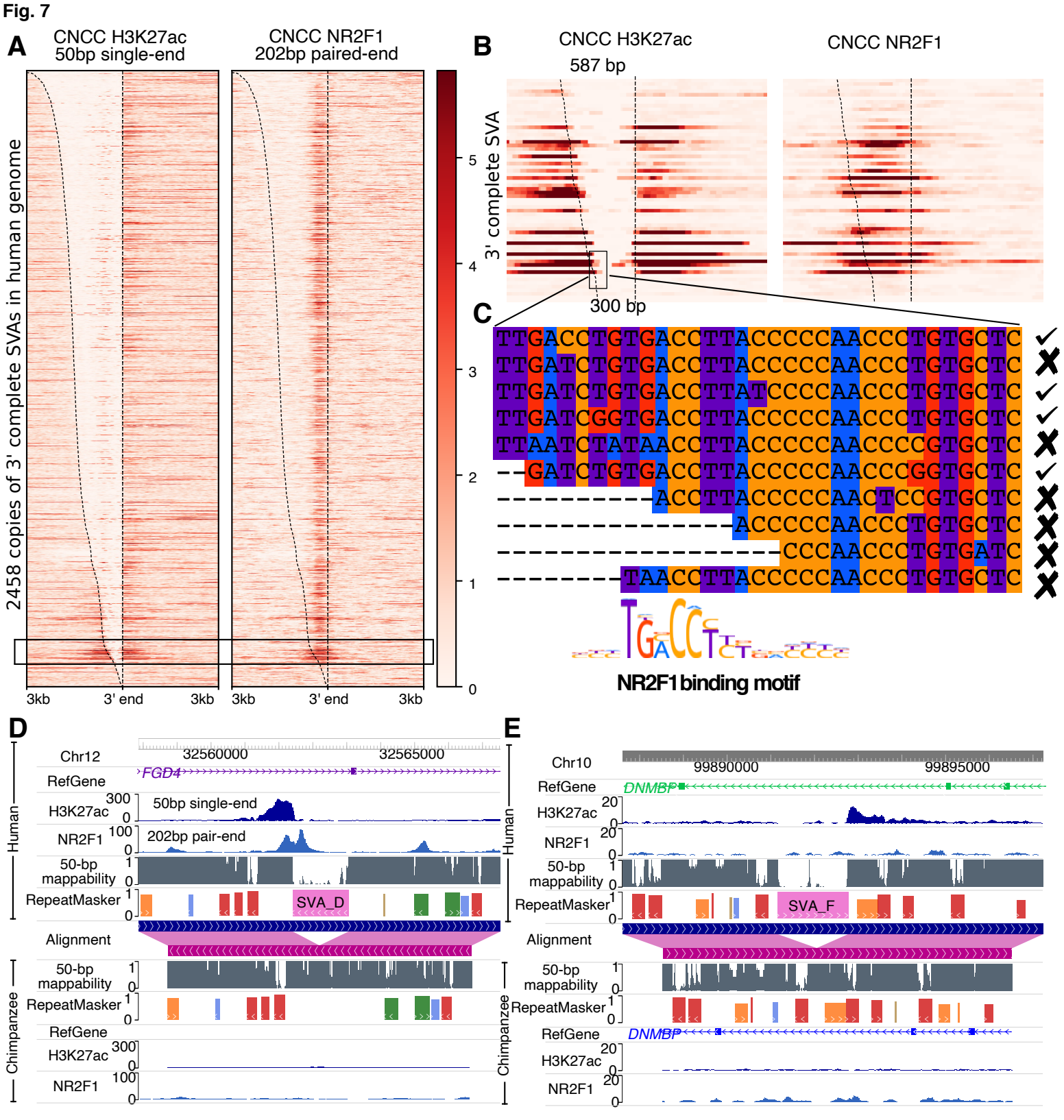


Fig. 7


A
CNCC H3K27ac
50bp single-end
CNCC NR2F1
202bp paired-end
2458 copies of 3' complete SVAs in human genome
3kb
3' end
3kb
3kb
3' end
3kb
5
4
3
2
1
0
B
CNCC H3K27ac
CNCC NR2F1
587 bp
3' complete SVA
300 bp
C
TTGACCTGTGACCTTACCCCCAACCCTGTGCTC
TTGATCTGTGACCTTACCCCCAACCCTGTGCTC
TTGATCTGTGACCTTATCCCCAACCCTGTGCTC
TTGATCGGTGACCTTACCCCCAACCCTGTGCTC
TTAATCTATAACCTTACCCCCAACCCCGTGCTC
GATCTGTGACCTTACCCCCAACCCGGTGCTC
ACCTTACCCCCAACTCCGTGCTC
ACCCCCAACCCTGTGCTC
CCCAACCCTGTGATC
TAACCTTACCCCCAACCCTGTGCTC
NR2F1 binding motif
D
Chr12
32560000
32565000
RefGene
FGD4
H3K27ac
300
0
50bp single-end
NR2F1
100
0
202bp pair-end
50-bp mappability
1
0
RepeatMasker
1
0
SVA_D
Alignment
Human
Chimpanzee
50-bp mappability
RepeatMasker
RefGene
H3K27ac
NR2F1
E
Chr10
99890000
99895000
RefGene
DNMBP
H3K27ac
20
0
NR2F1
50-bp mappability
RepeatMasker
SVA_F
Alignment
Human
Chimpanzee
50-bp mappability
RepeatMasker
RefGene
DNMBP
H3K27ac
NR2F1