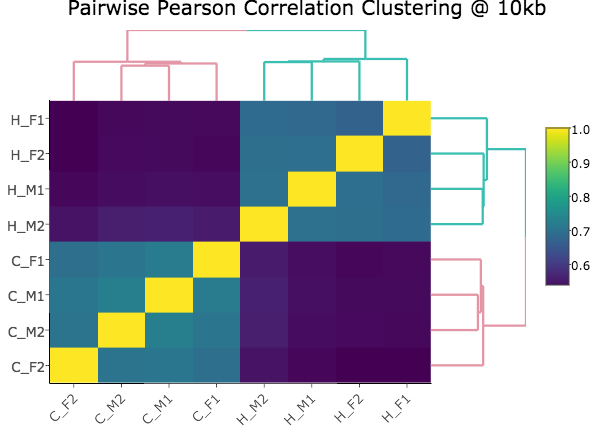
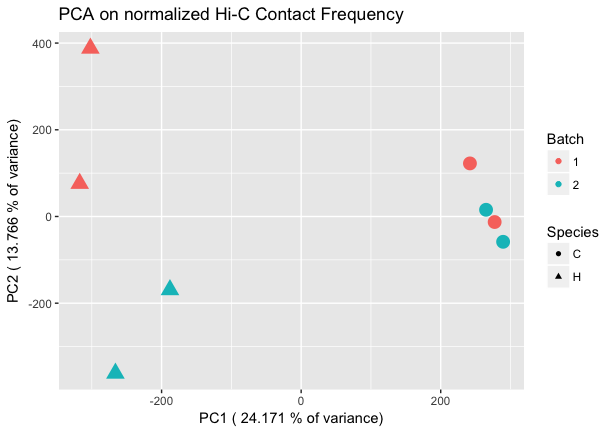
Between these two figures, have the text explaining the difference between the naïve version of looking at mere overlaps vs looking at the different values across all 8 individuals here to avoid issues of incomplete power. Identifying independently in each individual, but then looking at differences in actual values across everyone.



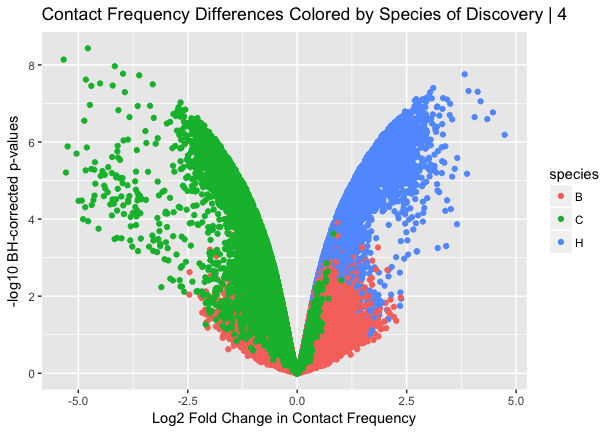
**Figure 1. Regulatory landscapes cluster by species.**

**Left panel: Principal components analysis (PCA)** of Homer-normalized interaction frequencies for the union of all significant hits between humans (H) and chimpanzees (C). PCA is a dimensional reduction technique where the first principal component (PC1) encapsulates as much of the variability in the data as possible, and subsequent PCs capture less. Here, PC1 captures ~24% of the variance, and appears to be highly correlated with species, separating the humans (H) in triangles from the chimps (C) in circles. Neither PC2 nor any of the other lower PCs showed strong correlations with batch or sex, suggesting that species is a strong driving factor in determining interaction frequency. This result was robust to filtering out hits where liftOver introduced genomic distance or bin size differences.

**Right panel: Unsupervised hierarchical clustering** at 10 kbof the pairwise Pearson correlations between Homer-normalized interaction frequencies. Pairwise Euclidean distances between samples are calculated, followed by complete agglomeration to produce hierarchical clustering results. Here, humans and chimps cluster separately, with chimp interaction frequencies being more highly correlated with other chimp interaction frequencies, and the same being true for humans. The first letter in the labels demarcates the species as above, and the following symbols indicate sex (male, M or female, F) and batch (1 or 2). This clustering result was also found to be robust to filtering out hits where liftOver introduced genomic distance or bin size differences. Scale on the right represents Pearson correlation values.

Figures and number reporting have to be one specific analysis, in this case I am more inclined to juicer since newer.

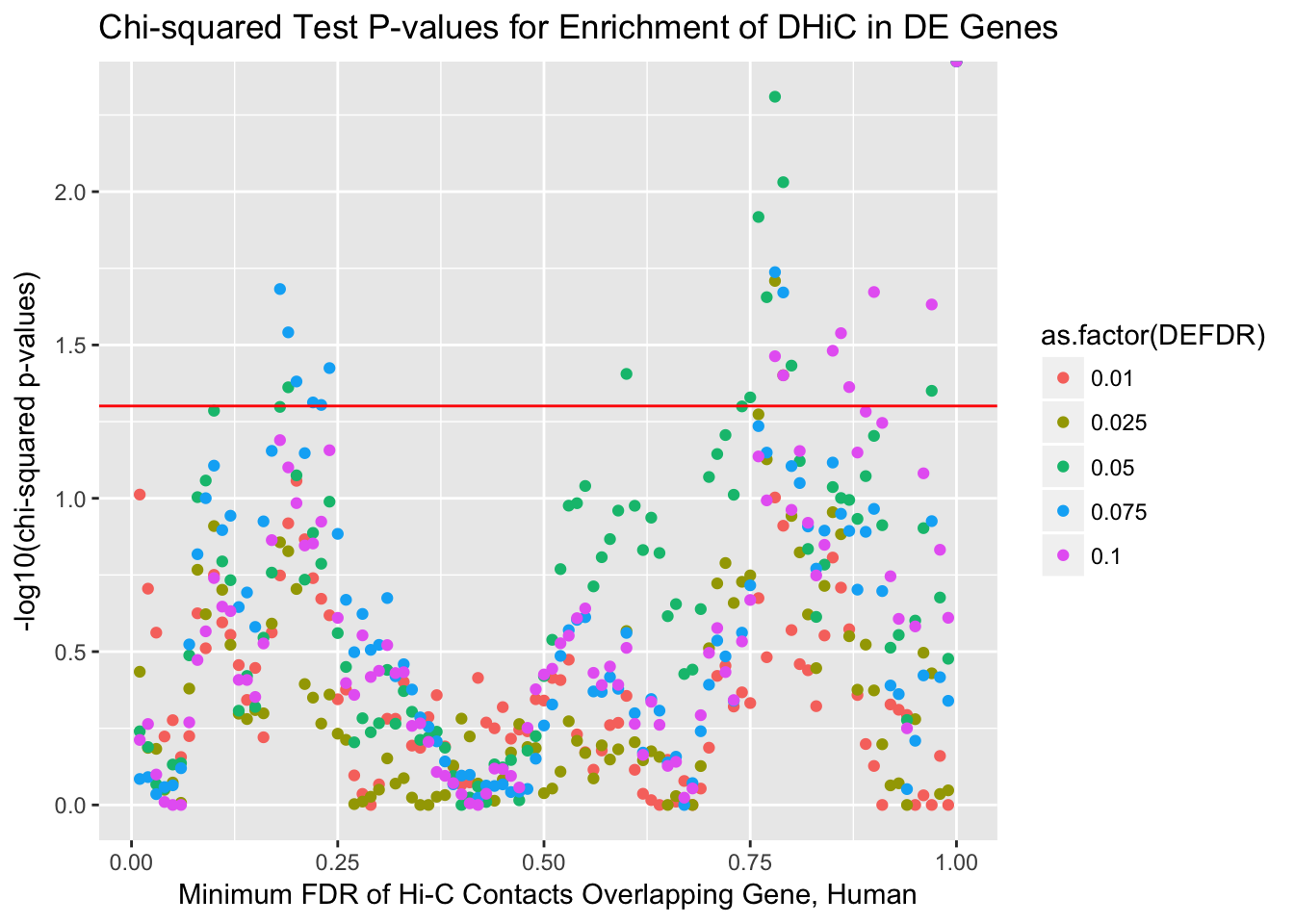
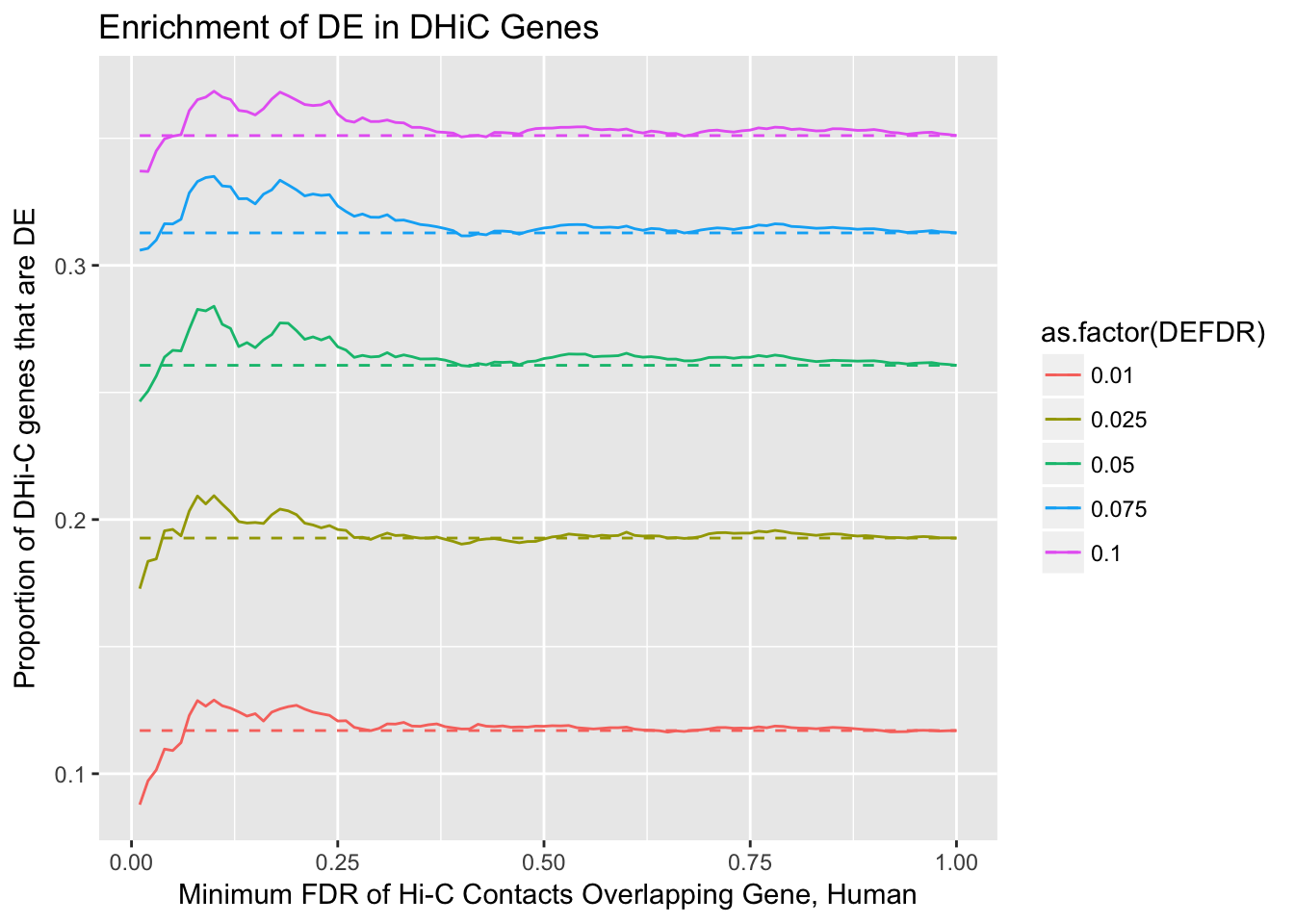
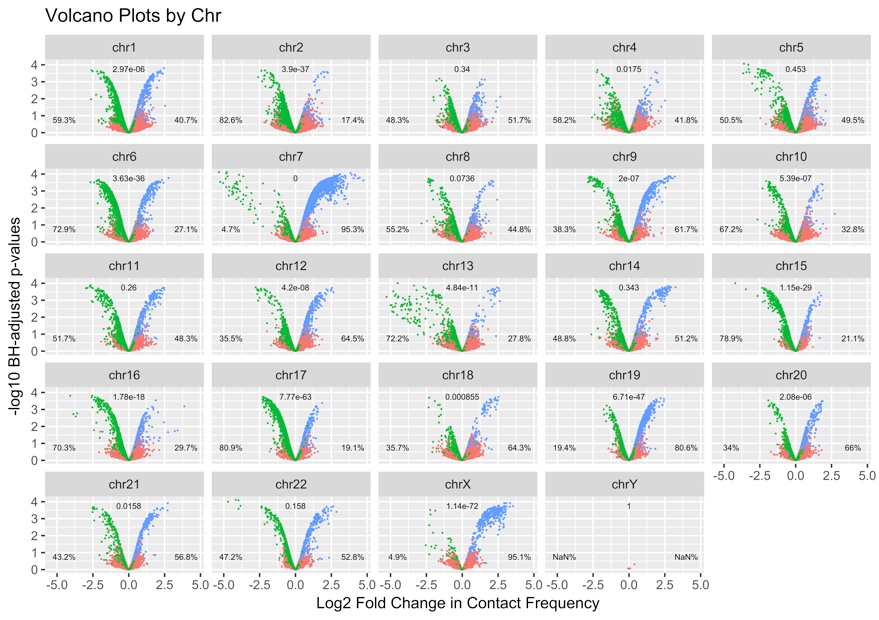
In between this and next figure, look at enrichments of functional categories in these hits! Prepare that for next week.



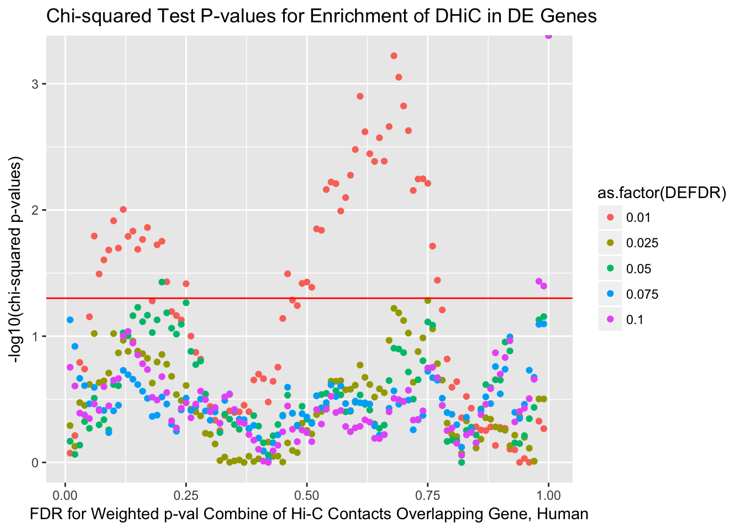
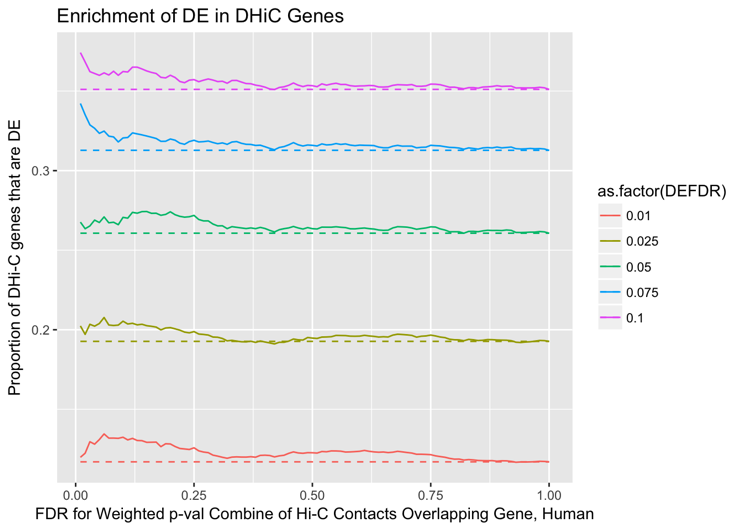
**Figure 2. Linear modeling reveals large-scale chromosomal differences in regulation.**

**Left panel: Volcano plot** showing log2 fold change in interaction frequency on the x-axis, and Benjamini-Hochberg corrected –log10 p-values on the y axis. Hits are colored by the species in which they were discovered significant in Homer (B-both, C-chimpanzees, H-humans). Humans are coded as a 1 in the linear model, meaning hits on the left side of the plot represent pairs with marked decrease in contact frequency in humans as compared to chimps, and the inverse on the right side. Reassuringly, the majority of hits that show significantly higher contact frequency in humans were discovered as significant by Homer in humans, and the same was true for chimps, as expected.

**Right panel: Chromosome-by-chromosome volcano plot** as in left panel.Initial creation of the volcano plot before filtering out more hits (supplementary Fig. 3) showed strong asymmetry, with many more strong effect size hits representing a decrease in contact frequency in humans compared to chimps (i.e. a buildup of points on the left side of the plot). This observation suggested doing a chromosome-by-chromosome analysis, with the expectation being that asymmetry would be uniformly distributed if it were a global feature of contact differences in humans and chimpanzees. Here, we see a subset of specific chromosomes showing distinct asymmetry patterns, suggesting large-scale chromosomal differences in regulation. Interestingly, some of these chromosomes are known to have undergone large-scale changes between the human and chimpanzee lineages, such as the putative fusion event at chromosome 2, or the pericentric inversions in chromosomes 4, 5, 15, 16, 17, and 18.67-72



**A**



**B**

**Figure 4. Differentially expressed genes show enrichment for differential Hi-C contacts.**

**A:** Enrichment of differential expression (DE) in genes with differential Hi-C contact (DHiC). Top panel shows the proportion of DHiC genes that are also DE on the y-axis, across a range of FDRs for Hi-C linear modeling on the x-axis. Differently colored lines indicate different thresholds for DE, and dashed lines indicate the expected proportion of DHiC genes that would be DE based on conditional probability alone. A spike in the proportion of DHiC genes that are DE at a low Hi-C FDR, as seen here, indicates that species-specific contacts are indeed enriched for species-specific expression. Bottom panel shows –log10 chi-squared test p-values for enrichment of DHiC in DE genes on the y-axis across a range of Hi-C FDRs on the x-axis. Colors once again indicate different DE thresholds, and a horizontal red line is drawn where p=0.05. For two of the less stringent DE FDRs (5% and 7.5%), statistically significant enrichment of DHiC in DE genes is seen around 20% Hi-C FDR. (I shouldn’t be trying to explain the statistically significant enrichments at higher DHiC FDRs, right? Makes me wonder if the x-axis here should be shrunk to not see them at all in the paper). Because a given Hi-C bin overlapping a gene often makes contact with multiple other bins in the cross-species union of significant hits dataset, some kind of summary must be chosen to assess DHiC for a gene. Here, the contact with the minimum FDR value for species from Hi-C linear modeling was chosen, based on the assumption that these contacts should more accurately represent regulatory landscape divergence between species.

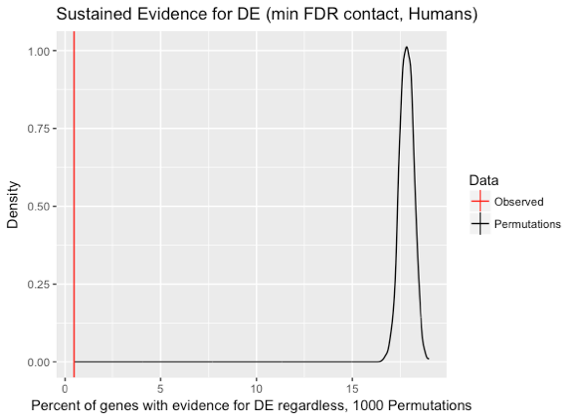
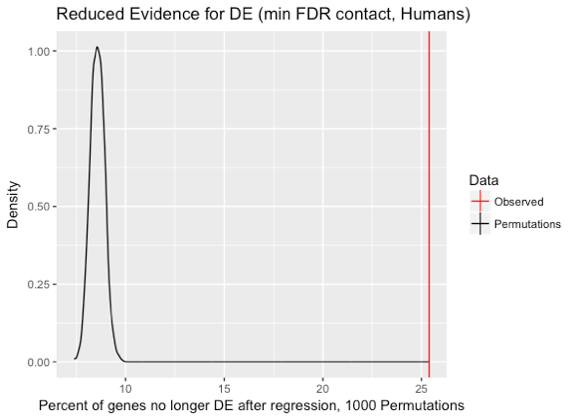
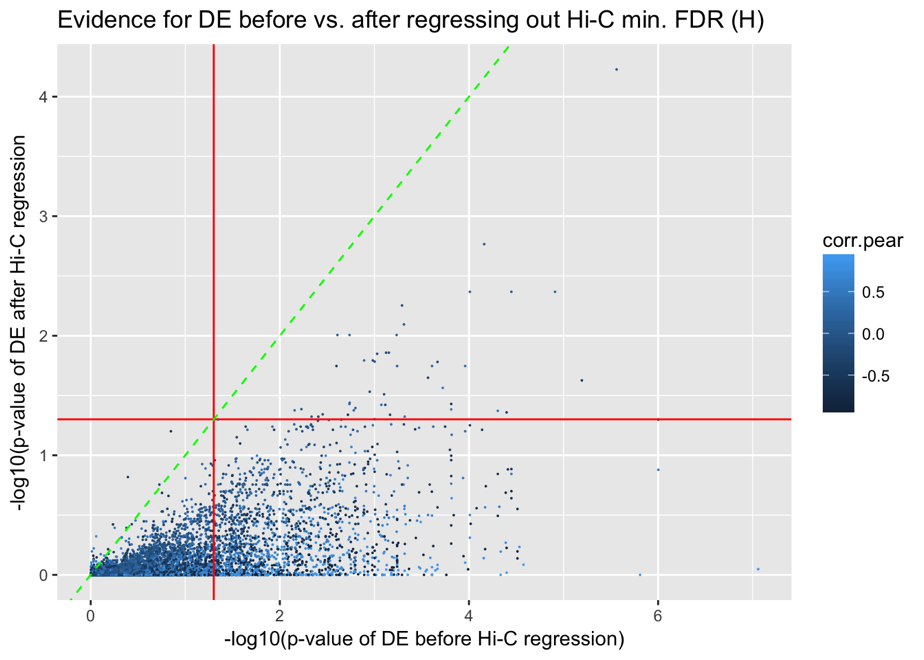
**B:** Same as in A, but this time, a weighted p-value combination techniqueref was used to integrate each Hi-C bin’s FDR across all its contacts. As a result, a more persistent enrichment is seen as Hi-C FDR increases. (this sentence won’t be necessary if we only go to 0.3 on x-axis or something in paper)

**Figure 5. Inter-species expression differences are no longer significant when accounting for inter-species regulatory differences.**

**A:** Scatter plot of Benjamini-Hochberg corrected p-values obtained by testing the null hypothesis of no differences in gene expression levels between human and chimpanzee before (x-axis) and after (y-axis) regressing out each gene’s Hi-C contact strength. As in Fig. 4, the contact with the minimum FDR from linear regression of the Hi-C data on species is used. Solid red lines correspond to a 5% FDR threshold. Hits in the bottom right quadrant represent instances of decreased evidence for DE after regressing out Hi-C contact frequency, hits in the top left represent instances of increased evidence for DE after regression, and hits in the top right quadrant represent genes where equivalent evidence was seen for statistically significant DE before and after regression. (Do we want to include something on bottom left; should the figure in the actual paper be just on the DE genes to begin with as in Pai’s, b/c this is on entire gene dataset? And which of the permutations actually matter to show? Based on Pai paper it would just be the one in (B)).

**B:** Density plot of distribution, based on 1000 permutations, of the percentage of genes for which evidence for inter-species gene expression differences is expected to be reduced beneath statistical significance after correcting for Hi-C contact probability, by chance alone. Black shows the distribution for the permutations, red line represents observed percentage seen in the data.

**C:** Same as in (B), but this time looking at genes that show sustained evidence for a statistically significant difference in expression across species regardless of Hi-C correction.

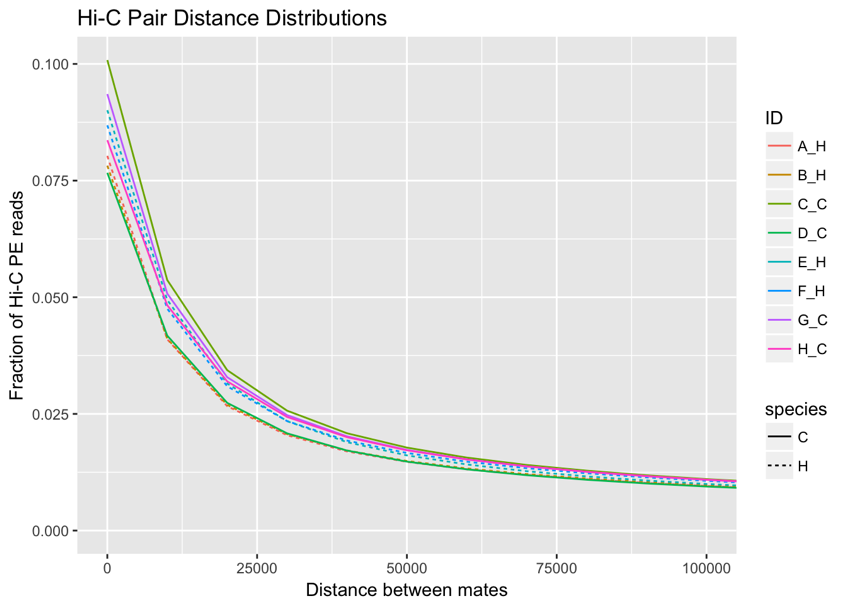
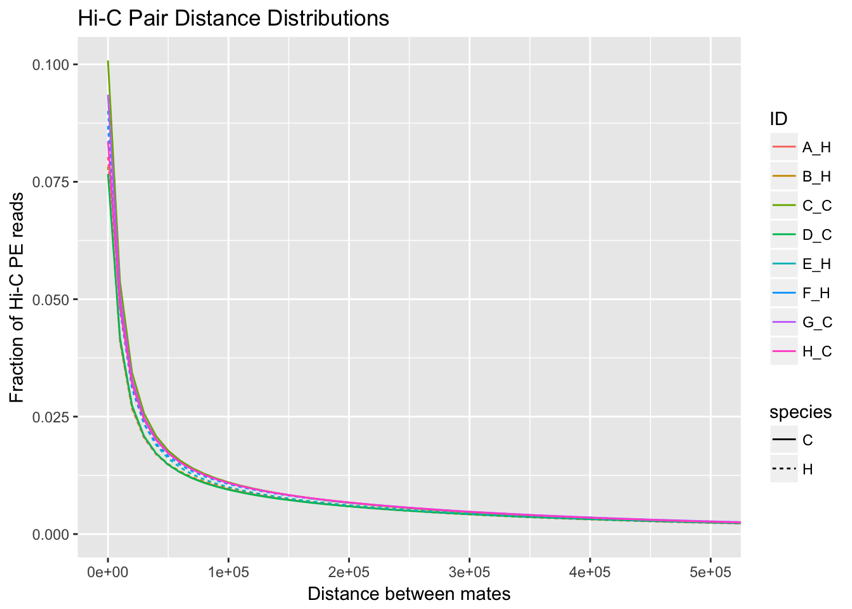


**B**

**C**

**A**

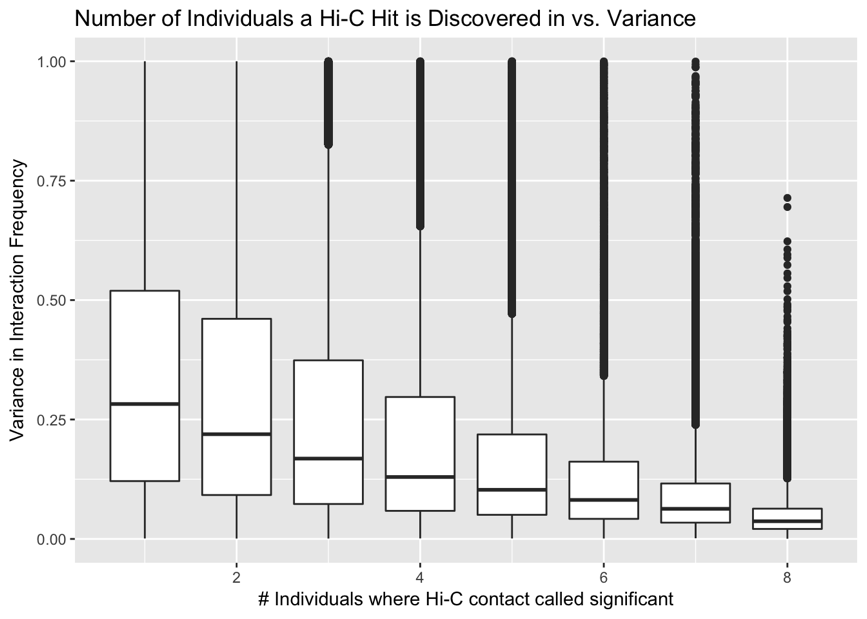
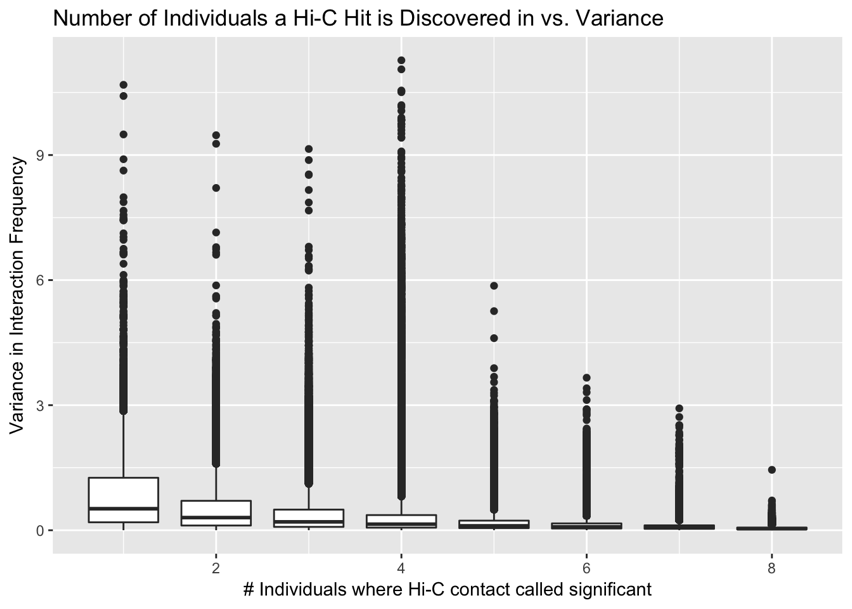
Would be nice to grab an example from up above to visualize, like 4C in Pai.



**Supplementary Figure 1.** **Distance distribution of paired-end reads from each library.**

**Left panel:** Fraction of paired-end (PE) reads representing mate pairs at different distances, in 10kb bins, shown colored differently for each individual library (ID). Dotted lines represent humans, and solid lines represent chimps. Reassuringly, we see no clear trend separating the species. Individual libraries from each species appear to have fractions of their PE reads at a given distance show up as both lower and higher than other individuals in the same or a different species.

**Right panel:** Zoom of left panel on x-axis to show degree of separation between libraries on a smaller scale. Again, human libraries and chimp libraries do not seem to be clustering separately here.



**Supplementary Figure 2. Variance in interaction frequency as a function of the number of individuals in which a significant interaction is independently discovered.**

**Left panel**: Boxplots of variance in contact frequency across all 8 individuals on the y-axis, binned by the number of individuals in which an interaction is independently called significant on the x-axis.

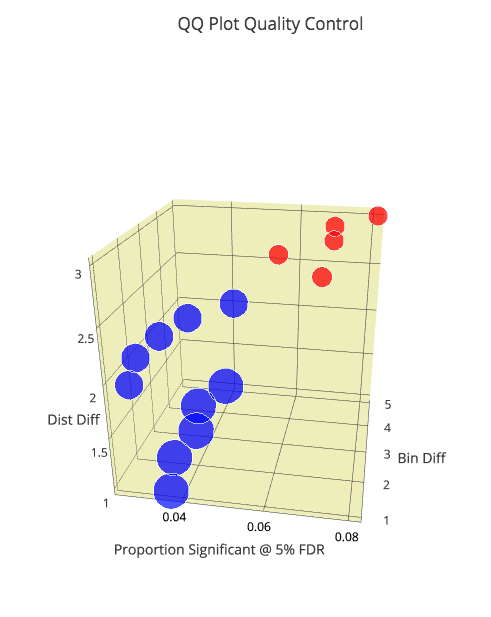
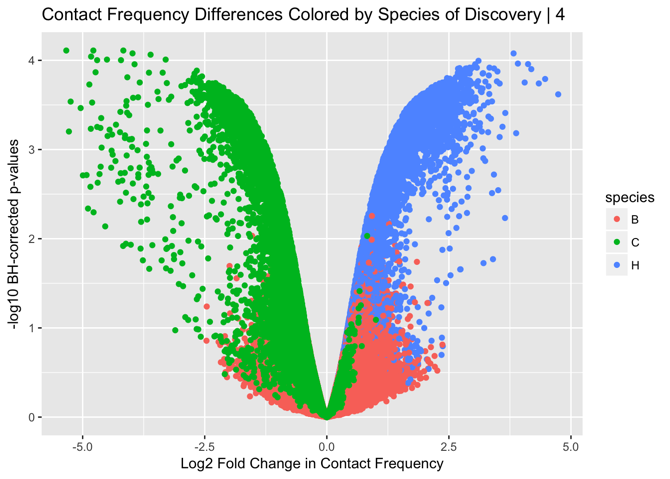
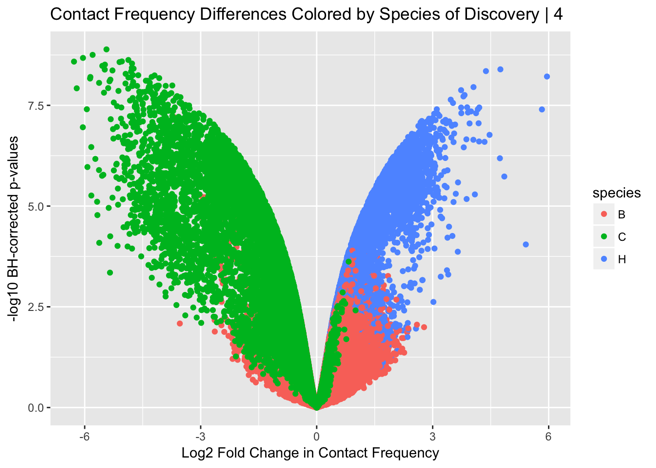
**Right panel:** Same as in left panel, but zoomed in on the y-axis to visualize finer-scale variation. A distinct monotonic trend is observed, with reductions in variance as an interaction is independently discovered in more individuals. This reduction appears to have diminishing returns after independently calling a Hi-C contact as statistically significant in more than four individuals, so a cutoff of discovery in at least four individuals was chosen to filter the data.

**Supplementary Figure 3. Volcano plot asymmetry quality control.**

**A:** Volcano plot showing log2 fold change in interaction frequency on the x-axis, and Benjamini-Hochberg corrected –log10 p-values on the y axis. Hits are colored by the species in which they were discovered significant in Homer (B-both, C-chimpanzees, H-humans). This plot shows data only filtered for independent discovery in at least 4 individuals. Strong asymmetry is observed, with many more inter-species significantly different interaction frequencies showing up on the left side, indicating overall increased contact in chimpanzees as compared to humans. Since this makes little biological sense, we sought to find technical differences that may explain this asymmetry.

**B:** 3-dimensional plot of sets of Hi-C contacts, with proportion of contacts significant at 5% FDR from linear modeling of interaction frequency on species on the x-axis. Contacts are binned by mate-pair distance differences on the y-axis, and bin size differences on the z-axis. Both of these issues may arise when using liftOver to convert the genomic coordinates of significant Hi-C contacts in one species to genomic coordinates for the other species. While changes in bin size appear to make marginal differences, large changes in the distance between mate pairs (≥20kb) created a noticeable inflation in the proportion of contacts significant at 5% FDR for linear modeling of interaction frequency on species. As this proportion clearly exceeded 5%, these hits were filtered out (red circles). Circle size indicates size of the set of Hi-C contacts falling into the given criteria.

**C:** Volcano plot as in A, but after removing contacts with large mate-pair distance differences across the species induced by liftOver. Notice the stark asymmetry seen in (A) has now disappeared. (I’m now also noticing this is on a drastically different scale on the y-axis as compared to the volcano plots before filtering—something must be wrong with that, I’ll check it out).



**B**

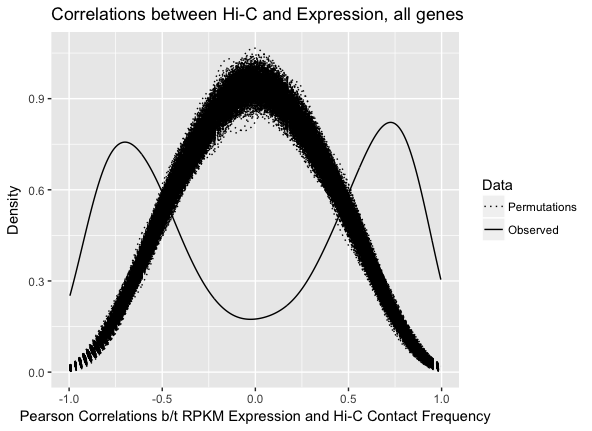
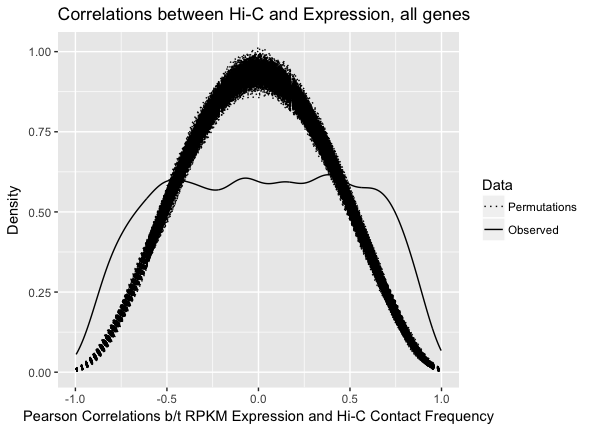
**C**

**A**

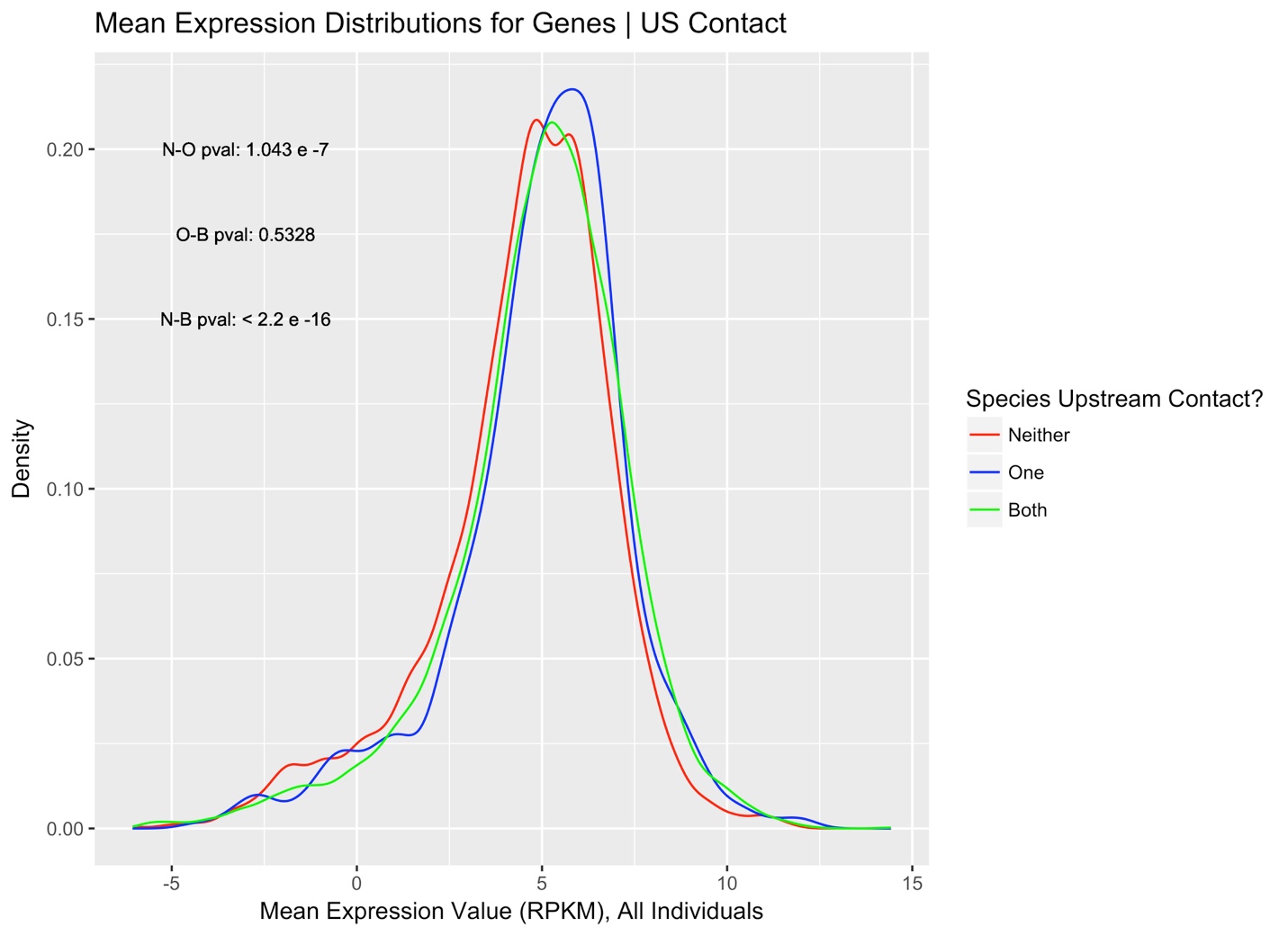
**Supplementary Figure 4. Correlations between Hi-C and Expression.**

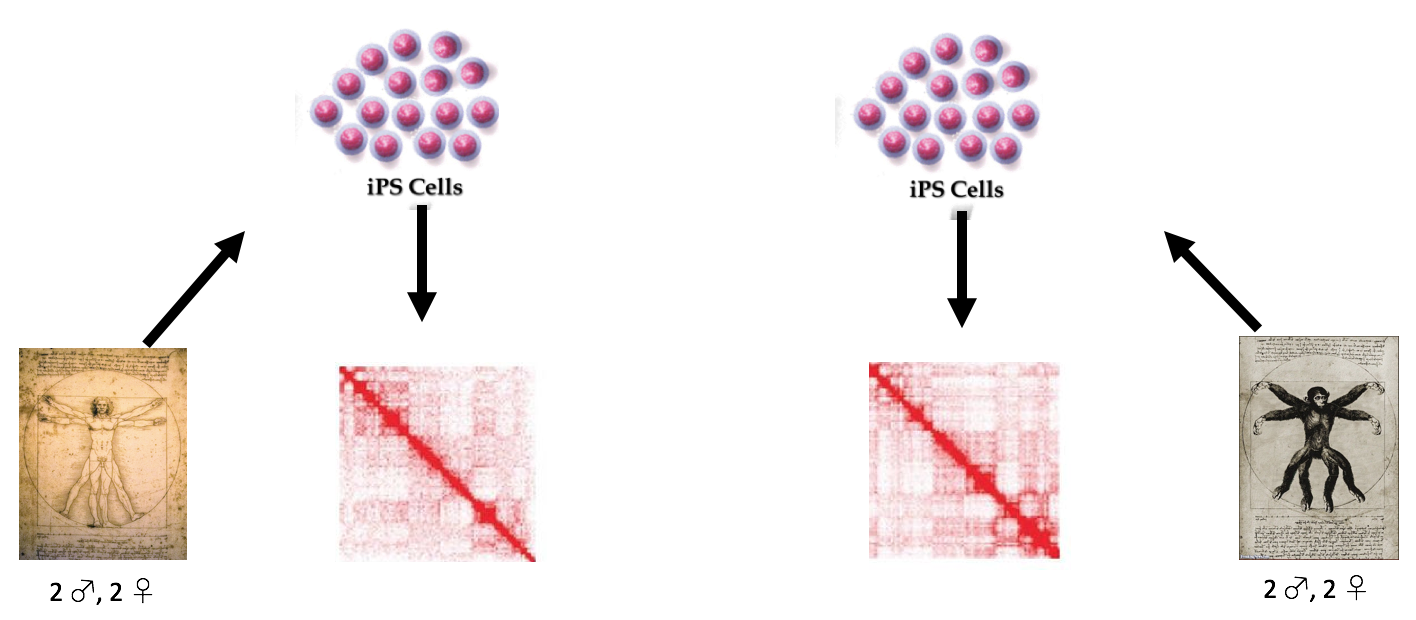
**Left:** Distribution of Pearson correlations between RPKM expression values and log2 Homer-normalized contact frequencies are shown along the x-axis, with density on the y-axis. Solid line indicates observed data, dotted lines are for 1000 permutations. The Hi-C contact frequency being used here is once again that which represents the minimum FDR contact from linear modeling of contact frequency on species. Reassuringly, correlations for the permuted data have a stronger peak at zero than those for the observed data.

**Right**: Similar to panel on left, but this time only for differentially expressed (DE) genes. A strong bimodal distribution of correlations between expression and contact suggests many instances where a contact difference between the species can lead to an increase (enhancer) or decrease (suppressor) of expression in the species where the contact is stronger. Once again, permuted data show a strong peak at zero, indicating this is not merely a technical artifact. (yes I know the title for this panel is still mis-labeled)



**Supplementary Figure 5. Presence of a statistically significant upstream contact is associated with increased expression.** Shown along the x-axis are mean RPKM expression values across all eight individuals, with density displayed on the y-axis. Genes are stratified by color to indicate whether they were found to have a statistically significant Hi-C upstream contact in neither, one, or both species. Pairwise p-values for a t-test of significantly different means between the distributions are shown in the top left (N-O, neither and one; O-B, one and both; N-B, neither and both). The observation of very small p-values for comparisons of genes without any statistically significant upstream contact (neither) to those with such a contact in one or both species suggests a that having a strong upstream contact is associated with higher expression levels. This idea is corroborated by the insignificant p-value found between genes that have an upstream contact in only one species and those that have an upstream contact in both species.





**Figure 1. Study design.**

A schematic of the study design. Two sex-balanced groups of four independent iPSC lines, one from humans and the other from chimpanzees, were subjected to Hi-C. Cell cultures were grown and libraries were prepared in two batches balanced for both sex and species, mitigating any potential batch effects. RNA-seq data previously collected from the same individuals was overlaid with the Hi-C data to interrogate the impact of divergence in regulatory landscape on expression divergence between the species.

It’s nice to have this when it shows the batches, as it stands this isn’t adding any info or helping people grasp this visually more quickly—should show that as a study design figure, but in the supplement since it’s pretty easy to figure out.