Title: Evolutionary insights into primate skeletal gene regulation using a comparative cell culture model

Authors and Affiliations: Genevieve Housman^{1*}, Emilie Briscoe¹, and Yoav Gilad^{1,2}

¹Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, United States

²Department of Human Genetics, University of Chicago, Chicago, United States

*Corresponding Author: ghousman@uchicago.edu (G.H.)

Running Title: Gene regulation during primate skeletal cell differentiation

Key Words: primate, gene regulation, evolution, osteogenic, iPSC

Competing interests: The authors declare that no competing interests exist.

Author Contributions:

All authors were involved in drafting or critical revision of this article, and all authors approved the final version to be published. G.H. had full access to the study data and takes responsibility for the integrity of these data and the accuracy of the analyses.

Study conception and design: G.H., Y.G.

Acquisition of data: G.H., E.B.

Analysis and interpretation of data: G.H.

Abstract:

The evolution of complex skeletal traits in primates was likely influenced by both genetic and environmental factors. Because skeletal tissues are notoriously challenging to study using functional genomic approaches, they remain poorly characterized even in humans, let alone across multiple species. The challenges involved in obtaining functional genomic data from the skeleton, combined with the difficulty of obtaining such tissues from nonhuman apes, motivated us to consider an alternative *in vitro* system with which to comparatively study gene regulation in skeletal cell types. Specifically, we differentiated six human and six chimpanzee induced pluripotent stem cell lines (**iPSCs**) into mesenchymal stem cells (**MSCs**) and subsequently into osteogenic cells (bone cells). We validated differentiation using standard methods and collected single-cell RNA sequencing data from over 100,000 cells across multiple samples and replicates at each stage of differentiation. While most genes that we examined display conserved patterns of expression across species, hundreds of genes are differentially expressed (**DE**) between humans and chimpanzees within and across stages of osteogenic differentiation. Some of these interspecific DE genes show functional enrichments relevant in skeletal tissue trait development. Moreover, topic modeling indicates that interspecific gene programs become more pronounced as cells mature. Overall, we propose that this *in vitro* model can be used to identify interspecific regulatory differences that may have contributed to skeletal trait differences between species.

Author Summary:

Primates display a range of skeletal morphologies and susceptibilities to skeletal diseases, but the molecular basis of these phenotypic differences is unclear. Studies of gene expression variation in primate skeletal tissues are extremely restricted due to the ethical and practical challenges associated with collecting samples. Nevertheless, the ability to study gene regulation in primate skeletal tissues is crucial for understanding how the primate skeleton has evolved. We therefore developed a comparative primate skeletal cell culture model that allows us to access a spectrum of human and chimpanzee cell types as they differentiate from stem cells into bone cells. While most gene expression patterns are conserved across species, we also identified hundreds of differentially expressed genes between humans and chimpanzees within and across stages of differentiation. We also classified cells by osteogenic stage and identified additional interspecific differentially expressed genes which may contribute to skeletal trait differences. We anticipate that this model will be extremely useful for exploring questions related to gene regulation variation in primate bone biology and development.

Introduction

The skeleton is a biologically and evolutionarily important organ system that consists of several tissues, including bone and cartilage. The skeletal system serves a variety of functions, most notably supporting body weight and facilitating locomotion. While these broad functions are conserved across vertebrates, different species have developed distinct skeletal morphologies, which enable differential use of skeletal elements. For instance, certain bony feature shapes and sizes enable efficient bipedal locomotion in humans, while others enable efficient quadrupedal locomotion in certain nonhuman primates [1,2]. Primates also vary in their susceptibilities to different skeletal disorders, such as osteoarthritis [3–7] and osteoporosis [8–10].

The emergence of conserved and divergent skeletal phenotypes within the primate lineage is not fully resolved. Clarifying the mechanisms that contribute to such complex traits will improve our understanding of skeletal evolution and development. As with all complex traits, skeletal traits are affected by both genetic [11–15] and environmental factors [16–20], and these effects may be mediated, at least in part, through gene expression changes. While the contribution of environmental factors to skeletal differences has been widely studied in the fields of comparative anatomy, forensics, and paleoanthropology, molecular variation in skeletal tissues is not well characterized, especially among primates.

Studying gene expression in skeletal tissues is challenging, as accessing bone and cartilage requires invasive procedures. In addition to the ethical and experimental challenges of collecting skeletal samples from living primates, the poor storage conditions of most skeletal remains that are available, make such samples unusable for most functional genomic applications, including the collection of gene expression data. Further, when well-preserved samples are available, the high cellular heterogeneity of tissues limits data interpretations. Perhaps because of these considerations, even large human transcriptomics consortia like GTEx [21] do not include data from bone and cartilage. Indeed, studies of human skeletal transcriptomics are limited to more targeted efforts to understand skeletal disease, and focus primarily on cartilage tissues and chondrogenic cell types [22–24]. Hence, while comparative primate functional genomics is a growing area of research [25], only a few groups have examined gene regulation in primate skeletal tissues [26–29]. Due to preservation issues, these studies predominantly focus on DNA methylation patterns as opposed to gene expression patterns.

As an alternative to *in vivo* skeletal tissues, induced pluripotent stem cell (iPSC) derived cell culture systems provide a new way to explore molecular variation in skeletal cell types. Previous studies have differentiated primate iPSCs into cranial neural crest cells (CNCCs), which are precursor cells that develop into a variety of tissues in the skull [29,30]. Additionally, protocols that differentiate iPSCs into osteoblasts [31–33], which are the primary cells in bone, do exist. However, most research utilizing skeletal cell differentiation schemes have only applied it to 1-2 cell lines from humans or other model organisms, and often do not account for the purity of primary or differentiated cell cultures. Such study designs limit the evolutionary perspective and interpretation of the data generated.

To examine skeletal gene expression among primates more effectively, we have established a comparative primate skeletal cell culture model that includes a large number of human and chimpanzee iPSCs. Using this system, we collected and characterized single-cell RNA sequencing (scRNA-seq) data from different stages along the differentiation towards osteogenic cells. Our study design allowed us to identify interspecific differences in gene expression, which may contribute to skeletal trait divergence between species.

Results

Comparative primate study design, data collection, and preprocessing

To characterize primate skeletal gene expression patterns, we differentiated previously characterized and validated iPSCs from six humans and six chimpanzees [34–39] through an intermediary mesenchymal stem cell (MSC) state, and subsequently into osteogenic cells, which are the primary cells in bone (**Figure 1, Methods, Tables S1-S8, Figure S1.1, Figure S1.2, Supplemental Text**). Using the 10X Genomics platform, we measured single-cell gene expression patterns from each cell line at each major stage along this differentiation trajectory – in pluripotent cells (Time 0), mesenchymal cells (Time 1), and osteogenic cells (Time 2) (**Figure 1A, Methods, Table S9, Supplemental Text**). To examine technical reproducibility, we also collected single-cell gene expression data from one human replicate and one chimpanzee replicate at the same stages of differentiation (**Figure 1C**). For each 10X collection, one human cell line and one chimpanzee cell line from the same stage of differentiation were pooled together to ensure that species and batch were not confounded (**Figure S2.1**).

Our scRNA-seq approach targeted equal numbers of cells from each species, individual, and replicate. After data processing and filtering (**Methods, Supplemental Text**), we retained high-quality data from 101,000 cells (**Figure 2A**), with an average of 7,214 cells per individual replicate, a median of 16,157 UMI counts per cell, and a median of 3,929 genes per cell. We assigned data from individual cells to their species of origin using a modified version of Cell Ranger [40] (**Methods, Figure S2.2, Figure S2.3, Supplemental Text**) and confirmed that scRNA-seq data are balanced across species (**Figure 2B-D**). A Uniform Manifold Approximation Projection (UMAP) [41] plot indicates that species are fairly well integrated within each stage of differentiation (**Figure S2.4A**), as expected. Cell counts, UMI counts per cells, and genes per cell for osteogenic cells are slightly lower than those for pluripotent cells and mesenchymal cells, which is likely due to the increased adhesion properties of osteogenic cells. Indeed, osteogenic cells required more potent dissociation reagents (**Methods, Supplemental Text**) which resulted in lower cell viability (**Figure S2.1B-C**) and an increased production of multiplets during 10X runs (**Figure S2.1E**).

Using data from the replicate samples described above, we found that scRNA-seq data recovery from our cell differentiations is highly reproducible. Specifically, gene expression patterns recovered from different technical replicates are strongly correlated within each stage of differentiation (**Figure 2E-G, Methods, Supplemental Text**), and we observed similar trends when examining data from the two species independently (**Figure S2.4B**). Lastly, for downstream analyses, we integrated data from all cells using reciprocal PCA in Seurat [42,43] (**Methods, Supplemental Text**).

iPSC-based system effectively models primate osteogenesis

Using standard methods, we validated that iPSCs successfully differentiated into MSCs and osteogenic cells. In accordance with the International Society for Cellular Therapy’s minimum criteria for MSCs [44], we visually confirmed that MSCs developed elongated morphologies and were plastic-adherent (**Figure 1D, Figure S1.1A**). We also used flow cytometry to ensure that MSCs began expressing known cell surface markers (**Figure S1.1B-C, Supplemental Text**). Further, we verified the multipotent differentiation potential of MSCs by performing osteogenic, adipogenic, and chondrogenic differentiations in each cell line (**Figure S1.1D-E, Supplemental Text**). Finally, for osteogenic differentiations, we validated that cells produced extracellular matrices containing calcium deposits, similar to those in bone tissues, using standard Alizarin Red staining (**Figure 1D, Figure S1.2, Supplemental Text**).

Using our processed and integrated scRNA-seq data, we confirmed that cells collected at different stages of differentiation display expected gene expression patterns for known pluripotent, mesenchymal, and osteogenic marker genes. In both species we found that the pluripotency marker *POU5F1* is highly expressed in pluripotent cells, the mesenchymal marker *CD44* is highly expressed in mesenchymal cells and tapers off in osteogenic cells, and the osteogenic marker *COL1A1* is expressed in mesenchymal cells and increases its expression in osteogenic cells (**Figure S2.4C**).

More broadly, we ensured that cells collected at different stages of differentiation display divergent transcriptomes. First, cell groupings in a UMAP dimensional reduction plot of our scRNA-seq data show that cells collected at different stages of differentiation generally form distinct groups, with some spillover between mesenchymal cells and osteogenic cells (**Figure 2A**). Second, we observed similar transcriptome deviations when performing pairwise correlations between the whole transcriptome pseudobulk of each stage of differentiation (**Figure S2.5, Supplemental Text**). Finally, we identified positively expressed marker genes for cells at each stage of differentiation (**Methods**). Testing for the enrichment of gene ontology (**GO**) functional categories in these marker gene sets (**Methods**) revealed expected and biologically relevant functional enrichments, including somatic stem cell population maintenance in pluripotent cells, focal adhesion and extracellular matrix organization in mesenchymal cells, and collagen fibril organization and extracellular matrix production in osteogenic cells (**Figure 2H-J**).

We also annotated cells using unsupervised clustering and ad hoc assignment methods (**Methods, Supplemental Text**). These alternative classification schemes produced similar groupings of pluripotent, mesenchymal, and osteogenic cells that displayed similar transcriptomic changes across cell types (**Figure S2.5, Figure S2.6, Figure S2.7, Figure S3.1, Figure S3.2**). Thus, we continued using pluripotent, mesenchymal, and osteogenic cell classifications in downstream analyses.

Identification of differentially expressed genes in our system

Having established our iPSC-derived cell culture system as a reasonable model for studying primate osteogenesis, we then sought to understand how and to what extent the process of osteogenic differentiation differs between humans and chimpanzees. To do this, we first generated pseudobulk data by consolidating single-cell data originating from the same individual, replicate, and cell classification (**Methods**). Using the framework of a linear mixed model to account for the effects of species and cell line (**Methods, Equation 1**), we analyzed pseudobulk data to identify differentially expressed genes between humans and chimpanzees (interspecific DE genes) across each stage of differentiation. Initially, we defined differentiation stage relatively broadly, labeling cells as either pluripotent, mesenchymal, or osteogenic, as described above. Later, we took advantage of our single-cell data to study osteogenesis at a higher resolution, using two alternative approaches – an ad hoc candidate gene-based cell classification approach and a topic modeling strategy.

We discovered hundreds of interspecific DE genes within each stage of differentiation (**Figure 3**). Using standard DE analyses of pseudobulk data (**Methods, Table S10**), we detected 2,098 interspecific DE genes in pluripotent cells, 904 in mesenchymal cells, and 446 in osteogenic cells at a false discovery rate (**FDR**) < 0.01 (**Figure 3A, Figure S3.3**). We considered the overlap between interspecific DE genes identified at different stages, initially by performing a pairwise comparisons (**Figure S3.4**). Although this pairwise approach is straightforward, it is not designed to capture dependence among multiple experimental conditions, and it is not ideal for detecting genes that are consistently DE but have small effect sizes. To address these issues, we also used Cormotif [45] to implement a Bayesian clustering approach capable of capturing the major patterns of correlation between interspecific DE genes identified at different stages of differentiation (**Methods, Table S10**). Two common temporal expression patterns (or correlation

motifs) best fit our pseudobulk data (**Figure 3B**). One motif notes a high degree of interspecific DE genes shared across all stages of differentiation, while a second contains interspecific DE genes unique to pluripotent cells. Using Cormotif, we detected 6,822 interspecific DE genes in pluripotent cells, 4,523 in mesenchymal cells, and 4,020 in osteogenic cells with a posterior probability > 0.65 (**Figure 3B, Figure S3.5, Figure S3.6**). However, due to the high degree of sharing across stages of differentiation, only 2,759 are unique to pluripotent cells, 164 are unique to mesenchymal cells, and none are unique to osteogenic cells (**Figure S3.5**). Overall, there is a decrease in interspecific DE genes as cells mature.

After measuring the degree to which interspecific DE genes are shared across the three stages of differentiation, we used Cormotif [45] to identify genes that are stage-specific; that is, genes that are DE between subsequent stages of differentiation. We then asked whether stage-specific genes are generally conserved among humans and chimpanzees. We identified four correlation motifs that best fit our pseudobulk data (**Figure 3C**) and found that the majority (about 89%) of stage-specific DE genes are conserved between species. Of the remaining divergent stage-specific DE genes, we detected 20 between pluripotent and mesenchymal cells and 266 between mesenchymal and osteogenic cells, which was somewhat unexpected given that there are substantially fewer stage-specific DE genes at later stages of differentiation than at earlier stages of differentiation. Additionally, this finding was surprising given our previous observation that there are fewer interspecific DE genes in osteogenic cells, which had initially suggested to us that gene expression patterns are more conserved in osteogenic cells than in early-stage cells. Another possible explanation for the observed decrease in interspecific DE over time is that DE is more difficult to detect in late-stage cells due to increased levels of gene expression variation.

Indeed, examining a total of 11,579 genes, we found a significant increase in gene expression variance between pluripotent cells and mesenchymal cells and between mesenchymal cells and osteogenic cells – a pattern that is maintained across species (**Figure 3D, Figure S3.7**). However, since there is no *a priori* reason to expect mesenchymal and osteogenic cells to have higher gene expression variance than other cell types, we hypothesized that the high variance we observe is more likely due to increased cell heterogeneity over the course of differentiation. Thus, our subsequent analyses address this property of the data.

Interspecific differential expression across five substages of osteogenesis

Our single-cell data allowed us to explore how cellular heterogeneity changes throughout differentiation. Although osteogenic differentiations were designed to push cells toward later stages of osteogenesis, we did not expect that all cells would develop into mature osteocytes. Rather, we anticipated that cells would reach variable stages of osteogenesis at the time of collection, either because differentiation started earlier or later in different cells, or because differentiation occurred at different rates in different cells. Indeed, increased gene expression variance in osteogenic cells (**Figure 3D, Figure S3.7**) hints at such cellular heterogeneity. Thus, we explored several methods to more precisely define each cell's position along the course of osteogenesis (**Supplemental Text**). We were particularly interested in whether a different method of classifying cells would impact the classification of interspecific DE; namely, we asked: would classifying cells more precisely reveal differences in the speed of osteogenesis between species, and would doing so help us identify more interspecific DE genes?

Five distinct stages of osteogenesis can be distinguished using the expression levels of marker genes. Compared to osteogenic precursors, osteogenic cells are broadly characterized as having increased expression of collagen I and alkaline phosphatase genes. More specific cell types emerge throughout osteogenesis, beginning with preosteoblast progenitors, transitioning to cells embedding themselves in collagen matrix and subsequently mineralizing that matrix, and finally, once encased in bony matrix,

maturing into osteocytes [46]. Using these known stages of osteogenesis that have designated candidate genes [46], we classified cells using an ad hoc approach (**Figure 4, Methods, Supplemental Text**), which we found produced more biologically meaningful groups of cells than standard clustering methods (**Figure S4.1, Figure S4.2, Supplemental Text**). Using our ad hoc annotation scheme, we detected 2,004 preosteoblasts, 3,878 osteoblasts, 3,019 embedding osteoblasts, 4,473 mineralizing osteoblasts, and 2,443 maturing osteocytes (**Figure 4A**) – assignments that are compatible with the results of our standard cell differentiation staining validations (**Figure 1D, Figure S1.2**). Although these stages of osteogenesis are unified by general osteogenic gene expression patterns, their pseudobulk transcriptomic profiles become subtly but increasingly divergent the longer osteogenesis proceeds (**Figure S4.3**). Additionally, because we assigned cells to specific osteogenic stages using marker gene expression (**Figure 4B**), they do retain their characteristic cell-specific expression patterns (**Figure S4.1C**).

To determine whether the timing of osteogenic differentiation differs between humans and chimpanzees, we compared the distribution of cell counts at each stage of osteogenesis between species. Interestingly, although humans and chimpanzees have similar numbers of preosteoblasts, the distribution of cell counts at other stages of osteogenesis differs between species. We observed an accumulation of intermediate-stage cells (osteoblasts, embedding osteoblasts, and mineralizing osteoblasts) in chimpanzees. In contrast, humans have a reduction of cells at intermediate stages (embedding osteoblasts) and an increase of cells at later stages of osteogenesis (mineralizing osteoblasts and maturing osteocytes) (**Figure S4.1B**). These cell count distributions reflect the patterns we observed using our more traditional validation methods. For example, on average, chimpanzee cells produced less calcium deposits than human cells (**Figure S1.2**), indicating that chimpanzees had a larger proportion of osteogenic precursor cells than humans. This may also suggest that human cells and chimpanzee cells transition between stages of osteogenesis at different speeds.

To assess the extent of interspecific DE between these more nuanced stages of osteogenesis, we performed DE analyses of pseudobulk data (**Methods, Table S10**). Using standard methods, we identified 144 interspecific DE genes in preosteoblasts, 242 in osteoblasts, 142 in embedding osteoblasts, 329 in mineralizing osteoblasts, and 115 in maturing osteocytes with an FDR < 0.01 (**Figure 4C, Figure S3.4**). After accounting for overlaps, this results in a total of 644 interspecific DE genes – many more than were detected in bulk osteogenic cells using our earlier three-stage classification method. As before, we used Cormotif [45] as a second method of identifying interspecific DE genes to better assess sharing across stages of osteogenesis (**Methods**). We found that 2 correlation motifs best fit our pseudobulk data (**Figure 4D**). One motif notes a high degree of interspecific DE gene sharing across stages of osteogenesis, while a second contains genes that show no DE across stages of osteogenesis. Using Cormotif, we detected 3,287 interspecific DE genes in preosteoblasts, 3,287 in osteoblasts, 3,289 in embedding osteoblasts, 3,299 in mineralizing osteoblasts, and 3,287 in maturing osteocytes with a posterior probability > 0.65 (**Figure 4D, Figure S3.5**). Almost all of these interspecific DE genes are shared across stages of osteogenesis, with unique interspecific DE genes found only in mineralizing osteoblasts (n=10) (**Figure S3.5**). Overall, this more nuanced approach to classifying cells, which was based on known marker genes, helped us to identify additional interspecific DE genes.

A continuous model of interspecific differential expression throughout osteogenesis

All of our analyses thus far have implicitly considered cell type to be a discrete phenomenon, despite the fact that cell types are known to be continuous (e.g., distributed from the pluripotent state across intermediate developmental events into mature osteoblasts). Specifically, our methods have relied on the expression levels of a small number of marker genes to classify cells into types. While this is a standard

approach in the field, we felt it was also important to examine continuous variation in our data using information from the whole transcriptome.

We accomplished this with topic modeling, which is an unsupervised classification approach that finds recurring patterns of gene expression within a dataset and then summarizes the expression profile of a cell as a mixture of these identified gene programs (or topics). Because this method allows each cell to have grades of membership in multiple topics simultaneously, we were able to examine both discrete and continuous variation between cells, which is not possible in standard clustering methods (**Figure 5**). We applied topic modeling (**Methods**) at a range of resolutions, identifying 3, 4, 5, 6, and 7 topics in our data. Functional enrichment of topics revealed that some topics correspond closely to unsupervised clusters (**Figure 5**), as well as other cell classification schemes that we examined in this study. For example, in the 3-topic analysis, pluripotent and mesenchymal cells display distinct gene programs, but this distinction becomes less defined as they transition to osteogenic cells. The dominant topic loadings for each cell classification are enriched in functions that are similar to those identified previously (**Figure 2H-J**). Interestingly, though, humans and chimpanzees differ in grades of membership as differentiation progresses, suggesting that interspecific gene programs become more pronounced as cells mature. This pattern becomes more striking when we consider differentiation at a more precise scale (or, at increasing values of k).

Concordance of interspecific DE genes with previous studies of differential gene regulation between humans and chimpanzees

Previous studies have identified genes that are differentially regulated between humans and chimpanzees in a number of tissue types [29,36,47,48]. We performed gene set enrichment tests to assess concordance between the interspecific DE genes we identified in this study and previously identified genes. For these analyses, we focus on the interspecific DE genes identified using Cormotif. First, we considered the interspecific DE genes we identified using the three-stage classification approach (**Table S11, Figure 3E-F, Figure S5.1**). As expected, genes previously identified in iPSCs as DE between species [36,48] are enriched among genes we identified as interspecific DE in pluripotent cells (all $P < 0.002$) but not among the interspecific DE genes we identified in mesenchymal or osteogenic cells (all $P > 0.66$). By and large, genes previously identified as DE between species in non-pluripotent, non-mesenchymal, and non-osteogenic cell types and tissues [36,47] are not enriched among genes we identified as DE between species. Still, we found that interspecific DE genes previously identified in kidney tissue [47] are enriched among genes we identified as interspecific DE in pluripotent cells ($P < 0.006$) – an enrichment that persists regardless of the posterior probability threshold used to classify significant interspecific DE genes (**Figure S3.6**). Lastly, while there are no previous studies of gene expression differences between humans and chimpanzees in mesenchymal and osteogenic cells, there has been research on gene expression in related CNCCs [48] and on differentially methylated regions (**DMRs**) in bone tissues [29]. Although we observed overlap of these external gene sets with our interspecific DE genes, we did not identify any significant enrichments, even when examining all interspecific DE genes identified in osteogenic cells regardless of sharing (all $P > 0.65$; **Figure S5.2**). However, previously identified interspecific DE genes in CNCCs are slightly enriched among genes we identified as interspecific DE in later stages of differentiation ($P < 0.06$).

Next, we performed gene set enrichment tests to assess concordance between previously annotated external DE gene sets [29,36,47,48] and the interspecific DE genes we identified across five stages of osteogenesis (**Figure 4E-F, Figure S5.1, Table S11**). As expected, genes previously identified as DE between species in iPSCs [36,48] and in alternative cell types and tissues [36,47] are not enriched among the interspecific DE genes we identified across stages of osteogenesis. Conversely, genes previously found to be differentially regulated between humans and chimpanzees in skeleton-related cell types and tissues

do overlap our interspecific DE genes. That is, interspecific DE genes identified in iPSC-derived CNCCs [48] and interspecific DMR-associated genes identified in bone [29] are slightly enriched among interspecific DE genes shared across all stages of osteogenesis ($P < 0.03$ and $P < 0.01$, respectively). Again, it appears that there was a benefit to using a higher-resolution osteogenic cell classification system, as we were able to identify more skeletally relevant functional enrichments when we grouped cells by osteogenic stage rather than bulking them together by collection timepoint.

Functional enrichments among interspecific DE genes

To ascertain the biological functions of interspecific DE genes, we examined their overlap with GO categories and relevant GWAS hits, again performing these analyses separately for the two main cell classification strategies used in this study and focusing on the interspecific DE genes identified using Cormotif. First, we considered interspecific DE genes identified using our bulk classification system. We did not find enrichment of interspecific DE genes among relevant GO categories (**Figure S5.3, Figure S5.4, Table S12, Table S13**). Overlap of skeletal trait and disease-related genes among interspecific DE genes was also limited, and we found no significant overlap between GWAS hits for hip geometry [49] or bone area [50] among interspecific DE genes. Because there are very few genes associated with GWAS hits for hip geometry [49] and bone area [50] that also overlap with genes tested in this study ($n=7$ and 8 respectively; **Supplemental Text**), we also examined overlapping genes individually. *PPP6R3* and *GAL*, which are associated with hip geometry [49], are interspecific DE genes that are shared across all stages of differentiation. Additionally, of the bone area-associated genes [50] that we tested, *BCKDHB*, *COL11A1*, *CTDSP2* and *SOX9* are interspecific DE genes that are shared between all stages of differentiation; and *BCKDHB*, *COL11A1*, *DYM*, *HHIP*, and *SH3GL3* are interspecific DE genes that are shared between early stages of differentiation (pluripotent and mesenchymal cells). We also examined overlap between interspecific DE genes and broader skeletal phenotypes. We detected significant enrichments of height GWAS hits [51] among interspecific DE genes identified in pluripotent cells (Time 0; $P < 0.01$), osteoporosis GWAS hits [52] among interspecific DE genes shared across mesenchymal and osteogenic cells ($P < 0.003$), and osteoarthritis-related loci [53] among interspecific DE genes identified in mesenchymal cells ($P < 0.002$).

Finally, we ascertained the biological functions of the interspecific DE genes we identified when we grouped cells into five stages of osteogenesis, again examining their overlap with GO categories and relevant GWAS hits (**Figure S5.3, Figure S5.4, Table S12, Table S13**). We observed that interspecific DE genes shared across all stages of osteogenesis are enriched in functional categories related to skeletal trait development and maintenance, including embryonic skeletal system development ($P < 0.01$), extracellular matrix ($P < 0.007$), and collagen-containing extracellular matrix ($P < 0.009$) (**Table S12**). We also found that interspecific DE genes unique to mineralizing osteoblasts are enriched in functions regulating cell morphogenesis ($P < 0.009$) and cell morphogenesis involved in differentiation ($P < 0.009$) (**Table S12**). However, overlap between skeletal trait and disease-related genes among interspecific DE genes was quite limited overall. We found no significant overlap between hip geometry [49], bone area [50], height [51], osteoporosis [52], or osteoarthritis [53] related loci among interspecific DE genes identified in this study. Again, because there are very few overlapping genes associated with GWAS hits for hip geometry [49] and bone area [50] ($n=6$ and 6 respectively; **Supplemental Text**), we examined these genes individually. Of the hip geometry genes [49] that were tested in our data, *PPP6R3* is an interspecific DE gene that is shared across all stages of osteogenesis. Of the bone area genes [50] that were tested in our data, we identified *COL11A1*, *CTDSP2*, and *SOX9* as interspecific DE genes that are shared between all stages of osteogenesis. When we considered skeletal phenotypes more broadly, we found a slight enrichment of osteoporosis GWAS hits [52] among interspecific DE genes unique to mineralizing osteoblasts ($P < 0.07$). Altogether, using a higher-resolution classification of osteogenic cells does improve

our ability to detect biologically meaningful interspecific DE genes. These osteogenic interspecific DE genes hint at some skeletal trait associations, but overall, they may only have a moderate impact on skeletal trait differences between species.

Discussion

This study presents a comparative primate skeletal cell culture model that we established to study gene expression in the skeleton. Using this model, we differentiated a panel of human and chimpanzee iPSCs through an intermediary MSC state and subsequently into osteogenic cells. To our knowledge, this is the largest panel of iPSC-derived osteogenic cell types in any species to date. Other groups have used similar differentiation protocols [31] as well as more directed differentiation protocols [32] in a small number of model organism cell lines, but this study is the first to use a larger, cross-species panel. In comparative primate functional genomics more broadly, there has been research on iPSC-derived CNCCs [30,48], which are precursor cells that contribute to the development of several tissues in the skull, including skeletal tissues. However, we were interested in specifically studying bone cells and opted for a differentiation strategy that more closely resembles the production of bone cells throughout life.

We characterized gene expression patterns at a single-cell resolution in cell types along the differentiation trajectory from iPSCs to osteogenic cells. We chose to collect single-cell data as opposed to bulk data because cell differentiation does not always produce cell types of interest at 100% purity. While some cell types, like cardiomyocytes, have standard purification steps [54] and good flow cytometry markers to estimate cell purity, osteogenic cells do not. Additionally, degrees of cell heterogeneity can vary substantially across different cell lines. Thus, collecting single-cell data allowed us to avoid these issues and instead examine gene expression in specific cell types of interest, while also providing an opportunity to examine gene expression data through the lens of continuous variation.

Overall, we were able to successfully recover high-quality scRNA-seq data from pluripotent cells, mesenchymal cells, and osteogenic cells from both species. Data were reproducible and balanced across species and cell types, indicating that this is a reliable model for studying gene expression patterns. Moreover, using standard methods alongside scRNA-seq analyses, we validated the successful differentiation of mesenchymal cells and osteogenic cells in both species, which further signals that this is an effective model for specifically studying molecular changes in skeletal cell types. While not unexpected, it is interesting that differentiation efficiency varied across individuals within both species. However, teasing apart individual-specific genetic effects that impact differentiation potential will require further exploration in larger sample sets.

Regardless of cell classification, we observed a decrease in interspecific DE as differentiation progresses, which suggests that expression patterns are more conserved in osteogenic cells than in earlier progenitor cells. Nevertheless, a surprising trend in our interspecific DE genes over the course of osteogenesis is that the greatest number of interspecific DE genes are detected in mineralizing osteoblasts, which are the transition cell type between mid-stage embedding osteoblasts and late-stage maturing osteocytes. Interspecific DE genes unique to mineralizing osteoblasts are enriched in functions related to regulating cell morphogenesis in the context of differentiation, so this may be a biological difference between human and chimpanzee skeletal cells that might contribute to complex skeletal trait differences. We considered the possibility that some of the regulatory differences we observed may be due to differences in the number of human and chimpanzee cells that were classified as mineralizing osteoblasts; However, randomly subsampling these to equal numbers of cells before performing DE tests did not change the observed patterns. Looking more closely at the number of cells assigned to each cell type, it is clear that

chimpanzees have an accumulation of cells at intermediate stages of osteogenesis as compared to humans, which have increased numbers of late-stage osteogenic cells. This pattern is present in both our scRNA-seq data and standard staining validation results. It is possible that human cells transition from a mineralizing osteoblast state to a maturing osteocyte state more quickly than chimpanzee cells. Such a phenomenon could have broad phenotypic consequences, and the interspecific DE genes that we identified in mineralizing osteoblasts may contribute to these effects. However, in order to determine whether osteogenesis timing is indeed regulated by gene regulation, we would need to examine differentiation trajectories at a higher temporal resolution.

We anticipated that identifying and characterizing DE genes at different stages of osteogenesis would be of interest because it is possible for changes in gene expression along the course of osteogenesis to influence differential skeletal trait development. Although we detected more interspecific DE genes when using a higher-resolution cell classification scheme, we only found slight increases in skeletally relevant functional enrichments among these osteogenesis interspecific DE genes as compared to bulk osteogenic interspecific DE genes. That is, interspecific DE genes in iPSC-derived CNCCs [48] and interspecific DMR-associated genes identified in bone tissue [29] are enriched among genes that we identified as interspecific DE across all stages of osteogenesis but not among bulk osteogenic interspecific DE genes. These enrichments are functionally relevant in light of previous work, which has found relationships between these external gene sets and divergent skeletal phenotypes between humans and chimpanzees, including differential vocal cord positioning, facial protrusion, and others [29,48]. Overlaps with our datasets, which are more apparent in our high-resolution osteogenesis cell types, suggest that perhaps similar skeletal-trait associations exist in our data. Of note, this prior comparative skeletal gene regulation research [29,48] focused on only a handful of genes that may be phenotypically important, including *EVC2*, *NFIX*, *XYLT1*, *ACAN*, *COL2A1*, and *SOX9*. We similarly found that *EVC2*, *XYLT1*, and *SOX9* are interspecific DE genes shared across stages of osteogenesis, which further supports a connection between our interspecific DE patterns contributing to phenotypic differences. However, *EVC2* and *SOX9* are also interspecific DE genes share across all stages of differentiation (pluripotent, mesenchymal, and osteogenic cells). This does not dampen the potential impact of DE on resultant phenotypes, but rather, it suggests that this relationship is not isolated to skeletal cells.

We were surprised that regardless of whether we used a broad or precise method of classifying cells, we did not identify enrichments of skeletal-disease related genes among our osteogenic interspecific DE genes. Although not unique to osteogenic cells, we did identify an enrichment of osteoporosis GWAS hits [52] among interspecific DE genes shared across later stages of differentiation (mesenchymal and osteogenic). Osteoporosis, which is characterized by a decrease in bone mineral density that can lead to bone fractures, has been observed in nonhuman primates, and manifestations of this disease appear to vary somewhat across different primate species [8–10]. This functional enrichment is reasonable because *in vivo* mesenchymal cells serve as precursor cells for osteogenic cells. In particular, this enrichment may signal important differences between humans and chimpanzees regarding bone turnover and susceptibility to osteoporosis.

We also found an enrichment of osteoarthritis-related loci [53] among interspecific DE genes identified in mesenchymal cells. Osteoarthritis is characterized by the degradation of cartilage and the underlying bone in joints, and humans and chimpanzees vary in their susceptibilities to this skeletal disorder. In particular, osteoarthritis is prevalent in humans [4], and while it is similarly prevalent in other nonhuman primates like baboons [3], the prevalence of osteoarthritis in wild [5] and captive [6,7] chimpanzees is very low. It is interesting that this enrichment was found among interspecific DE genes in mesenchymal cells, as these cells are not the primary cells in tissues directly affected during osteoarthritis progression (e.g., cartilage

and bone). However, because mesenchymal cells serve as precursor cells for these skeletal tissues, this enrichment may signal important differences between humans and chimpanzees regarding tissue development and maintenance, as well as susceptibility to osteoarthritis. Overall, given these enrichments in mesenchymal cells, along with those identified in cell types transitioning from mesenchymal to osteogenic states, it is worth looking at earlier stages of mesenchymal cell differentiation and osteogenesis more systematically to better identify skeletal trait related DE genes. Additionally, future research incorporating disease-related perturbations will allow us to explore this evolutionarily divergent trait in greater depth and detail.

In conclusion, we have established a novel comparative primate skeletal cell culture model that can be used to examine gene expression and various aspects of primate skeletal gene regulation. This includes regulatory mechanisms such as histone modifications and DNA methylation, regulatory responses to environmental treatments, and more. Our findings in this study reveal novel information about the evolutionary changes in gene expression patterns across primate skeletal cell development, but there are many research avenues yet to be explored which this system is poised to address.

Methods

Ethics Statement

iPSC lines used in this study were previously generated [38] from fibroblasts collected from human participants with written informed consent under the University of Chicago IRB protocol 11–0524 and from fibroblasts collected from chimpanzees at the Yerkes Primate Research Center of Emory University under protocol 006–12, in full compliance with IACUC protocols and prior to the September 2015 implementation of Fish and Wildlife Service rule 80 FR 34499.

Human and Chimpanzee iPSC Lines

This study included six iPSC lines from humans and six iPSC lines from chimpanzees (**Figure 1B, Table S1, Table S2**), which is a sufficient number to identify interspecific gene expression differences [35–39]. Technical replicates (independent MSC and osteogenic differentiations) from one human cell line and one chimpanzee cell line were used to examine experimental reproducibility. iPSC lines were derived from fibroblasts using the same experimental design and episomal reprogramming protocol as previously described [38], and pluripotency was previously characterized [34–39]. In addition to species having matched cell type of origin and reprogramming method, biological replicates within each species comprise an equal sampling of both sexes (**Figure 1B, Table S1, Table S2**).

Cell Differentiation Protocols

Feeder-free iPSCs were maintained on Matrigel Growth Factor Reduced Matrix (354230, Corning, Bedford, MA, USA) at a 1:100 dilution and in mTeSR™1 Medium (85851, STEMCELL Technologies, Vancouver, Canada) supplemented with 1% Penicillin/Streptomycin (30-002-CI, Corning). Cells were cultured at 37°C in 5% CO₂ and atmospheric O₂ and passaged every 3–4 days using a dissociation reagent (0.25mM EDTA, 150mM NaCl in PBS).

iPSC-derived MSCs and osteogenic cells from both species were differentiated using protocols modified from [31] (**Figure 1A, Figure 1C**). After iPSCs were cultured for 15–30 passages as described above, they were seeded at approximately 30% confluency in Matrigel-coated culture dishes and cultured in mTeSR™1 (85851, STEMCELL Technologies) until cells fully adhered to the plate (at least 3 hours). Following this, the medium was replaced with MSC medium, which consisted of Dulbecco's Modified Eagle Medium (**DMEM**) (10567-014 or 11330-032, Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 20% fetal

bovine serum (FBS) (FB5001, Thomas Scientific, Swedesboro, NJ, USA) and 1% Penicillin/Streptomycin (30-002-CI, Corning) (Table S2). Cells were cultured at 37°C in 5% CO₂ and atmospheric O₂ with daily MSC medium changes until 80-100% confluent (2-5 days). Cells were subsequently detached from the Matrigel-coated culture dishes using 0.05% Trypsin (25-052-CI, Corning) and passaged to uncoated polystyrene culture dishes. Cells continued to be cultured at 37°C in 5% CO₂ and atmospheric O₂ with MSC medium changes every 2-3 days, and cells were sub-cultured at an approximately 1:3 ratio until at least passage 4, when cells began to display characteristic MSC morphologies (3-9 weeks) (Table S3) [44]. At this point, cells were classified as iPSC-derived MSCs and cryopreserved in cryomedium consisting of 80% FBS, 10% MSC medium, and 10% Dimethyl Sulfoxide (25-950-CQC, Corning).

For osteogenic differentiations, iPSC-derived MSCs were cultured to at least passage 6 and seeded at 4.2 x 10⁴ cell/cm² in culture dishes (4-10cm diameter) coated in 5ug/cm² Type I Collagen (sc-136157, Santa Cruz Biotechnology, Dallas, TX, USA). After culturing cells for 1 day in MSC medium, this medium was replaced with osteogenic medium that contained DMEM (11965-092, Thermo Fisher Scientific) supplemented with 10% FBS (FB5001, Thomas Scientific), 1% Penicillin/Streptomycin (30-002-CI, Corning), 50ug/mL Vitamin C (A4034 or A8960, Sigma-Aldrich, St Louis, MO, USA), 100nM Dexamethasone (D4902 or D1756, Sigma-Aldrich), 10mM β-glycerophosphate (G9422, Sigma-Aldrich), and 1uM Vitamin D (D1530, Sigma-Aldrich). Cells were cultured at 37°C in 5% CO₂ and atmospheric O₂, and medium was changed every 2-3 days for 3 weeks.

For all data collections, cells were cultured in discrete batches so that species and cell culture batches were not confounded (Table S2). In each batch, data collections from iPSCs and MSCs were performed after cells were cultured for at least 1 passage following cryostock thawing, and data collections from osteogenic cells were performed immediately after completing the differentiation protocol described above (Tables S3-S4). iPSC-derived MSCs also underwent chondrogenic and adipogenic differentiations using protocols modified from [31] for cell type validation purposes (Tables S5-S8, Supplemental Text).

Single-Cell RNA Sequencing

scRNA-seq data were obtained from all cell lines at three stages of differentiation – pluripotent cells (Time 0), mesenchymal cells (Time 1), and osteogenic cells (Time 2) – using the 10X Genomics Chromium Single Cell 3' Reagent kit (chemistry v3) (Figure 1A) with a pooling strategy that ensured that collection batch and species were not confounded (Figure S2.1A, Table S9). First, cells were cultured as described above and then dissociated from adherent conditions into single-cell suspensions (Supplemental Text). Cell counts and viability measures were performed for each sample separately (Figure S2.1B, Table S9). Before single-cell sequencing, samples were pools such that each pool contained one human sample and one chimpanzee sample at the same stage of differentiation at equal proportions. After cell counts and viability measures were performed for each pooled sample (Figure S2.1C, Table S9), they were loaded into separate wells of 10X chips with a target collection of 5,000 cells per pooled sample (50% human cells and 50% chimpanzee cells). A total of 21 wells across 21 10X chips were used. Single-cell cDNA libraries were prepared using the 10X Genomics Chromium Single Cell protocol [40]. Libraries were multiplexed into three batches and sequenced to 100 base pairs, paired-end across 17 lanes on the Illumina HiSeq4000 at the University of Chicago Genomics Core Facility (Table S9). These newly reported data have been made available on NCBI's Gene Expression Omnibus and are accessible under accession numbers GSE181744.

Single-Cell Data Processing

Sequencing read quality was confirmed using FastQC, and raw scRNA-seq reads were processed using standard 10X Genomics Cell Ranger 3.1.0 pipelines (Cell Ranger) [40] (Figure 2.1D), with the exception that reads were aligned once to the human genome (hg38) and once to the chimpanzee genome

(panTro6) and that a curated set of orthologous exons [55] was used for transcriptome alignment (**Supplemental Text**). Briefly, 10X cell barcodes and UMIs were extracted from reads, and the remaining reads were mapped to genes using the specified reference genomes.

10X cell barcodes were also bioinformatically reassigned to their species of origin using standard Cell Ranger pipelines (**Figure S2.1E**). As before, reads were aligned to both human (hg38) and chimpanzee (panTro6) genomes, the same curated set of orthologous exons [55] was used. Additional modifications to the assignment protocol were also incorporated (**Supplemental Text**). Briefly, in the standard pipeline, Cell Ranger aligns reads to each of two genomes, retaining only those reads that specifically align to one genome and discarding reads that align to both genomes. It then assigns human and chimpanzee cells based on which genome has more aligned UMI counts and assigns cells as multiplets when the human UMI counts and the chimpanzee UMI counts exceed the 10th percentile of each species' distribution. Instead of using these default species assignments, cells were assigned as either human or chimpanzee based on the ratio of human-aligned UMIs to the total number of aligned UMIs within each cell (**Figure S2.2, Figure S2.3**). Specifically, cells with a ratio greater than or equal to 0.9 were assigned as human, and cells with a ratio less than or equal to 0.1 were assigned as chimpanzee (**Figure S2.1E**).

Once processed, matrices containing gene counts per 10X barcode were imported into R using the Seurat package (v3.1.2) [42,43], and cells assigned as multiplets were removed. Of note, gene counts for cells assigned as human were determined using read alignments to the human genome, and cells assigned as chimpanzee were determined using read alignments to the chimpanzee genome. Cells were filtered if fewer than 1000 UMIs were detected, if fewer than 700 genes were detected, and if more than 25% of reads mapped to the mitochondria. This resulted in a total of 101,000 cells that were used in subsequent analyses.

Single-Cell Data Integration

After processing and filtering, scRNA-seq data were integrated using Seurat (**Supplemental Text**). Briefly, cells derived from individual cell lines were treated as individual datasets (n=14). Data were log normalized, and datasets were integrated across individuals and collection batches using the reference-based, reciprocal PCA method in Seurat [42,43] (**Supplemental Text**). Integration was performed across all genes that had non-zero UMI counts (n=18,482). Following integration, dimensional reduction was performed using a UMAP [41] of all principal components that explain more than 0.1% of data variance (n=13) (**Figure 2A**), and unwanted variation due to UMI counts and the percent of mitochondria-mapped reads was regressed out from the data.

Single-Cell Data Annotation

Cell type classifications were defined and assigned to single-cell data using several methods (**Supplemental Text**). The main methods reported here include the stage of differentiation at which cells were collected, unsupervised clustering, and ad hoc assignments, although further modifications to parameters in each of these classification schemes were also examined (**Supplemental Text**). The stage of differentiation at which cells were collected was known from the experimental design, but all other annotation methods involved additional analytical steps.

Unsupervised clustering was performed in Seurat. Specifically, the 40 nearest neighbors of each cell were determined using the FindNeighbors function and all principal components that explain more than 0.1% of data variance (n=13). Using this nearest neighbor information, clusters were then identified using the FindClusters function with resolutions of 0.05, 0.25, and 0.50.

Ad hoc assignments were determined using candidate gene expression patterns (**Supplemental Text**). A cell was defined as positively expressing a particular gene if the expression level for that gene within the cell was greater than the mean expression level of that gene across all cells. Expression levels were based on scaled integrated data that had confounding variables regressed out as described above. First, general ad hoc assignments of pluripotent, mesenchymal, and osteogenic were determined for cells (**Supplemental Text**), and then more specific ad hoc assignments for different stages of osteogenesis were determined for osteogenic cells. Different stages of osteogenesis include preosteoblasts, osteoblasts, embedding osteoblasts, mineralizing osteoblasts, and maturing osteocytes, and candidate genes known to vary in expression levels across these stages of osteogenesis were examined [46]. Based on additional analyses (**Supplemental Text**), the final set of candidate genes for defining osteogenic ad hoc assignments included *RUNX2*, *BGLAP*, *PHEX*, and *MEPE* (**Figure 4B**).

Topic modeling of Single-Cell Data

In addition to annotating single-cell data into discrete cell classifications, data structure was further examined using topic modeling. In this method, major patterns in gene expression (or topics) within the data are learned, and each cell is modeled as a combination of these topics with different grades of membership in each topic. First, raw scRNA-seq counts were filtered to remove genes with 0 counts across all cells and batch corrected for collection and replicate using the BatchCorrectedCounts function in the R package CountClust [56]. Unsupervised topic modeling was then performed using the R package FastTopics [57]. The fit_poisson_nmf function with default parameters was used to fit a Poisson non-negative matrix factorization model with 3, 4, 5, 6, or 7 ranks to the data. To convert these Poisson non-negative matrix factorization models into a topic model, the fitted loadings matrices were rescaled to a total sum of 1 across each cell barcode and defined as topic probabilities, and the factors matrices were rescaled to a total sum of 1 across each gene and defined as the word probabilities in the resulting topic model.

Identifying Marker Genes

For cell annotations, marker genes were identified for each classification using the FindMarkers function in Seurat. In this method, only genes that were detected in at least 25% of cells within a given classification and that had an average log fold change of at least 0.25 were tested. Wilcoxon Rank Sum tests between specific cell classifications (e.g., Time 0) and all remaining cells were performed to determine marker genes, and only those genes that had elevated expression levels in the specific cell classification as compared to all remaining cells were retained. For topic modeling, the top 100 genes defining each topic were identified using the ExtractTopFeatures function in the R package CountClust [56] in order to interpret the cellular functions captured by each topic.

Differential Expression using Standard Methods

For cell annotations, DE between species was performed using the dream package in R [58]. First, pseudobulk expression values were calculated for each unique grouping of cell line, replicate, and cell classification (e.g., cells from C1, replicate 2, that were collected at pluripotent Time 0). Pseudobulk expression values were defined as the sum of gene counts for a particular group. Mitochondrial genes and ribosomal genes were filtered from data, along with genes that had an average log2 transformed counts per million (CPM) less than or equal to zero (**Table S10**). Pseudobulk expression values were then TMM-normalized. In order to account for the mean-variance relationship in these data [59], weighted gene expression values were estimated using the voomWithDreamWeights function. A linear mixed model (**Equation 1**) in which species was modeled as a fixed effect and individual cell line was modeled as a random effect was then fit to the data using the dream function. Multiple testing was corrected using the Benjamini-Hochberg FDR [60], and genes were designated as significantly DE between species if they had

an FDR < 0.01. This cutoff was chosen as it optimized expected patterns of gene set enrichments (Supplemental Text, Figure S3.3).

$$\text{Equation 1: } Y \sim \beta_0 + \beta_{\text{species}} * X_{\text{species}} + \beta_{\text{cell line}} * X_{\text{cell line}} + \varepsilon$$

Differential Expression using Cormotif

For cell annotations, the Cormotif package in R [45] was used as a secondary method of identifying DE between species and the primary method for identifying DE across differentiation trajectories. This method is ideal as it helps to overcome issues of incomplete power that affect naïve pairwise DE comparisons and to account for dependency in data coming from different stages of differentiation. Specifically, Cormotif implements a Bayesian clustering approach that identifies common temporal expression patterns (or correlation motifs) that best fit the given data and designated pairwise differential tests. First, pseudobulk expression values were calculated as described above, and mitochondrial genes, ribosomal genes, and genes that had an average log2 transformed CPM less than or equal to zero were removed from the data (Table S10). Of note, Cormotif was initially used to analyze microarray data, so standard analyses simply use log2 transformed CPM values without accounting for the mean-variance relationship in these data. However, this is necessary for RNA sequencing data [59], so Cormotif functions were modified to first TMM-normalize raw pseudobulk expression values and estimate weighted gene expression values using the voom function in the R package limma [61] before identifying correlation motifs. Genes with posterior probability > 0.65 were defined as significantly DE. This cutoff was chosen as it optimized patterns of external gene set enrichments (Supplemental Text, Figure S3.6).

Additionally, when examining the conservation of cell classification DE genes across species, DE genes were classified as conserved if the absolute difference in posterior probability between humans and chimpanzees was less than or equal to 0.3, as previously reported [36].

Gene Expression Variance

For cell annotations, the variance in gene expression across cell classifications was examined. This was done separately for each species. First, pseudobulk expression values were calculated as described above, and mitochondrial genes, ribosomal genes, and genes that had an average log2 transformed CPM less than or equal to zero were removed from the data (Table S10). For each gene, mean expression levels and variance in expression were calculated across cells from a given species and cell classification. Finally, two-sided t-tests were performed separately for the distribution of mean values and the distribution of variance values between cell classifications that are adjacent along a given differentiation trajectory (e.g., Time 0 vs. Time 1, Time 1 vs. Time 2, etc.). Plotted variance values were log2 transformed.

Concordance and Functional Enrichment Tests

The enrichment of GO categories was assessed among various sets of genes identified in this study using the enrichGO function in the clusterProfiler v3.12.0 package in R [62]. This method uses a hypergeometric test to evaluate the enrichment of GO categories in a set of genes as compared to the total genes examined. Enrichment was classified as significant using a p-value cutoff of 0.01 and a q-value cutoff of 0.05. For cell annotation marker genes, only the top 100 marker genes based on average log fold change values were tested for GO enrichment.

The enrichment of external gene sets was also assessed in our DE gene sets using a Fisher's exact test. In all enrichment tests, the background gene set consisted of all genes that were tested for DE (Table S10). Several external gene sets were manually compiled from papers, including genes previously identified as differentially expressed between humans and chimpanzee [29,36,47,48], skeletal phenotype-related

genes [49–51], and genes associated with osteoarthritis and osteoporosis [52,53] (**Table S11**). Genes previously identified as DE between humans and chimpanzee [29,36,47,48] were primarily used to confirm consistent and biologically meaningful DE. One dataset comprised DE genes in human and chimpanzee iPSCs differentiated towards endoderm [36], another comprised DE genes in human and chimpanzee heart, kidney, liver, and lung tissue [47], and a final comprised DE genes in human and chimpanzee parental and hybrid iPSCs differentiated towards CNCCs [48]. In each of these datasets, only DE genes that are unique to each cell or tissue type examined within a given study were considered. An additional dataset comprising DMRs identified between chimpanzees, anatomically modern humans, and archaic hominins in bone tissues [29] was also considered, since DMRs can impact gene expression patterns. In particular, only those genes with DMRs that differentiate anatomically modern humans and chimpanzees were considered. General skeleton-related and skeletal phenotype-related genes [49–51] and genes associated with the skeletal diseases osteoarthritis and osteoporosis [52,53] were used to assess whether DE genes in this study may influence specific skeletal traits. Datasets included genes associated with hip geometry GWAS loci [49], bone area GWAS loci [50], height GWAS loci [51], and osteoporosis GWAS loci [52], as well as genes that have cross-omics evidence for involvement in osteoarthritis progression [53].

Data Availability

All computational scripts and analysis pipelines can be found on GitHub at <https://github.com/ghousman/human-chimp-skeletal-scRNA>. All scRNA-seq data have been deposited in the NCBI's Gene Expression Omnibus under the SuperSeries accession number GSE181744, which contains the SubSeries GSE167240 and GSE18174.

Acknowledgements

This work was supported by the National Institutes of Health (NIH R01GM122930 to Y.G.; NIH F32AR075397 to G.H.). We thank members of the Gilad lab for helpful discussions, particularly Kenneth Barr, Michelle Ward, and Anthony Hung. We also thank Dr. Heike E. Daldrup-Link and Olga Lenkov at Stanford University for training during the initial optimization of our differentiation protocols, as well as Natalia Gonzales for help editing the paper. The computational resources were provided by the University of Chicago Research Computing Center.

Figure 1. Comparative skeletal cell culture model.

Schematic of the differentiation protocol and the stages at which single-cell RNA-seq data were collected (A), along with descriptions of the human and chimpanzee cell lines used (B), a diagram of the overall study design (C), and cell images from one human cell line and one chimpanzee cell line at each stage of differentiation (D). Pluripotent cells (Time 0) and mesenchymal cells (Time 1) are phase contrast images at 4X magnification, and osteogenic cells (Time 2) are stained with Alizarin Red and zoomed out to display the entire cell culture well. Silhouette images were adapted from <http://phylopic.org/>.

Figure 2: scRNA-seq data recovery is similar across species and reproducible across replicates.

UMAP dimensional reduction plot of scRNA-seq data with cells labeled by the stage of differentiation at which they were collected (A), along with a bar plot depicting the number of chimpanzee and human cells collected at each stage of differentiation (B), violin plots displaying the distribution of UMI counts per cell (C) and gene counts per cell (D), and the correlation of average gene expression patterns between technical replicates collected in pluripotent cells (Time 0) (E), mesenchymal cells (Time 1) (F), and osteogenic cells (Time 2) (G). Enrichment of GO functional categories among marker genes for cells collected at each stage of differentiation (H-J). The top 5 GO functions identified in biological processes (BP), cell components (CC), and molecular functions (MF) are displayed along with the adjusted p-value (p-adjust), the number of marker genes overlapping a GO function (Count), and the ratio of overlapping to non-overlapping marker genes for a given GO function (GeneRatio).

Figure 3: Interspecific DE across three stages of osteogenic differentiation.

Bar plot showing the number of interspecific DE genes identified for each stage of differentiation using standard methods (A). Correlation motifs based on the probability of differential expression between species for each stage of differentiation (B) and correlation motifs based on the probability of differential expression across stages of differentiations for each species (C) with the number of genes assigned to each motif shown in the bar plot on the right and the posterior probability that a gene is DE shown by the shading of each box. Box plots of the log2 transformed gene expression variance values for cells collected at each stage of differentiation for each species (***) $p < 0.001$ (D). Enrichment of external DE gene sets among Cormotif interspecific DE genes identified for each stage of differentiation for validation (E) and functional interpretation (F) with the p-value (p.value), the number of DE genes overlapping an external gene set (DE.Interest), and the ratio of overlapping to non-overlapping DE genes for a given external gene set (GeneRatio) denoted.

Figure 4: Interspecific DE across five substages of osteogenesis.

UMAP dimensional reduction plot of scRNA-seq data with cells labeled by the stage of osteogenesis to which they were assigned (A), along with a simplified schematic of the osteogenic ad hoc assignment method. Bar plot showing the number of interspecific DE genes identified for each stage of osteogenesis using standard methods (C). Correlation motifs based on the probability of differential expression between species for each osteogenic ad hoc assignment with the number of genes assigned to each motif shown in the bar plot on the right and the posterior probability that a gene is DE shown by the shading of each box (D). Enrichment of external DE gene sets among Cormotif interspecific DE genes identified for each stage of osteogenesis for validation (E) and functional interpretation (F) with the p-value (p.value), the number of DE genes overlapping an external gene set (DE.Interest), and the ratio of overlapping to non-overlapping DE genes for a given external gene set (GeneRatio) denoted.

Figure 5: Examining cell types at different resolutions using discrete and continuous perspectives.

UMAP dimensional reduction plot of scRNA-seq data with cells labeled by the unsupervised cluster (resolutions of 0.05, 0.25, and 0.50) to which they were assigned (A), and structure plots showing the results of topic modelling at k=3, k=4, k=5, k=6, and k=7 with each row representing the gene expression profile from one cell, each colored bar representing a topic, and the grade of membership in each topic depicted by the length of the bar along the x-axis (B). In the structure plots, cells are grouped by their species of origin and collection timepoint. In both sets of plots, the key notes the top GO category enrichment of marker genes for a given cluster or topic.

References

1. Fleagle JG. Primate Adaptation and Evolution. New York: Academic Press; 1999.
2. Ankel-Simons F. Primate Anatomy. 3rd ed. New York: Academic Press; 2007.
3. Cox LA, Comuzzie AG, Havill LM, Karere GM, Spradling KD, Mahaney MC, et al. Baboons as a Model to Study Genetics and Epigenetics of Human Disease. *ILAR J.* 2013;54: 106–121. doi:10.1093/ilar/ilt038
4. O'Connor MI. Osteoarthritis of the hip and knee: sex and gender differences. *Orthop Clin North Am.* 2006;37: 559–568. doi:10.1016/j.ocl.2006.09.004
5. Rothschild BM, Rühli FJ. Etiology of reactive arthritis in *Pan paniscus*, *P. troglodytes troglodytes*, and *P. troglodytes schweinfurthii*. *Am J Primatol.* 2005;66: 219–231. doi:10.1002/ajp.20140
6. Magden ER, Haller RL, Thiele EJ, Buchl SJ, Lambeth SP, Schapiro SJ. Acupuncture as an Adjunct Therapy for Osteoarthritis in Chimpanzees (*Pan troglodytes*). *J Am Assoc Lab Anim Sci.* 2013;52: 475–480.
7. Videan EN, Lammey ML, Lee DR. Diagnosis and Treatment of Degenerative Joint Disease in a Captive Male Chimpanzee (*Pan troglodytes*). *J Am Assoc Lab Anim Sci.* 2011;50: 263–266.
8. Smith S y., Jolette J, Turner C h. Skeletal health: primate model of postmenopausal osteoporosis. *Am J Primatol.* 2009;71: 752–765. doi:10.1002/ajp.20715
9. Gunji H, Hosaka K, Huffman MA, Kawanaka K, Matsumoto-Oda A, Hamada Y, et al. Extraordinarily low bone mineral density in an old female chimpanzee (*Pan troglodytes schweinfurthii*) from the Mahale Mountains National Park. *Primates.* 2003;44: 145–149. doi:10.1007/s10329-002-0030-8
10. Sumner DR, Morbeck ME, Lobick JJ. Apparent age-related bone loss among adult female Gombe chimpanzees. *Am J Phys Anthropol.* 1989;79: 225–234. doi:10.1002/ajpa.1330790210
11. Joganic JL, Willmore KE, Richtsmeier JT, Weiss KM, Mahaney MC, Rogers J, et al. Additive genetic variation in the craniofacial skeleton of baboons (genus *Papio*) and its relationship to body and cranial size. *Am J Phys Anthropol.* 2017;165: 269–285. doi:10.1002/ajpa.23349
12. Ritzman TB, Banovich N, Buss KP, Guida J, Rubel MA, Pinney J, et al. Facing the facts: The *Runx2* gene is associated with variation in facial morphology in primates. *Journal of Human Evolution.* 2017;111: 139–151. doi:10.1016/j.jhevol.2017.06.014
13. Valdes AM, Spector TD. Genetic epidemiology of hip and knee osteoarthritis. *Nature Reviews Rheumatology.* 2011;7: 23–32. doi:10.1038/nrrheum.2010.191
14. Reynard LN, Loughlin J. Insights from human genetic studies into the pathways involved in osteoarthritis. *Nature Reviews Rheumatology.* 2013;9: 573–583. doi:10.1038/nrrheum.2013.121
15. Zengini E, Hatzikotoulas K, Tachmazidou I, Steinberg J, Hartwig FP, Southam L, et al. Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nature Genetics.* 2018;50: 549–558. doi:10.1038/s41588-018-0079-y

- 855 16. Lewton KL. The effects of captive versus wild rearing environments on long bone articular surfaces
856 in common chimpanzees (*Pan troglodytes*). *PeerJ*. 2017;5: e3668. doi:10.7717/peerj.3668
- 857 17. Lewton KL, Ritzman T, Copes LE, Garland T, Capellini TD. Exercise-induced loading increases ilium
858 cortical area in a selectively bred mouse model. *American Journal of Physical Anthropology*. 2018;
859 1–9. doi:10.1002/ajpa.23770
- 860 18. Cooper C, Snow S, McAlindon TE, Kellingray S, Stuart B, Coggon D, et al. Risk factors for the
861 incidence and progression of radiographic knee osteoarthritis. *Arthritis & Rheumatism*. 2000;43:
862 995–1000. doi:10.1002/1529-0131(200005)43:5<995::AID-ANR6>3.0.CO;2-1
- 863 19. Blagojevic M, Jinks C, Jeffery A, Jordan KP. Risk factors for onset of osteoarthritis of the knee in
864 older adults: a systematic review and meta-analysis. *Osteoarthritis and Cartilage*. 2010;18: 24–33.
865 doi:10.1016/j.joca.2009.08.010
- 866 20. Macrini TE, Coan HB, Levine SM, Lerma T, Saks CD, Araujo DJ, et al. Reproductive status and sex
867 show strong effects on knee OA in a baboon model. *Osteoarthritis Cartilage*. 2013;21: 839–848.
868 doi:10.1016/j.joca.2013.03.003
- 869 21. Consortium Gte. Genetic effects on gene expression across human tissues. *Nature*. 2017;550: 204–
870 213. doi:10.1038/nature24277
- 871 22. Zhang X, Huang N, Huang R, Wang L, Ke Q, Cai L, et al. Single-cell RNA seq analysis identifies the
872 biomarkers and differentiation of chondrocyte in human osteoarthritis. 2020 [cited 16 Jun 2020].
873 doi:10.21203/rs.3.rs-25051/v1
- 874 23. Chou C-H, Jain V, Gibson J, Attarian DE, Haraden CA, Yohn CB, et al. Synovial cell cross-talk with
875 cartilage plays a major role in the pathogenesis of osteoarthritis. *Scientific Reports*. 2020;10:
876 10868. doi:10.1038/s41598-020-67730-y
- 877 24. Rai MF, Cheverud JM, Schmidt EJ, Sandell LJ. Genetic correlations between cartilage regeneration
878 and degeneration reveal an inverse relationship. *Osteoarthritis and Cartilage*. 2020;28: 1111–1120.
879 doi:10.1016/j.joca.2020.04.013
- 880 25. Housman G, Gilad Y. Prime time for primate functional genomics. *Current Opinion in Genetics &*
881 *Development*. 2020;62: 1–7. doi:10.1016/j.gde.2020.04.007
- 882 26. Housman G, Quillen EE, Stone AC. Intraspecific and interspecific investigations of skeletal DNA
883 methylation and femur morphology in primates. *American Journal of Physical Anthropology*.
884 2020;173: 34–49. doi:10.1002/ajpa.24041
- 885 27. Housman G, Quillen EE, Stone AC. An evolutionary perspective of DNA methylation patterns in
886 skeletal tissues using a baboon model of osteoarthritis. *Journal of Orthopaedic Research*. 2021;39:
887 2260–2269. doi:10.1002/jor.24957
- 888 28. Housman G, Havill LM, Quillen EE, Comuzzie AG, Stone AC. Assessment of DNA Methylation
889 Patterns in the Bone and Cartilage of a Nonhuman Primate Model of Osteoarthritis. *Cartilage*.
890 2019;10: 335–345. doi:10.1177/1947603518759173

- 891 29. Gokhman D, Nissim-Rafinia M, Agranat-Tamir L, Housman G, García-Pérez R, Lizano E, et al.
892 Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nature*
893 *Communications*. 2020;11: 1–21. doi:10.1038/s41467-020-15020-6
- 894 30. Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, et al. Enhancer Divergence
895 and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell*. 2015;163: 68–83.
896 doi:10.1016/j.cell.2015.08.036
- 897 31. Nejadnik H, Diecke S, Lenkov OD, Chapelin F, Donig J, Tong X, et al. Improved Approach for
898 Chondrogenic Differentiation of Human Induced Pluripotent Stem Cells. *Stem Cell Rev*. 2015;11:
899 242–253. doi:10.1007/s12015-014-9581-5
- 900 32. Kanke K, Masaki H, Saito T, Komiyama Y, Hojo H, Nakauchi H, et al. Stepwise Differentiation of
901 Pluripotent Stem Cells into Osteoblasts Using Four Small Molecules under Serum-free and Feeder-
902 free Conditions. *Stem Cell Reports*. 2014;2: 751–760. doi:10.1016/j.stemcr.2014.04.016
- 903 33. Hynes K, Menicanin D, Gronthos S, Bartold MP. Differentiation of iPSC to Mesenchymal Stem-Like
904 Cells and Their Characterization. *Induced Pluripotent Stem (iPS) Cells*. Humana Press, New York,
905 NY; 2014. pp. 353–374. doi:10.1007/7651_2014_142
- 906 34. Burrows CK, Banovich NE, Pavlovic BJ, Patterson K, Romero IG, Pritchard JK, et al. Genetic
907 Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in
908 iPSCs. *PLOS Genetics*. 2016;12: e1005793. doi:10.1371/journal.pgen.1005793
- 909 35. Ward MC, Zhao S, Luo K, Pavlovic BJ, Karimi MM, Stephens M, et al. Silencing of transposable
910 elements may not be a major driver of regulatory evolution in primate iPSCs. Wittkopp PJ, editor.
911 *eLife*. 2018;7: e33084. doi:10.7554/eLife.33084
- 912 36. Blake LE, Thomas SM, Blischak JD, Hsiao CJ, Chavarria C, Myrthil M, et al. A comparative study of
913 endoderm differentiation in humans and chimpanzees. *Genome Biology*. 2018;19: 162.
914 doi:10.1186/s13059-018-1490-5
- 915 37. Pavlovic BJ, Blake LE, Roux J, Chavarria C, Gilad Y. A Comparative Assessment of Human and
916 Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. *Sci Rep*. 2018;8: 1–14.
917 doi:10.1038/s41598-018-33478-9
- 918 38. Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Zhou X, Ward MC, Banovich NE, et al. A panel
919 of induced pluripotent stem cells from chimpanzees: a resource for comparative functional
920 genomics. *eLife Sciences*. 2015;4: e07103. doi:10.7554/eLife.07103
- 921 39. Ward MC, Gilad Y. A generally conserved response to hypoxia in iPSC-derived cardiomyocytes from
922 humans and chimpanzees. Regev A, Roy S, Danko CG, editors. *eLife*. 2019;8: e42374.
923 doi:10.7554/eLife.42374
- 924 40. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital
925 transcriptional profiling of single cells. *Nat Commun*. 2017;8: 14049. doi:10.1038/ncomms14049

926 41. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for
927 Dimension Reduction. arXiv:1802.03426 [cs, stat]. 2020 [cited 13 Sep 2021]. Available:
928 <http://arxiv.org/abs/1802.03426>

929 42. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data
930 across different conditions, technologies, and species. *Nature Biotechnology*. 2018 [cited 27 Apr
931 2018]. doi:10.1038/nbt.4096

932 43. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive
933 Integration of Single-Cell Data. *Cell*. 2019;177: 1888-1902.e21. doi:10.1016/j.cell.2019.05.031

934 44. Dominici M, Le Blanc K, Mueller I, Slaper-Cortenbach I, Marini F, Krause D, et al. Minimal criteria
935 for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy
936 position statement. *Cytotherapy*. 2006;8: 315–317. doi:10.1080/14653240600855905

937 45. Wei Y, Tenzen T, Ji H. Joint analysis of differential gene expression in multiple studies using
938 correlation motifs. *Biostatistics*. 2015;16: 31–46. doi:10.1093/biostatistics/kxu038

939 46. Dallas SL, Prideaux M, Bonewald LF. The Osteocyte: An Endocrine Cell ... and More. *Endocr Rev*.
940 2013;34: 658–690. doi:10.1210/er.2012-1026

941 47. Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, et al. A comparison of gene
942 expression and DNA methylation patterns across tissues and species. *Genome Res*. 2020;30: 250–
943 262. doi:10.1101/gr.254904.119

944 48. Gokhman D, Agolia RM, Kinnebrew M, Gordon W, Sun D, Bajpai VK, et al. Human–chimpanzee
945 fused cells reveal cis -regulatory divergence underlying skeletal evolution. *Nature Genetics*. 2021;
946 1–10. doi:10.1038/s41588-021-00804-3

947 49. Hsu Y-H, Estrada K, Evangelou E, Ackert-Bicknell C, Akesson K, Beck T, et al. Meta-Analysis of
948 Genomewide Association Studies Reveals Genetic Variants for Hip Bone Geometry. *Journal of Bone
949 and Mineral Research*. 2019;34: e3698. doi:10.1002/jbmr.3698

950 50. Styrkarsdottir U, Stefansson OA, Gunnarsdottir K, Thorleifsson G, Lund SH, Stefansdottir L, et al.
951 GWAS of bone size yields twelve loci that also affect height, BMD, osteoarthritis or fractures.
952 *Nature Communications*. 2019;10: 2054. doi:10.1038/s41467-019-09860-0

953 51. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-
954 wide association studies for height and body mass index in ~700000 individuals of European
955 ancestry. *Hum Mol Genet*. 2018;27: 3641–3649. doi:10.1093/hmg/ddy271

956 52. Morris JA, Kemp JP, Youlten SE, Laurent L, Logan JG, Chai RC, et al. An atlas of genetic influences on
957 osteoporosis in humans and mice. *Nat Genet*. 2019;51: 258–266. doi:10.1038/s41588-018-0302-x

958 53. Steinberg J, Southam L, Butterfield NC, Roumeliotis TI, Fontalis A, Clark MJ, et al. Decoding the
959 genomic basis of osteoarthritis. *bioRxiv*. 2020; 835850. doi:10.1101/835850

54. Tohyama S, Hattori F, Sano M, Hishiki T, Nagahata Y, Matsuura T, et al. Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. *Cell Stem Cell*. 2013;12: 127–137. doi:10.1016/j.stem.2012.09.013
55. 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. Coop G, Barkai N, editors. *eLife*. 2021;10: e62548. doi:10.7554/eLife.62548
56. Dey KK, Hsiao CJ, Stephens M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLOS Genetics*. 2017;13: e1006599. doi:10.1371/journal.pgen.1006599
57. Carbonetto P, Luo K, Dey K, Hsiao J, Stephens M. fastTopics: fast algorithms for fitting topic models and non-negative matrix factorizations to count data. R package version 04-11. Available: <https://github.com/stephenslab/fastTopics>
58. Hoffman GE, Roussos P. Dream: powerful differential expression analysis for repeated measures designs. Gorodkin J, editor. *Bioinformatics*. 2021;37: 192–201. doi:10.1093/bioinformatics/btaa687
59. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014;15: R29. doi:10.1186/gb-2014-15-2-r29
60. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57: 289–300.
61. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl Acids Res*. 2015;43: e47. doi:10.1093/nar/gkv007
62. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS*. 2012;16: 284–287. doi:10.1089/omi.2011.0118