Genome Biology

RESEARCH

Open Access

The relationship between regulatory changes in *cis* and *trans* and the evolution of gene expression in humans and chimpanzees

Kenneth A. Barr[1†], Katherine L. Rhodes[1†] and Yoav Gilad[1,2*]

†Kenneth A. Barr and Katherine L. Rhodes contributed equally to this work.

*Correspondence: gilad@uchicago.edu

[1] Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[2] Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Abstract

Background: Comparative gene expression studies in apes are fundamentally limited by the challenges associated with sampling across different tissues. Here, we used single-cell RNA sequencing of embryoid bodies to collect transcriptomic data from over 70 cell types in three humans and three chimpanzees.

Results: We find hundreds of genes whose regulation is conserved across cell types, as well as genes whose regulation likely evolves under directional selection in one or a handful of cell types. Using embryoid bodies from a human-chimpanzee fused cell line, we also infer the proportion of inter-species regulatory differences due to changes in *cis* and *trans* elements between the species. Using the *cis*/*trans* inference and an analysis of transcription factor binding sites, we identify dozens of transcription factors whose inter-species differences in expression are affecting expression differences between humans and chimpanzees in hundreds of target genes.

Conclusions: Here, we present the most comprehensive dataset of comparative gene expression from humans and chimpanzees to date, including a catalog of regulatory mechanisms associated with inter-species differences.

Keywords: Comparative genomics, Gene regulation, Gene expression, Chimpanzee

Introduction

Comparative functional genomic studies in primates reveal insight into the evolution of gene regulation and help us identify genes and pathways that are associated with species-specific traits [1–8]. In particular, studies of gene expression in humans and other apes allow us to identify regulatory changes that may be associated with human-specific adaptations [6, 9–11]. Comparative inter-species studies also allow us to identify patterns of regulatory variation that are consistent with the action of natural selection. Such patterns suggest functional importance and point to regulatory phenotypes that may affect fitness. The inference of fitness-related molecular function is profoundly important, not only to studies that attempt to understand the mechanisms of evolutionary change, but also to studies of the genetic and gene regulatory basis for complex traits and diseases in humans [5, 8, 11–17].

One goal of comparative genomic studies in primates has been to characterize gene regulatory variation across a wide range of tissue types. Comparative genomic studies are not typically hypothesis-driven; rather, they aim to build comprehensive comparative catalogs that can be used to infer function, explore the evolution of different regulatory mechanisms, and establish hypotheses based on causal inference between genotypes and phenotypes [2–5, 7–9, 11]. Yet, more than two decades after the first genome-scale comparison of gene expression in humans and chimpanzees [18], comparative genomic data from apes remains limited to just a handful of tissue types [6].

For obvious ethical and practical reasons, only a small number of tissue types can be accessed from live humans, and direct experimentation in vivo is impossible. While in vivo studies are permitted in some primates, ethical considerations forcefully apply to studies of apes, particularly to studies involving chimpanzees — our closest extant evolutionary relatives. Comparative studies in apes must therefore rely on opportunistic collection of post-mortem tissues. Efforts to collect a broad array of post-mortem tissues from humans have been quite successful [19], but comparative genomic studies in humans and chimpanzees have only examined about a dozen different tissues. Moreover, the sample size in nearly all comparative studies in apes is quite modest: most studies include just 4–6 donors from each non-human species [4, 5, 7, 17, 20–33], and in some cases, only a single sample from the non-human species was available [18, 34–36]. In the handful of comparative studies that examine more than one tissue in apes, multiple tissues are rarely sampled from the same individuals [18, 23, 25, 27, 30, 33, 35, 37], resulting in a severe confounding of tissue and individual effects on the observed variation in gene regulation (a confounding effect that is particularly severe due to the small sample size available from each tissue [7]).

Even if we had unlimited access to post-mortem frozen tissues from apes, we would still be unable to study dynamic gene regulation, let alone study how different tissues and cell types respond to environmental exposures. Induced pluripotent stem cells (iPSCs) from apes provide a way to study gene regulatory dynamics during differentiation or in response to external perturbations [38], but the inefficiency of differentiation protocols limits the number of individuals, cell types, and contexts that can be queried in a single experiment. Thus, the scope of comparative studies in apes has remained narrow: we are unable to study gene regulation in more than a few tissues from apes; we are generally unable to collect samples from enough individuals to map and study regulatory quantitative trait loci (QTLs) in non-human apes; we are unable to study the dynamics of gene regulation during development; and we are unable to study regulatory responses to evolutionarily and clinically relevant exposures.

If we are ever to unleash the power of the comparative paradigm in apes, we require an in vitro system that allows us to dynamically study gene regulation in dozens of different cell types and individuals, and in hundreds of different contexts. Here, we show that embryoid bodies (EBs) can potentially fulfill these requirements. EBs are dynamic, iPSC-derived organoids that contain spontaneously and asynchronously differentiating cells from all three germ layers [39].

We recently used single-cell RNA-sequencing (scRNA-seq) to characterize cellular and gene expression heterogeneity in a panel of human EBs, which revealed the presence of dozens of differentiating and terminal cell types from multiple different tissues [40]. In the present study, we applied scRNA-seq to EBs from three humans and three chimpanzees. Despite the small size of our sample, we were able to assemble a comparative catalog of gene expression data from more than 70 different cell types per species, including differentiating cell types that have never been accessible from frozen tissues. This catalog contains nearly an order of magnitude more cell types than have been examined before, and three times more comparative data from humans and chimpanzees than all previous studies combined.

Using gene expression data from EBs, we identify hundreds of genes that are differentially expressed across dozens of cell types, as well as genes that are broadly conserved. We analyze tissue-restricted patterns of differential expression between the species, implicating a handful of candidate genes that may underlie functional differences between humans and chimpanzees. In addition, we quantify the contribution of *cis* and *trans* changes to inter-species differences in gene expression using EBs derived from a fused human-chimpanzee iPSC line.

Results

Generation of human and chimpanzee embryoid bodies

We generated EBs from three human YRI (Yoruba in Ibadan, Nigeria) and three western chimpanzee iPSC lines that were previously validated [38, 41]. From each line, we established EBs in three independent replicates (Fig. 1A and Additional file 1: Figure S1; see the "Materials and methods" section). We also included a single replicate of EBs from three additional human lines of European ancestry to confirm that inter-species differences are not population-specific.

We formed EBs in micropatterned plates and cultured them for 2 days before transferring them to suspension culture. On day 21, we dissociated EBs into single-cell suspensions and collected single-cell transcriptomes using 10x Genomics scRNA-seq. To prevent systematic biases in transcriptomes due to differences in genome annotation, we aligned reads to a curated set of orthologous exons in the human and chimpanzee genomes. After filtering (see the "Materials and Methods"), we retained data from 115,269 cells (a median of 5385 cells from each replicate) in which we detected the expression of a median of 3149 genes (Additional file 1: Table S1).

We first verified the presence of cells from the three germ layers in all samples by considering the expression of *SOX17*, *HAND1*, and *PAX6* as markers for endoderm, mesoderm, and ectoderm, respectively (Additional file 1: Fig. S2A). In all samples, we also identified cells expressing *POU5F1* (Oct4), indicating the presence of undifferentiated cells (iPSCs). At low resolution, we identified 11 clusters of cells using Louvain clustering in Seurat (Additional file 1: Figure S2B). In all clusters, we found cells from both species, with most clusters showing a similar proportion of human and chimpanzee cells (Additional file 1: Figure S2B and Table S2). We confirmed that cell identity is the major driver of variation in expression (Additional file 1: Figure S2C).

To identify cell types within EBs, we considered human and chimpanzee gene expression data together with previously generated reference data containing cells from EBs,

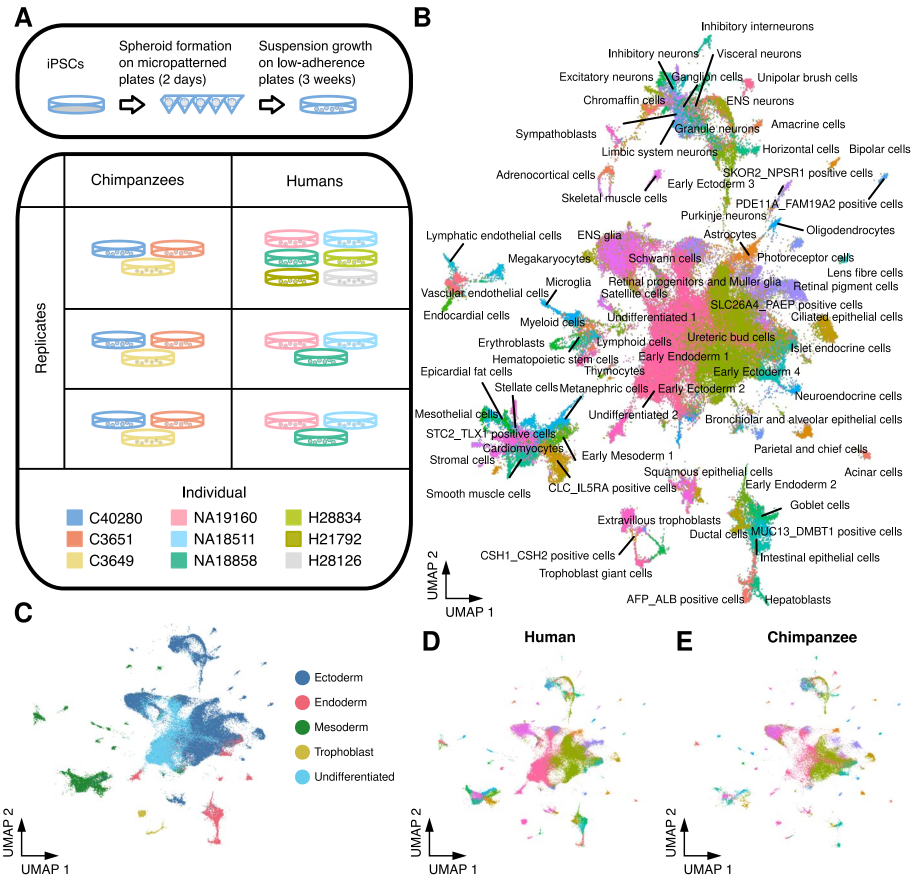

Fig. 1 Human and chimpanzee embryoid bodies are composed of 86 cell types from all three germ layers. **A** Overview of experimental design. EBs were generated in three replicates of six to nine human and chimpanzee iPSC lines. After two days in micropatterned plates, EBs were grown in low-adherence plates for three weeks before dissociating to single cells and sequencing with 10x scRNA-seq. **B** UMAP projection of all human and chimpanzee single cells after reference alignment. Cells are colored and labeled with the reference-assigned cell type. **C** The UMAP in **B** colored by germ layer of each cell type. The UMAP in **B** with only **D** human or **E** chimpanzee cells

human embryonic stem cells (hESCs), and primary fetal tissues [40, 42, 43]. We labeled each of our cells based on its relative position within this reference (Fig. 1B), which allowed us to partition cells into meaningful groups for subsequent analyses. In total, we identified 86 cell types present in both human and chimpanzee EBs (77 labels transferred from the reference and nine cell types identified de novo; see Additional file 1: Table S3). These represent undifferentiated cells, cells from each of the three germ layers, and cells that closely resemble extra-embryonic cell types (Fig. 1C). Cells from both species are well-distributed across all cell types (Fig. 1D, E).

Differential expression between humans and chimpanzees in 72 cell types

Next, we characterized differential expression (DE) between humans and chimpanzees in individual cell types. We excluded cell types that did not contain at least five cells in at least two replicates. We generated pseudobulk expression data for each of the remaining 72 cell types and performed DE analysis using a linear mixed model implemented in

DREAM [44]. In addition to modeling the effect of species, we included random effects for replicate and individual. At a 5% false discovery rate (FDR), we identified DE genes in 70 cell types. Across all 70 cell types, we identified 59,940 instances of DE genes between humans and chimpanzees; 10,457 (72% of all tested) genes are DE between the species in at least one cell type, with a range of two (chromaffin cells) to 3712 (early ectoderm) DE genes in individual cell types (Additional file 1: Table S4 and Additional file 2: Data S1). Effect sizes (log2 fold-change in expression) for DE genes range from -8.9 to 8.4, with a mean absolute effect size of 1.4 (Additional file 1: Figure S3 and Figure S4).

We explored the extent to which genes show inter-species DE in more than one cell type. To account for incomplete power to detect DE (given that we have only three individuals per species), we used Cormotif [45] to jointly analyze data from all 72 cell types simultaneously. This allowed us to share information across cell types, improving the power and accuracy of DE analysis. We identified 13 correlation motifs (cormotifs) of DE across cell types (Additional file 1: Table S5 and Additional file 2: Data S1). At a posterior probability of 95%, we identified between 27 and 7095 DE genes in each cell type, with a median of 2718. 10,783 genes (74%) were DE in at least one cell type (Additional file 1: Table S4). Compared to the underpowered single-tissue analysis, the mean number of cell types in which each gene was classified as DE increased from 4.1 to 14.7.

Examining the Cormotif results (Fig. 2A), we found that the two largest classes of genes are those that are DE between species in all cell types (cormotif 11) and those that are not DE in any cell type (cormotif 1). These patterns of DE could indicate different degrees of selective constraint. Indeed, we found that the number of cell types in which a gene is classified as DE between species is significantly associated with coding sequence similarity between humans and chimpanzees (Fig. 2B; $p = 2.7 \times 10^{-6}$). The ratio of non-synonymous to synonymous coding mutations across mammals (dN/dS) and the loss-of-function observed/expected upper bound fraction (LOEUF) [46] are also positively associated with the number of cell types in which a gene is classified as DE (Fig. 2C, D; dN/dS $p = 10^{-13}$; LOEUF $p < 10^{-15}$). We also found that the average effect size correlates positively with LOEUF and dN/dS and negatively with coding identity (Additional file 1: Figure S5). These observations are consistent with the notion that broad gene expression differences between humans and chimpanzees are generally associated with relaxation of evolutionary constraint. For a subset of broadly DE genes, it may be the case that directional selection has favored a new regulatory pattern. In support of this, we found that genes targeted by human accelerated regions (HARs) are also DE in a greater number of cell types (Fig. 2E; $p = 3.7 \times 10^{-5}$).

Genes with conserved expression levels across all cell types are of particular interest, as they may shed insight on core functions in both species. Yet, the identification of these genes is confounded by statistical power. To confidently classify these genes, we filtered non-DE genes by their absolute expression levels (Fig. 2F). We classified hundreds of non-DE genes across all cell types, even at relatively stringent filters. For example, non-DE genes with at least 100 UMIs in at least 10 cell types in both species are enriched for core cellular processes, including protein transport and mRNA processing, splicing, and transport (Additional file 1: Table S6). As an alternative method of identifying non-DE genes across species, we assessed the probability that each gene has an effect size smaller than 0.5 ("Materials and methods" section; Additional file 2: Data S1). Again, we

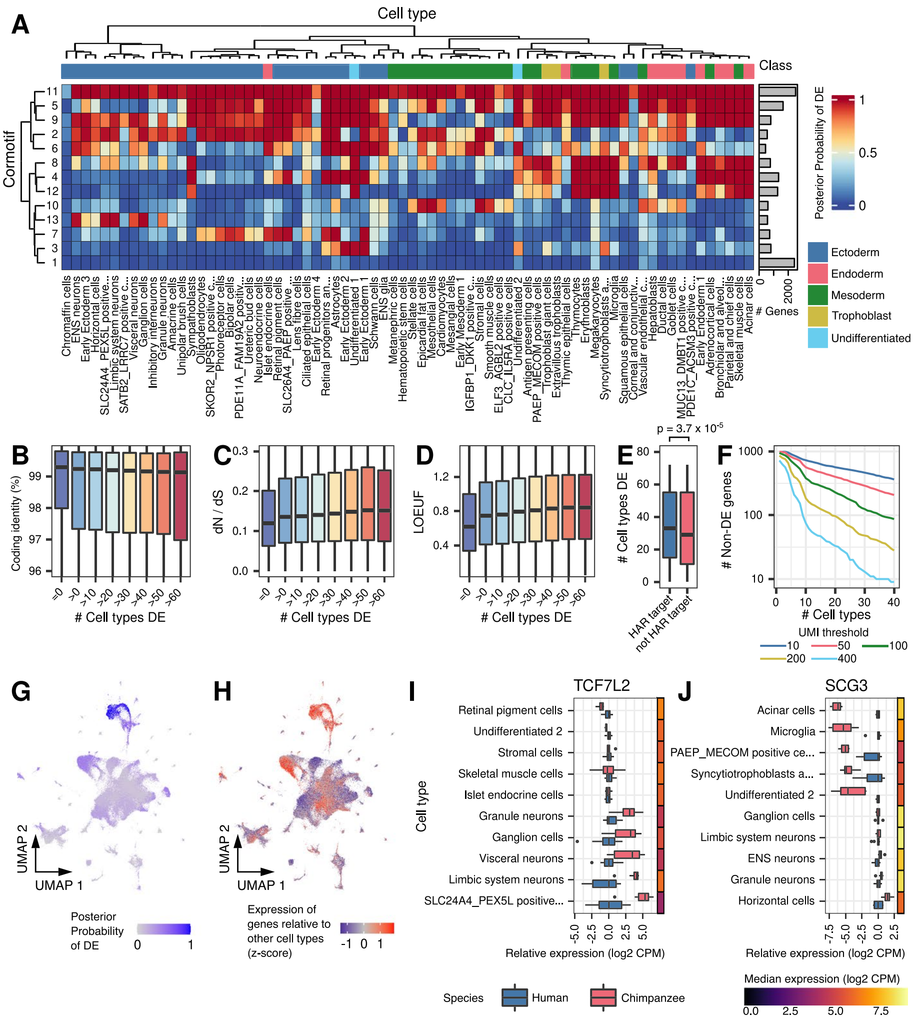

Fig. 2 Sharing of differential expression across cell types. **A** Heatmap of correlation motifs across cell types. Cells are shaded according to the posterior probability of differential expression for each cormotif and cell type. Columns are clustered by mean gene expression. Inferred germ layers are given in Data S1. Boxplots of **B** Percent coding identity, **C** dN/dS, or **D** LOEUF in genes that are DE in zero or increasingly large numbers of cell types at posterior probability 0.5. **E** Boxplot showing the distribution of the number of cell types in which genes are DE for HAR targets and non-HAR target genes. **F** The number of non-DE genes remaining, requiring the specified absolute expression cutoffs in at least the given number of cell types. UMAP from Fig. 1B is colored to show the **G** posterior probability of DE for genes in cormotif 13 or **H** the expression of genes in cormotif 13 relative to other cell types (*z*-score, "Materials and methods" section). Boxplots show expression relative to humans in select cell types for **I** *TCF7L2* or **J** *SCG3*. Adjacent heatmaps show the median expression across all samples from both species

identified hundreds of genes with minimal changes in expression between species across dozens of cell types (Additional file 1: Figure S6).

Beyond genes that are DE in no cell types or all cell types, many of the remaining DE cormotifs have tissue-specific patterns. Inter-species DE in specific tissues could be driven by tissue-specific expression patterns, where genes may be DE everywhere they are expressed. In other cases, a gene may have acquired DE in a restricted set of cell types, despite being expressed in additional cell types. To examine these possibilities, we considered the expression levels of the genes in each cormotif across all cell types and compared them to patterns of DE (Fig. 2G, H and Additional file 1: Figure S7). We focused on cormotifs exhibiting tissue-specific DE where the genes are highly expressed but conserved in other cell types. We then identified genes with cell type-restricted acquisition of DE (see the "Materials and methods" section). We identified 3906 genes (27%) that are broadly expressed but show cell-type-specific DE between humans and chimpanzees (Additional file 3: Data S2).

Tissue-restricted DE could underlie important functional differences between humans and chimpanzees. For instance, the transcription factor (TF) TCF7L2 is a WNT modulator associated with human psychiatric and metabolic disorders [47]. *TCF7L2* is broadly expressed in EBs, but inter-species DE is restricted to neuronal subtypes, where it may underlie inter-species differences in neurodevelopment (Fig. 2I and Additional file 1: Figure S8A-B). Another example is *SCG3*, an obesity-associated gene [48] for which inter-species DE is restricted to a few cell types, including acinar (pancreatic) cells (Fig. 2J and Additional file 1: Figure S8C-D). This cell-type-specific pattern could underlie inter-species differences in digestive activity.

Generation of EBs from tetraploid human-chimpanzee hybrids

Next, we sought to further characterize mechanisms of inter-species DE. To quantify the contribution of *cis* versus *trans* regulatory divergence to inter-species differences in gene expression, we sequenced single EB cells from a human-chimpanzee fused stem cell line [13, 49]. We generated 65,247 fused single-cell transcriptomes from three replicates. Upon mapping these cells to the reference data, we observed cells from all 86 previously identified cell types (Fig. 3A). We identified DE genes in each cell type using a paired Wilcoxon signed-rank test. At a 5% FDR, we observed DE in 67 cell types, with 13,842 genes classified as DE in at least one cell type (Additional file 1: Table S4).

Fused tetraploid cells contain the complete genome of both the human and chimpanzee parental lines; thus, genes from both species share the same *trans* environment. We estimated the proportion of gene expression divergence due to *cis* by comparing the change in gene expression observed between human and chimpanzee alleles in the tetraploid cells, to the change in gene expression observed between the original human and chimpanzee diploid cells (Fig. 3B). Inter-species gene expression differences that cannot be explained by *cis* are assumed to be driven by changes in *trans*. Thus, our measurement of changes in *trans* includes differences in the *intra*- and *extra*-cellular environment, as well as noise and measurement error. Systematic differences in cell composition between species could potentially create different signaling and metabolic environments and confound our *trans* estimates. To test this, we fit a linear mixed model including covariates for cell type proportion in each sample ("Materials and methods" section). We found that cell type composition explains only a small percent of variation and does not significantly affect our estimates of effect size (Additional file 1: Figure S9A-B).

On average, we estimated that 70% of inter-species changes in gene expression are due to changes in *trans* elements (Fig. 3C), though as we pointed out, measurement error is expected to contribute to this estimate. Indeed, when we restrict this analysis to the 1000 tests with the lowest error in the estimate of effect size, the proportion of *trans* regulation is decreased, as expected, to 62% (Fig. 3C). We found that changes in *trans* elements are associated with larger effect sizes (Fig. 3D; $p < 10^{-15}$).

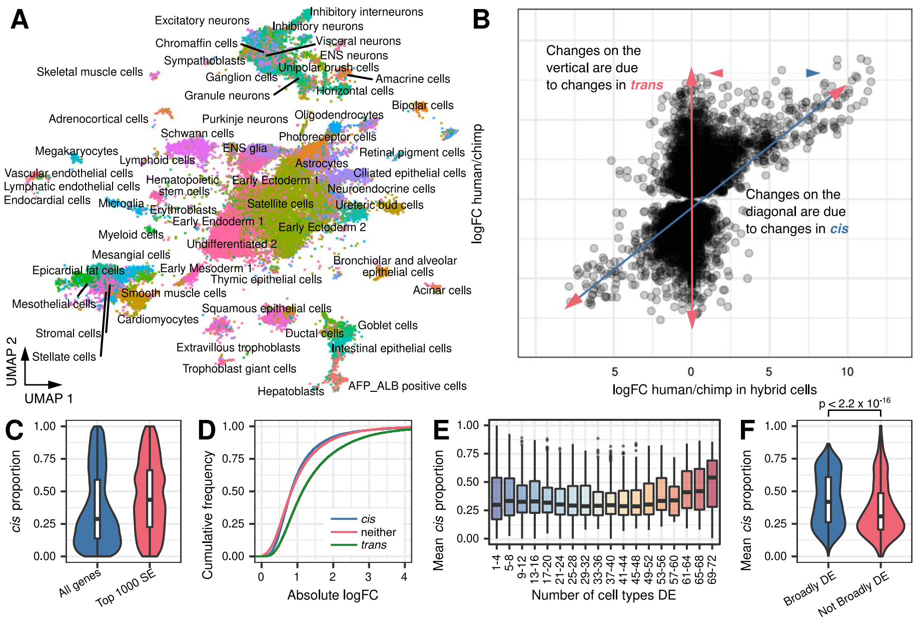

Fig. 3 Allele-specific expression in tetraploid hybrid cells partitions differential expression into *cis* and *trans* components. **A** UMAP of hybrid cells after projection onto the space of Fig. 1B. **B** DE gene effect sizes in the hybrid vs. the effect sizes across humans and chimpanzees (for genes DE at $p < 0.001$). **C** Boxplot of estimated *cis* proportion in the top 1000 tests with the lowest estimate of standard error versus all other genes. **D** Empirical cumulative distributions of effect size (log2 fold-change from DREAM) for tests with high *cis* proportion (> 0.75), low *cis* proportion (< 0.25), or neither. **E** Boxplot of distribution of *cis* proportion for genes DE in increasingly large sets of cell types at posterior probability > 95%. **F** Boxplot of *cis* proportion for genes DE in more than 80% of tested cell types (at posterior probability > 95%) or all other genes

We generally expect *trans* changes — particularly pleiotropic changes that affect multiple cell types — to be deleterious, and therefore, rare. However, a *trans* change that affects only one cell type is more likely to persist. Consistent with this expectation, we found that *trans* changes are depleted from genes that are broadly DE between the species — that is, the number of cell types in which a gene is classified as inter-species DE is negatively correlated with the proportion of DE that is explained by differences in *trans* (Fig. 3E, F; $p < 10^{-15}$).

We proceeded by arbitrarily classifying DE genes as *cis* if the mean estimated *cis* proportion across DE cell types was greater than 75%, or *trans* if the estimated *cis* proportion was less than 25%. We found that *cis* genes are associated with higher coding sequence divergence (Fig. 4A; $p = 2 \times 10^{-6}$), higher dN/dS ratios (Fig. 4B; $p = 0.01$), and higher LOEUF (Fig. 4C; $p = 10^{-15}$) relative to *trans* genes. *Trans* genes are significantly enriched for functions related to development and morphogenesis (Additional file 1: Table S7), and for many miRNA and TF binding motifs (Additional file 1: Table S8), the latter of which suggests a link between changes in expression and upstream regulators. For instance, *CLDN4* has a *trans* DE proportion of 93% and is a predicted target of the TF *CEBPA*. Consistently, the pattern of DE across cell types for these two genes is highly correlated (Fig. 4D, E and Additional file 1: Figure S10; $\rho = 0.88$, $p = 0.007$).

In order to more broadly explore the link between changes in expression and upstream regulators, we identified a set of 179 sequence motifs (of 610 from [50]) that are enriched in the promoters of *trans* genes (hypergeometric test; FDR ≤ 0.05). These 179 motifs

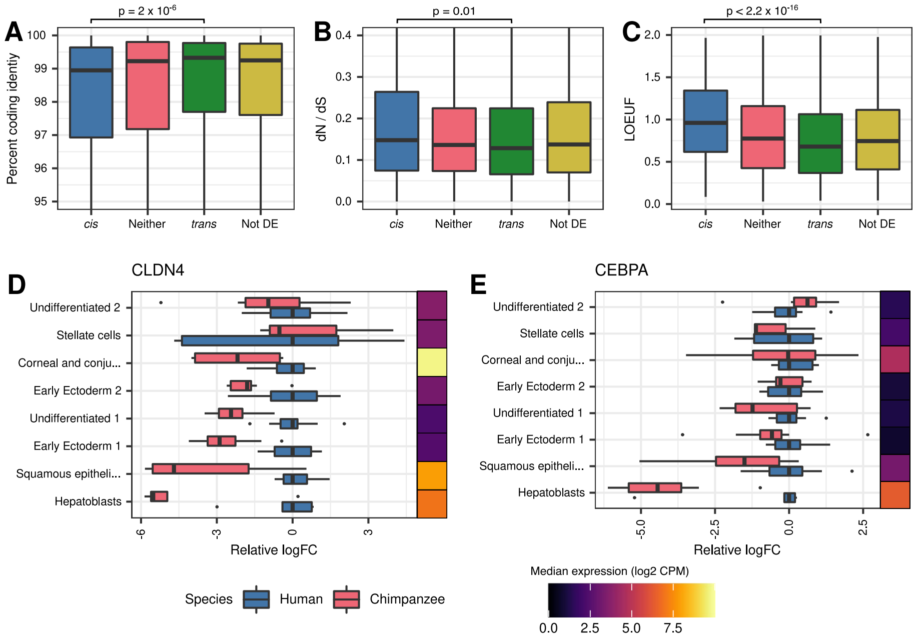

Fig. 4 Genes with high *cis* proportion show evidence of relaxation of constraint. Boxplot showing the distribution of **A** percent coding identity **B** dN/dS or **C** LOEUF for genes with a high, moderate, or low *cis* proportion. Boxplots showing diploid expression relative to mean human expression in select cell types for **D** *CLDN4* or **E** *CEPBA*. Adjacent heatmaps show the median diploid expression across all samples from both species

correspond to binding sites for 127 TFs, 66 of which are expressed in at least three cell types in EBs, and are DE between the species in at least one cell type (Additional file 4: Data S3). In cases where these TFs drive inter-species DE, we expect the magnitude of DE of the downstream gene target to be correlated with that of the upstream TF gene. Indeed, when we correlated the pattern of inter-species DE of each TF with every gene that contains the respective binding motif in its promoter, 54 TFs showed significant enrichment of smaller p-values than expected (Pearson r; Fisher's combined test; FDR $\leq$ 0.05; Additional file 4: Data S3). The inter-species DE pattern of these 54 TFs showed a nominally significant correlation with 3580 putative targets (Pearson r; $p \leq 0.05$; Additional file 4: Data S3). At a more conservative 10% FDR, there are still 511 genes whose inter-species DE pattern significantly correlates with 44 TFs (Additional file 4: Data S3).

Discussion

We used EBs to create the most comprehensive comparative catalog of gene expression levels in humans and chimpanzees to date, including single-cell data from 86 developing and mature cell types across multiple germ layers and tissues. Importantly, the EB model system allowed us to generate each of these cell types from every human and chimpanzee individual using uniform processing, such that we did not confound cell type with individual, species, or processing method in our analysis. Our balanced study design allowed us to explore gene regulatory divergence at an unprecedentedly high cellular and temporal resolution.

We confirmed that gene regulatory divergence is ubiquitous across cell types, even in closely related species. More than two thirds of tested genes are DE in at least one cell type, and in the majority of cell types, there are hundreds to thousands of DE genes between the species. We identified thousands of genes that are DE in a subset of the cell types in which these are expressed, which suggests that there are quite a few functional changes in tissue-specific gene regulation between the two species.

We were somewhat surprised to find a large subset of genes whose expression is conserved in all of the cell types we were able to study. We did not expect such a large fraction of genes to be regulated in a similar way in humans and chimpanzees in over 70 different cell types, as this requires that the regulation of these genes is of sufficient functional importance in all of these cell types to evolve under relatively strong constraints. Of course, it is possible that with data from more individuals, we would find that a subset of the "generally conserved" genes are in fact DE between the species in some of the cell types; nevertheless, this observation is intriguing and suggests that we should reconsider our assumptions regarding the degree of context-specific gene regulation.

Turning our attention to regulatory mechanisms, the use of fused cell lines allowed us to assess the relative contribution of *cis* and *trans* changes to regulatory evolution – an approach that has not been feasible in primates until recently. We found that *cis* changes are common and can impact any number of cell types, as expected given the potential for spatial and temporal specificity of *cis* elements. Our observation that *trans* changes are associated with inter-species DE in fewer cell types than *cis* changes may seem counterintuitive, because one expects *cis* changes to be more specific than changes in *trans*. However, *trans* changes are more likely to be associated with pleiotropic deleterious outcomes than changes in *cis*. Thus, while *trans* changes are likely more exposed to negative selection, they may persist if their impact is restricted to a small number of cell types and thus drive the observed correlation between the breadth of DE and the proportion of *trans*.

We also found that fused lines help to identify putative regulatory relationships between genes. By combining correlated patterns of DE across cell types with binding motif enrichment in the promoters of *trans*-regulated DE genes, we demonstrated that *CEBPA* may drive tissue-specific differential expression of *CLDN4*. We extended this analysis to identify 54 differentially expressed TFs that may have highly pleiotropic effects. These TFs have tissue-specific DE patterns that are nominally correlated with target genes, representing approximately a third of all DE genes. Of these 54 TFs, however, only 11 have an estimated *cis* proportion greater than 50%. This indicates that there are more upstream drivers of DE that are not detected through this approach. While this analysis was restricted to promoters, future studies including chromatin accessibility data could extend the motif search into *cis* regulatory elements. Similarly, this work was restricted to counts of polyadenylated transcripts, and thus cannot assess the role of alternative regulatory mechanisms such as splicing or microRNA expression. Future studies in hybrids could explore the contribution of these alternative mechanisms to *cis* and *trans* differences in expression.

Collectively, our work validates EBs as a powerful model system for comparative studies of gene expression in humans and chimpanzees. The EB system confers several advantages over existing models. First, the use of EBs sidesteps the many pitfalls of working with frozen tissues, which have severely limited the scope of comparative studies in primates. Frozen tissues from non-human apes must be collected opportunistically, making it difficult to sample broadly across tissues, obtain samples from more than a handful of individuals per species, and design unbiased studies that are balanced with respect to species, tissue, and individual [7]. Moreover, frozen tissues cannot be used to study how gene regulation changes over time or in response to external perturbations. While iPSC-derived cell types and organoids provide a way to explore gene regulatory dynamics in vitro [12, 14, 16, 49, 51], EBs can be used to efficiently generate multiple cell types with a fraction of the effort it takes to differentiate a single iPSC line. Indeed, EBs permit the study of gene regulation across dozens of cell types, individuals, developmental time points, and environmental exposures, which will allow future studies to explore gene regulatory divergence in apes in greater detail than has been possible thus far.

In particular, gene expression response phenotypes have not been widely explored in apes, even using in vitro models. Understanding how human and chimpanzee cells respond to external perturbations can reveal how gene regulatory divergence affects transcriptional responses that are relevant to human-specific traits and diseases and may reveal functionally important genes and pathways that affect fitness. In a previous study, we examined the immune response to infection in primary monocytes from humans, chimpanzees, and rhesus macaques [29]. We observed that while the regulatory response to bacterial infection is largely conserved between species, the regulatory response to viral infection is often lineage-specific. For example, we found chimpanzee-specific immune signaling pathways to be enriched for HIV-interacting genes, which could explain why HIV-infected chimpanzees exhibit relatively strong resistance to AIDS progression [52].

More recently, we used iPSC-derived cardiomyocytes from humans and chimpanzees to examine inter-species differences in response to hypoxia, a symptom of myocardial ischemia [12]. Despite differences in myocardial ischemia risk, heart development, and other cardiovascular traits between the two species, we observed that the response of human and chimpanzee cardiomyocytes to hypoxia was largely conserved. We also found that hypoxia-induced transcription factors bind more frequently to conserved hypoxia response genes than to chimpanzee-specific hypoxia response genes, suggesting that changes to transcription factors or their binding sites might underlie inter-species differences in hypoxia (and possibly, differences in myocardial infarction risk). In each of these studies, we identified functionally important genes and mechanisms by comparing the transcriptional responses of a single cell type to a disease-relevant exposure. The EB model system makes it possible to perform similar studies in dozens of cell types at once, including transient, developmental cell types that are not accessible from adult tissues.

In summary, EBs are a powerful in vitro model system that can be used to address a number of outstanding questions in primate comparative genomics. Our initial exploration of human and chimpanzee EBs resulted in the first comparative database of *cis* and *trans* regulatory divergence between the species, which includes gene expression data from more than 70 human and chimpanzee cell types. Future studies will take full advantage of this model by including additional human and chimpanzee cell lines, by studying how *cis* and *trans* changes interact to drive adaptation and divergence, and by exploring divergent transcriptional responses to biomedically relevant perturbations between the two species.

Materials and methods

Samples

In this study, we included six human and three chimpanzee iPSC lines. The three human lines were derived from unrelated individuals from the Yoruba population in Ibadan, Nigeria (YRI) as part of the International HapMap Collection. These lines were reprogrammed from lymphoblastoid cell lines and characterized in [41]. Two of these lines are female (18,511, 18,858) and one is male (19,160). We included three additional fibroblast-derived human iPSC cell lines characterized in [38]. Two of these lines are female (H28834, H21792) and one is male (H28126). We included three fibroblast-derived chimpanzee iPSC lines, also characterized in [38]. These also included two females (C40280, C3651) and one male (C3649). We obtained the tetraploid human/chimpanzee hybrid iPSC line (HLI-25) from Dr. Hunter Fraser.

iPSC maintenance

We maintained feeder-free iPSC cultures on Matrigel Growth Factor Reduced Matrix (CB-40230, Thermo Fisher Scientific) with StemFlex Medium (A3349401, Thermo Fisher Scientific) and Penicillin/Streptomycin (30002Cl, Corning). We grew cells in an incubator at 37 °C, 5% CO_2, and atmospheric O_2. Every 3–5 days thereafter, we passaged cells to a new dish using a dissociation reagent (0.5 mM EDTA, 300 mM NaCl in PBS) and seeded cells with ROCK inhibitor Y-27632 (ab120129, Abcam). Cells have been routinely screened for mycoplasma (ATCC 30-1012 K and Thermo Fisher M7006).

Embryoid body formation and maintenance

We formed EBs using the STEMCELL AggreWell™400 protocol according to the manufacturer's directions. Briefly, we generated a single-cell suspension of iPSCs by incubating them for 5 to 13 min with 300 mM NaCl in PBS, followed by gentle pipetting. We seeded iPSCs into an AggreWell™400 24-well plate at a density of 1000 cells per microwell. We cultured the aggregates in AggreWell™ EB Formation Medium (05893, STEMCELL) with ROCK inhibitor Y-27632 for 24 h. After 24 h, we replaced 1 mL of the media with fresh EB Formation Medium. After another 24 h, we harvested the EBs and placed them on an ultra-low attachment 6-well plate (CLS3471-24EA, Sigma) in E6 media (A1516401, Thermo Fisher Scientific). We replaced the media every other day for the next 19 days, for a total of 21 days of EB culture.

We formed EBs on three separate days. On the first day, we included all six human lines and all three chimpanzee lines. For the next two experimental replicates, we included all three chimpanzee lines, but only the three YRI human lines.

Embryoid body dissociation

On day 21 post-formation, we dissociated EBs into a single-cell suspension. We washed the cells in PBS (Corning 21–040-CV) and then incubated them in 1 mL of AccuMax dissociation reagent (STEMCELL 7921) for 10 min at 37 °C. After 10 min, we gently pipetted the EBs for 30 s using a clipped p1000 tip. We repeated this every 5 min until we achieved a single-cell suspension, up to a maximum of 35 min. Next, we stopped dissociation by adding 5 mL of ice-cold E6 media. We then filtered cells using a 40 μm strainer (Fisherbrand 22–363-547). We washed the cells three times using ice-cold PBS + 0.04% BSA. Prior to running the cells on the 10X Chromium controller, we mixed the cells together, including an equal number of cells from each individual.

Library preparation and sequencing

We generated scRNA-seq libraries using the 10X Genomics 3′ scRNA-seq v3.0 kit. We targeted 10,000 cells per lane of the 10X chip. In the first replicate, with nine individuals, we initially collected nine 10X lanes (~ 90,000 cells). For the remaining two replicates we collected four 10X lanes (~ 40,000 cells). In total, we generated nine libraries from the first replicate and four from each of the next two replicates. We performed all sequencing using 100 bp paired-end sequencing on a HiSeq 4000 at the University of Chicago Functional Genomics Core Facility. For replicate 1 libraries, we initially pooled libraries 1–2 and 6–9 and sequenced on one lane, and pooled libraries 1 and 3–5 on another lane. Our preliminary analysis indicated that library 2 was of poor quality and we excluded it from further analyses. Libraries 3, 4, and 5 contained fewer cells than expected, but otherwise appeared to be good quality. We re-pooled the eight remaining libraries using half the amount of libraries 3–5 to account for lower library complexity. We sequenced this pool on eight lanes of the HiSeq 4000. We also pooled and sequenced libraries from replicates 2 and 3 on one lane each of the HiSeq 4000. After preliminary analysis, we performed an additional six and four lanes of sequencing on the HiSeq400 for replicate 2 and replicate 3 libraries, respectively. For each replicate, the number of sequencing lanes was chosen such that we obtained 50% saturation as assessed by 10X cellranger software.

Orthologous exon reference generation

In order to make quantitative comparisons across species that are not biased by differences in mapping annotation quality across species, we generated a database of orthologous exons. We followed the general approach of [28, 53] but updated this database to use the most recent version of the chimpanzee genome assembly (Clint_PTRv2/panTro6). Briefly, we mapped the human genome annotation (hg38/Ensembl98) onto the chimpanzee genome using three separate BLAT steps. First, we used BLAT on every protein-coding human exon to identify the ortholog in chimpanzees. We kept exons that had at least 80% sequence identity. For exons with multiple locations at greater than 80% sequence identity, we retained the exon that shared the most neighboring exons (within 100 kb) with the original reference. We also removed exons with indels greater than 20% of the exon length. Next, we used BLAT to map these chimpanzee exons back to the human genome. We retained exons mapped back to their original location in humans. Finally, we used BLAT to map the chimpanzee exons to the chimpanzee genome. We retained exons that map back to the original location in the chimpanzee. The resulting reference contains 1,077,831 orthologous exons from 19,787 genes.

Alignment, species assignment, demultiplexing, filters, and normalization of human and chimpanzee libraries

We compiled three separate references using the cellranger *mkref* command from 10 × Genomics: one human, one chimpanzee, and one combined human and chimpanzee reference. Each reference was generated using the appropriate species annotation from the orthologous exon database described above. We used cellranger count to map reads to each of these three references. We used the combined genome reference to assign species to each barcode. We assigned a cell barcode to a species if greater than 90% of uniquely mapping reads mapped to that species' genome.

We used demuxlet to assign each droplet to an individual [54]. We used genotypes for NA19160, NA18858, and NA18511 from the 1000 genomes project [55]. Genotypes for human and chimpanzee individuals H21792, H28126, C3649, C3651, and C40280 were obtained from Dr. Gregory Wray. We generated genotype calls for the remaining individual, H28834, using RNA-sequencing data from [20] and the GATK best-practices pipeline for calling genotypes from sequencing data.

We filtered the data to remove droplets that had fewer than 1,000 genes detected. We also removed droplets where mitochondrial reads represented more than 20% of total reads or fewer than 0.1% of total reads. We processed the data using the SCTransform pipeline from Seurat [56] using 5000 variable genes and obtained Pearson residuals for all genes.

Reference integration and cell type assignment of human and chimpanzee cells

We have previously reported a joint reference containing cells from human EBs [40] as well as cells from a fetal cell atlas [42] and human embryonic stem cells (hESCs) [43]. We mapped the human and chimpanzee cells onto this reference. Briefly, we embedded the human and chimpanzee cells in the principal component (PC) space of the joint reference. Next, we used Harmony [57] to correct the PCs of the human and chimpanzee datasets while holding the reference PCs fixed.

In order to assign cell type identities from the fetal cell reference, we first assigned cluster midpoints (by taking the mean of each harmony-corrected PC) for each cell type in the fetal cell atlas as well as hESCs. We then assigned a cell identity to each cell in the human and chimpanzee EB dataset by finding the closest reference cluster midpoint by Euclidean distance. This resulted in a substantial proportion of cells (47%) that were assigned to hESCs, meaning they more closely resembled undifferentiated cells than anything within the fetal reference. These cells represent both undifferentiated cells and immature or transient developing cell types not present in either reference. To assign identities to these cells, we performed unsupervised clustering using the Louvain algorithm (in Seurat, using resolution 0.1). This resulted in 9 new cell types. We then reassigned every human and chimpanzee EB cell using the fetal cell reference midpoints, as well as midpoints for each of these 9 de novo cell types.

Alignment, filters, normalization, and cell type assignment of hybrid libraries

We performed two separate alignments of the tetraploid human-chimpanzee hybrid cell lines using cellranger count. We aligned all of the data to the standard human reference

(2020-A, obtained from 10 × Genomics) as well as the combined orthologous exon reference described above. We used the human-aligned data to filter, normalize, integrate with the reference, and assign cell identities. We removed droplets that had fewer than 2000 genes detected, as well as droplets with fewer than 0.1% or more than 20% of reads aligning to mitochondrial genes. We processed the data using SCTransform with 5000 variable genes, returning residuals for all genes. Using the procedure described above, we then aligned our data with the joint reference, including all human and chimpanzee data generated in this study. We assigned each hybrid cell the identity of the nearest reference cell by Euclidean distance in Harmony-corrected PC space.

We used data aligned to the combined human and chimpanzee reference for all quantitative comparisons. By default, cellranger discards sequencing reads that do not uniquely map to the genome. Therefore, reads that remain after alignment represent fragments that can be uniquely assigned to the allele of one species. We retained droplets that passed filters in the human-aligned data. We also transferred relevant metadata, including cell type identities, from the human-aligned data.

Differential expression between humans and chimpanzees with DREAM

In order to perform differential expression (DE) analysis, we generated pseudo-bulk expression by taking the sum of counts for all cells that came from the same individual, replicate, and cell type. We then removed any samples with 5 or fewer cells. For each cell type, we filtered genes using the *filterByExpr* function from the R package edgeR [58], setting the variable *min.count* to 5. We used VariancePartition to assess the effect of technical factors such as replicate, sex, and donor cell type (YRI LCL vs CEU fibroblast). We found minimal contributions from sex and donor cell type (Figure S2C). We then performed DE analysis using DREAM [44] separately in each cell type, setting species as a fixed effect and individual and replicating as random effects. We estimated size factors using the trimmed mean of M-values (TMM) with singleton pairing and we performed cyclic loess normalization.

Estimating the contribution of cell type proportion

To assess the contribution of cell type composition to changes in gene expression across species, we performed differential expression testing including PCs for cell type composition. For each sample, we generated a matrix that contained the proportion of cells that fell into each of the 86 previously defined cell types. We then performed PCA on this matrix. We found that the first two PCs explained 66% and 22% of the variation in cell type composition, respectively. Since these two PCs explained the vast majority (88%) of cell type composition differences, we included them as fixed effects in the linear mixed model we used to test for differential expression. We ran this analysis for 65 of the 72 cell types; the remaining seven cell types could not be analyzed due to high collinearity.

Next, we tested whether cell type composition is a major source of variation in genes that we previously identified as DE (5% FDR). We found that PC1 and PC2 explained a median of 2% and 1.3% of the variation in gene expression, respectively (Additional file 1: Figure S9A). We also found that the effect size estimates for these DE genes did not change significantly when the PCs were included in the model (Additional file 1: Figure S9B).

Joint differential expression with Cormotif

The package Cormotif [45] uses *t*-statistics from LIMMA [59], performed in multiple conditions, to share information and improve power of DE analysis. The software takes normalized gene expression measurements as input. This may be inappropriate for single-cell RNA-sequencing projects, where the number of cells per sample may vary by more than an order of magnitude. We modified the Cormotif package to accept UMI counts as input. To generate *t*-statistics, our modified package runs limma-voom [60] in order to apply precision weights to each sample. We used the same gene filters as we used for DREAM. We also used TMM and cyclic loess normalizations. The modified Cormotif package that implements this is available on github at https://github.com/kennethabarr/CormotifCounts.

Identification of genes effect size less than 0.5 across species

We calculated the confidence that any particular effect size is less than 0.5 using summary statistics from the linear mixed model implemented in DREAM. Using the t-statistic, standard error, and residual degrees of freedom from every test, we calculated the probability that the effect size was in the range (-0.5, 0.5) using the *pt* function in R. When we defined genes with conserved expression across multiple cell types, we excluded genes that have effect sizes greater than 0.5 in at least one cell type with probability 0.95.

Identification of cell type restricted differential expression

To visualize the expression patterns of genes within correlation motifs, we generated a cormotif score for each cormotif in each cell. For each cormotif, we first took the product of the SCTtransform residuals for each gene and the probability that each gene is a member of that cormotif. We then summed these cormotif membership-weighted SCTransform residuals for all genes. We then computed *z*-scores of this metric to visualize distributions of cormotif expression.

To identify single genes with tissue-restricted patterns of DE, we first filtered for genes with greater than 95% posterior probability of DE in at least one cell type. Within this set, we identified genes that had a posterior probability of DE of less than 5% in at least one cell type. To ensure that differential power could not explain the lack of DE in the conserved cell type, we required that the conserved cell type have at least 100 total observed UMIs for that gene in both humans and chimpanzees.

Differential expression in tetraploid hybrid cells

For every gene and cell type, we performed a paired Wilcoxon signed rank test to compare the within-cell difference between counts of reads assigned to human alleles vs chimpanzee alleles. We used the row_wilcoxon_paired function from the matrixTests package to implement this [61]. We then adjusted p-values within each cell type using the Benjamini–Hochberg procedure [62].

Estimates of cis proportions

To estimate the relative contribution of allele-specific changes (*cis* changes) to human and chimpanzee gene expression divergence, we compared the effect size between humans and chimpanzees to the effect size within tetraploid hybrid cells where the *trans* environment is fixed. We used the procedure outlined in [13, 49]. Briefly, we compute the *cis* component as the absolute value of the log2 fold change between humans and chimpanzees measured in hybrid cells, which we denote abs(hybrid.logFC). The *trans* and non-genetic component is the absolute value log2 fold change between humans and chimpanzees measured in human and chimpanzee cells, minus the *cis* component. For this, we used the effect size estimate computed by DREAM. We denote the *trans* and non-genetic components as abs(dream.logFC – hybrid.logFC). Thus, the *cis* proportion is given by abs(hybrid.logFC)/(abs(hybrid.logFC) + abs(dream.logFC – hybrid.logFC)).

External data sources

We obtained the percent coding identity between humans and chimpanzees from BioMart [63]. Mammalian dN/dS was obtained from a study of 29 mammalian genomes [64]. LOEUF was obtained from a study of 141,456 human exomes [46]. Human accelerated regions and their targets were obtained from a recently compiled set of 3,171 regions [65].

Gene ontology and gene set enrichment

We performed gene ontology analysis using the R package topGO [66]. We used the Fisher exact test as the statistic and set the nodeSize parameter to 10. We assigned each gene to the cormotif for which it had the highest posterior probability of membership. We performed gene set enrichments using the package *fgsea* [67], with the eps parameter set to 0. We obtained gene sets from the Molecular Signature Database (MSigDB v7.5.1) [68]. For gene set enrichment of genes generally regulated in *trans*, we ranked genes based on the mean *trans* proportion across all cell types in which each gene was DE. We provided this ranking as input to *fgsea* and used the positive score type.

Identification of transcription factors putatively driving broad changes in gene expression

We obtained the motifs and targets from Xie et al. (2005) from MSigDB v7.5.1 as the TFT legacy subset [50]. We filtered the targets to include only the 14,487 genes that were tested for DE in this work. Using *phyper* in R, we performed a hypergeometric test to test for enrichment of each target set in our set of *trans*-DE genes (genes with a mean *trans*-proportion > 75% across all cell types in which that gene is DE at Cormotif posterior probability > 95%). 179 motifs were enriched at 5% FDR. Using their description on msigDB, we manually annotated this set of 179 motifs to identify the TF that binds to each motif. In cases where the motif is bound by an entire family of TFs, we included each member of the family. After annotation, we filtered the set to include TFs that were (1) DE in at least one cell type in EBs and (2) expressed in at least three cell types in EBs. This resulted in 66 TFs for further analysis.

We then tested the correlation in tissue-specific DE patterns between TFs and their putative targets. For each TF, we ran a Pearson correlation between the logFC of the TF (computed by DREAM) and the logFC of each putative target that is DE in at least once cell type in EBs. To identify which TFs show statistical enrichment for correlated DE with their targets, we used Fisher's method of combining *p*-values, implemented by the *sumlogp* function in the metap package.

Statistics

For all correlations reported in the main text, we computed p-values using Spearman's rank-order correlation test, implemented in R with the function corr.test with option method = 's'. For all comparisons of the difference between distributions, we computed p-values using the two-sided Mann–Whitney U test, computed in R with the function wilcox.test.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03019-3.

Additional file 1. Supplementary figures, tables, and detailed description of Additional files 2, 3, and 4.

Additional file 2. Data S1. A csv file with summary statistics for all DE tests and calculated *cis*-proportion. Columns are described in Additional file 1.

Additional file 3. Data S2. A Raw text file containing the gene symbol of each gene with cell-type-restricted differential expression. One gene is listed per line.

Additional file 4. Data S3. Statistics for tests of enrichment in TFs in promoters, and correlated DE between TFs and putative target genes. Columns are described in Additional file 1.

Additional file 5. Peer review history.

Acknowledgements
We thank all members of the Gilad lab for helpful discussions, the Functional Genomics Core Facility at the University of Chicago for sequencing the RNA-seq libraries, and Natalia Gonzales for editing the manuscript.

Review history
The review history is available as Additional file 5.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
Conceptualization: KAB, KLR, YG. Methodology: KAB, KLR. Software: KAB. Validation: KLR, KAB. Formal analysis: KAB. Investigation: KAB, KLR. Resources: YG. Data curation: KLB. Writing – original draft: KAB, YG. Writing – review and editing: KAB, YG. Visualization: KAB. Funding acquisition: YG. Project administration: YG. Supervision: YG.

Authors' Twitter handles
Twitter handles: @Y_Gilad (Yoav Gilad).

Funding
National Institutes of Health grant R01HG010772 (YG).

Availability of data and materials
All raw and processed sequencing data generated in this study have been submitted to NCBI's Gene Expression Omnibus (GSE201516 [69]). This repository includes an h5Seurat object for interactive exploration of this data. Code generated for this project is available on GitHub (https://github.com/kennethabarr/CormotifCounts; https://doi.org/10.5281/zenodo.8092177 and https://github.com/kennethabarr/HCPipeline; https://doi.org/10.5281/zenodo.8092175) under BSD-3-Clause license.

Declarations

Ethics approval and consent to participate
Human fibroblast samples for the generation of iPSC lines were collected under the University of Chicago IRB protocol 11–0524. Informed consent to participate in the study was obtained from all human participants. The original chimpanzee fibroblast samples for the generation of iPSC lines were obtained from Yerkes Primate Research Center of Emory University under protocol 006–12, in full accordance with IACUC protocols.

Competing interests
The authors declare that they have no competing interests.

Received: 22 November 2022 Accepted: 21 July 2023
Published online: 11 September 2023

References

1. Anderson JA, Vilgalys TP, Tung J. Broadening primate genomics: new insights into the ecology and evolution of primate gene regulation. Curr Opin Genet Dev. 2020;62:16.
2. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nat Ecol Evol. 2018;2:152–63.
3. Klein JC, Keith A, Agarwal V, Durham T, Shendure J. Functional characterization of enhancer evolution in the primate lineage. Genome Biol. 2018;19:99.
4. Eres IE, Luo K, Hsiao CJ, Blake LE, Gilad Y. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. PLoS Genet. 2019;15:e1008278.
5. Mittleman BE, Pott S, Warland S, Barr K, Cuevas C, Gilad Y. Divergence in alternative polyadenylation contributes to gene regulatory differences between humans and chimpanzees. Elife. 2021;10:e62548.
6. Housman G, Gilad Y. Prime time for primate functional genomics. Curr Opin Genet Dev. 2020;62:1.
7. Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, et al. A comparison of gene expression and DNA methylation patterns across tissues and species. Genome Res. 2020;30:250–62.
8. Fair BJ, Blake LE, Sarkar A, Pavlovic BJ, Cuevas C, Gilad Y. Gene expression variability in human and chimpanzee populations share common determinants. Elife. 2020;9:e59929.
9. Danko CG, Choate LA, Marks BA, Rice EJ, Wang Z, Chu T, et al. Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. Nat Ecol Evol. 2018;2:537–48.
10. O'Bleness M, Searles V, Varki A, Gagneux P, Sikela JM. Evolution of genetic and genomic features unique to the human lineage. Nat Rev Genet. 2012;13:853–66.
11. Pizzollo J, Nielsen WJ, Shibata Y, Safi A, Crawford GE, Wray GA, et al. Comparative Serum Challenges Show Divergent Patterns of Gene Expression and Open Chromatin in Human and Chimpanzee. Genome Biol Evol. 2018;10:826–39.
12. Ward MC, Gilad Y. A generally conserved response to hypoxia in iPSC-derived cardiomyocytes from humans and chimpanzees. Elife. 2019;8:e42374.
13. Gokhman D, Agoglia RM, Kinnebrew M, Gordon W, Sun D, Bajpai VK, et al. Human-chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. Nat Genet. 2021;53:467–76.
14. Pollen AA, Bhaduri A, Andrews MG, Nowakowski TJ, Meyerson OS, Mostajo-Radji MA, et al. Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. Cell. 2019;176:743–756.e17.
15. Benito-Kwiecinski S, Giandomenico SL, Sutcliffe M, Riis ES, Freire-Pritchett P, Kelava I, et al. An early cell shape transition drives evolutionary expansion of the human forebrain. Cell. 2021;184:2084–2102.e19.
16. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. 2012;8:e1002789.
17. Housman G, Briscoe E, Gilad Y. Evolutionary insights into primate skeletal gene regulation using a comparative cell culture model. PLoS Genet. 2022;18:e1010073.
18. Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, Giavalisco P, et al. Intra- and interspecific variation in primate gene expression patterns. Science. 2002;296:340–3.
19. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
20. Blake LE, Thomas SM, Blischak JD, Hsiao CJ, Chavarria C, Myrthil M, et al. A comparative study of endoderm differentiation in humans and chimpanzees. Genome Biol. 2018;19:162.
21. Hernando-Herraez I, Prado-Martinez J, Garg P, Fernandez-Callejo M, Heyn H, Hvilsom C, et al. Dynamics of DNA methylation in recent human and great ape evolution. PLoS Genet. 2013;9:e1003763.
22. Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science. 2013;342:1100–4.
23. Perry GH, Melsted P, Marioni JC, Wang Y, Bainer R, Pickrell JK, et al. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. Genome Res. 2012;22:602–10.
24. Iskow RC, Gokcumen O, Abyzov A, Malukiewicz J, Zhu Q, Sukumar AT, et al. Regulatory element copy number differences shape primate expression profiles. Proc Natl Acad Sci U S A. 2012;109:12656–61.
25. Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. PLoS Genet. 2011;7:e1001316.
26. Cain CE, Blekhman R, Marioni JC, Gilad Y. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics. 2011;187:1225–34.
27. Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genet. 2008;4:e1000271.
28. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage-specific alternative splicing in primates. Genome Res. 2010;20:180–9.
29. Barreiro LB, Marioni JC, Blekhman R, Stephens M, Gilad Y. Functional comparison of innate immune signaling pathways in primates. PLoS Genet. 2010;6:e1001249–e1001249.
30. De la Cruz O, Blekhman R, Zhang X, Nicolae D, Firestein S, Gilad Y. A signature of evolutionary constraint on a subset of ectopically expressed olfactory receptor genes. Mol Biol Evol. 2009;26:491–4.
31. Lemos B, Meiklejohn CD, Cáceres M, Hartl DL. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution. 2005;59:126–37.
32. Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature. 2006;440:242–5.
33. Cáceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, et al. Elevated gene expression levels distinguish human from non-human primate brains. Proc Natl Acad Sci U S A. 2003;100:13030–5.

34. Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Res. 2005;15:674–80.
35. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478:343–8.
36. Hernando-Herraez I, Heyn H, Fernandez-Callejo M, Vidal E, Fernandez-Bellon H, Prado-Martinez J, et al. The interplay between DNA methylation and sequence divergence in recent human evolution. Nucleic Acids Res. 2015;43:8204–14.
37. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338:1587–93.
38. Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Zhou X, Ward MC, Banovich NE, et al. A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. Elife. 2015;4:e07103.
39. Brickman JM, Serup P. Properties of embryoid bodies. Wiley Interdiscip Rev Dev Biol. 2017;6(2):e259.
40. Rhodes K, Barr KA, Popp JM, Strober BJ, Battle A, Gilad Y. Human embryoid bodies as a novel system for genomic studies of functionally diverse cell types. eLife. 2022;11:e71361.
41. Banovich NE, Li YI, Raj A, Ward MC, Greenside P, Calderon D, et al. Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res. 2018;28:122–31.
42. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. Science. 2020;370:eaba7721.
43. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature. 2020;581:303–9.
44. Hoffman GE, Roussos P. Dream: powerful differential expression analysis for repeated measures designs. Bioinformatics. 2021;37:192–201.
45. Wei Y, Tenzen T, Ji H. Joint analysis of differential gene expression in multiple studies using correlation motifs. Biostatistics. 2015;16:31–46.
46. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
47. del Bosque-Plata L, Hernández-Cortés EP, Gragnoli C. The broad pathogenetic role of TCF7L2 in human diseases beyond type 2 diabetes. J Cell Physiol. 2022;237:301–12.
48. Tanabe A, Yanagiya T, Iida A, Saito S, Sekine A, Takahashi A, et al. Functional Single-Nucleotide Polymorphisms in the Secretogranin III (SCG3) Gene that Form Secretory Granules with Appetite-Related Neuropeptides Are Associated with Obesity. J Clin Endocrinol Metab. 2007;92:1145–54.
49. Agoglia RM, Sun D, Birey F, Yoon S-J, Miura Y, Sabatini K, et al. Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. Nature. 2021;592:421–7.
50. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, et al. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature. 2005;434:338–45.
51. Elorbany R, Popp JM, Rhodes K, Strober BJ, Barr K, Qi G, et al. Single-cell sequencing reveals lineage-specific dynamic genetic regulation of gene expression during human cardiomyocyte differentiation. PLoS Genet. 2022;18:e1009666.
52. Greenwood EJD, Schmidt F, Kondova I, Niphuis H, Hodara VL, Clissold L, et al. Simian Immunodeficiency Virus Infection of Chimpanzees (Pan troglodytes) Shares Features of Both Pathogenic and Non-pathogenic Lentiviral Infections. PLoS Pathog. 2015;11:e1005146.
53. Pavlovic BJ, Blake LE, Roux J, Chavarria C, Gilad Y. A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. Sci Rep. 2018;8:15312.
54. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2018;36:89–94.
55. Consortium T 1000 GP. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
56. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296.
57. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
58. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
59. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.
60. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
61. Koncevičius K. matrixTests: Fast Statistical Hypothesis Tests on Rows and Columns of Matrices. 2021. Available from: https://CRAN.R-project.org/package=matrixTests. Cited 2022 Apr 14.
62. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc: Ser B (Methodol). 1995;57:289–300.
63. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). 2011;2011:bar030.
64. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478:476–82.
65. Girskis KM, Stergachis AB, DeGennaro EM, Doan RN, Qian X, Johnson MB, et al. Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. Neuron. 2021;109:3239–3251.e7.

66. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. Bioconductor version: Release (3.14); 2022. Available from: https://bioconductor.org/packages/topGO/. Cited 2022 Apr 1.
67. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. bioRxiv; 2021. p. 060012. Available from: https://www.biorxiv.org/content/10.1101/060012v3. Cited 2022 Apr 1.
68. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.
69. Barr KA, Rhodes KL, Gilad Y. The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees. GSE201516. Gene Expression Omnibus. 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201516.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.