**Abstract**

Over the last several decades, a growing body of evidence has suggested that variation in gene expression plays a crucial role in both speciation and tissue differentiation. However, a comprehensive functional understanding of the mechanisms regulating variance in gene expression remains elusive. Numerous studies have suggested that much of this regulation is driven by cis-regulatory elements (CREs), relatively short stretches of DNA that interface with transcription factors to make contact with gene promoters, thereby affecting expression. While chromatin conformation capture technologies have enabled a genome-wide quantification of these CRE-promoter contacts, relatively little interspecies research has been done. Of the few comparative studies examining the 3D genome, most focus on distantly related species and differences in broad-scale genomic structures, such as topologically associating domains (TADs). A higher-resolution evaluation of 3D genome divergence between more closely related species is necessary to understand how much differential contacts affect differential expression.

To address these issues, we probed 3D regulatory divergence between humans and chimpanzees by performing Hi-C on induced pluripotent stem cells (iPSCs) from both species. Initial analysis of Hi-C data in iPSCs revealed that contacts were most different between humans and chimpanzees on chromosomes with large-scale structural rearrangements between the species. In order to assess how much variance in CRE-gene contacts is concomitant with gene expression divergence between species, we integrated our data with orthogonal RNA-seq data from the same individuals. Analyzing this joint dataset, we found that differentially contacting loci and differentially expressed genes were significantly more likely to be involved in a contact that crosses TAD boundaries in one species but not the other. We also found that as much as 12% of the interspecies variance seen in gene expression could be explained by interspecies variance in CRE-gene contacts. In addition, we quantified the overlap between species-divergent Hi-C contacts and published human iPSC histone mark data. We observed strong enrichment for both active and repressive marks in loci involved in contacts that were species specific, that overlapped a differentially expressed gene, or that were both. Overall, our data demonstrates that, as expected, 3D genome reorganization is key to explaining regulatory evolution in primates.

**Introduction (refs here from F31)**

Over the last several decades, a growing body of evidence has suggested that variation in gene expression plays a crucial role in phenotypic divergence between species [1-5],especially in the primate lineageREFs. Numerous studies have characterized gene expression divergence between primate species, finding many differentially expressed genesREFs. In some select instances, these gene expression differences may even be drivers of phenotypic variation between primatesREFs. The suggestion that expression divergence may explain phenotypic divergence between primates is further corroborated by the high degree of coding-sequence conservation observed between humans and chimpanzeesREFsK&W.

However, we still lack a comprehensive functional understanding of the mechanisms that regulate variation in gene expression between primates. Only recently have studies assessed epigenetic variation between primate species, attempting to individually partition different epigenetic features’ relative contributions to gene expression divergence.PAI10-18/ZHOU23-25 While numerous associations have been found between inter-species differences in epigenetic profiles and variation in gene expression between species, these effects are generally small, and a complete picture of different epigenetic elements’ relative evolutionary significance is still lacking. PAI10-18/ZHOU23-25,39

Research in mice, flies, and yeast has revealed that expression divergence between species is often largely driven by mutations in cis-regulatory elements (CREs), rather than trans elements (e.g. transcription factors).22 Whereas the latter may operate broadly across the genome, the former typically act in cis on the same chromosome, as DNA modules that interface with transcription factors and associated proteins to contact gene promoters and thus affect expression.6 CREs are an essential component of regulatory variation underlying tissue differentiation and phenotypic divergence between species.7-11 Classification of CREs into different types of regulators with different effects on expression (e.g. enhancers, silencers) has improved tremendously thanks to epigenetic techniques assaying both chemical modifications to and the accessibility of chromatin.29-31 Perhaps more important than identifying and categorizing CREs, however, is discovering which gene(s) a given CRE actually regulates. Determining CREs’ targets is of particular importance both because CREs act in a distance-independent manner, and because many CREs are tissue-specific in their activity.7-8, 29-33 Numerous epigenetic mechanisms have been proposed to explain this tissue specificity, with most somehow altering CRE accessibility.8,42 Ultimately, CREs’ effects on gene expression are likely to be principally determined by which gene promoter(s) they contact, which is a function of the three-dimensional (3D) structure of the genome.

The 3D conformation of chromosomes is known to affect how genes are expressed within a cell.Babei2015(1-6)Rao2017CohesinlossFulai2013(Dixonlab) Previous studies have shown that 3D genome structure may bring linearly distant loci into close proximity, connecting genes with CREs.Babei7,8;Lieberman-Aiden2009(1-5) Expressed genes have even been observed spatially colocalizing with CREs in numerous 3D FISH experiments.REFS(Babei 7,9?) Chromosome conformation capture (3C) techniques allow for an examination of CRE-gene contacts by quantitatively assaying locus-locus contacts *in vivo*.46 The latest 3C-based technique, Hi-C, pairs the original method’s proximity-based ligation with high throughput next-generation sequencing to find DNA-DNA contacts genome-wide.47 By coupling it with statistical inference, Hi-C can ultimately yield a comprehensive map of the 3-dimensional structure of an entire genome at unprecedented resolution,48 connecting CREs with the gene(s) they actually regulate.

Previous studies utilizing Hi-C data have usually not looked across species. Most have primarily focused on examining variation in chromatin contact frequencies genome-wide within a species, often between different transformed or cancerous cell culture lines that may not faithfully represent endogenous phenotypes.55-57 Performing Hi-C in a model system more analogous to endogenous tissues, such as induced pluripotent stem cells (iPSCs),39-41 would be desirable for assessing variation in 3D genome structure in a developmental biology framework. What few interspecies Hi-C studies do exist typically draw comparisons between distantly related species, and tend to focus on variation in larger-scale structure, rather than fine-scale interactions (i.e. CRE-gene promoter contacts)50-51,plants one. Dixon *et al.* [] first applied Hi-C across human and mouse cell types and discovered highly conserved, megabase-scale self-interacting regions of the genome termed topologically associating domains (TADs). Looking across Hi-C contact maps in four different mammalian species, Rudan *et al.* [] observed strong conservation in TADs, along with evidence of intra-domain interspecies variance. Given the high complexity of Hi-C libraries, it has only recently become fiscally feasible to sequence deep enough to resolve intra-domain variation, and thus to directly assay CRE-gene promoter contacts genome-wide.

Here, we attempt to address these limitations by applying Hi-C to human and chimpanzee iPSCs, assessing variance in 3D genome conformation between the species. We also overlay these data with previously-collected RNA-seq expression data from the same cell lines and publicly accessible human iPSC histone mark data, assessing the extent to which interspecies variance in 3D genome structure and epigenetic profiles affects gene expression divergence between humans and chimpanzees.

**Results**

**3D Genome Locus-Locus Contact Profiles Show Marked Variation Between Humans and Chimpanzees and Reveal Chromosomal Evolutionary Dynamics**

We performed *in situ* Hi-C as previously described [Lieberman-Aiden09] on a sex-balanced panel of four human and four chimpanzee integration-free iPSCs previously generated by the Gilad lab [BJP2015]. We then utilized HiCUP [62] and HOMER [63] (see Methods) to generate genome-wide Hi-C contact maps at 10 kb resolution for all eight individuals, with each map containing approximately one billion pairwise contacts. Since there is no clear gold standard for Hi-C normalization and significance calling, we performed both aspects of the analysis utilizing HOMER, but note that our results are robust with respect to other choices of normalization schemes (ICE, KR) or significance callers (Hiccups, FitHiC). We utilized HOMER at a 5% FDR rate to find locus-locus contacts with more reads connecting them than would be expected by chance, independently discovering an average of ~800,000 significant contacts in each individual.

One natural approach to a comparative analysis might be to liftOver significant contacts from each individual into the other species, subsequently calling contacts found in both as shared and those not as species-specific. However, this approach suffers from incomplete power due to the two different species’ genomes, and limits the scope of the analysis to a simple question of overlap. Instead, we chose to liftOver all significant contacts from each individual into the other species to create a cross-species union list of significant contacts, subsequently extracting the quantitative, HOMER-normalized log2(observed/expected) interaction frequencies between all pairs of loci on the list for each individual.

The distributions of these values were remarkably similar across species (Figure S1), but we nonetheless used pairwise cyclic loess normalization to minimize any induced technical variation. In an effort to be conservative and find biologically meaningful contacts, we also examined variance distributions binned by the number of individuals in which a contact was independently called significant (Figure S2). We observed a monotonic decrease in variance with increasing number of individuals, and identified diminishing returns in reduction of variance after four individuals, consequently filtering out any significant contacts not discovered independently in at least four individuals (regardless of species). This filtering process left 347,206 significant contacts’ normalized interaction frequencies, which we initially analyzed with unsupervised hierarchical clustering and principal components analysis. As expected, both techniques were found to starkly separate samples by species (Figure 1A & 1B).

In an effort to quantitatively assess species-specific contacts, we next modeled the interaction frequencies from each pair of loci, using a linear model with fixed effects for species, sex, and processing batch. We then tested each locus pair for differences in interaction frequency between humans and chimpanzees with moderated t-tests under the limma linear modeling framework (see Methods) [LIMMA REF]. At an FDR≤0.05, we discovered 7,469 and 6,103 differentially contacting (DC) regions specifically enriched in chimpanzees and humans, respectively.

Initial visualization of the linear modeling results for the species term revealed a stark asymmetry of significant hits in a volcano plot of log2 fold change against multiple testing corrected p-values (Figure S3), suggesting much stronger contacts globally in chimpanzees than in humans. Since this did not make biologically meaningful sense, we looked for technical factors that could explain such asymmetry. We partitioned our species-significant linear modeling hits into categories based on how much liftOver to the other species’ genome altered each locus pair, both in terms of changing the distance between mates and changing the sizes of the individual loci involved in the interaction. When interrogating these metrics against species-term significance (Figure S4), we found that changes in locus sizes had little effect, whereas liftOver-induced changes in mate pair distances ≥ 20 kb tended to produce sets of hits with higher than 5% enrichment at 5% FDR. To be conservative, we filtered out such hits from our analysis and visualized the species-specific significant contacts again, and observed the stark species asymmetry in significant hits disappear (Figure 2A).

In the process of interrogating the contacts with liftOver-induced changes between the species, we broke down our DC linear modeling results on a chromosome-by-chromosome basis, under the reasoning that, if there truly were an overall decrease in contact strength in humans as compared to chimps, we would expect it to be uniformly distributed across chromosomes. Both before and after our liftOver filtering, we observed specific chromosomes had much stronger asymmetry in the number of significant contacts stronger in chimps or humans (Figure S5 and 2B). Interestingly, some of the chromosomes displaying the strongest species asymmetry in significant contacts are those that are known to have undergone large-scale rearrangements between the human and chimpanzee lineages [67-72] (see Discussion).

**Comparison of Higher-order Chromosomal Structure Between Humans and Chimpanzees Exhibits Both Strong Conservation and Surprising Differences**

In an effort to understand interspecies variance in higher-order chromosomal structure, we next characterized topologically associating domains (TADs) in each species. We used TopDom [Shin et al. 2016] to independently identify TADs across several window sizes in each individual sample’s HOMER-normalized Hi-C matrix. In a similar fashion to how we compiled significant Hi-C contacts, we used a reciprocal best hits liftOver approach (see Methods) on both sets of TAD boundaries across all individuals in both species. We then merged these to generate a union list of mappable TAD boundaries between the genomes, assessing what proportion of these boundary calls were shared across species. With no filtering for synteny whatsoever and stringent criteria of direct overlap between boundaries, we found

**Differential Contacts are Enriched for Differentially Expressed Genes and May Contribute to Inter-Primate Expression Divergence**

We next examined the connection between inter-species variation in 3D genome structure and variation in gene expression between humans and chimpanzees. Previous work in the Gilad lab generated RNA-seq gene expression data on the same cell lines we collected Hi-C data on, allowing for a quantitative assessment of the relationship between the two data types. Using an in-house curated set of orthologous genes between the species[BJP?], we overlapped individual significant Hi-C contact loci with gene promoters. It should be noted that, while the gene expression data is one-dimensional (each gene has a single expression value per individual), the Hi-C data is not (a given locus may, and often does, make contact with many different loci). We thus decided to summarize a given gene’s contacts simply as the single contact which appeared to be most divergent between species (i.e. the contact that had the minimum FDR for the species term from our linear modeling on contact frequencies). Unless otherwise noted, all joint downstream analyses on both datasets are performed with this summary of contact for each gene.

After filtering out lowly expressed genes, we once again used limma [LIMMA REF] to test for differential expression on 11,946 genes, 7,764 of which had promoters directly overlapping a significant Hi-C contact locus. Reassuringly, we found a wide range of correlations between contact frequency and gene expression in the set of all genes, but a distinctively bimodal correlation structure with a negative and a positive peak in just the set of 1,537 differentially expressed (DE) genes (Figure S6).

Given the strong correlation structure seen in the DE genes, we next sought to assess whether low-FDR differential contacts that overlapped promoters were enriched for differential expression. If species-specific 3D genome conformation is a driver of species-specific expression, we may expect that strongly species-specific contacts would be enriched for differentially expressed genes between humans and chimpanzees. We calculated the proportion of DC loci that were also DE across a range of FDRs from our Hi-C linear modeling results, and found slight enrichment for DE (Figure 3A). Although most of the comparisons do not show up as statistically significant in a chi-squared test, the marked spike in the proportion of genes that are DE at lower Hi-C FDRs still suggests some enrichment for DE amongst species-specific contacts. We found relatively similar results, with somewhat more statistical significance, when we repeated the analysis using a p-value combinatorial method [REF] to summarize the significance of each gene’s contacts (Figure 3B).

We next moved beyond simple measures of correlation and enrichment, seeking to quantitatively estimate the extent to which interspecies DC underlies gene expression differences between the species. To accomplish this, we analyzed the Hi-C and expression data in a linear modeling paradigm as was done in Pai *et al.* []. Briefly, the idea is to use linear modeling to initially assess differential expression in the RNA-seq data, and then to do so again after regressing expression against contact (i.e. to model the residuals of Hi-C data’s predictive power for expression as “Hi-C corrected” expression values). To test if DC may be contributing to DE, we analyzed the difference in effect sizes from the two different models in an empirical Bayes adaptive shrinkage framework [FDRStephens]. We found that the vast majority of DE genes (1530/1537) showed a statistically significant difference in effect size between the two models, with confidence in the direction of their effects (s-values < 0.05) (Figure 4A). In contrast, repeating the same analysis on only the set of non-DE genes resulted in far fewer (2089/6227) genes with a statistically significant difference in effect sizes and confidence in direction of the effect (Figure 4B). Using an orthogonal approach under the same modeling paradigm, but assessing significance (rather than effect size) before and after “correcting” expression levels without empirical Bayes shrinkage[], we also found that the vast majority of DE genes lose their DE status after Hi-C correction (Figure S7). The results from both these methods suggest that interspecies differences in 3D locus-locus contacts may contribute to gene expression divergence between humans and chimpanzees.

**Differential Contacts and Differentially Expressed Genes Have Distinct Epigenetic Profiles**

Finally, we were interested in characterizing the differential contacts, and especially those that overlap differentially expressed genes, on the basis of their epigenetic marks. We hypothesized that loci involved in species-specific contacts, species-specific expression, or both would be more likely to represent active, functional regulatory elements. To test this assumption, we assessed the overlap between our Hi-C loci and publicly-available 15-state chromHMM data from human embryonic stem cells (hESCs) [ENCODE/Broad], as well as histone mark data collected in human iPSCs [ROADMAP].

In our initial analysis we assigned each Hi-C locus to a given epigenetic state based off its maximum base pair overlap with 15-state chromHMM annotations. Since the resolution of our Hi-C analysis is 10 kb, and because the peak length of chromHMM state assignments is quite variable, we weighted each chrom-HMM-Hi-C locus overlap by the reciprocal of the mean base pair overlap of that element amongst our Hi-C bins. This weighting had the desired effect of increasing the number of Hi-C loci assigned to shorter-peak states, particularly those associated with transcriptional and enhancer activity (Figure S8). We note that many of the results we find below were robust to whether or not this weighting scheme was applied, but applying the weighting scheme created a starker visual difference.

After assigning each Hi-C locus to a chromHMM state, we first examined the proportion of different states across our linear modeling Hi-C FDR amongst contacts that did or did not interact with a promoter bin. We observed marked differences in the chromHMM assignments, particularly at low Hi-C FDR, in the set of Hi-C contacts involving a promoter vs. the set that do not involve a promoter (Figure 5A). Compared to contacts not involving a promoter, promoter-involved contacts displayed higher proportions of chromHMM states associated with transcriptional and enhancer activity, and lower proportions of heterochromatic and quiescent states. As expected, we also observed analogous differences between the set of Hi-C contacts that are DE vs. those that are not DE when we subset the analysis to only promoter-involved contacts (Figure 5B).

To thoroughly assess differences in epigenetic state between subsets of differential contacts, we also examined the overlap between our Hi-C contacts and publicly-accessible human iPSC histone mark data, including H3K27ac, H3K4me1, and H3K4me3. In order to also assess a repressive mark and speak to chromatin accessibility, we additionally examined overlaps with publicly-accessible H3K27me3 data and DNase I hypersensitivity site (DHS) data collected in h1-hESCs. For all marks, we used t-tests of the mean to quantitatively assess differences in overlap length with Hi-C loci of different classes (i.e. overlapping a promoter, differentially contacting, differentially expressed).

As expected, contacts involving a promoter showed statistically significant, much stronger overlap with DHS peaks than contacts not involving a promoter (Figure S9A). Among promoter-involved contacts, those that were differential for both expression and contact between the species had the largest overlap with DHS peaks, followed by contacts differential in only one category, with contacts not differential in either category showing the lowest DHS overlap (Figure S9B). These results suggest that loci involved in contacts with promoters, and particularly loci with divergent contact frequencies and/or gene expression levels between species, are much more likely to be in open chromatin (as assayed by DHS).

Remarkably, across all histone marks examined, Hi-C loci involved in both differential expression and differential contact between the species had the highest extent of base pair overlap (Figure 6). Compared to loci differential for both data types, loci with differential contact between species consistently showed slightly lower overlap with regulatory marks, followed by loci involved with differential expression, and more distantly trailed by loci not showing any statistically significant differences between humans and chimpanzees. We note that many of the comparisons between these overlap distributions on active histone marks are statistically significant (Figure 6A-C), while the comparisons done on H3K27me3 (a heterochromatic, repressive marker) are not (Figure 6D). Despite a lack of significance in the H3K27me3 comparison, the relationships between different classes of Hi-C hits remain consistent with those observed in all other histone mark analyses. We note that it is possible this lack of significance is due to the fact that H3K27me3 was the only histone mark obtained from data collected in hESCs, rather than iPSCs.

Taken together, these enrichments for chromHMM states and histone marks suggest that the loci identified as differential between species in both contact and expression represent evolutionarily relevant sequences of the genome. The stark epigenetic differences amongst different Hi-C contact classes (DE, DC, etc.) confirm that these loci are functional regulatory elements that may be drivers of divergence between humans and chimpanzees.

**Discussion**

**Interspecies Divergence in 3D Genome Structure**

We interrogated the extent to which higher-scale 3D genome structure and fine-scale locus-locus contacts are conserved between humans and chimpanzees by collecting Hi-C data from iPSCs in both species. One critical element of such an analysis, briefly mentioned in the results section, is ensuring that differences observed are due to true, meaningful biological effects, rather than differences in genome quality and read mappability between the species. For this reason, we restricted the comparisons we made in both fine-scale locus-locus contacts and higher-order structure to only look at orthologous regions between the species. We independently called significant contacts and TADs in each individual, but then only kept significant contacts/TADs that were retained when run through a reciprocal best hits liftOver between the species’ genomes. This ensures that only regions which could be mapped to the other species’ genome, and then re-mapped back to their original genome, are compared. These filtering steps and the described quantitative methodologies used afterwards have the added benefit of sidestepping issues of incomplete power and differential genome qualities between the species.

As we previously noted, however, use of the reciprocal best hits liftOver (RBHLO) method also induced some changes across the species in terms of individual Hi-C locus size and distance between mates of a contact pair. While we filtered some of these contacts that appeared to be inflated for statistical significance of the species term in our linear modeling (Figure S4), we note that it is impossible to ascertain the relative biological and/or technical relevance of the differences seen in these contacts. We thus took a conservative approach and removed all these contacts from our downstream analyses to minimize false positives, accepting that some of the interspecies differences we observe may actually be underestimated.

As might be expected from previous comparative primate studies of other regulatory phenotypes [PAI and more], our filtered Hi-C data strongly separated the species in both unsupervised hierarchical clustering and principal components analyses (Figures 1A&B). The Pearson correlations we observe between samples from the same species are relatively low those found between replicates from previous studies, however, we calculated these correlations on only the subset of contacts that showed up as significant in the HOMER analysis, whereas previous correlation coefficients were calculated on the entire Hi-C matrix [HiCRep, (Imakaev et al. 2012; Hu et al. 2012; Gorkin et al. 2014; Rao et al. 2014; Ay and Noble 2015; Servant et al. 2015; Dixon et al. 2015, Dixon 2012?]. Furthermore, only recently have robust methods emerged to assess within-species inter-individual variability in 3D genome structure [BNBC, Kipper Fletez-Brant Preprint, HiCRep]. Some of these studies note the weaknesses of using whole-matrix Pearson or Spearman correlations between biological replicates, as correlation values for these metrics can often be higher between unrelated cell types than between two replicates of the same cell type. The strong separation between species and the relatively low within-species correlations we observe in our hierarchical clustering make sense in light of these observations, and the fact that we were only comparing contacts that were independently called significant by HOMER across four separate individuals (regardless of species).

Our comparison of fine-scale locus-locus contact differences between humans and chimpanzees is the first of its kind, as the few other existing comparative Hi-C studies have typically focused on higher-order structure and not undertaken a genome-wide quantitative assessment of conservation [PLANTS one, comparative one]. Consistent with the close evolutionary relatedness of humans and chimpanzees, we found the vast majority of orthologous, HOMER-significant contacts to have no statistically significant difference in contact frequency between the species. We were quite surprised to find a non-uniform distribution of species asymmetry in statistically significant differential contacts across the chromosomes (Figure 2B). While we note that some of the more asymmetric chromosomes are ones known to have undergone large-scale changes between the human and chimpanzee lineages (e.g. fusions, inversions, duplications) [67-72], our relatively limited understanding of structural variation makes it difficult to determine to what extent these are biologically meaningful differences vs. technical artifacts due to differences in genome build qualities.

TAD analysis paragraph

Overall, our interspecies comparative Hi-C analysis reveals intra-TAD variation in contacts, as well as overall TAD structure conservation (or not). No other studies have undertaken this kind of direct, quantitative, genome-wide approach to characterizing interspecies variation in 3D genome structure, but we hope the paradigms established here could be useful in future comparative Hi-C studies. Such studies would also help provide a good benchmark of the efficacy of these comparisons, with the expectation that many of the results we observed here would be more divergent/less conserved between more distantly related species.

**Contribution of Variation in 3D Genome Structure to Expression Divergence**

We overlaid our Hi-C data with RNA-seq gene expression data previously collected in the Gilad lab on the same cell lines [BJP/IGR], and assessed the extent to which variation in 3D genome contacts contributes to gene expression divergence between species. Previous studies using Hi-C data as well as 3D FISH have frequently found spatial co-expression of genes to be associated with their chromatin interaction profiles [Schoenfelder 2010, Homouz 2013, Dong 2010, Zhana Duren 2017, Babei 2015]. Some studies have even found DE genes between conditions to be enriched for differences in 3D Hi-C contacts between conditions [Haiming Chen 2017, Francois Le Dily 2014]. These past findings, and the strong correlation structure we observe between expression levels of DE genes and Hi-C contact interaction frequencies (Figure S6), initially suggested that our Hi-C contacts may have good predictive power for explaining differential expression.

Partially because of these correlations, we were surprised to find relatively weak and often not statistically significant enrichments of DE genes in DC contacts (Figure 3). We do still observe a spike above expectation in proportion of contacts that overlap a DE gene at lower Hi-C FDRs, indicating that species-specific contacts play an important role in species-specific expression. The weak signal observed could be due to differences in resolution between the two datasets; as our Hi-C loci were constrained to be 10 kb in size, a given contact locus often overlapped more than a single gene, adding some noise to our estimated proportions of DE genes amongst DC contacts. The aforementioned difference in dimensions between the datasets could also be to blame for the weak enrichment signal, as a given locus has many opportunities to be called differentially contacting with the various loci it comes into contact with, while a given gene has only one “chance” at being differentially expressed with its single expression value per individual. It is also possible that we observe weak enrichment signal because the expression data were collected from different cell culture instances of the same cell lines, although we believe this to be a less problematic issue as both the Hi-C and RNA-seq experiments were done on bulk samples of millions of cells, and should thus represent steady-state levels of contact frequencies and gene expression levels, respectively.

Most of the prior studies described interrogating gene expression and 3D genome conformation were correlative, did not examine fine-scale contacts, and did not directly attempt to model gene expression as a function of locus-locus contacts genome-wide [Everything from 2 paragraphs above]. Rao *et al*. [Rao 2017] observed relatively modest effects on gene expression after degradation of cohesin, one of the proteins involved in maintaining TAD boundaries and large-scale loops. However, a number of other studies previously found correlations between contact loops and high-fold changes in expression of genes involved in the loop [Rao 2014, Kagey 2010, Lupianez 2015]. These data, combined with the results presented here, lend credence to the idea that the locus-locus contacts created as a result of DNA loops affect transcription activity of loci involved in contact. Our analyses specifically reveal that interspecies differences in contact frequencies between loci on a finer-scale basis (i.e. within TADs) may indeed be drivers of expression divergence between humans and chimpanzees. This idea is supported by our finding a much higher proportion of significance and confidence in direction of effect in the set of DE genes vs. the set of non-DE genes (Figure 4) when we ascertained the effect of Hi-C on expression with two linear models followed by empirical Bayes shrinkage [Pai, STEPHENSFDR]. As noted, these results are further corroborated by a somewhat orthogonal assessment looking only at the significance of differential expression between the original expression model and the “Hi-C corrected” expression values [Pai], where we found that the vast majority of DE genes lose DE status after correction (Figure S7).

Given the resolution of our Hi-C analysis, it is possible that we have underestimated the contribution of interspecies variation in 3D genome structure to gene expression divergence between species. Future Hi-C studies will hopefully be able to sequence deeply enough to obtain lower, sub-kilobase resolutions, allowing researchers to resolve variation in contact frequency at even smaller scales and match genes more accurately with their contact probabilities, thus enabling better predictive power. It is worth noting that, even under an assumption of causality, the Hi-C interaction frequencies alone are not capable of predicting all inter-species variation in gene expression levels. Numerous other studies have found contributions of a wide variety of different regulatory phenotypes to explaining inter-primate differences in gene expression levels [Pai and more?]; 3D genome conformation is merely one of the (seemingly) upstream factors in the large cascade of gene expression regulation.

**Functional Annotation Enrichments in DC and DE Hi-C Contact Loci**

In order to make one last orthogonal assessment of our differentially contacting loci, we overlapped our Hi-C data with publicly-accessible chromHMM epigenetic states and a number of different histone marks. Previous studies have shown that the 3D genome maps produced by Hi-C can be accurately recapitulated by integrating various 1-dimensional epigenetic markings [Di Pierro 2017, Zhu 2016]. Other publications have also found enrichments for a variety of chromatin accessibility and histone marks among interactions predicted by chromosome conformation capture data [Roy 2015, Ron 2017].

Our results corroborate and expand upon some of these findings, specifically interrogating differences in epigenetic profiles of locus-locus contacts in humans and chimpanzees. The differences we observe in chromHMM state assignment dynamics in our comparisons—namely, more active and less repressive states in promoter-involved contacts and contacts overlapping DE genes as compared to contacts not involving a promoter and contacts overlapping non-DE genes—insinuate that sequences functionally relevant for divergence between species have a distinct set of epigenetic marks. We note that these differences could be even more discrete, with the potential to sharply define evolutionarily functional classes of locus-locus contacts, if the chromHMM algorithm is refined and applied to epigenetic mark data collected from the same cell lines as higher-resolution Hi-C data are.

Although all the epigenetic mark datasets we used came only from human cell lines, we observed consistent enrichments in their base pair overlap with the Hi-C loci involved in differential contact, loci overlapping genes with differential expression, or loci that did both. It is not surprising that the Hi-C loci with stronger orthogonal signals of divergence between the species also showed stronger enrichments for regulatory marks, both active and repressive. We interpret these findings as evidence for these loci representing “evolutionary opportunities” between the species. The majority of the genome does not appear to be modified by epigenetic marks, suggesting sequences enriched for any regulatory modifications (active *or* repressive) are more likely to be functionally relevant. It is impossible to tell from our dataset whether differences in epigenetic marks in orthologous regions between humans and chimpanzees could be drivers of contact and/or expression divergence across species. Future studies integrating similar data types could address this issue by assaying epigenetic marks across species, enabling researchers to polarize the regulatory differences in orthologous sequences between humans and chimpanzees.

In conclusion, we have provided some of the first steps in understanding how interspecies differences in fine-scale locus-locus contacts affect variation in gene expression across species. Orthology issues and the relative youth of chromatin conformation capture technologies have precluded the possibility of many comparative Hi-C studies, but hopefully the framework presented here can provide a structural outline for how such research could proceed. More generally, 3D genome conformation is one of many proposed mechanisms involved in regulation of gene expression[Pai28,31,52?]. Our findings indicate that many differences at the smaller scales of 3D genome conformation may affect gene expression differences between humans and chimpanzees, and suggest variation in locus-locus contacts is a pervasive feature of regulatory evolution in primates.

**Materials and Methods**

**Induced Pluripotent Stem Cells (iPSCs)**

Previous work in the Gilad lab generated large panels of both human58 and chimpanzee18 iPSCs via episomal reprogramming as described.59 To ensure their quality, the iPSCs from both species were extensively validated as pluripotent at high passages (>10). Quality control checks included an embryoid body assay confirming ability to differentiate into all three germ layers, qPCR of endogenous transcription factors associated with pluripotency, PCR to confirm the absence of exogenous pluripotency genes (both from residual episomal plasmid or genomic integration), and PluriTest,60 a bioinformatics classifier that assesses pluripotency based on gene expression data. To avoid batch effects, all cell lines were grown in the same incubator in two passage-matched batches, which were also balanced across species and sex.

**In-situ Hi-C Library Preparation and Sequencing**

We performed *in situ* Hi-C as previously described56 on the iPSCs from both species. This version of Hi-C has the distinct advantage of being performed on intact nuclei, reducing the frequency of spurious DNA-DNA contacts due to random proximity ligation in solution. Cells were grown in feeder-free conditions61 to approximately 80% confluence before adding formaldehyde to crosslink the proteins mediating DNA-DNA contacts. Pellets of 5 million cells were flash-frozen and later exposed to the *in situ* Hi-C protocol56 utilizing the restriction enzyme MboI. MboI cut the DNA at each of its 4 base pair recognition sites across the genome, and the resultant proximal fragments were ligated with T4 DNA ligase before being isolated to yield chimeric DNA molecules representing two distinct loci. Libraries were created in two balanced batches identical to the cell growth batches, and were then sequenced (100bp paired-end) on an Illumina Hi-Seq 4000 at the University of Chicago Genomics Core Facility. In order to avoid batch effects, libraries were sequenced across three lanes each on separate flow cells balanced for species.

**Hi-C Read Mapping, Filtering, and Normalization**

We preprocessed, mapped, and filtered the resulting FastQ sequence files using HiCUP.62 The pipeline was used to truncate the reads at ligation junctions, thereafter using bowtie2 to map the two mates of paired-end sequences independently to either the hg38 or panTro5 genomes. We carried out HiCUP filtering as described based on an *in silico* genomedigest in order to remove experimental artifacts.62 We then used HOMER, a foundational statistical analysis suite for Hi-C data,63 to tile the genome into a matrix of 10kb bins and assign reads to their appropriate intersecting bins. We subsequently used HOMER to normalize Hi-C contact bins as previously described63, accounting for known technical biases in Hi-C data. Finally, we called statistically significant interactions independently in each individual using HOMER, based on a null expectation of read counts falling into bins in a cumulative binomial distribution.63

**Creation and Filtering of a Union List of Hi-C Contacts Across Species using liftOver**

In order to ensure the contact frequencies we compared across species were from representative orthologous sequences in humans and chimpanzees, we used liftOver64 with a reciprocal best hits method65 to transfer interaction bin coordinates across genomes. We then extracted the HOMER-normalized interaction frequencies for all contacts in this union list from each individual’s 10 kb Hi-C matrix. We applied a conservative filter upon observing higher variance in hits discovered independently by Homer in fewer than 4 individuals (Figure S2), leaving a total of 347,206 interactions.

**Linear Modeling of Hi-C Interaction Frequencies**

In an effort to quantify species-specific differences in the Hi-C interaction frequency values, we parameterized a linear model:

Y=β0 + βsps + βsxx + βbtcb + εi

Where Y is the response variable Hi-C interaction frequency and β0 is the intercept. βsp, βsx, and βbtc are effect sizes for species, sex, and batch, respectively, with their classifier categorical variables s, x, and b, and an error term εi. We used the R package limma66 to run mediated t-statistic tests on this model, resulting in 54,273 Hi-C contact pairs where the species term is significant (p<=0.05), more than 3 times as many as expected by chance. We applied Benjamini-Hochberg multiple testing correction to find 13,572 interaction pairs significant at 5% FDR. One final filtering step removed interaction pairs that showed a difference in distance between mates across species of >20kb, based upon observing an inflated proportion of these hits significant at 5% FDR (Figure S4).

**Identification of Topologically Associating Domains (TADs)**

We used TopDom [] with search window sizes corresponding to 50 kb, 100 kb, and 200 kb to call topologically associating domains (TADs) independently in each individual sample. To obtain robust sets of boundaries found in at least two, three, or four individuals from within the same species, we did pairwise bedtools overlaps on the domain and boundary elements output by TopDom. As we did before with the Hi-C contacts, we then used a reciprocal best hits liftOver method to ensure we only compared domains and boundaries that could be orthologously mapped across both species. We then calculated our estimates of percentage of conservation by looking at the shared overlap of orthologous domains and boundaries found independently in each species, and dividing this by the total number of unique domains and boundaries (respectively) in the final union list created across species. To assess differences in total number and size of boundaries and domains between the species, we used a t-test of difference in the means, grouping the samples within each species into a single distribution.

**Differential Expression Analysis**

Previous work in the Gilad lab generated RNA-seq expression data on the iPSC lines in this study [18]. We filtered out lowly expressed genes based on a log RPKM normalization cutoff of 0.4 that produced a similar proportion of retained genes when using a log CPM normalization cutoff of 1.5. This filtering process retained 11,946 genes, which we analyzed with limma [66] under a simplistic linear modelling framework that only includes a species term. Using this method we found 2,268 differentially expressed genes <5% FDR after Benjamini-Hochberg multiple testing correction.

**Broad Integration of Hi-C and Gene Expression Data**

We obtained the overlap between our gene expression data and our Hi-C data by using bedtools overlap on the Hi-C loci and the first exon of each gene. We utilized an in-house curated file of orthologous gene coordinates between humans and chimpanzees, extracting a one-base-pair interval at the beginning of each first exon as a proxy for transcription start sites (TSSs). Given the large size of our Hi-C loci (10 kb), we were not concerned that the overlapping contacts would not actually be representative of the genes’ TSSs, but we nonetheless checked where our one-base-pair intervals fell along the length of the Hi-C loci and found a relatively uniform distribution (not shown).

As we described above, the difference in dimensionality between the two datasets also presented a challenge. While every gene has only one expression value per individual, a given Hi-C locus can and frequently does make contact with many other loci. When a given gene overlapped a Hi-C locus making multiple contacts, we chose the contact with the lowest species term FDR from our linear modeling (i.e. the most species-specific contact) as that gene’s corresponding interaction frequency value. We also used the FDR-adjusted p-value from that contact as that gene’s corresponding differential contact significance.

**Enrichment of Differential Expression in Differential Contacts**

We cumulatively tabulated proportions of genes with both differential contact and differential expression continuously across differential contact FDRs and across a discrete range of differential expression FDRs (1%, 2.5%, 5%, 7.5%, and 10%). We then compared these proportions at each FDR to those that would be expected based on conditional probability alone (dotted lines in Figure 3). Finally, we utilized Pearson’s chi-squared test to assess the statistical significance of the enrichment at each FDR (Bottom panel of Figure 3).

**Assessing the Quantitative Contribution of Hi-C Contact Frequencies to Gene Expression Levels**

In order to quantify the effect of Hi-C contact interaction frequencies on RPKM gene expression values, we parameterized and compared two separate linear models as was done prior in the Gilad lab [Pai]. The first model is identical to the one described above for the RNA-seq data: a simple linear model attempting to predict expression levels with only a species term. We then ran a linear model attempting to predict gene expression levels with corresponding Hi-C interaction frequencies as the only predictor variable. We treated the residuals from this model as expression values “corrected” for Hi-C interaction frequencies, and input them into the same initial differential expression model as before (with species as the only term).

For each gene, we then computed the difference in effect sizes between the two models (one run on RPKM expression values and one run on “Hi-C corrected” expression values), as well as the standard error of the difference. We then used an empirical Bayes approach to shrink both variance and effect sizes with the R packages vash and ash, respectively [refs, NEW DEAL STEPHENS]. We further utilized ash to test all genes for significant differences in their effect sizes between the models. Rather than using q-values, however, we called genes’ expression statistically significantly affected by Hi-C contacts if they had an s-value of <0.05 [FDR New Deal]. Using the s-values, rather than q-values, not only takes significance into account but also has the added benefit of assessing our confidence in the direction of the effect.

**Integration with Epigenetic Annotations**

We obtained chromHMM 15-state model peak calls in human iPS-18C cells from ENCODE (Supplementary Table 1). We subsequently found the overlap between our Hi-C contact loci and the chromHMM peak calls, quantifying the extent of base pair overlap for each locus with all chromHMM peaks overlapping it. Because we wanted to speak broadly to epigenetic profile dynamics of our Hi-C loci, we sought to assign each individual locus a single chromHMM annotation based on whichever peak had the highest base pair overlap with that locus. However, the distribution of overlaps of different chromHMM annotation peaks with our Hi-C bins were quite variable in size. To account for this, we normalized each annotation’s overlap length in each locus by multiplying it by the reciprocal of its mean base pair overlap across all our bins (Figure S8). After removing duplicate Hi-C loci, we then assigned individual loci to chromHMM annotations based on these normalized base pair overlaps. To characterize epigenetic profiles of our Hi-C loci, we then took the top ten most differentially contacting loci (i.e. the ten lowest FDR loci from our Hi-C linear modeling), and tabulated proportions of which annotations were represented amongst them. We then iteratively added the next-lowest FDR contact (i.e. two Hi-C loci at a time) to this tabulation, re-calculating proportions on the new set of contacts. We ran this same cumulative proportions analysis separately on contacts not overlapping promoters, contacts overlapping promoters, contacts overlapping promoters of DE genes, and contacts overlapping promoters of genes that were not DE (Figure 5).

We also obtained data on H3K4me1, H3K4me3, and H3K27ac collected in human iPS-18A cells, and data on H3K27me3 and DNase hypersensitivity sites collected in H1-hESCs, all from ENCODE (Supplementary Table 1). We used bedtools intersect to find the base pair overlaps of each of these different marks with our Hi-C contact loci. We then removed duplicate Hi-C loci from the dataset and used a standard t-test of difference in the means in a pairwise fashion to assess the statistical significance of differences in the overlap distributions amongst different sets of Hi-C classes (based on differential contact and differential expression, Figure 6).

**References**

1. Britten, R.J. and E.H. Davidson (1971) Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty. The Quarterly Review of Biology 46(2): 111-138. doi:10.1086/406830
2. King, M., and A. Wilson. *Evolution at Two Levels in Humans and Chimpanzees.* Science, 1975. **188**(4184): p. 107-16.
3. Carroll, Sean B. *Evolution at Two Levels: On Genes and Form.* PLoS Biology, 2005. **3**(7): p. 1159-1166
4. Gilad, Y., et al. *Expression profiling in primates reveals a rapid evolution of human transcription factors.* Nature, 2006. **440**(7081): p. 242-5.
5. Wray, G.A., *The evolutionary significance of cis-regulatory mutations.* Nature Reviews Genetics, 2007. **8**(3): p. 206-16.