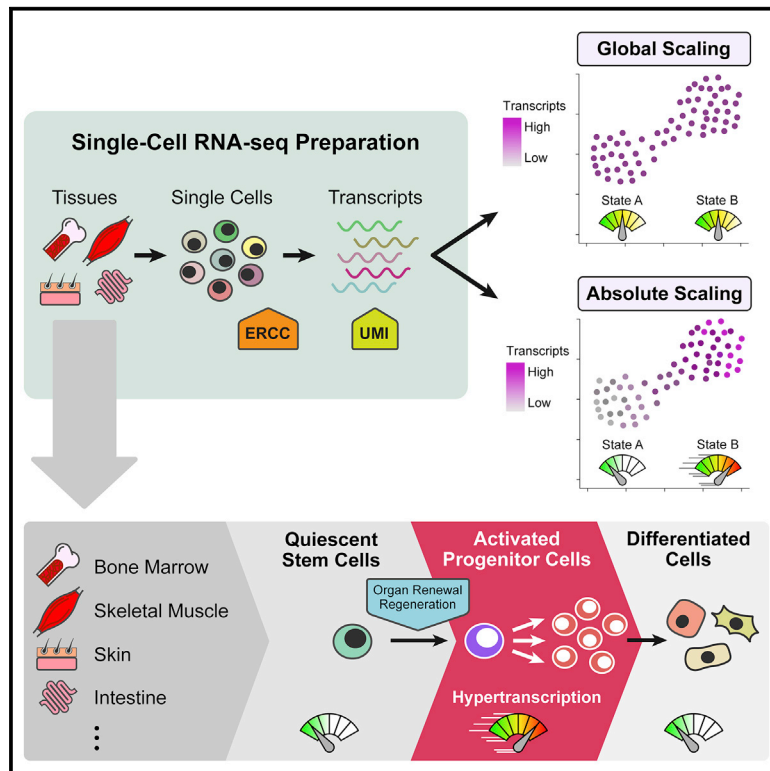# Absolute scaling of single-cell transcriptomes identifies pervasive hypertranscription in adult stem and progenitor cells

Graphical abstract

Authors

Yun-Kyo Kim, Brandon Cho, David P. Cook, Dan Trcka, Jeffrey L. Wrana, Miguel Ramalho-Santos

Correspondence

yun-kyo.kim@sickkids.ca (Y.-K.K.), mrsantos@lunenfeld.ca (M.R.-S.)

In brief

Using absolute scaling to estimate transcript content in single-cell RNA-seq data, Kim et al. identify hypertranscription within adult stem/progenitor cell types of high-turnover organs. Adult hypertranscription states are dynamically regulated during tissue differentiation and regeneration and share molecular features with hypertranscription previously reported in embryonic systems.

Highlights

- Absolute scaling detects global transcriptional output in single-cell RNA-seq data

- Adult cell types display a wide range of transcriptional output

- Hypertranscription is pervasive in activated stem/progenitor cells across organs

- Hypertranscription is induced during regeneration prior to proliferation

Cell Reports

Article

Absolute scaling of single-cell transcriptomes identifies pervasive hypertranscription in adult stem and progenitor cells

Yun-Kyo Kim,[1,2,3,*] Brandon Cho,[2,3] David P. Cook,[2] Dan Trcka,[2] Jeffrey L. Wrana,[2,3] and Miguel Ramalho-Santos[2,3,4,*]

[1]Program in Developmental and Stem Cell Biology, Hospital for Sick Children, Toronto, ON M5G 0A4, Canada
[2]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON M5T 3L9, Canada
[3]Department of Molecular Genetics, University of Toronto, Toronto, ON M5G 1X5, Canada
[4]Lead contact
*Correspondence: yun-kyo.kim@sickkids.ca (Y.-K.K.), mrsantos@lunenfeld.ca (M.R.-S.)
https://doi.org/10.1016/j.celrep.2022.111978

SUMMARY

Hypertranscription supports biosynthetically demanding cellular states through global transcriptome upregulation. Despite its potential widespread relevance, documented examples of hypertranscription remain few and limited to early development. Here, we demonstrate that absolute scaling of single-cell RNA-sequencing data enables the estimation of total transcript abundances per cell. We validate absolute scaling in known cases of developmental hypertranscription and apply it to adult cell types, revealing a remarkable dynamic range in transcriptional output. In adult organs, hypertranscription marks activated stem/progenitor cells with multilineage potential and is redeployed in conditions of tissue injury, where it precedes bursts of proliferation during regeneration. Our analyses identify a common set of molecular pathways associated with both adult and embryonic hypertranscription, including chromatin remodeling, DNA repair, ribosome biogenesis, and translation. These shared features across diverse cell contexts support hypertranscription as a general and dynamic cellular program that is pervasively employed during development, organ maintenance, and regeneration.

INTRODUCTION

Cells dynamically regulate their biosynthetic capacities to fulfill the requirements of demanding cell-state transitions.[1] At the level of transcription, cells can meet these demands by entering a state of relative hypertranscription, which is characterized by a global upregulation of nascent transcriptional output.[2] This global shift powers biological phenomena requiring substantial increases in total biomass, such as rapid proliferation, secretion, and cell activation.[1,2] Hence, hypertranscription has been proposed to play major roles in developmental transitions, adult organ homeostasis, and tumorigenesis.[2] Our group and others have reported evidence supporting the occurrence and functional relevance of hypertranscription in mouse embryogenesis and select human cancer cell types.[3–8] However, the occurrence of hypertranscription in adult physiology remains largely unexplored.

Despite the paucity of data in adult tissues, work in mouse embryos and mouse embryonic stem cells (mESCs) has identified five recurring "hallmarks" of hypertranscription. First, hypertranscribing cells display a transcriptome-wide increase in gene expression, including at housekeeping genes and ribosomal RNA (rRNA).[4,9] These changes occur at the levels of both steady-state and nascent transcripts and result in measurable increases in total cellular RNA content.[4] Second, hypertranscrip-

tion is dependent on a decondensed and distinctly permissive chromatin landscape, maintained by the activity of euchromatic remodeling factors.[10,11] One notable factor is the ATP-dependent remodeler Chd1, which binds specifically to H3K4me3 and is required for the transcriptional output of proliferating epiblast cells.[4] Third, hypertranscribing cells upregulate protein synthesis/translational machinery, which is necessary for the continuous translation and steady-state maintenance of Chd1 and several other unstable euchromatic regulators.[10] Fourth, hypertranscribing cells endogenously accumulate promoter-proximal double-strand breaks (DSBs), thus displaying a heightened dependency on DNA repair factors.[12] Fifth, hypertranscription is thought to be mediated by the expression of general and "universally amplifying" transactivators, of which the Myc family of transcription factors is the best characterized.[3,13]

A critical requirement for detecting hypertranscription is the ability to distinguish absolute differences in transcript expression between samples or cell states. Despite substantial modern advancements in bulk and single-cell transcriptomic profiling, such measurement is complicated by the standard use of between-transcriptome normalization procedures.[14] For instance, normalization to read depth in RNA-sequencing (RNA-seq) data or housekeeping genes in qRT-PCR assumes similar amounts of cellular RNA content between samples of interest. Analysis of single-cell RNA-seq (scRNA-seq) commonly employs similar

global scaling approaches, which are used to minimize the impact of stochastic technical effects, including capture inefficiencies, amplification biases, and variable sequencing depths.[14–16] However, the global scaling tools used to eliminate these technical biases also suppress genuine biological differences in mRNA content, thereby masking the detection of global shifts driven by hypertranscriptive states.[14] Thus, the analysis of hypertranscription at single-cell resolution requires development of alternative methodologies.

In previous work investigating hypertranscription, we circumvented the problem of between-sample normalization by developing cell-number-normalized (CNN) profiling methods, which accurately reproduce absolute cellular transcript levels.[3,4,9] Here, we apply an analogous approach to single-cell transcriptomic data by leveraging unique molecular identifiers (UMIs) or External RNA Controls Consortium (ERCC) spike-in sequences, two tools commonly included in various scRNA-seq protocols.[17–19] We use these tools to perform absolute scaling of scRNA-seq data, which we validate as being able to faithfully capture hypertranscription in ground-truth settings and developmental cases of hypertranscription previously demonstrated using bulk CNN RNA-seq. We apply this methodology to multiple scRNA-seq datasets in heterogeneous adult tissues and report the identification of progenitor cell lineages displaying hallmarks of hypertranscription during homeostasis and regeneration. These results support a model wherein hypertranscription acts as a general mechanism to facilitate dynamic regulation of cellular biosynthetic capacity during development, organ homeostasis, and regeneration.

RESULTS

Absolute scaling accurately estimates transcript content in ground-truth data

To estimate absolute cellular transcript abundances in scRNA-seq data, we focused on protocols using UMIs or ERCCs for normalization and quality control (Figure 1A). ERCC spike-in sequences are typically used in flow cytometric methods (e.g., Smart-seq2, CEL-seq) and are added to lysis buffer solutions at known concentrations. Similar to our previous bulk CNN RNA-seq methodology, the equal distribution of spike-ins prior to capture allows scaling factors to be generated using the ERCC fraction of reads.[3] These scaling factors normalize for downstream technical effects while also retaining biological differences in total transcript abundance.[14] In contrast, UMIs are typically used in droplet and split-pooling methods (10x Genomics, sci-RNA-seq3) and uniquely label each transcript molecule during mRNA capture. Although raw UMI counts retain variation due to per-cell capture efficiency and dropout, we reasoned that they can still provide an acceptable estimate of transcript abundance when analyzed across large enough samples of cells. Collectively, we refer to these approaches to normalization as *absolute scaling*, in contrast to the *global scaling* performed in typical normalization of scRNA-seq data (Figure 1B).

To evaluate whether UMIs and ERCCs can be used for absolute scaling, we explored two ground-truth scenarios. First, we assessed whether using raw UMI counts can distinguish differences in total transcript abundance from UMI-containing libraries gener-

ated from singlets vs. doublets of cells. As doublet libraries are expected to contain roughly twice the mRNA of singlets, they represent one scenario of hypertranscription with mostly uniform amplification across the transcriptome. We used three publicly available 10x Genomics datasets with experimental doublet annotation: in two datasets based on peripheral blood mononuclear cells (Cell Hashing, Demuxlet), we found that doublets contained roughly 1.8 to 1.9 times the total transcripts as singlets, consistent with previous reports (Figure 1C).[20,21] These doublets also displayed similar increases in the expression of ribosomal and housekeeping genes (Figures S1B–S1D). In mixtures of different cell types (10x Mixture), such as human and mouse, transcript content of singlets was reflective of known cell size differences (17,946.0 UMIs in 293T vs. 10,522.0 UMIs in 3T3).[22] Importantly, transcripts in doublets were consistent with the aggregation of these two cell types (mean of 26,222.5 UMIs), indicating accurate capture of total abundance (Figure 1C).

Second, we assessed whether raw UMIs and ERCCs can recapitulate mRNA differences in artificially generated pseudocell libraries. We took advantage of the previously published "scRNA-seq Mixology" experiment, which generated UMI- and ERCC-containing CEL/Sort-seq libraries using extracted mRNA over a four-step dilution series.[23] This series contained material from three different human cell lines and spanned an order of magnitude, sufficiently representing the variation found in hypertranscribing cells.[3,4] As the mRNA content of each single-cell library was known, we evaluated the reproduction of transcript abundance using either raw total UMIs or ERCC-derived per-cell size factors. Compared with global scaling, absolute scaling using either method more accurately reproduced ground-truth differences in pseudocell transcript levels (Figures 1D and 1F). Importantly, changes in transcript abundance were seen to be largely uniform across detected genes, analogous to what would be observed in biological hypertranscription (Figures 1E and 1G). We found that, regardless of platform, ERCC normalization slightly outperformed raw UMI counts and generated lower inter-cell variability in transcript abundance, particularly at higher-ground-truth RNA amounts (Figures 1D–1G, S1E, and S1F). Furthermore, subsampling analysis demonstrated that absolute scaling retains accurate reproduction of ground-truth values even at low cell numbers (Figure S1G). Together, these results document the utility of ERCCs and UMIs in absolute scaling and recovering global transcriptomic differences in scRNA-seq data.

Validation of absolute scaling using embryonic hypertranscription data

We next sought to test absolute scaling in previously established contexts of hypertranscription. The early mouse embryo is especially well characterized in this regard and can be modeled *in vitro* using mESCs.[24] Specifically, serum/LIF-grown mESCs are hypertranscriptional and represent the rapidly proliferating early post-implantation epiblast, while mESCs under dual GSK/MEK inhibition (2i medium) are comparatively hypotranscriptional and represent a pre-implantation-like state.[9,24] We performed absolute scaling on an ERCC-spiked mESC dataset generated using Fluidigm C1 containing both serum and 2i cells.[25] While globally scaled data showed only a modest
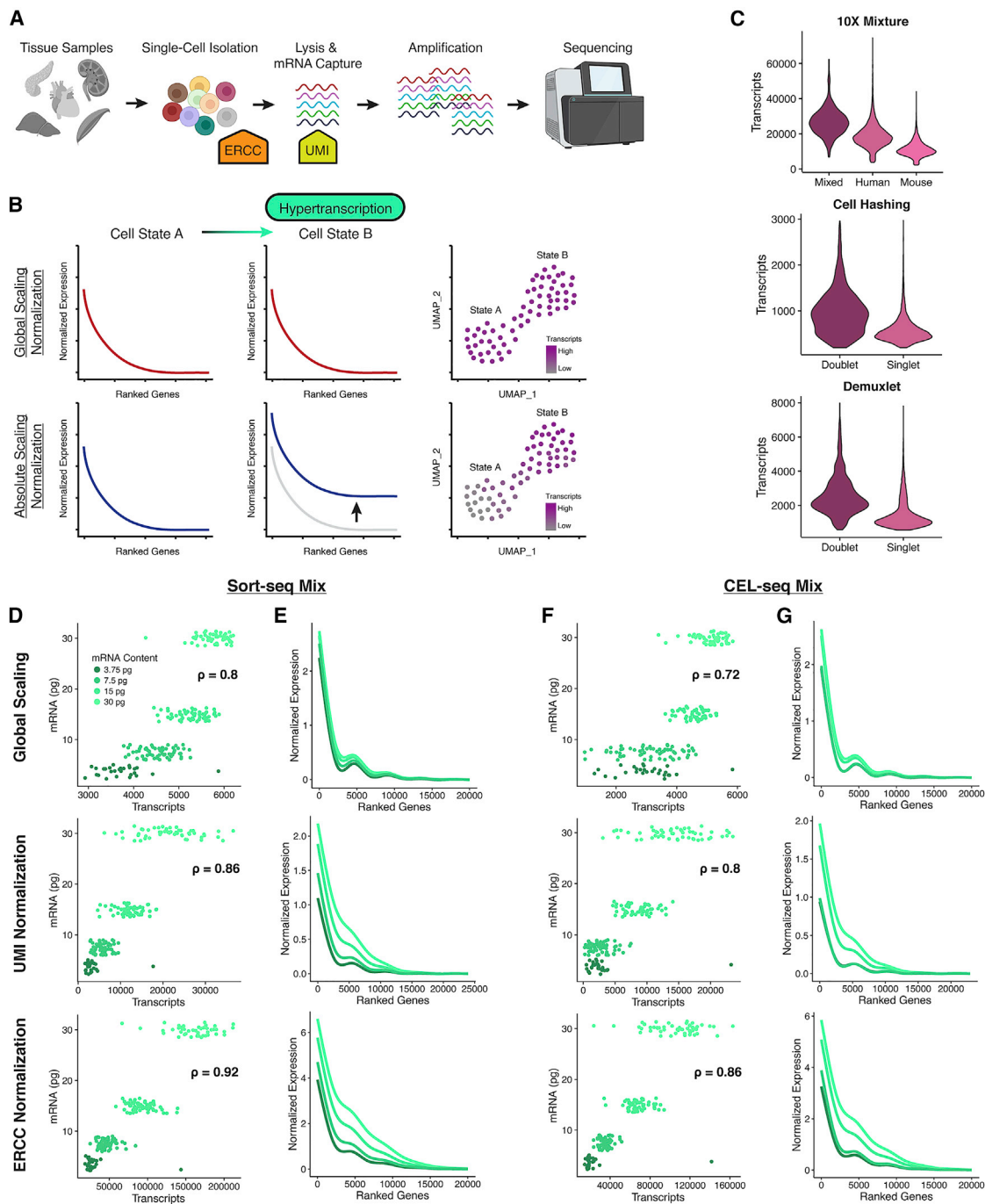
Figure 1. Absolute scaling reproduces ground-truth transcript content differences

(A) Schematic of ERCC and UMI tools within common scRNA-seq protocols. Tissue samples are disassociated into single cells and lysed to capture mRNA transcripts. ERCC spike-in sequences are added to lysed contents and captured alongside endogenous mRNA species. UMIs are added during the capture and reverse transcription steps to uniquely label transcripts.

(B) Model of hypertranscription in scRNA-seq datasets. Absolute scaling preserves large-scale transcriptomic shifts that are typically masked in global scaling approaches due to assumptions of mRNA content parity between cells of a dataset.

(C) Transcript abundance in singlets and doublets within ground-truth datasets. Transcripts represent raw UMIs.

(D and F) Transcript abundances per pseudocell vs. ground-truth mRNA content under different normalization methods. Each point represents a single artificially generated pseudocell transcriptome using defined amounts of cell-line mRNA. ρ values represent Spearman's coefficient.

(E and G) Transcriptome curves depicting gene expression across top 25,000 genes in pseudocells. Individual genes are ranked using combined log2 expression between all pseudocell ground-truth conditions.

(legend on next page)

1.02-fold change between the two cell states, ERCC normalization revealed that serum mESCs contain 2.37-fold higher transcript counts than 2i mESCs, in concordance with findings from bulk CNN RNA-seq data (Figures 2A–2F).[9]

To confirm whether this increase in total transcript abundance is reflective of hypertranscription, we turned to evaluating known hallmarks. We found that under absolute scaling, mESCs display global transcriptomic upregulation across both highly and lowly expressed genes (Figures 2G and 2H). These upregulated genes include chromatin remodelers, DNA repair factors, ribosomal genes, housekeeping genes, and *Myc* (Figures S2A and S2B). Performing Gene Ontology (GO) analysis using differentially expressed genes in serum vs. 2i mESCs extracted from globally scaled data revealed an enrichment for terms associated predominantly with metabolic pathways, mainly highlighting the metabolic differences between the cell states (Figure 2I). In contrast, GO analysis starting with absolute scaling revealed a high enrichment of terms related to translational/ribosomal processes and biosynthesis (Figure 2J). These findings are in strong agreement with our previous bulk-CNN characterizations of hypertranscribing serum vs. 2i mESCs.[9]

We then looked to validate our absolute scaling methodology using data from cells *in vivo*, rather than cultured cells. We have previously shown that during mouse embryogenesis, the midgestation primordial germ cell (PGC) lineage undergoes Myc-driven hypertranscription relative to the surrounding soma.[3] To assess whether hypertranscription can also be captured during this period in scRNA-seq data, we used a 10x Genomics dataset generated from whole mouse gonads at seven time points between E11.5 and P5.[26] When identified by the expression of *Oct4*, globally scaled PGCs appear as having minimal difference in total transcript counts compared with soma (Figures 2K, 2M, and 2O). In contrast, an analysis of raw UMIs revealed that PGC transcript abundances are 1.62- and 2.17-fold increased over soma at E12.5 and E14.5, respectively (Figures 2L, 2N, and 2P). These elevations in UMI counts correspond precisely to the time points of PGC mitotic expansion during development and are accompanied by increases in G2/M- and S-phase cell-cycle scores (Figure 2Q).[26,27]

Similar to serum mESCs, the transcriptomes of E12.5 and E14.5 PGCs normalized by absolute scaling display a high overrepresentation of genes related to hypertranscription hallmarks (Figures S2C and S2D). Importantly, these differential expression changes are largely absent prior to and following the burst in PGC proliferation, in agreement with the restriction of hypertranscription to midgestation development.[3] Together, these data indicate that absolute scaling, whether based on UMI or ERCC-spiked data, accurately reproduces embryonic hypertranscription in scRNA-seq data both in culture and *in vivo*, prompting us to apply it to complex adult cell datasets.

Transcriptional content heterogeneity across adult cell lineages

We previously hypothesized that the rapid proliferation required for turnover in select organs systems may be facilitated in part by hypertranscription.[2] We looked to explore this possibility by applying the absolute scaling methodology described above to existing single-cell atlases of mouse organs. Large-scale scRNA-seq studies have begun to assemble compendia of cellular diversity, which serve as general resources of organism-wide cell expression profiles.[28–30] We chose to focus on the Tabula Muris atlas of 20 mouse organs, which was particularly amenable to absolute scaling due to its inclusion of both ERCC spike-ins and UMIs.[31] Within the atlas, these inclusions are divided between datasets generated using Smart-seq2 (fluorescence-activated cell sorting [FACS]) and 10x Genomics (Droplet) platforms, respectively. In the following analysis, we focused primarily on FACS datasets due to the wider organ selection and greater read depth, as well as the better performance of ERCC normalization for absolute scaling (see Figures 1 and S1).[14] Where relevant, we corroborated significant findings using data based on raw UMI counts.

To apply absolute scaling to the Tabula Muris, we first ensured linear correlations of ERCC expression between organ datasets, as well as validating that individual ERCC species were present at expected frequencies (Figure S3). In addition, we performed filtering for ERCC levels, library size, and doublets for all datasets (Figure S3H, and see STAR methods).[32] These steps resulted in a total of 42,047 ERCC-containing and 45,568 UMI-containing libraries, which upon absolute scaling showed an interquartile range (IQR) of 1,119,112 transcripts and 6,377 UMIs, respectively (Figures S3A–S3D). Where available, FACS and Droplet datasets of the same organ were correlated at the levels of both gene expression and median transcripts per cell (Figures S3J and S3K).

Given the substantial IQR of cellular transcript content in both FACS and Droplet datasets, we next looked to assess whether transcript content differed significantly at the level of individual

---

Figure 2. Detection of embryonic hypertranscription by absolute scaling

(A and B) Uniform approximation and projection method (UMAP) visualization of serum/2i mESCs under global/absolute scaling. Plots depict dimensionality reduction performed under the indicated scaling type.

(C–F) Total transcript counts in serum/2i mESCs under global/absolute scaling. Counts in (C) and (E) represent sums of globally scaled transcripts following log2 scaling, while counts in (D) and (F) represent total transcripts following ERCC normalization. Lines correspond to median values. Statistical test performed was the Wilcoxon rank-sum test.
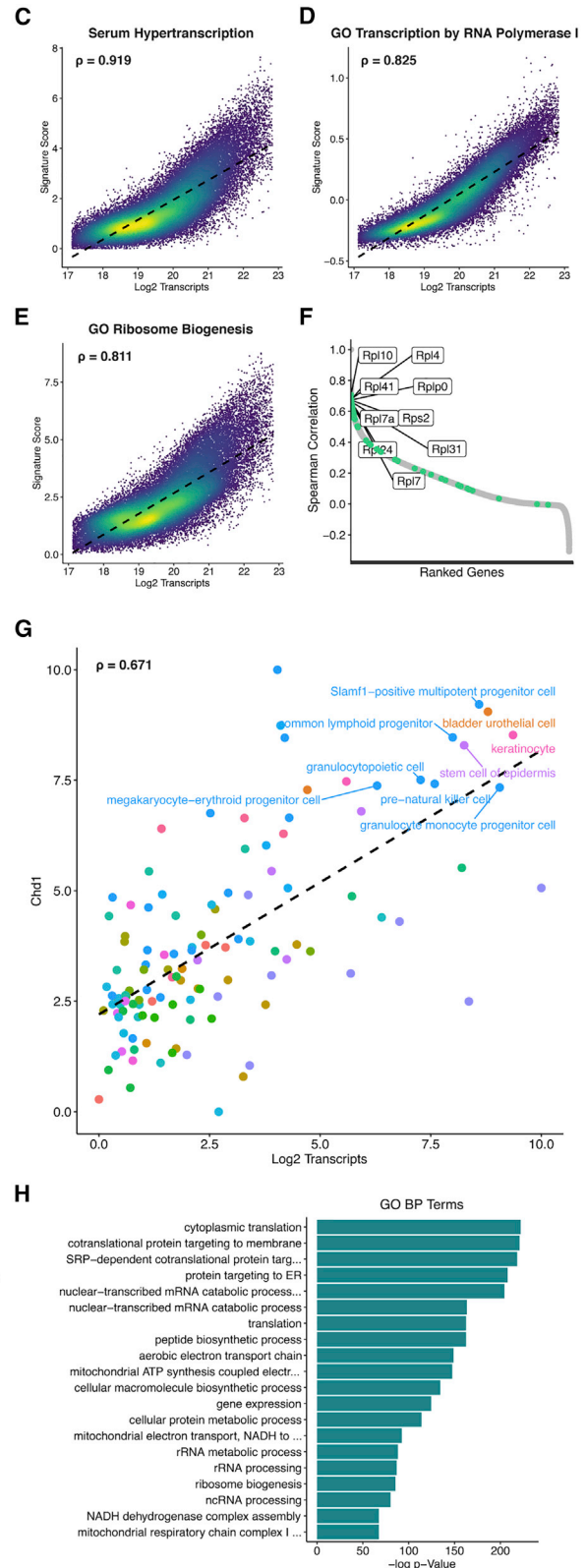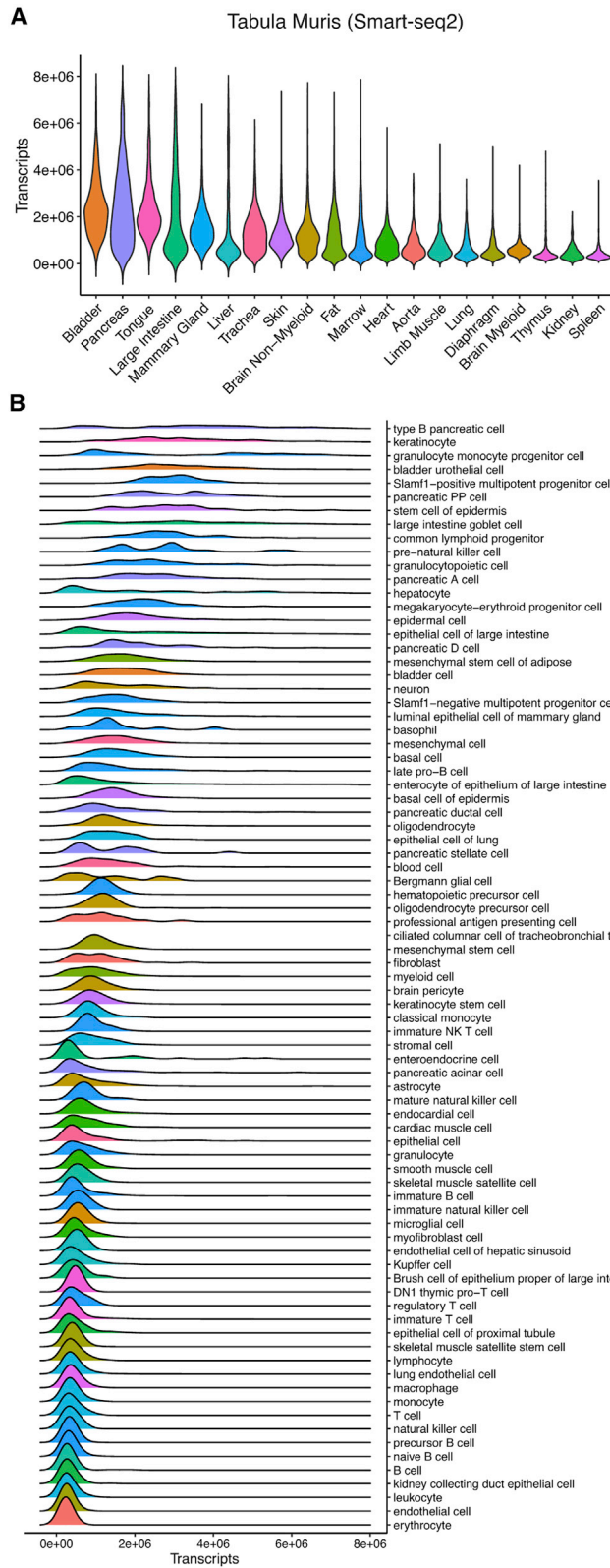
(G and H) Cumulative distribution plots depicting gene expression across top 25,000 genes in serum/2i mESCs. Distributions were computed using log2 expression in both conditions. $p < 2.2 \times 10^{-16}$ between serum/2i (Kolmogorov-Smirnov test) for both absolute and global scaling.

(I and J) Enriched GO terms from ≥2-fold differentially upregulated genes in serum mESCs under global or absolute scaling conditions.

(K and L) UMAP visualization of E12.5 gonad scRNA-seq data under global/absolute scaling containing Oct4$^+$ PGCs and Oct4$^-$ soma. Plots depict dimensionality reduction performed under the indicated scaling type.

(M–P) Transcript abundance in gonad scRNA-seq data under global/absolute scaling. Counts in (M) and (O) represent sums of globally scaled UMIs following log2 scaling, while counts in (N) and (P) represent raw UMIs. Lines correspond to median values and are colored according to cell type.

(Q) Cell-cycle scoring of gonad cells. Phase scores were calculated using cell-cycle markers from globally scaled gene expression data. Lines correspond to median values and are colored according to cell type.

A — Tabula Muris (Smart-seq2)

B

C — Serum Hypertranscription — ρ = 0.919

D — GO Transcription by RNA Polymerase I — ρ = 0.825

E — GO Ribosome Biogenesis — ρ = 0.811

F

G — ρ = 0.671

H — GO BP Terms

(legend on next page)

organ systems. Remarkably, we found that the Tabula Muris displays substantial variation in median transcript counts per organ, spanning a 6.63-fold difference in FACS data and a 3.78-fold difference in 10x data (Figures 3A and S4A; Table S1). The distribution of transcript abundances within organs also varies substantially, with the highest variance in the large intestine (FACS IQR = 2,454,958) or bladder (10x IQR = 10,987) and lowest variance in the spleen (Smart-seq2 IQR = 199,384; 10x IQR = 2,174) (Figures 3A, S4A, and S4B; Table S1). Interestingly, we noticed that datasets derived from the hematopoietic, intestinal, and integumentary organs—systems classically associated with continuous turnover[33]—rank highly among other organs in transcript content and are bimodally distributed (see below).

We next looked to visualize the distribution and heterogeneity of transcript abundance in the Tabula Muris dataset. As absolute scaling changes the structure of the data compared with global scaling (Figure S5), we performed dimensionality reduction under global scaling to maintain the identification of cell types using standard methods, but overlaid transcript counts generated under absolute scaling for visualization. We found that even under these conditions of RNA content parity, the distribution of transcript counts is heterogeneous and a defining characteristic of many cell clusters (Figure S4C). In addition, we observed a widespread presence of transcript abundance gradients between highly separated groups of cells, suggesting intermediate states of cellular transcript content. To further explore these differences, we leveraged cell-type annotations by the Tabula Muris Consortium, which revealed that transcript counts between different cell types span a remarkable 15.79-fold difference in FACS data (over 82 cell types) and an 8.29-fold difference in 10x data (over 56 cell types) (Figures 3B and S4B; Table S1).[31] Interestingly, we found that a large proportion of the highest transcript-count cells are associated with an adult stem cell or progenitor cell identity, particularly from the bone marrow, intestine, and skin datasets. These findings are supported by the observation of higher levels of uridine incorporation in intestinal crypts, where stem and progenitor cells reside (see Jao and Salic[34] and below). In contrast, we found that cell types with the lowest transcript counts are often associated with more highly differentiated or terminal cell states (Figures 3B and S4B; Table S1). Some notable exceptions to this association between elevated transcript counts and stem/progenitor cell states are discussed below. The accuracy of these differences was supported by transcript abundances in T cells, B cells, macrophages, and endothelial cells, which together display minimal variation while spanning multiple organ datasets (Figure S6).

Together, these data reveal the extent of inter- and intraorgan variation with regard to transcript content per cell. As a corollary, these findings suggest that the application of global-scaling approaches in complex tissues, while useful for identifying outlier modules of differential expression, is insufficient to capture important biological information with regard to transcriptional output per cell.

Relationship between elevated transcriptional output and the cell cycle

We next looked to evaluate whether high-content cell types were possible candidates for hypertranscribing populations. As hypertranscription is thought to facilitate rapid proliferation in embryonic contexts, we assessed the relationship between cell cycling and transcript abundance.[2] Using globally scaled expression data, we performed cell-cycle staging as previously described[35] and evaluated the proportion of cycling and noncycling cells within each annotated cell identity. Surprisingly, when identities were ranked by the proportion of non-cycling cells, we observed a wide distribution with regard to total transcript content (Figure S7). Cell types canonically associated with rapid turnover, including progenitor cells of the bone marrow, skin, and intestines, display high proportions of G2/M- and S-phase cells alongside elevated transcript counts. In contrast, several non-cycling G1 cell types of the liver and bladder urothelium also display heightened transcript abundance.

We hypothesized that the properties of these latter cell types might in part be driven by alternative mechanisms, including somatic polyploidy. Polyploidy has been well documented in both hepatocytes and superficial urothelial cells and arises through either incomplete cytokinesis or endoreplication.[36–41] Although we lacked DNA content information to directly measure these processes, we took advantage of previous studies identifying markers specific for polyploid cell populations. Using the Tabula Muris FACS liver dataset, we found that cell clusters with elevated transcript abundance are also enriched in markers of 4n hepatocyte *Mixipl*, *Lifr*, and *Nr1i3*[42] (Figures S8A–S8D). Similarly, we found that high-transcript cells within the FACS bladder dataset co-express *Krt20* and *Upk2* while lacking *Trp63*, a combination specific for 4n + 4n superficial cells[36] (Figures S8E–S8O). Thus, these data suggest that estimation of mRNA abundance using absolute scaling is consistent with expected differences driven by polyploidy.[38]

A further exception to the general association between high transcript content and cycling stem/progenitor cells appears to be secretory terminally differentiated cells, such as goblet and

---

Figure 3. Absolute scaling reveals transcriptional content heterogeneity in the Tabula Muris

(A and B) Distribution of cellular transcript content between FACS organ datasets under absolute scaling, ranked by median total transcripts. Transcripts represent ERCC-normalized reads.

(C–E) Single-cell correlation of transcription signatures with log2 transcript abundance. Signature scores were determined using VISION with absolute-scaled FACS data across the entire atlas. Each point represents a single cell, with color scale depicting plotting density. ρ values represent Spearman's coefficient.

(F) Ranking of genes by Spearman coefficient to log2 transcript abundance across the entire FACS atlas. Highlighted genes represent all expressed Rpl and Rps genes.

(G) Correlation between *Chd1* expression and log2 transcript abundance. Each point is representative of a single averaged cell type, with colors matching organs in (A). ρ values represent Spearman's coefficient.

(H) Enrichment analysis for GO biological process terms using top 3,000 highly correlated genes (to log2 transcript abundance) under absolute scaling.

Paneth cells within the intestinal epithelium (see below). We had previously speculated that the biosynthetic demand of secretory cells may require features of hypertranscription.[2] This possibility is supported by the present analysis but warrants further investigation. Taken together, these analyses indicate that elevated transcriptional counts are generally associated with cycling stem/progenitor cells across multiple organs, with important exceptions observed in terminally differentiated cells that may be in part explained by polyploidy or high secretory output.

Adult cells with high transcript content display hallmarks of hypertranscription

We next investigated the relationship between transcript content and hypertranscription using properties previously defined in bulk RNA-seq studies of embryonic cell populations (see the introduction and Percharde et al.[2]). First, we chose a global approach by evaluating the atlas-wide expression of various transcriptional signatures representative of relevant biological processes. At the level of single cells, we found that transcript abundance is robustly correlated with GO hallmark gene sets previously associated with hypertranscription, such as chromatin organization, RNA polymerase II (Pol II) activity, ribosome biogenesis, DNA repair, and cell division (Figures 3C–3E and S9A–S9C). We also found that a signature of embryonic hypertranscription generated using the earlier Kolodziejczyk et al.[25] dataset is highly correlated with transcript content in adult cells (Figure 3C). These associations are continuous, indicating that cells do not endogenously rest at discrete levels of transcription. Rather, these data suggest that transcriptional output is fine-tuned across a wide spectrum by euchromatin regulation, RNA stability, translation, and other related processes.

We next sought to assess whether these hallmarks are reflected at the level of the expression of key individual regulators of hypertranscription. We found that the absolute-scaled expression of *Chd1*, transcription factors of the *Myc* family (see below), several DSB repair factors, mTOR, and ribosomal proteins is robustly correlated with single-cell transcript content (Figures 3F, 3G, and S9D–S9I; Table S1). In addition, performing enrichment analysis on the set of genes with correlation coefficients of ≥ 0.75 to transcriptional content per cell revealed strong enrichment for GO terms related to translation, transcription, and metabolic processes (Figure 3H; Table S1).

A key hypothesis underlying the mechanistic basis of hypertranscription is the activity of universally amplifying transcription factors, which bind and regulate large portions of the transcriptome. Experimental evidence for both Myc- and Yap-driven hypertranscription has been reported in embryonic contexts.[3,7,8] To explore whether the role of these transcription factors is retained in adult cells, we first used chromatin immunoprecipitation (ChIP) enrichment analysis (ChEA) on the Tabula Muris using genes with a ≥ 0.75 Spearman's coefficient to transcript abundance. ChEA performs gene-set enrichment using protein-DNA interactions derived from ChIP-sequencing and DamID studies.[43] Interestingly, we found that genes with strong correlations to transcript content are highly enriched in Myc and N-Myc ChEA interactions across several biological contexts (Figure S9J; Table S1). In contrast, correlated genes display weaker enrichment in Yap interactions (Table S1). These differences are re-

flected in the overall higher correlation of per-cell transcript content with *Myc* over *Yap* (Figures S9G–S9I). Thus, the Myc family of transcription factors, more so than Yap, is strongly associated with hypertranscription in adult stem/progenitor cells. This analysis further reveals other transcription factors that have not to date been studied in the context of hypertranscription and display chromatin binding patterns strongly linked to transcript content in adult stem/progenitor cells (Figure S9J; Table S1). These transcription factors, such as Eklf, E2f1, and Zfx, deserve to be revisited for potential roles in hypertranscription.

Taken together, these data demonstrate a strong congruity of hypertranscription hallmarks between established embryonic models and stem/progenitor cells of adult tissues.[3–5,7–10,25] These similarities indicate that adult cycling stem/progenitor cells redeploy an embryonic hypertranscriptional program to meet the biosynthetic demands of organ renewal.

Hypertranscription is detected in activated progenitor cells with multilineage potential in adult organs

The presence of hypertranscription hallmarks within adult multipotent stem/progenitor cells led us next to probe the association between transcript content and lineage progression. We first asked whether cellular transcript counts are correlated with any previously established methods of inferring differentiation trajectories in scRNA-seq data. Differentiation progression has been captured using feature counts (number of expressed genes) or quantile polarization scores of principal components of the data (possibly reflecting polarized biological activity), providing the framework for the CytoTRACE and VECTOR trajectory inference tools, respectively.[44,45] Across individual cells of each organ dataset, we found that transcript levels, but not the variance in these levels, are well correlated with both metrics, displaying particularly strong associations with feature counts (Figures S10 and S11). It should be noted, however, that while CytoTRACE and VECTOR focus on defining hierarchies of differentiation, the purpose of absolute scaling is to detect global transcriptomic shifts that occur in hypertranscribing cells. Nonetheless, these data suggest that transcriptional output, alongside transcript diversity, decreases along differentiation trajectories concomitant with a loss of multipotent capacity.

To explore this concept in the context of a well-defined differentiation model, we focused specifically on hematopoiesis within the bone marrow, the progression of which has been extensively characterized.[46] We found that ranking cell types within the Tabula FACS bone marrow dataset by absolute transcript abundance approximates the known hematopoietic hierarchy from multipotent stem cells to terminal differentiated mononuclear cell types (Figures 4A–4C). Importantly, hematopoietic multipotent stem/progenitor cells display the previously explored hallmarks of hypertranscription, including upregulation of chromatin remodelers, DNA repair factors, ribosomal genes, and housekeeping genes (Figure 4D), as well as a substantial enrichment of the embryonic stem cell serum hypertranscription signature (Figure 4E). Moreover, we found that hematopoietic multipotent stem/progenitor cells display higher levels of both steady-state and nascent (intronic) reads, in agreement with the notion that hypertranscription is primarily regulated at the level of nascent transcription[2] (Figure 4F).
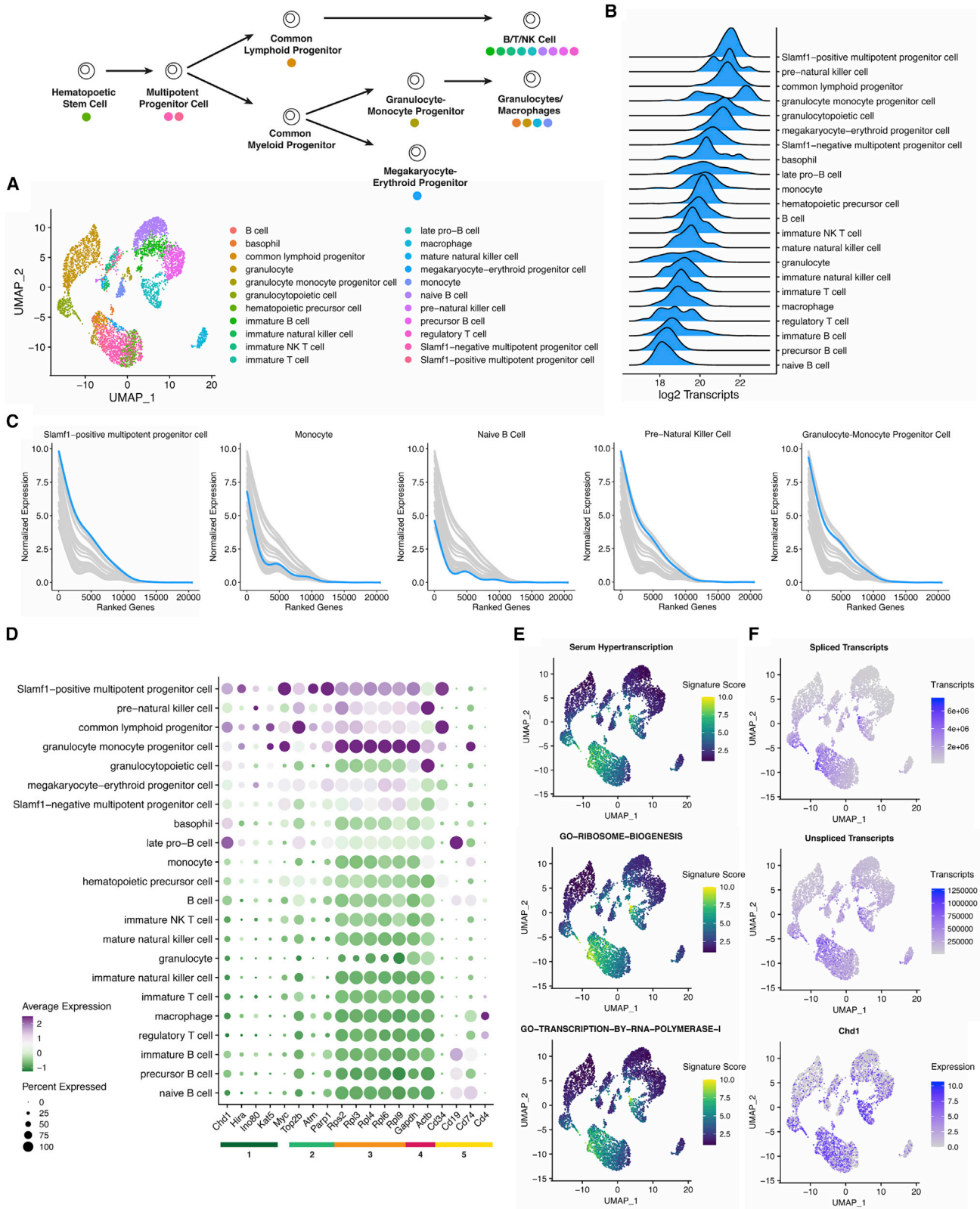
We independently validated these observations in the hematopoietic system by generating an original dataset of single-cell transcriptomes from whole bone marrow on the 10x Genomics platform. After standard filtering, we obtained deeply sequenced transcriptome datasets from 2,454 cells with a mean of 143,767 reads and 19,640 UMIs per cell. Unsupervised clustering and expression-based marker annotation revealed 17 cell types, covering both mature and progenitor populations of the lymphoid/myeloid lineages (Figure S12A). Consistent with bone marrow data in the Tabula Muris FACS and 10x datasets, we identified enrichment of transcript content in multipotent progenitor populations compared with terminally differentiated cell types, as well as corresponding upregulation of hallmarks of hypertranscription (Figures S12B–S12F). Of note, our findings in the scRNA-seq datasets from both Tabula Muris and our lab (Figure S12) are further orthogonally validated by previous measurements of total RNA levels and spike-in normalized RNA-seq in bulk hematopoietic populations.[47,48] Thus, these data support the utility of absolute scaling of scRNA-seq data for uncovering hypertranscription within complex tissues.

Consistent with the model that hypertranscription acts in adult progenitors to facilitate renewal, we identified concordant results in stem cells of the skin and adult large intestine. In the interfollicular epidermis, *Top2a*+ cycling basal stem cells contain elevated transcript abundance compared with superficial epidermal cells (Figure 5).[31] In the intestine, multipotent *Lgr5*+ stem cells in crypt base display elevated steady-state and nascent transcripts compared with the more terminally differentiated enterocyte and brush cell populations (Figures S13A–S13E). These stem cells, which rapidly replenish the colonic epithelium during homeostatic turnover, widely display the previously explored hallmarks of hypertranscription (Figures S13F–S13H).[49] Independent support for our findings is provided by Jao and Salic, who reported that intestinal crypts, where stem/progenitor cells reside, have high levels of incorporation of 5-ethynyl uridine (EU), a marker of nascent transcription.[34]

Interestingly, we also observed a substantial enrichment of transcripts in goblet cells, which are required for continuous mucin production (Figure S13D). To explore the hypertranscriptional status in other secretory cell populations, we turned to examining the gastric pits of the mouse corpus and pylorus.[50] In addition to identifying high transcript content within actively cycling *Mki67*+ isthmus stem cells, we also observed that several slow-dividing secretory cell populations, such as acid-secreting parietal cells and zymogen-secreting chief cells, also display elevated transcript content (Figure S14). Importantly, cells of the gastric epithelium have not been reported to contain high

levels of polyploidy, unlike in other exocrine organs such as the liver or pancreas.[39,51] The presence of high transcript counts in these populations adds further support to a possible role for hypertranscription in driving secretory cell output and is deserving of further investigation.

Dynamic modulation of hypertranscription during adult cell differentiation

We next looked to characterize dynamic changes in hypertranscription through the course of differentiation. We performed pseudotemporal ordering using Monocle on isolated hematopoietic cell trajectories representing the differentiation pathways of granulocytes, erythroblasts, and monocytes.[52] In granulocytes, we found that transcript abundance follows a gradual decrease throughout differentiation toward terminal identities (Figures 6A–6C). To further probe transcriptional dynamics during this differentiation path, we used k-means clustering to organize differentially expressed genes into six groups (Figures 6D–6I). Despite some overlap, these groups are enriched in unique biological processes, such as translational components in group 1 and cell-cycle genes in group 5. Remarkably, we found that only group 1 genes are downregulated substantially faster than the average total downregulation of transcriptional output over pseudotime, while all other groups track with the average or lag behind (colored lines in Figures 6D–6I). These results suggest that in the context of differentiation, the exit from hypertranscription may be primarily initiated via the downregulation of ribosome biogenesis and translational output. The preferential protein instability of euchromatin and transcriptional activators that we have previously reported in embryonic stem cells may similarly, in adult cells, make hypertranscription dependent on high translational capacity.[10]

In the erythroblast and monocyte trajectories, we observed similar results with progressive losses of total transcript content with pseudotime (Figure S15). In addition to this overall decline, we detected small populations of intermediate cells with transient increases in transcript content, reflected most strikingly in the levels of ribosomal and housekeeping genes (Figures S15J and S15K). Interestingly, these brief bursts of increased transcription coincide with the onset of key differentiation effector genes that rapidly rise to very high levels, such as the induction of β-2 hemoglobin in developing erythroblasts (Figure S15N). It is possible that these transient rises in transcriptional output support high levels of synthesis needed for the rapid production of differentiation effector proteins. Importantly, while global scaling would correctly identify the onset of these differentiation markers, it would entirely miss the remarkable dynamics of

---

Figure 4. Hematopoietic progenitors display hallmarks of hypertranscription

(A) UMAP visualization of cell types within the bone marrow dataset using dimensionality reduction performed under absolute scaling.

(B) Distribution of cellular transcript content between cell types of the bone marrow FACS dataset. Transcript counts represent log2 ERCC-normalized reads.

(C) Transcriptome curves depicting gene expression across top 20,000 genes in representative cell types of the bone marrow. Individual genes are ranked using combined log2 expression between all cell types. Indicated cell types correspond to the highlighted curve.

(D) Expression of select genes relevant to hallmarks of hypertranscription, including (1) chromatin remodelers, (2) DNA repair factors, (3) ribosomal genes, (4) housekeeping genes, and (5) hematopoietic markers.

(E) UMAP visualization of signature scores generated by VISION using absolute-scaled expression data.

(F) UMAP visualization of total spliced/unspliced transcripts and *Chd1* expression under absolute scaling. Unspliced or nascent transcripts were assigned by the detection of intronic sequences.
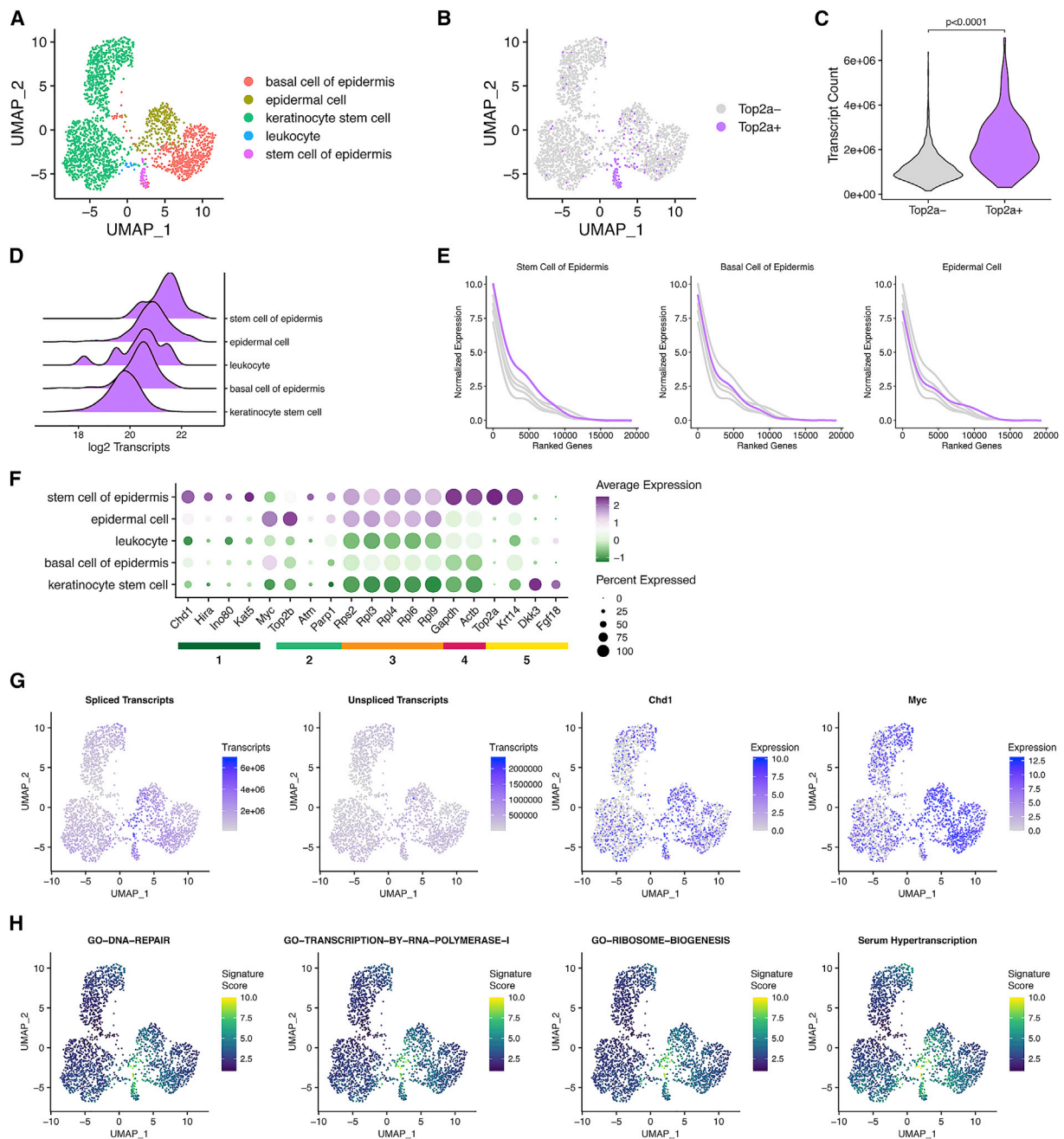
Figure 5. Epidermal progenitors display hallmarks of hypertranscription

(A) UMAP visualization of cell types within the skin FACS dataset.

(B) Visualization of cycling Top2a+ epidermal stem cells within the dataset. Plots in (A) and (B) depict dimensionality reduction performed under absolute scaling.

(C) Transcript content differences between Top2a+ and Top2a− cells.

(D) Distribution of cellular transcript content between cell types of the epidermis, ranked by median content. Transcript counts represent log2 ERCC-normalized reads.

(E) Transcriptome curves depicting gene expression across top 20,000 genes in representative cell types of the epidermis. Individual genes are ranked using combined log2 expression between all cell types. Indicated cell types correspond to the highlighted curve.

(F) Expression of select genes relevant to hallmarks of hypertranscription, including (1) chromatin remodelers, (2) DNA repair factors, (3) ribosomal genes, (4) housekeeping genes, and (5) epidermal compartment markers.

(G) UMAP visualization of total spliced/unspliced transcripts and Chd1 and Myc expression under absolute scaling.

(H) UMAP visualization of signature scores generated by VISION using absolute scaling.
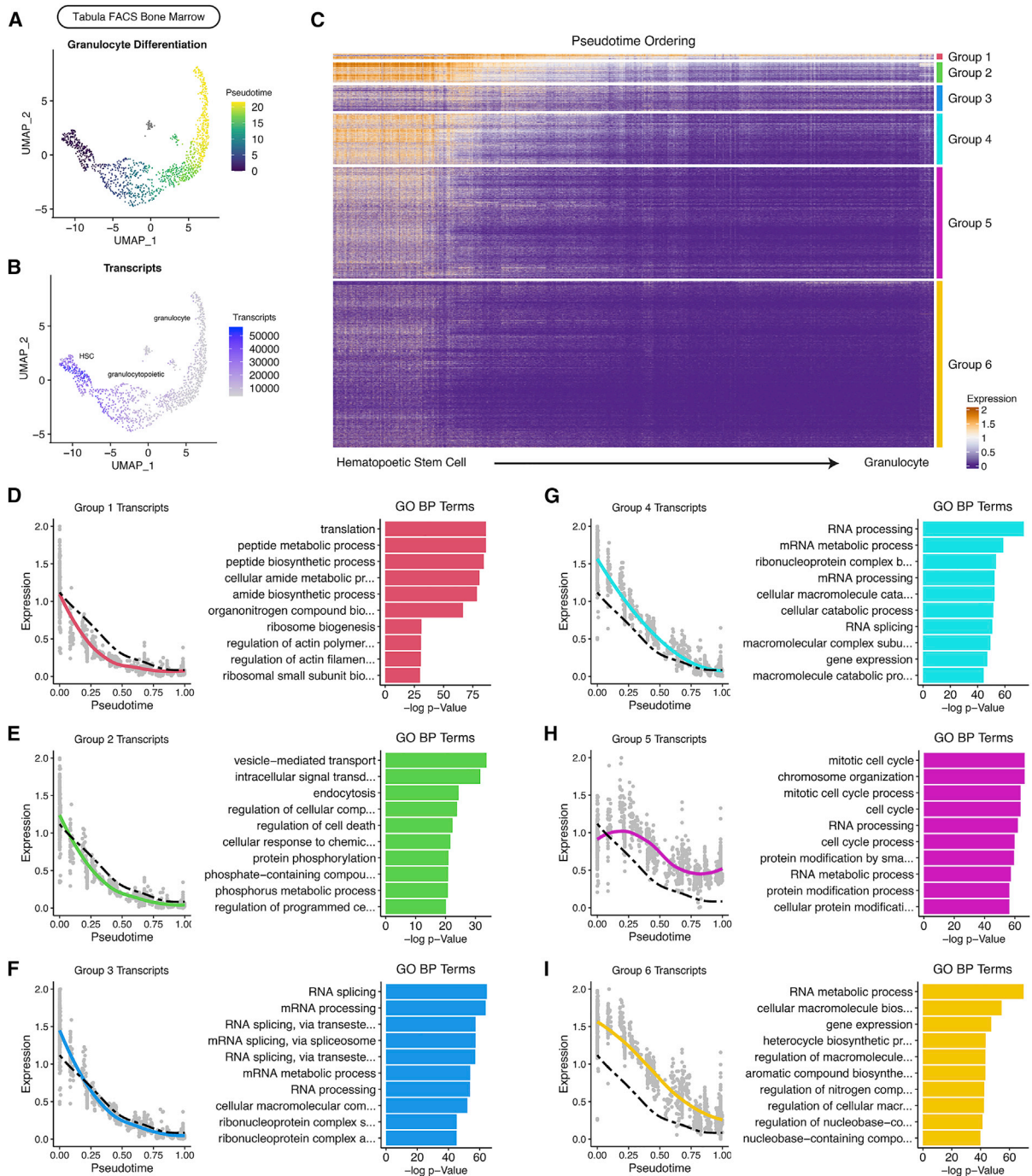
Figure 6. Hypertranscription marks developmental progress in hematopoiesis

(A and B) UMAP visualization of Monocle pseudotime scores and total transcript abundance under absolute scaling. Cd34[+]/cKit[+] hematopoietic stem cells were defined as root cells for trajectory inference analysis.

(C) Heatmap depicting transcriptomic expression changes across differentiation pseudotime from hematopoietic stem cell to granulocyte. Rows depict individual genes and columns depict single-cell transcriptomes ordered by pseudotime values. Group annotations represent k-means clusters generated under absolute scaling, with expression presented as Z scores.

(D–I) Transcriptome curves representing expression of clustered genes across progression of pseudotime. Dotted line represents average expression of all genes (as depicted in [C]). Right side of each column shows enriched GO terms derived from genes within each group. Expression in (C)–(I) represents ranged log2 expression of ERCC-normalized reads.

global transcription during lineage commitment, including large-scale shifts in the levels of the transcription and translation machineries (Figure S15O). Taken together, our analyses uncover a remarkably rich level of dynamics in global transcription at the single-cell level and point to a pervasive redeployment of hypertranscription beyond embryogenesis in adult stem/progenitor cells.

Rapid induction of hypertranscription during adult organ regeneration

The observation that hypertranscription is dynamically regulated during physiological renewal of adult organs led us to explore its status during regeneration. We first focused on the context of skeletal muscle repair, which is mediated by the rapid proliferation and differentiation of resident muscle satellite stem cells.[53] This context was of particular interest as steady-state satellite stem cells display relatively low transcript abundance in the Tabula Muris (see Figure 3 and Table S1). Remarkably, we found that the total transcript content of satellite stem cells ranks highest compared with other cell types across a dataset of cardiotoxin-induced tibialis muscle injury[54] (Figures 7A–7C). When stratified into specific days post-injury (DPI), we found that uninjured and 21 DPI satellite stem cells display comparably low transcript content (Figure 7D). In contrast, 0.5 DPI satellite stem cells display a notable 3.5-fold upregulation of transcriptional output, which peaks at 2 DPI and slowly returns to baseline over the next 19 days. This rapid elevation of transcription upon injury precedes a large increase in the numbers of satellite stem cells, which rise by nearly two orders of magnitude across a 3 day period (Figure 7D). In agreement with our data on hypertranscribing hematopoietic and epithelial stem/progenitor cells, we found that the genes most differentially expressed in 0.5 DPI vs. uninjured satellite stem cells are enriched for functions in ribosome biogenesis, translation regulation, and RNA processing (Figure 7E). Together, these data indicate that muscle satellite stem cells enter a state of injury-induced hypertranscription that precedes rapid expansion and subsequent regeneration.

In a separate model of organ damage, we explored the recovery of the intestinal epithelium after irradiation. Canonical intestinal stem cells, marked by Lgr5 expression, reside at the base of intestinal crypts and renew the epithelium during normal homeostasis.[55] However, recent work has demonstrated that a rare, separate type of cell, named revival stem cell and marked by Clu expression, is responsible for regeneration in response to irradiation.[56] Consistent with our analysis in the Tabula Muris (see Figures 3 and S13), Lgr5$^+$ intestinal stem cells display the highest transcript content among all cell types in the crypt compartment (Figures 7F–7H). Interestingly, upon irradiation-induced loss of Lgr5$^+$ intestinal stem cells, Clu$^+$ revival stem cells assume comparably high levels of transcript content, as well as both cycling and non-cycling compartments, as previously noted (Figure 7H).[56] These findings suggest that, as Clu$^+$ revival stem cells exit quiescence to replenish the crypt niche, they enter hypertranscription to support the biosynthetic requirements of regenerative proliferation and differentiation. Thus, our results indicate that hypertranscription is redeployed in adult stem/progenitor cells both during physiological organ renewal and in different contexts of regeneration upon injury.

DISCUSSION

We report here the development and validation of absolute-scaling approaches to estimate transcript abundance in scRNA-seq data. When applied to a variety of scRNA-seq datasets, absolute scaling accurately captures known cases of hypertranscription in embryonic cells and uncovers a rich level of global transcriptional dynamics in adult tissues. Our analyses indicate that hypertranscription is pervasively redeployed by stem/progenitor compartments across adult organ systems under both homeostatic and regenerative conditions. Thus, rather than being largely ignored, as it is by standard scRNA-seq analysis methods, we propose that hypertranscription is central to the biology of adult organs and needs to be considered in order to understand the mechanisms that underlie their maintenance and regeneration.

Our results document that absolute scaling is capable of capturing large-scale changes in transcriptional output in scRNA-seq data, as long as ERCC exogenous spike-ins or UMI barcodes are used. While both methods remain useful, our data slightly favor the use of ERCCs over UMIs in this context, likely because they are introduced farther upstream in sample processing. Our ground-truth tests and reproduction of embryonic hypertranscription results strongly validate the ability of absolute scaling to detect global changes in transcriptional output. Moreover, our insights on hypertranscription in adult hematopoietic progenitors are validated by an independently generated whole bone marrow dataset that displays strong consistency with both platforms of the Tabula Muris (10x and Smart-seq) and are further supported by data in bulk populations[47,48] (Figure S12; Table S1). Taken together, our results demonstrate that, despite its relatively simple methodology, absolute scaling analysis is highly informative for the study of hypertranscription and its associated features.

Our results highlight the notion that global variations in total cellular mRNA content represent a widely relevant but under-appreciated biological dimension, expanding the role of hypertranscription beyond select examples in development and tumorigenesis. This notion is also supported by early studies dating back to the 1930s, reporting evidence of elevated total RNA content within progenitor cells of the hematopoietic and intestinal systems.[34,57,58] Our work builds upon these studies by comprehensively exploring transcriptional heterogeneity and dynamics at the single-cell and transcriptome-wide levels. In the process, our analyses uncovered putative molecular drivers of hypertranscription in diverse contexts of adult organ homeostasis. We note that global scaling and absolute scaling are not mutually exclusive and can be applied to the same dataset, maximizing the information that is extracted from each experiment. Thus, absolute scaling provides a dimension with which to refine biologically heterogeneous cell populations that may otherwise express similar markers. The capacity to explore hypertranscription in single-cell data will enrich our understanding of many biological questions, notably in adult organ maintenance, regeneration, pathology, and malignancy.

We identify hypertranscription as a feature of active stem/progenitor cells that decays along lineage commitment. We note
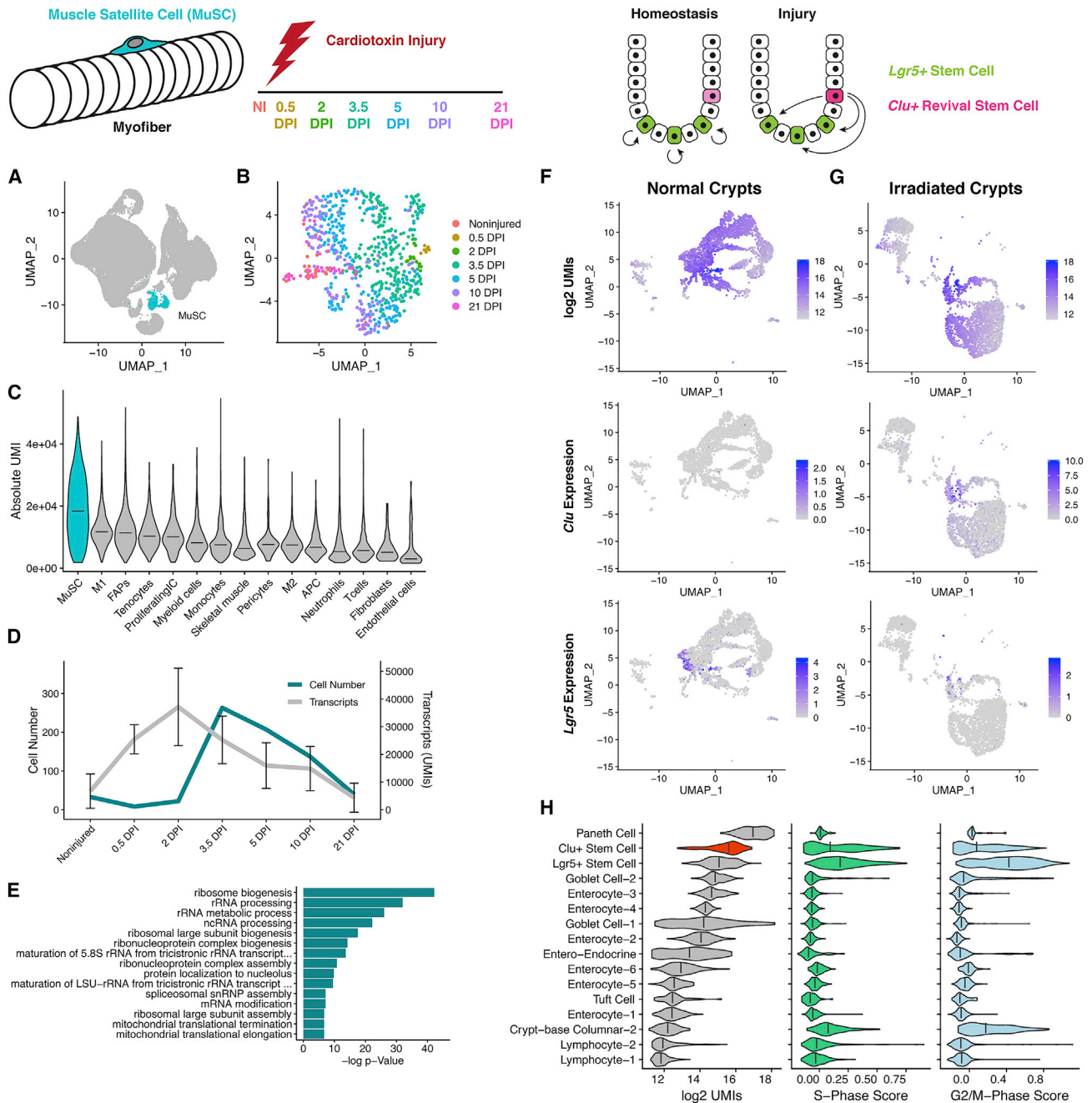
Figure 7. Hypertranscription is deployed during regenerative processes

(A) UMAP visualization of all cells in the muscle repair atlas. Highlighted cells represent annotated satellite cells across all time points. Plot depicts dimensionality reduction performed under absolute scaling.

(B) UMAP visualization of muscle satellite cells. Plot depicts dimensionality reduction performed under absolute scaling.

(C) Distribution of transcript content between cell types across all time points. Lines correspond to median values.

(D) Progression of satellite cell transcript content and cell number through injury and regeneration. Transcript counts represent raw UMIs, error bars represent standard deviation.

(E) GO terms enriched in >1.5-fold differentially upregulated genes in 0.5 DPI vs. uninjured satellite cells.

(F and G) UMAP visualization of transcript abundance, stem cell marker *Lgr5*, and revival stem cell marker *Clu* across all cells in normal/irradiated crypts. Plots depict dimensionality reduction performed under absolute scaling.

(H) Distribution of transcript abundance and cell-phase scores between cell types in normal and irradiated crypts. Cell types are ranked by median log2 UMIs with lines corresponding to median values.

that the transcriptome is not linearly amplified or repressed, but follows patterns such as the primacy of modulating regulators of ribogenesis and translation that are not fully understood and deserve further investigation.[9,10] Moreover, transcript content does not decline strictly linearly down differentiation pathways but rather includes transient fluctuations, possibly due to high synthesis of differentiation effector proteins or expansion of transit amplifying-like cells (see Figure S15). Dissection of these transcriptional dynamics is likely to further reveal aspects of the biology of stem cell compartments. The existence of high-transcript-content cells occupying adult stem cell and progenitor niches greatly expands the prevalence of hypertranscription beyond development and tumorigenesis.[2,13] Our initial analyses reveal many similarities between adult hypertranscription and its embryonic counterpart, but there are likely to be adult-specific and/or organ-specific aspects of the regulation of hypertranscription. It will be of interest to identify the primary drivers of the onset of and exit from hypertranscription in different organs, decipher the chromatin dynamics that underlie these changes, and dissect how they integrate with signaling from the stem/progenitor cell niche, both during physiological organ homeostasis and during regeneration after injury. Our work lays a foundation for the investigation of hypertranscription in adult organs at the single-cell level.

Early histochemical studies dating back to the 1950s reported evidence of elevated total RNA and protein content within renewing cells of planarians and salamanders, but these findings have essentially been forgotten in the genomics era.[59,60] Our results put a focus on hypertranscription and its molecular hallmarks that can now be applied to a plethora of regeneration paradigms being explored with scRNA-seq approaches, such as axolotl limb or mouse digit tip regeneration.[61–64] Of note, it has recently been shown that, during the conversion of fibroblasts to motor neurons, a combination of high proliferative capacity with hypertranscription identifies cells with the greatest probability of faithfully reprogramming.[65] We anticipate that further exploration of hypertranscription in the context of regeneration and cellular reprogramming paradigms will provide fundamental molecular insights that may have a broad impact in regenerative medicine.

Limitations of the study

Despite considerable validation of the absolute scaling methodology, several potential sources of technical variation remain. ERCC-based protocols are dependent on proper experimental aliquoting, and whether various technical and sequencing biases equally affect endogenous and extrinsic sequences remains debated.[14] The accuracy of UMIs, in addition to the aforementioned issues, can also be affected by fluctuations in sequencing depth.[14] As with all scRNA-seq studies, the use of higher cell numbers and greater read depths per cell attenuate sources of potential artifactual noise. We anticipate that these limitations will be aided by the development of improved single-cell protocols that more quantitatively capture the composition of total cellular RNA, including ribosomal and non-coding RNA species.

We report hypertranscription to be frequently deployed in activated stem/progenitor cell compartments. However, it is clear that hypertranscription is not a feature that can on its own be used to accurately identify stem/progenitor cells. Several exceptions are noted in our analyses, such as quiescent vs. activated muscle satellite stem cells (Figure 7). Another example is that of B and T cells, which enter hypertranscription when activated while maintaining their terminally differentiated cellular identity. Other methods, such as CytoTRACE and VECTOR, are better suited to identifying multipotency and cellular hierarchies.[44,45] The identification of hypertranscription as reported here may act best as a complementary approach to such methods, providing orthogonal insights into biosynthetic demand and transcriptional dynamics during differentiation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Mice
- METHODS DETAILS
  - Pre-processing of scRNA-seq data
  - Scaling and dimensional reduction
  - Cell cycle scoring and signature analysis
  - Quantile polarization scores and pseudotemporal ordering
  - Bone marrow sample preparation
  - Sample multiplexing for scRNA-seq
  - scRNA-seq library preparation
  - Processing of raw sequencing reads and sample demultiplexing
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.111978.

AUTHOR CONTRIBUTIONS

Y.-K.K. and M.R.-S. conceived of the project. Y.-K.K. performed all analyses. B.C., D.P.C., D.T., and J.L.W. generated the independent bone marrow scRNA-seq dataset. M.R.-S. supervised the project. Y.-K.K. and M.R.-S. wrote the manuscript with input from all the authors.

REFERENCES

1. Lane, A.N., and Fan, T.W.M. (2015). Regulation of mammalian nucleotide metabolism and biosynthesis. Nucleic Acids Res. *43*, 2466–2485. https://doi.org/10.1093/nar/gkv047.

2. Percharde, M., Bulut-Karslioglu, A., and Ramalho-Santos, M. (2017). Hypertranscription in development, stem cells, and regeneration. Dev. Cell *40*, 9–21. https://doi.org/10.1016/j.devcel.2016.11.010.

3. Percharde, M., Wong, P., and Ramalho-Santos, M. (2017). Global hypertranscription in the mouse embryonic germline. Cell Rep. *19*, 1987–1996. https://doi.org/10.1016/j.celrep.2017.05.036.

4. Guzman-Ayala, M., Sachs, M., Koh, F.M., Onodera, C., Bulut-Karslioglu, A., Lin, C.J., Wong, P., Nitta, R., Song, J.S., and Ramalho-Santos, M. (2015). Chd1 is essential for the high transcriptional output and rapid growth of the mouse epiblast. Development *142*, 118–127. https://doi.org/10.1242/dev.114843.

5. Koh, F.M., Lizama, C.O., Wong, P., Hawkins, J.S., Zovein, A.C., and Ramalho-Santos, M. (2015). Emergence of hematopoietic stem and progenitor cells involves a Chd1-dependent increase in total nascent transcription. Proc. Natl. Acad. Sci. USA *112*, E1734–E1743. https://doi.org/10.1073/pnas.1424850112.

6. Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. Cell *151*, 56–67. https://doi.org/10.1016/j.cell.2012.08.026.

7. Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D.R., Tessarollo, L., Casellas, R., et al. (2012). c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell *151*, 68–79. https://doi.org/10.1016/j.cell.2012.08.033.

8. Lavado, A., Park, J.Y., Paré, J., Finkelstein, D., Pan, H., Xu, B., Fan, Y., Kumar, R.P., Neale, G., Kwak, Y.D., et al. (2018). The hippo pathway prevents YAP/TAZ-Driven hypertranscription and controls neural progenitor number. Dev. Cell *47*, 576–591.e8. https://doi.org/10.1016/j.devcel.2018.09.021.

9. Bulut-Karslioglu, A., Biechele, S., Jin, H., MacRae, T.A., Hejna, M., Gertsenstein, M., Song, J.S., and Ramalho-Santos, M. (2016). Inhibition of mTOR induces a paused pluripotent state. Nature *540*, 119–123. https://doi.org/10.1038/nature20578.

10. Bulut-Karslioglu, A., Macrae, T.A., Oses-Prieto, J.A., Covarrubias, S., Percharde, M., Ku, G., Diaz, A., McManus, M.T., Burlingame, A.L., and Ramalho-Santos, M. (2018). The transcriptionally permissive chromatin state of embryonic stem cells is acutely tuned to translational output. Cell Stem Cell *22*, 369–383.e8. https://doi.org/10.1016/j.stem.2018.02.004.

11. Gaspar-Maia, A., Alajem, A., Polesso, F., Sridharan, R., Mason, M.J., Heidersbach, A., Ramalho-Santos, J., McManus, M.T., Plath, K., Meshorer, E., and Ramalho-Santos, M. (2009). Chd1 regulates open chromatin and pluripotency of embryonic stem cells. Nature *460*, 863–868. https://doi.org/10.1038/nature08212.

12. Bulut-Karslioglu, A., Jin, H., Kim, Y.K., Cho, B., Guzman-Ayala, M., Williamson, A.J.K., Hejna, M., Stötzel, M., Whetton, A.D., Song, J.S., and Ramalho-Santos, M. (2021). Chd1 protects genome integrity at promoters to sustain hypertranscription in embryonic stem cells. Nat. Commun. *12*, 4859. https://doi.org/10.1038/s41467-021-25088-3.

13. Wolf, E., Lin, C.Y., Eilers, M., and Levens, D.L. (2015). Taming of the beast: shaping Myc-dependent amplification. Trends Cell Biol. *25*, 241–248. https://doi.org/10.1016/j.tcb.2014.10.006.

14. Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J.C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat. Methods *14*, 565–571. https://doi.org/10.1038/nmeth.4292.

15. Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. *17*, 63. https://doi.org/10.1186/s13059-016-0927-y.

16. Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Res. *5*, F1000 Faculty Rev-182. https://doi.org/10.12688/f1000research.7223.1.

17. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.

18. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166. https://doi.org/10.1038/nmeth.2772.

19. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I., and Enard, W. (2018). Quantitative single-cell transcriptomics. Brief. Funct. Genom. *17*, 220–232. https://doi.org/10.1093/bfgp/ely009.

20. Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. *19*, 224. https://doi.org/10.1186/s13059-018-1603-1.

21. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol. *36*, 89–94. https://doi.org/10.1038/nbt.4042.

22. Chen, Y., Rivers-Auty, J., Cricǎ, L.E., Barr, K., Rosano, V., Arranz, A.E., Loret, T., Spiller, D., Bussy, C., Kostarelos, K., and Vranic, S. (2021). Dynamic interactions and intracellular fate of label-free, thin graphene oxide sheets within mammalian cells: role of lateral sheet size. Nanoscale Adv. *3*, 4166–4185. https://doi.org/10.1039/d1na00133g.

23. Tian, L., Dong, X., Freytag, S., Lê Cao, K.A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T.S., Seidi, A., Jabbari, J.S., et al. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat. Methods *16*, 479–487. https://doi.org/10.1038/s41592-019-0425-8.

24. Lim, P.S.L., and Meshorer, E. (2021). Organization of the pluripotent genome. Cold Spring Harbor Perspect. Biol. *13*, a040204. https://doi.org/10.1101/cshperspect.a040204.

25. Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C.H., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Bühler, M., Liu, P., et al. (2015). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell *17*, 471–485. https://doi.org/10.1016/j.stem.2015.09.011.

26. Niu, W., and Spradling, A.C. (2020). Two distinct pathways of pregranulosa cell differentiation support follicle formation in the mouse ovary. Proc. Natl. Acad. Sci. USA *117*, 20015–20026. https://doi.org/10.1073/PNAS.2005570117.

27. Tam, P.P.L., and Snow, M.H.L. (1981). Proliferation and migration of primordial germ cells during compensatory growth in mouse embryos. J. Embryol. Exp. Morphol. *64*, 133–147. https://doi.org/10.1242/dev.64.1.133.

28. Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309. https://doi.org/10.1038/s41586-020-2157-4.

29. Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the mouse cell atlas by microwell-seq. Cell *172*, 1091–1107.e17. https://doi.org/10.1016/j.cell.2018.02.001.

30. Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*, eaba7721. https://doi.org/10.1126/science.aba7721.

31. Tabula Muris Consortium (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367–372. https://doi.org/10.1038/s41586-018-0590-4.

32. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. *8*, 329–337.e4. https://doi.org/10.1016/j.cels.2019.03.003.

33. Sender, R., and Milo, R. (2021). The distribution of cellular turnover in the human body. Nat. Med. *27*, 45–48. https://doi.org/10.1038/s41591-020-01182-9.

34. Jao, C.Y., and Salic, A. (2008). Exploring RNA transcription and turnover in vivo by using click chemistry. Proc. Natl. Acad. Sci. USA *105*, 15779–15784. https://doi.org/10.1073/pnas.0808480105.

35. Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. *25*, 1860–1872. https://doi.org/10.1101/gr.192237.115.

36. Wang, J., Batourina, E., Schneider, K., Souza, S., Swayne, T., Liu, C., George, C.D., Tate, T., Dan, H., Wiessner, G., et al. (2018). Polyploid superficial cells that maintain the urothelial barrier are produced via incomplete cytokinesis and endoreplication. Cell Rep. *25*, 464–477.e4. https://doi.org/10.1016/j.celrep.2018.09.042.

37. Celton-Morizur, S., and Desdouets, C. (2010). Polyploidization of liver cells. Adv. Exp. Med. Biol. *676*, 123–135. https://doi.org/10.1007/978-1-4419-6199-0_8.

38. Martin, N.C., McCullough, C.T., Bush, P.G., Sharp, L., Hall, A.C., and Harrison, D.J. (2002). Functional analysis of mouse hepatocytes differing in DNA content: volume, receptor expression, and effect of IFNγ. J. Cell. Physiol. *191*, 138–144. https://doi.org/10.1002/jcp.10057.

39. Donne, R., Saroul-Aïnama, M., Cordier, P., Celton-Morizur, S., and Desdouets, C. (2020). Polyploidy in liver development, homeostasis and disease. Nat. Rev. Gastroenterol. Hepatol. *17*, 391–405. https://doi.org/10.1038/s41575-020-0284-x.

40. Wang, M.J., Chen, F., Lau, J.T.Y., and Hu, Y.P. (2017). Hepatocyte polyploidization and its association with pathophysiological processes. Cell Death Dis. *8*, e2805. https://doi.org/10.1038/CDDIS.2017.167.

41. Ni, V.V., Stein, G.I., Maitesyan, E.S., and Kudryavtsev, B.N. (1988). The gallocyanin-chromalum-RNA content in hepatocytes of different ploidy. Tsitologiya *30*, 354–358.

42. Richter, M.L., Deligiannis, I.K., Yin, K., Danese, A., Lleshi, E., Coupland, P., Vallejos, C.A., Matchett, K.P., Henderson, N.C., Colome-Tatche, M., and Martinez-Jimenez, C.P. (2021). Single-nucleus RNA-seq2 reveals functional crosstalk between liver zonation and ploidy. Nat. Commun. *12*, 4264. https://doi.org/10.1038/s41467-021-24543-5.

43. Keenan, A.B., Torre, D., Lachmann, A., Leong, A.K., Wojciechowicz, M.L., Utti, V., Jagodnik, K.M., Kropiwnicki, E., Wang, Z., and Ma'ayan, A. (2019). ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. *47*, W212–W224. https://doi.org/10.1093/nar/gkz446.

44. Zhang, F., Li, X., and Tian, W. (2020). Unsupervised inference of developmental directions for single cells using VECTOR. Cell Rep. *32*, 108069. https://doi.org/10.1016/j.celrep.2020.108069.

45. Gulati, G.S., Sikandar, S.S., Wesche, D.J., Manjunath, A., Bharadwaj, A., Berger, M.J., Ilagan, F., Kuo, A.H., Hsieh, R.W., Cai, S., et al. (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. Science *367*, 405–411. https://doi.org/10.1126/science.aax0249.

46. Jagannathan-Bogdan, M., and Zon, L.I. (2013). Hematopoiesis. Development *140*, 2463–2467. https://doi.org/10.1242/dev.083147.

47. Signer, R.A.J., Magee, J.A., Salic, A., and Morrison, S.J. (2014). Haematopoietic stem cells require a highly regulated protein synthesis rate. Nature *509*, 49–54. https://doi.org/10.1038/nature13035.

48. Signer, R.A.J., Qi, L., Zhao, Z., Thompson, D., Sigova, A.A., Fan, Z.P., Demartino, G.N., Young, R.A., Sonenberg, N., and Morrison, S.J. (2016). The rate of protein synthesis in hematopoietic stem cells is limited partly by 4E-BPs. Genes Dev. *30*, 1698–1703. https://doi.org/10.1101/gad.282756.116.

49. Barker, N., Van Es, J.H., Kuipers, J., Kujala, P., Van Den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P.J., and Clevers, H. (2007). Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature *449*, 1003–1007. https://doi.org/10.1038/nature06196.

50. Busslinger, G.A., Weusten, B.L.A., Bogte, A., Begthel, H., Brosens, L.A.A., and Clevers, H. (2021). Human gastrointestinal epithelia of the esophagus, stomach, and duodenum resolved at single-cell resolution. Cell Rep. *34*, 108819. https://doi.org/10.1016/j.celrep.2021.108819.

51. Pohl, M.N., and Swartz, F.J. (1979). Development of polyploidy in B-cells of normal and diabetic mice. Acta Endocrinol. *90*, 295–306. https://doi.org/10.1530/acta.0.0900295.

52. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502. https://doi.org/10.1038/s41586-019-0969-x.

53. Yin, H., Price, F., and Rudnicki, M.A. (2013). Satellite cells and the muscle stem cell niche. Physiol. Rev. *93*, 23–67. https://doi.org/10.1152/physrev.00043.2011.

54. Oprescu, S.N., Yue, F., Qiu, J., Brito, L.F., and Kuang, S. (2020). Temporal dynamics and heterogeneity of cell populations during skeletal muscle regeneration. iScience *23*, 100993. https://doi.org/10.1016/j.isci.2020.100993.

55. Zhu, G., Hu, J., and Xi, R. (2021). The cellular niche for intestinal stem cells: a team effort. Cell Regen. *10*, 1. https://doi.org/10.1186/s13619-020-00061-5.

56. Ayyaz, A., Kumar, S., Sangiorgi, B., Ghoshal, B., Gosio, J., Ouladan, S., Fink, M., Barutcu, S., Trcka, D., Shen, J., et al. (2019). Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. Nature *569*, 121–125. https://doi.org/10.1038/s41586-019-1154-y.

57. Thornell, B. (1948). Studies on the formation of cellular substances during blood cell production. J. Am. Med. Assoc. *137*, 1265. https://doi.org/10.1001/jama.1948.02890480085039.

58. Caspersson, T., and Schultz, J. (1939). Pentose nucleotides in the cytoplasm of growing tissues. Nature *143*, 602–603. https://doi.org/10.1038/143602c0.

59. Bodemer, C.W., and Everett, N.B. (1959). Localization of newly synthesized proteins in regenerating newt limbs as determined by radioautographic localization of injected methionine-S35. Dev. Biol. *1*, 327–342. https://doi.org/10.1016/0012-1606(59)90032-6.

60. Bodemer, C.W. (1962). Distribution of ribonucleic acid in the regenerating urodele limb as determined by autoradiographic localization of uridine-H3. Anat. Rec. *142*, 457–467. https://doi.org/10.1002/ar.1091420403.

61. Leigh, N.D., Dunlap, G.S., Johnson, K., Mariano, R., Oshiro, R., Wong, A.Y., Bryant, D.M., Miller, B.M., Ratner, A., Chen, A., et al. (2018). Transcriptomic landscape of the blastema niche in regenerating adult axolotl limbs at single-cell resolution. Nat. Commun. *9*, 5153. https://doi.org/10.1038/s41467-018-07604-0.

62. Li, H., Wei, X., Zhou, L., Zhang, W., Wang, C., Guo, Y., Li, D., Chen, J., Liu, T., Zhang, Y., et al. (2021). Dynamic cell transition and immune response landscapes of axolotl limb regeneration revealed by single-cell analysis. Protein Cell *12*, 57–66. https://doi.org/10.1007/s13238-020-00763-1.

63. Gerber, T., Murawala, P., Knapp, D., Masselink, W., Schuez, M., Hermann, S., Gac-Santel, M., Nowoshilow, S., Kageyama, J., Khattak, S., et al. (2018). Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. Science *362*, eaaq0681. https://doi.org/10.1126/science.aaq0681.

64. Storer, M.A., Mahmud, N., Karamboulas, K., Borrett, M.J., Yuzwa, S.A., Gont, A., Androschuk, A., Sefton, M.V., Kaplan, D.R., and Miller, F.D. (2020). Acquisition of a unique mesenchymal precursor-like blastema state underlies successful adult mammalian digit tip regeneration. Dev. Cell *52*, 509–524.e9. https://doi.org/10.1016/j.devcel.2019.12.004.

65. Babos, K.N., Galloway, K.E., Kisler, K., Zitting, M., Li, Y., Shi, Y., Quintino, B., Chow, R.H., Zlokovic, B.V., and Ichida, J.K. (2019). Mitigating antagonism between transcription and proliferation allows near-deterministic cellular reprogramming. Cell Stem Cell *25*, 486–500.e9. https://doi.org/10.1016/j.stem.2019.08.005.

66. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

67. Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. *5*, 2122. https://doi.org/10.12688/f1000research.9501.2.

68. DeTomaso, D., Jones, M.G., Subramaniam, M., Ashuach, T., Ye, C.J., and Yosef, N. (2019). Functional interpretation of single cell similarity maps. Nat. Commun. *10*, 4376. https://doi.org/10.1038/s41467-019-12235-0.

69. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. https://doi.org/10.1038/s41586-018-0414-6.

70. Tian, L., Su, S., Dong, X., Amann-Zalcenstein, D., Biben, C., Seidi, A., Hilton, D.J., Naik, S.H., and Ritchie, M.E. (2018). scPipe: a flexible R/Bio-conductor preprocessing pipeline for single-cell RNA-sequencing data. PLoS Comput. Biol. *14*, e1006361. https://doi.org/10.1371/journal.pcbi.1006361.

71. McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L., Murrow, L.M., Weissman, J.S., Werb, Z., et al. (2019). MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods *16*, 619–626. https://doi.org/10.1038/s41592-019-0433-8.

72. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97. https://doi.org/10.1093/nar/gkw377.

73. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414. https://doi.org/10.1038/s41587-020-0591-3.

74. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.

75. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

76. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinf. *14*, 128. https://doi.org/10.1186/1471-2105-14-128.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Mouse embryonic stem cell dataset	Kolodziejczyk et al.[25]	ArrayExpress: E-MTAB-2600
Single-cell mixology dataset	Tian et al.[23]	GEO: GSE118767
Demuxlet human PBMC dataset	Kang et al.[21]	GEO: GSE96583
Cash Hashing human PBMC dataset	Stoeckius et al.[20]	GEO: GSE108313
Mouse skeletal muscle regeneration dataset	Oprescu et al.[54]	GEO: GSE138826
Mouse gonad/ovary dataset	Niu et al.[26]	GEO: GSE136441
Mouse atlas (Tabula Muris) dataset	Schaum et al.[31]	GEO: GSE109774
Mouse intestine regeneration dataset	Ayyaz et al.[56]	GEO: GSE123516
Mouse stomach dataset	Busslinger et al.[50]	GEO: GSE157694
12K 1:1 HEK293T-NIH3T3 mixture dataset	10x Genomics	Chromium v2 Chemistry Demonstration
Mouse supplementary adult bone marrow dataset	This study	GEO: GSE221184
Code for all figures and scRNA-seq analysis	This study	Zenodo: https://doi.org/10.5281/zenodo.7408025
Experimental models: Organisms/strains		
C57Bl/6-*Vav-Cre*+/-;*Chd1*fl/+	Koh et al.[5]	N/A
Software and algorithms		
Seurat v4	Stuart et al.[66]	https://satijalab.org/seurat/
Scran v1.9	Lun et al.[67]	https://doi.org/10.18129/B9.bioc.scran
VISION v2.1.0	DeTomaso et al.[68]	https://github.com/YosefLab/VISION
Velocyto v0.16.7	La Manno et al.[69]	https://github.com/theislab/scvelo
scPipe v3.12	Tian et al.[70]	https://github.com/LuyiTian/scPipe
DoubletFinder v3	McGinnis et al.[32]	https://github.com/chris-mcginnis-ucsf/DoubletFinder
VECTOR v0.0.3	Zhang et al.[44]	https://github.com/jumphone/Vector
Monocle 3	Cao et al.[52]	http://cole-trapnell-lab.github.io/monocle-release/
Cellranger v6.1.1	10x Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation
Demultiplex v1.0.2	McGinnis et al.[71]	https://github.com/chris-mcginnis-ucsf/MULTI-seq
Enrichr v3.1	Kuleshov et al.[72]	https://maayanlab.cloud/Enrichr/

RESOURCE AVAILABILITY

Lead contact
Requests for further information, resources, and reagents should be directed to and will be fulfilled by the lead contact, Miguel Ramalho-Santos (mrsantos@lunenfeld.ca).

Materials availability
This study did not generate new, unique materials.

Data and code availability
- Single-cell RNA-seq data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice

Vav-Cre+/-;*Chd1*fl/+ mice in a C57Bl/6 background, a phenotypically normal strain, were generated in previously published work.[5] As part of a separate study, mice were fed a tamoxifen-supplemented diet (TD.130860, Envigo) for 9 days, followed by 5 days of standard mouse diet. All procedures involving animals were performed according to the Animals for Research Act of Ontario and the Guidelines of the Canadian Council on Animal Care. The Animal Care Committee of The Center for Phenogenomics, Toronto, reviewed and approved all procedures conducted on animals (Protocol 26-0331). Animals were randomly selected from available mice bred in our facility. Experiments in this study used female animals 6-16 weeks of age in good health.

METHODS DETAILS

Pre-processing of scRNA-seq data

Published count matrices with collapsed UMIs or mapped ERCC sequences were used where available. For CEL/Sort-seq data from Tian et al 2019, count matrices were generated using the scPipe pipeline using genomic references containing ERCC sequence data.[70] Spliced/unspliced count matrices for the bone marrow, intestine, and skin FACS datasets were generated using the scVelo pipeline for analysis of nascent reads.[73]

Seurat objects were used for processing of all scRNA-seq data. Low-quality cells in individual organ datasets were filtered by mitochondrial gene expression, ERCC fraction, and gene/transcript counts. To address the possibility that high-content cells were the result of libraries generated from doublets, we performed heterotypic doublet removal on Smartseq2 and 10x Genomics datasets using DoubletFinder.[32] The total doublet rate was estimated at 4%, and the homotypic doublet rate was derived from Tabula Muris cell-type annotations. All doublet removal was performed on globally-scaled data with standard processing.

Scaling and dimensional reduction

Absolute scaling analysis of ERCC datasets was performed using the *scran* package.[67] To ensure that transcript counts between organs were comparable, individual organ count matrices were merged and per-cell size factors were generated from spike-in data using the computeSpikeFactors function. This calculation is derived from the sum of all spike-in reads and generates a set of size factors with a mean of unity. Single-cell libraries were then normalized to size factors and log2 scaled. For absolute scaling analysis of UMI datasets, raw UMI counts were used following log2 scaling. Global scaling was performed on ERCC and UMI datasets by first removing spike-in data (if present) and normalizing for library size using the relative counts methods in Seurat's NormalizeData function. Results were subsequently log2 scaled and used for analysis.

Standard dimensional reduction was performed using Seurat functions with detailed parameters available in code. Briefly, highly variable features were scaled and used for principal component analysis, with the number of components determined using resampling tests.[74] Data was subsequently visualized using the Uniform Approximation and Projection Method (UMAP).

Cell cycle scoring and signature analysis

To determine cell phase, we used the CellCycleScoring method from Seurat with markers as previously published.[35] Briefly, individual cells were scored based on the expression of either G2/M or S phase cell markers. Cells lacking expression of either marker set were identified to be in a non-cycling G1 phase. For signature analysis, we used the VISION tool to calculate signature scores using gene sets derived from the Molecular Signatures Database.[68,75] Signatures were calculated using absolute-scaled data without internal transformation within VISION. For the "Serum Hypertranscription" signature, >2.5-fold differentially upregulated genes in serum mESCs (Kolodziejczyk et al 2015) were used to generate the gene set.

Quantile polarization scores and pseudotemporal ordering

Quantile polarization scores were calculated with VECTOR using the vector.getValue function. As previously described, we determined quantile polarization values on absolute-scaled data using the top 150 principal components.[44] For pseudotemporal analysis with Monocle 3, we converted pre-processed and absolute-scaled Seurat objects into Monocle3 cds objects. Cells were clustered on previously generated UMAP reductions using the Leiden algorithm. To construct single-cell trajectories, hematopoietic progenitors expressing *Cd34* and *Kit* were used as root cells. Principal graphs were constructed and used for pseudotemporal ordering.

Bone marrow sample preparation

Femur and tibiae were collected from 2 mice and flushed with 5mL of cold complete media (RPMI +Glutamax, 10% fetal bovine serum (FBS), 1% PenStrep) per mouse. Bone marrow was subsequently passed through a 70μm cell strainer, and collected cells were washed in 25mL of complete media and centrifuged at 600g for 4 minutes at 4°C. Red blood cell lysis was then performed in 3mL of ACK Lysis Buffer (A1049201, Gibco) for 3-5 minutes at room temperature. Cells were washed with 25mL complete media and cryopreserved at -80°C in freezing medium containing 15% DMSO, 20% FBS and 65% complete media at a final concentration of 1.0 x 10^7 cells/mL.

Sample multiplexing for scRNA-seq

Multiplexing was performed according to the MULTI-seq protocol.[71] Individual samples were thawed and both viability and cell number were assessed using trypan blue exclusion. 2 x 105 cells from each sample were pelleted in microcentrifuge tubes at 400g for 5 minutes. Cells were resuspended in 200uL of a 200nM lipid-modified anchor (kindly provided by Dr. Zev Gartner from University of California, San Francisco) and 200nM barcode solution in PBS. A unique 8nt barcode harboring a synthetic polyA tail was added to each sample to facilitate computational demultiplexing. Sample were incubated at room temperature (RT) for 10 minutes, mixed by pipetting approximately every 2.5-3 minutes to ensure homogeneous labelling. After 10 minutes, samples were supplemented with 20uL of a 20nM co-anchor solution to stabilize membrane residency of the barcode. Samples were incubated for an additional 5 minutes on ice, mixed by pipetting half-way through the incubation. The samples were then centrifuged at 400g for 5 minutes. The supernatant was removed, and the cell pellet was then resuspended in 500uL of fresh PBS + 1% BSA. The cells were then centrifuged again at 400g for 5 minutes. The supernatant was removed, and the cells were resuspended in 50uL of PBS + 1% BSA. Samples were then pooled together at a 1:1 ratio before assessing cell viability and concentration by trypan blue exclusion.

scRNA-seq library preparation

The pooled samples were diluted to a concentration of 1000 cells/uL. The single-cell suspension was processed using three reactions from the 10x Genomics Single Cell 3' v3 RNA-seq kit targeting a yield of 15,000 cells per reaction. Gene expression libraries were prepared according to the manufacturer's protocol. As per the MULTI-seq protocol, sample barcode libraries were isolated by collecting the supernatant during the first size exclusion step during cDNA library preparation.[71] A ubiquitous sequence in each sample barcode was used as a PCR handle, allowing for amplification of the barcode libraries. Final libraries were sequenced on a NovaSeq6000 S1 flow cell (Illumina) to achieve an average depth of >100,000 reads/cell.

Processing of raw sequencing reads and sample demultiplexing

Fastq files for both gene expression and barcode libraries from each reaction were generated using the mkfastq function in cellranger. A gene-by-cell transcript quantification matrix was produced for each of the gene expression libraries using cellranger count using default parameters. Using the R package deMULTIplex, valid cell barcodes called from the quantification were used to query the fastq files associated with the barcode libraries, quantifying the occurrence of each barcode. Each cell was then assigned to a given sample based on its positivity for specific barcodes. Cells with negative signal or positive signal for more than a single barcode were labelled as "Negative" or "Multiplet" and removed from all subsequent analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analysis was performed using functions implemented in R. Spearman correlation was used to identify signatures related to cellular transcript abundance and was calculated using cor. Transcriptome curves were drawn using geom smooth in ggplot2 using a "loess" model. Cumulative distributions plots for mESC transcriptomes were generated using ecdf, and statistical analyses were conducted with ks.test, wilcox.test, and pairwise.wilcox.test functions. For differential gene expression in scRNA-seq data, we used the FindMarkers functions with Seurat and performed gene ontology and ChEA enrichment analysis in Enrichr.[76]

Supplemental information

Absolute scaling of single-cell transcriptomes

identifies pervasive hypertranscription

in adult stem and progenitor cells

Yun-Kyo Kim, Brandon Cho, David P. Cook, Dan Trcka, Jeffrey L. Wrana, and Miguel Ramalho-Santos

Figure S1

Supplementary Figure 1: Absolute Scaling Pipeline and Validation for Single-Cell Datasets. Supplemental to Figure 1.

(A) Summary of bioinformatic tools used for absolute scaling in this study.

(B) Dot plots depicting expression of ribosomal/housekeeping genes under absolute scaling in PBMC ground truth datasets. Expression represents z-scores of log2-scaled raw UMIs.

(C) Average expression of RPL and RPS gene expression in Cell Hashing PBMC dataset.

(D) Average expression of RPL and RPS gene expression in Demuxlet PBMC dataset. Statistical test performed is the Wilcoxon rank sum test.

(E, F) Comparison of UMI vs ERCC absolute scaling. Ranged expression represents log2 raw UMIs or ERCC-normalized reads.

(G) Absolute-scaled transcript abundances per pseudocell vs ground truth mRNA content under random sub-sampled conditions. All absolute scaling was performed post subsampling. Values above plots represent Spearman coefficients between ground truth and absolute-scaled expression values. Comparable results obtained for CEL-seq mixology system (data not shown).
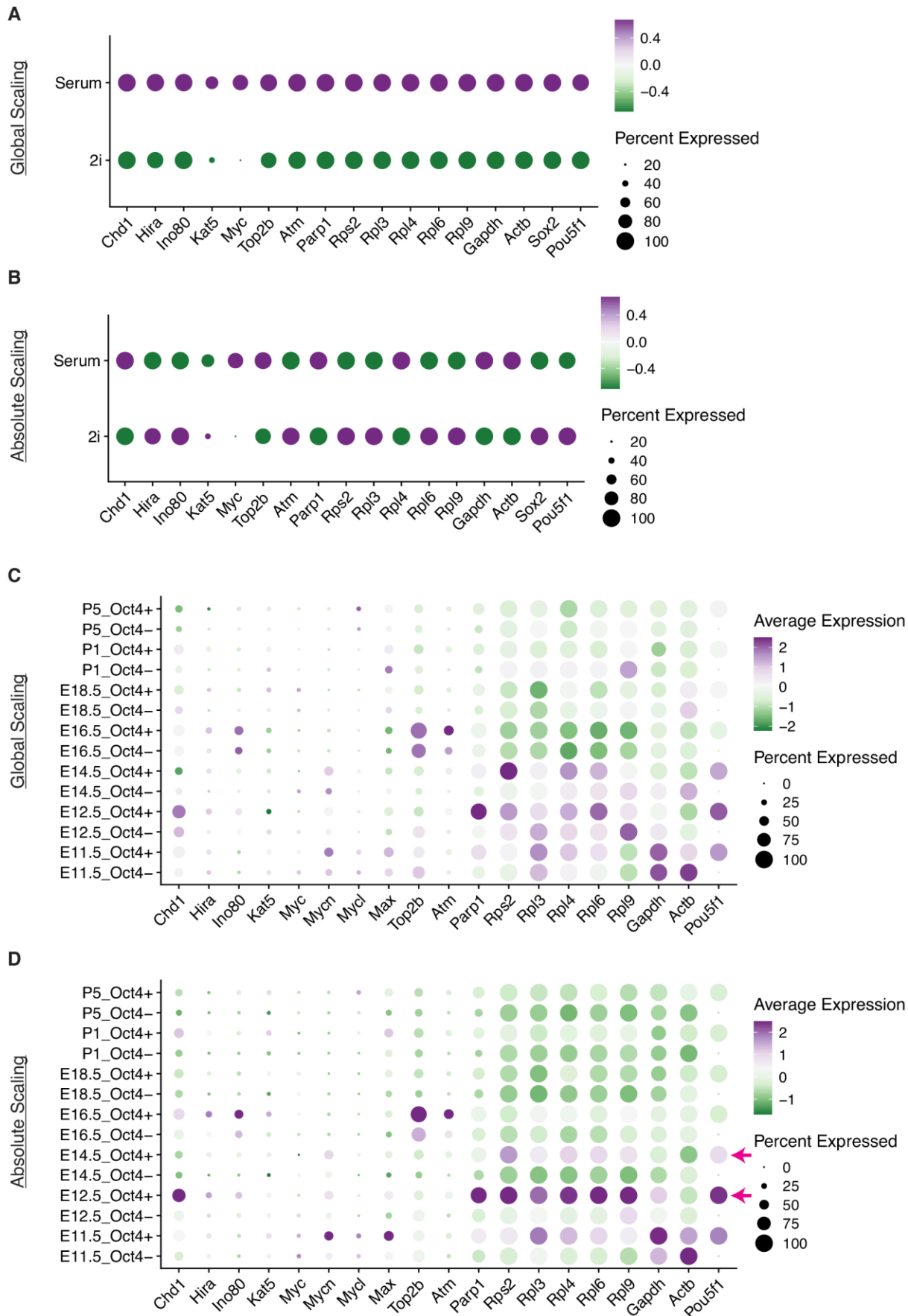
Figure S2

Supplementary Figure 2: Absolute Scaling Captures Hypertranscription Hallmarks in Embryonic Datasets. Supplemental to Figure 2.
(A-B) Expression of genes relevant to hypertranscription hallmarks in serum/2i mESCs under global vs absolute scaling.
(C-D) Expression of genes relevant to hypertranscription hallmarks in Oct4+ PGCs/Oct4- soma under global vs absolute scaling. Expression represents z-scores of log2-scaled transcripts.
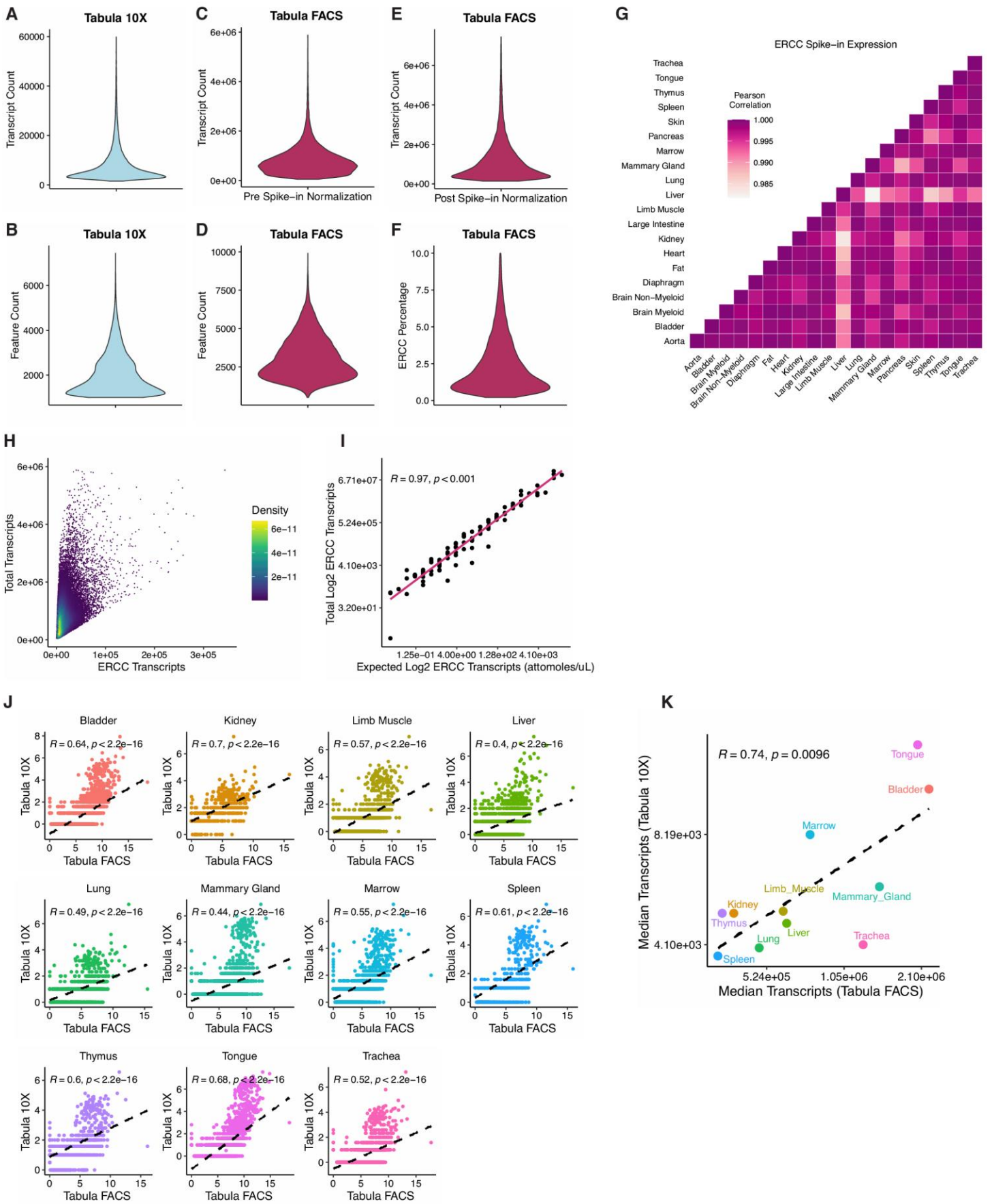
Figure S3

Supplementary Figure 3: Quality Control of Tabula Muris. Supplemental to Figure 3.

(A-F) Transcript and feature counts of Tabula Muris FACS or Droplet atlases.

(G) Linear correlations of ERCC spike-in expression between FACS organ datasets. Values represent Pearson coefficients.

(H) Representation of filtering in FACS datasets for transcript counts and ERCC percentage.

(I) Comparison of actual and expected ERCC species expression based on spike-in mix composition.

(J) Comparison of individual gene expression between FACS and Droplet datasets. Each point represents a single shared gene between FACS and Droplet dataset.

(K) Comparison of average transcript abundance in organ datasets between FACS and Droplet atlases. For **(J-K)**, transcripts in FACS represents log2 ERCC-normalized expression, while values in Droplet represent log2 absolute UMIs. Correlation values represent Pearson coefficients.
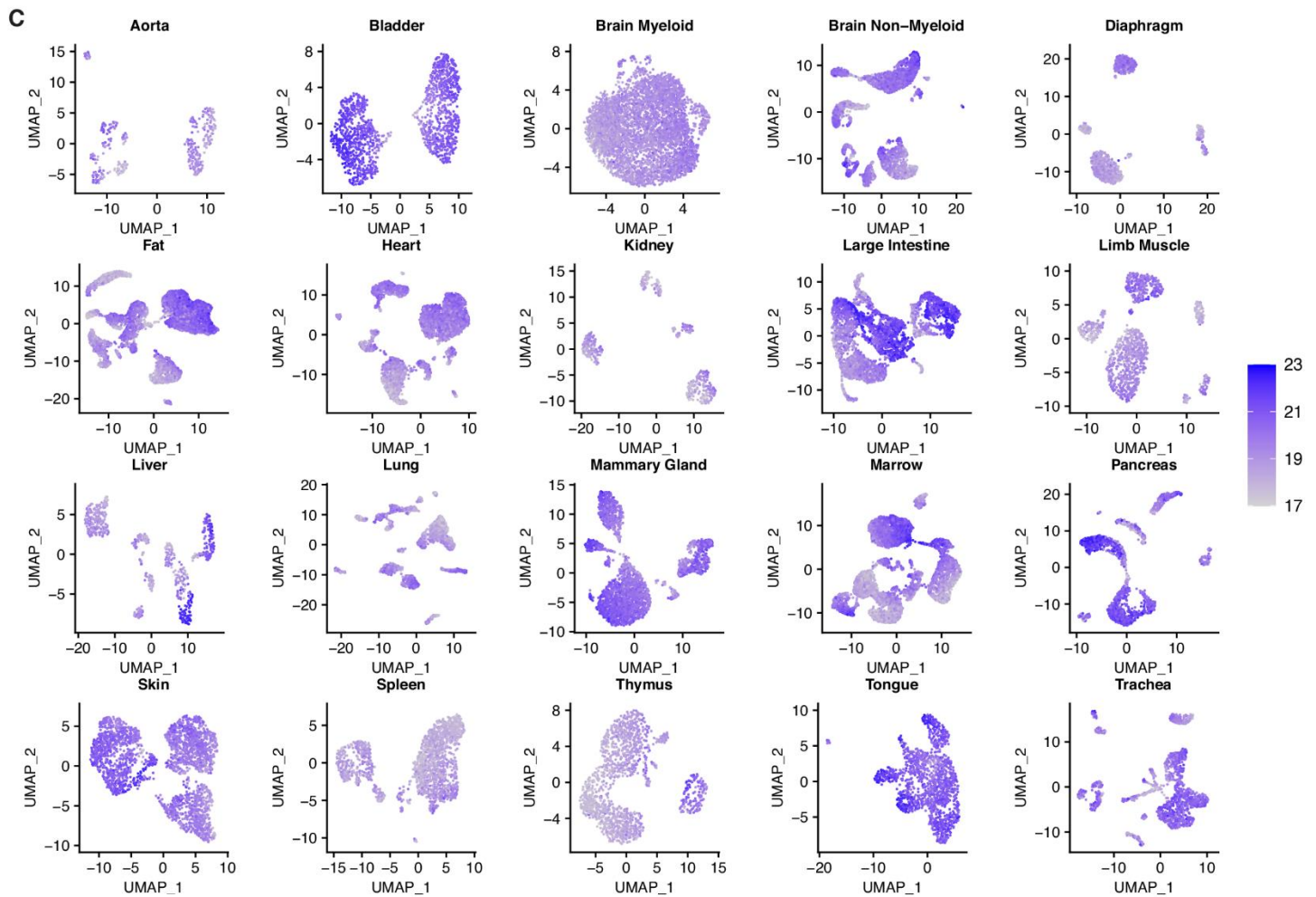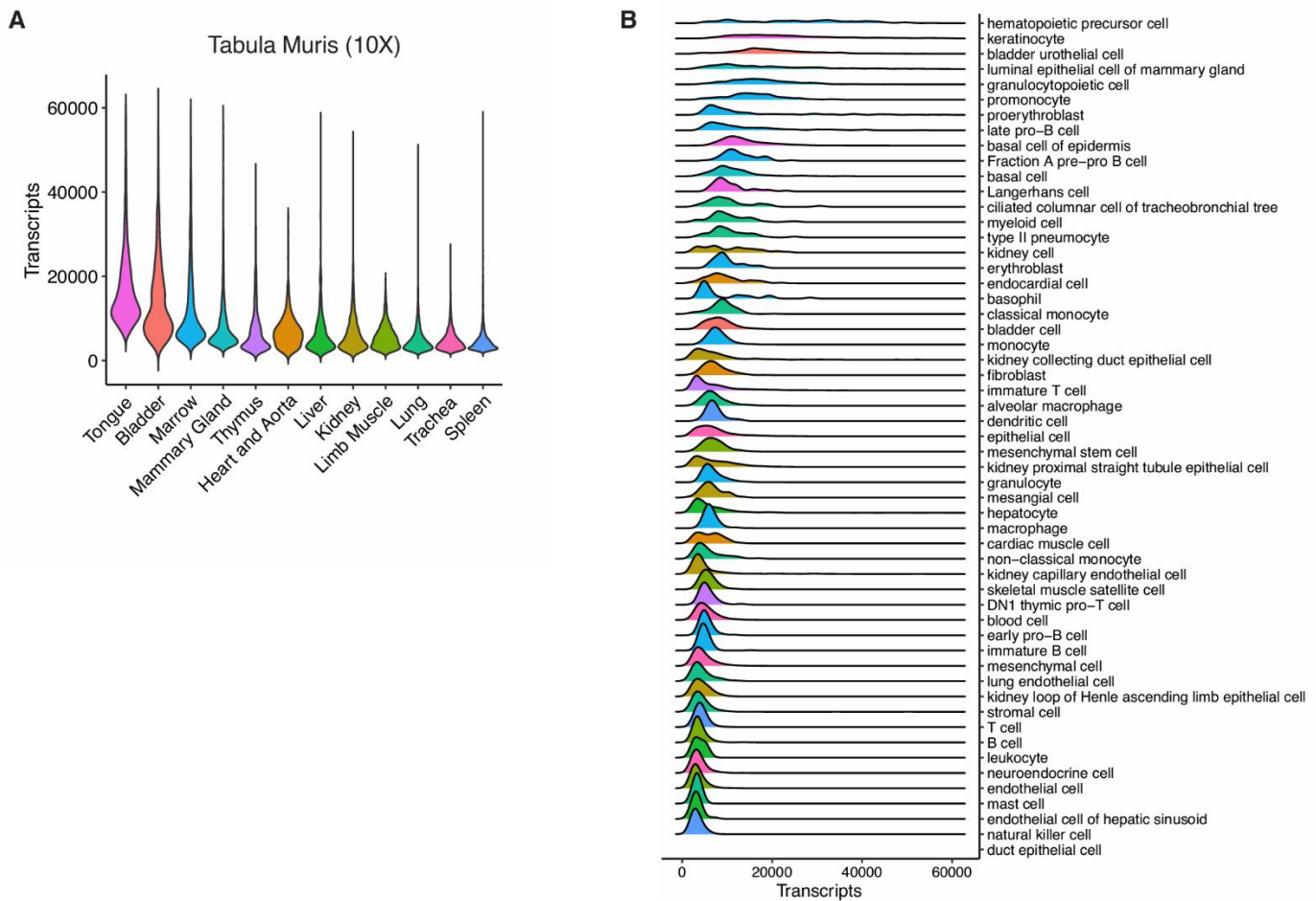
Figure S4

Supplementary Figure 4: Transcript Content Heterogeneity in Tabula Muris Datasets. Supplemental to Figure 3.
(A-B) Distribution of cellular transcript content between Droplet organ datasets under absolute scaling, ranked by median total transcripts. Transcripts represent raw UMIs.
(C) UMAP representations of FACS organ datasets using dimensionality reduction under globally-scaled gene expression. Transcript abundance represents log2 scaled ERCC-normalized reads
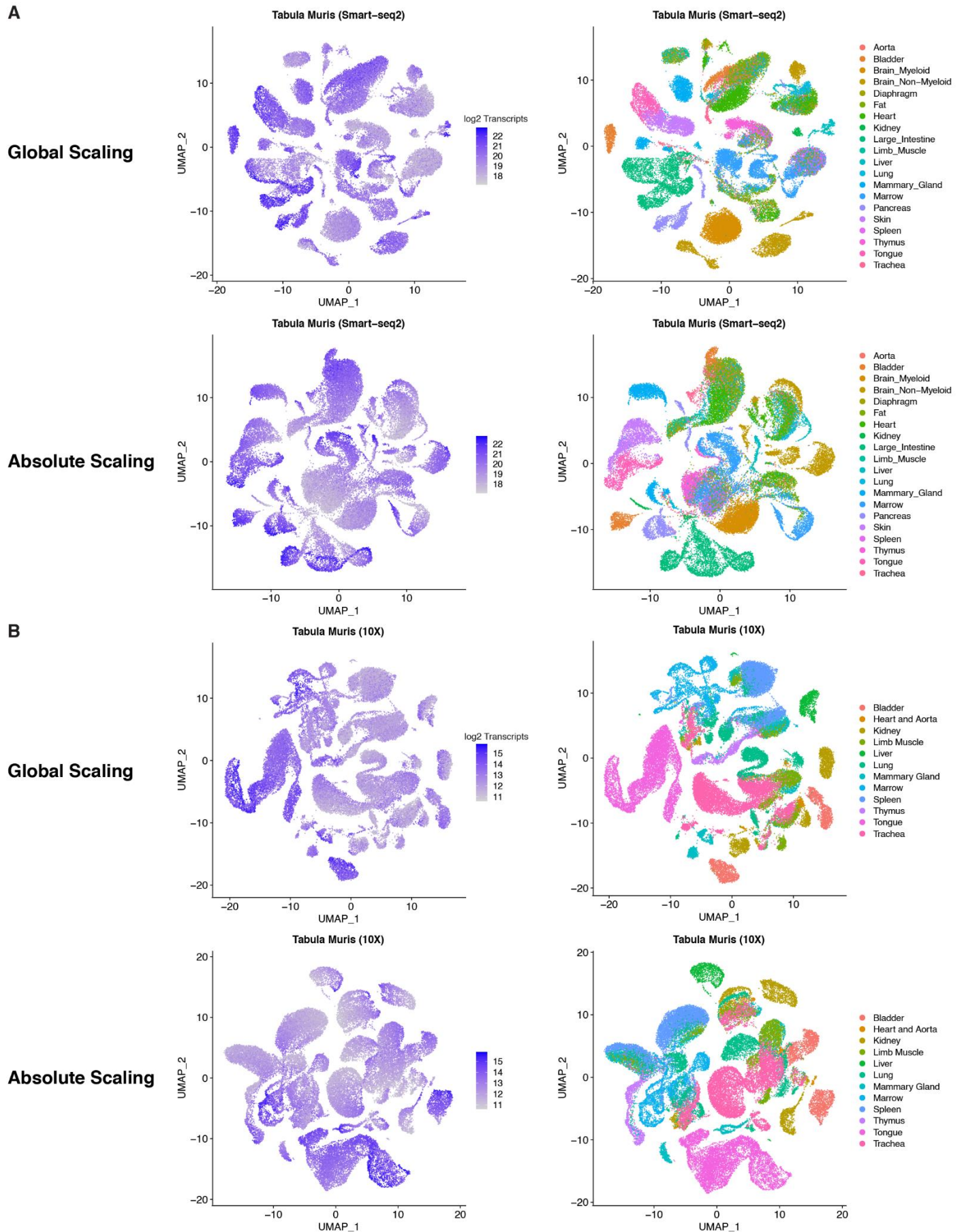
Figure S5

Supplementary Figure 5: Comparison of Structural Changes Between Global and Absolute Scaling. Supplemental to Figure 3.

(A) UMAP representations of FACS Tabula Muris Atlas with dimensionality reduction performed under global or absolute scaled gene expression. Transcript abundance represents log2 scaled ERCC-normalized reads.

(B) UMAP representations of 10X Tabula Muris Atlas with dimensionality reduction performed under global or absolute scaled gene expression. Transcript abundance represents log2 absolute UMIs.
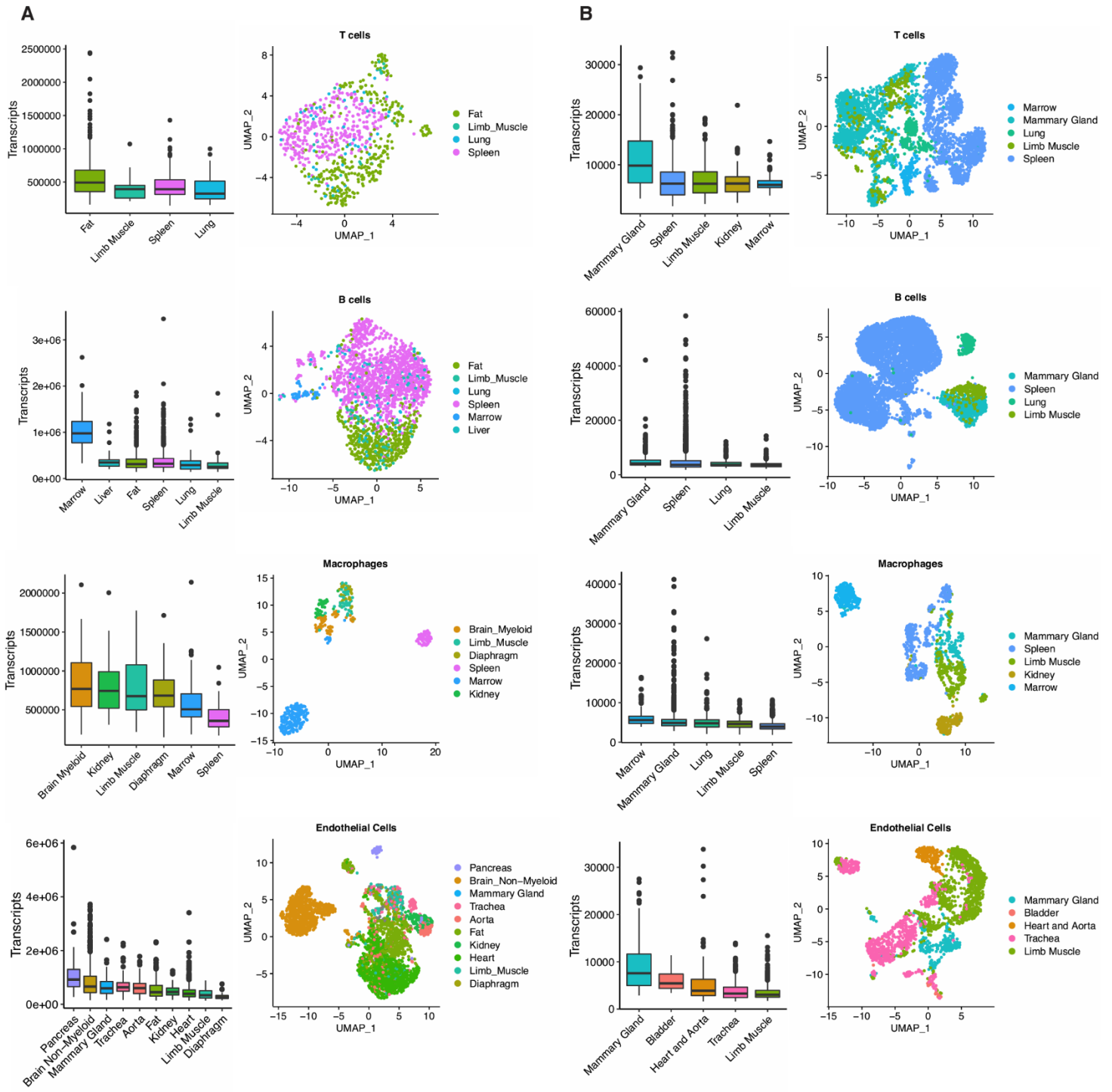
Figure S6

Supplementary Figure 6: Shared Cell Types Between Organ Datasets Display Similar Transcript Abundances. Supplemental to Figure 3.
(A-B) Boxplots of transcript content and UMAP visualizations of shared cell types in FACS **(A)** or Droplet **(B)** datasets. Cell types ranked by median transcript content. Values in **(A)** represent log2 scaled ERCC-normalized reads, while values in **(B)** represent log2 absolute UMIs.
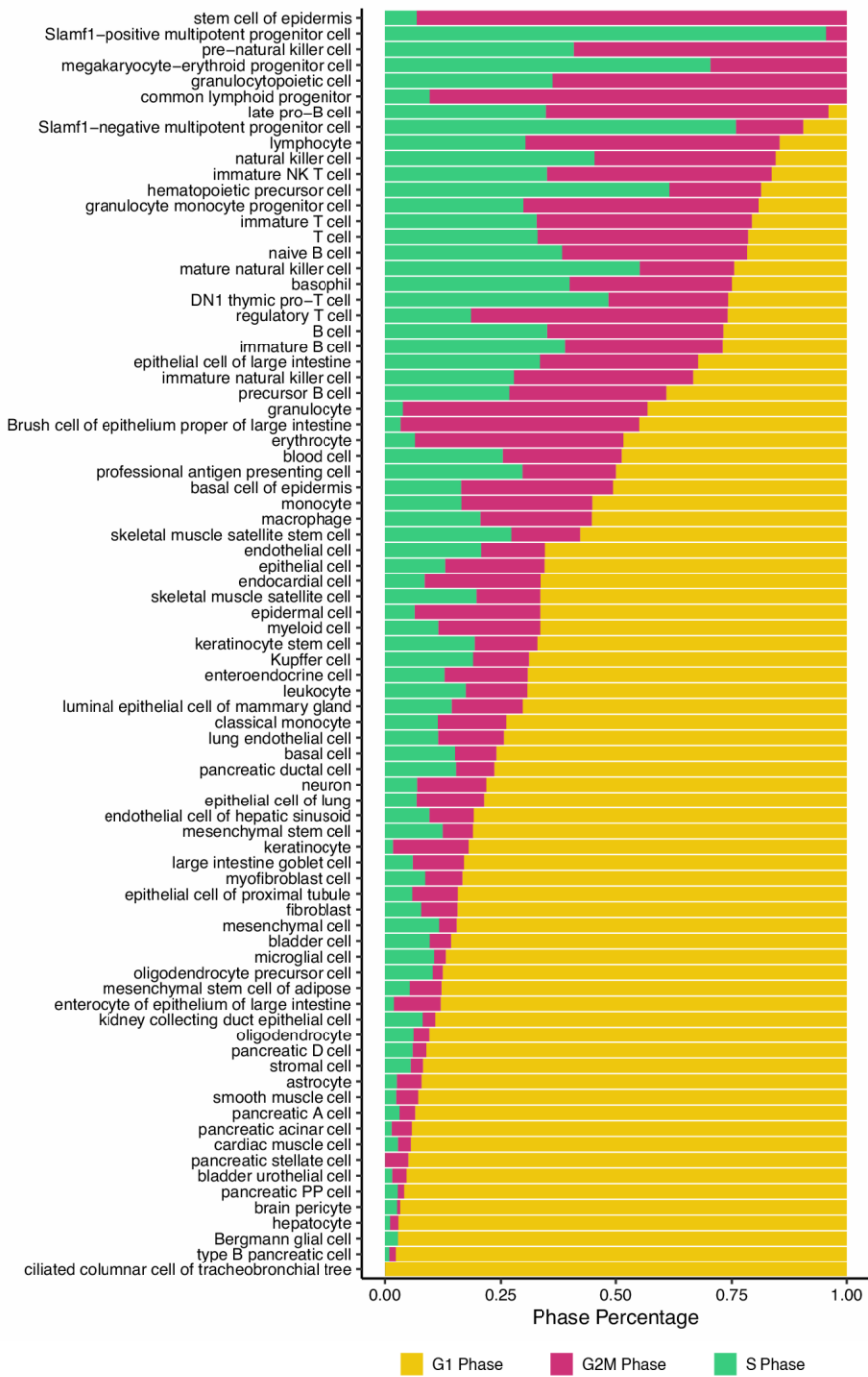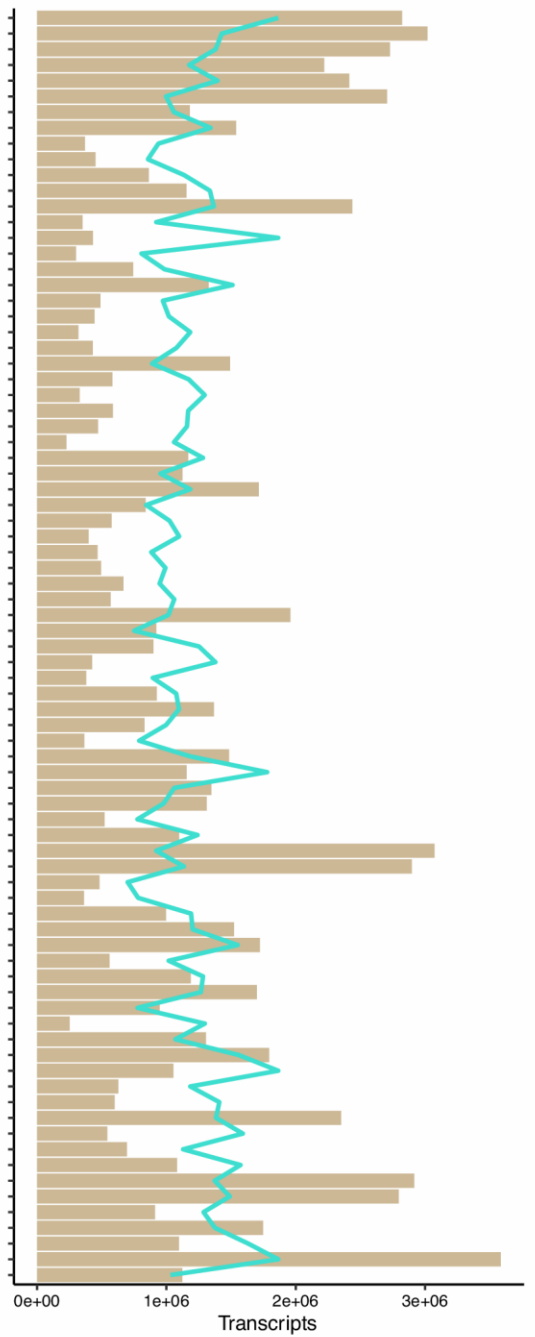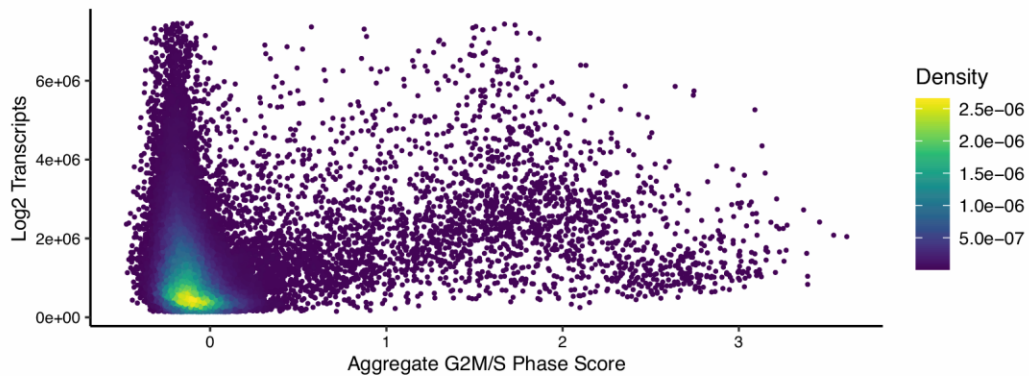
Figure S7

Supplementary Figure 7: Cell Cycle Phase Distribution in Across FACS Atlas. Supplemental to Figure 3.
(A) Cell cycle phase proportions across all cell types, sorted by percentage of G2/M and S cells. Phase scores were calculated using globally-scaled expression data.
(B) Transcript abundance across all cell types, sorted by percentage of G2M and S cells. Transcripts represent log2 scaled ERCC-normalized reads. Line represents moving average across 7 cell types.
(C) Comparison between aggregate G2M/S score and transcript abundance across all cells. Comparable results obtained for cell types within the 10X atlas (data not shown).

A Transcripts (Absolute Scaling)

B Mlxipl (Global Scaling)

C Lifr (Global Scaling)

D Nr1i3 (Global Scaling)

E Transcripts (Absolute Scaling)

F Bladder Epithelium

- Superficial
- Intermediate 1
- Intermediate 2
- Basal

G Transcripts

H Krt20 (Global Scaling)

I Krt20

J Trp63 (Global Scaling)

K Trp63

L Upk2 (Global Scaling)

M Upk2

N Krt14 (Global Scaling)

O Krt14
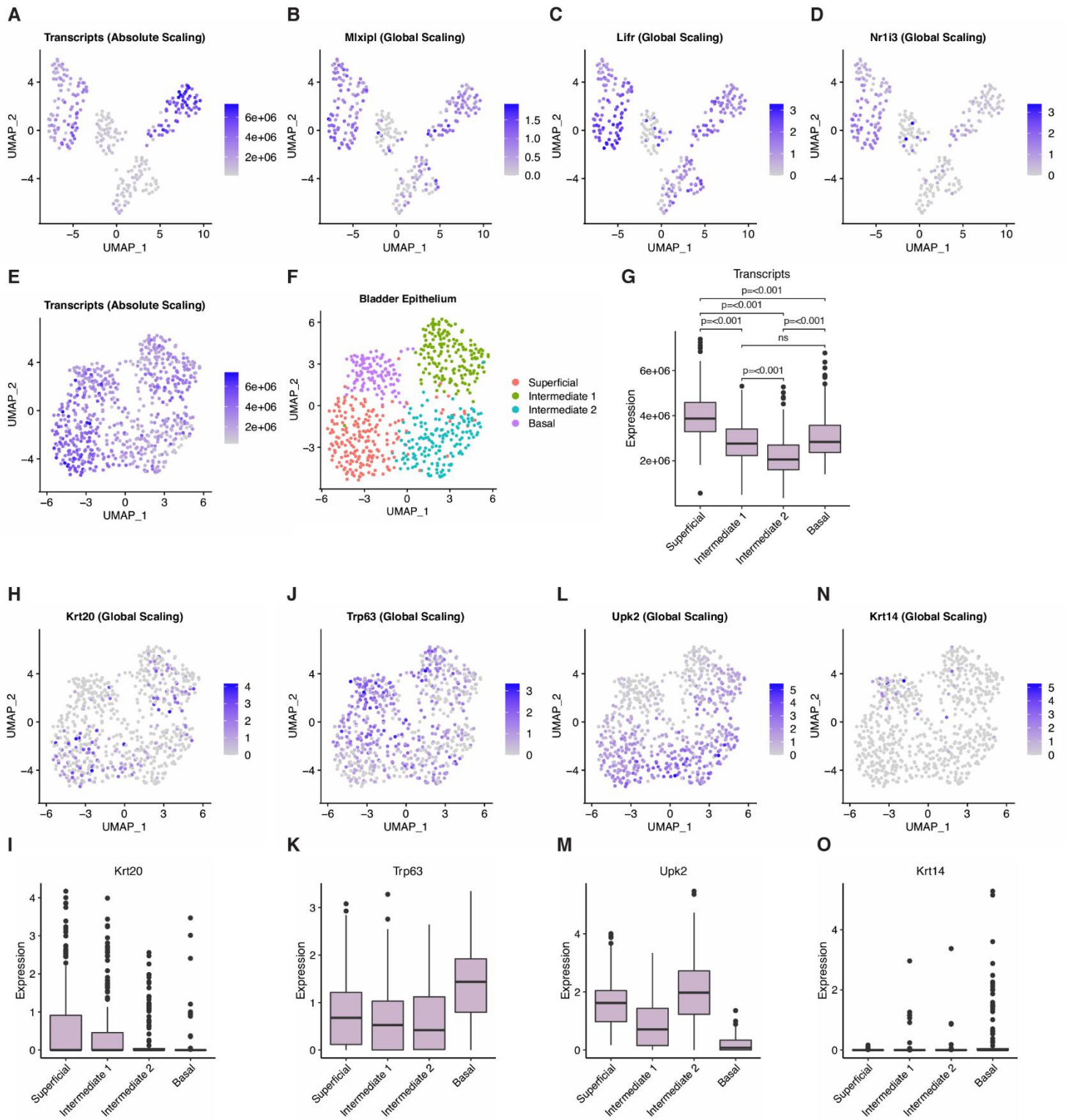
Figure S8

Supplementary Figure 8: High-Content Cells of the Liver and Bladder Display Polyploidy Markers. Supplemental to Figure 3.
(A-D) UMAP representation showing ERCC-normalized reads and log2 expression of liver polyploid cell markers. Plots depict dimensionality reduction performed under absolute scaling.
(E-O) ERCC-normalized transcript abundance and log2 expression of polyploid cell markers in bladder cells. Plots depict dimensionality reduction performed under the indicated scaling type. Statistical test performed is the pairwise Wilcoxon rank sum test with Bonferroni correction.
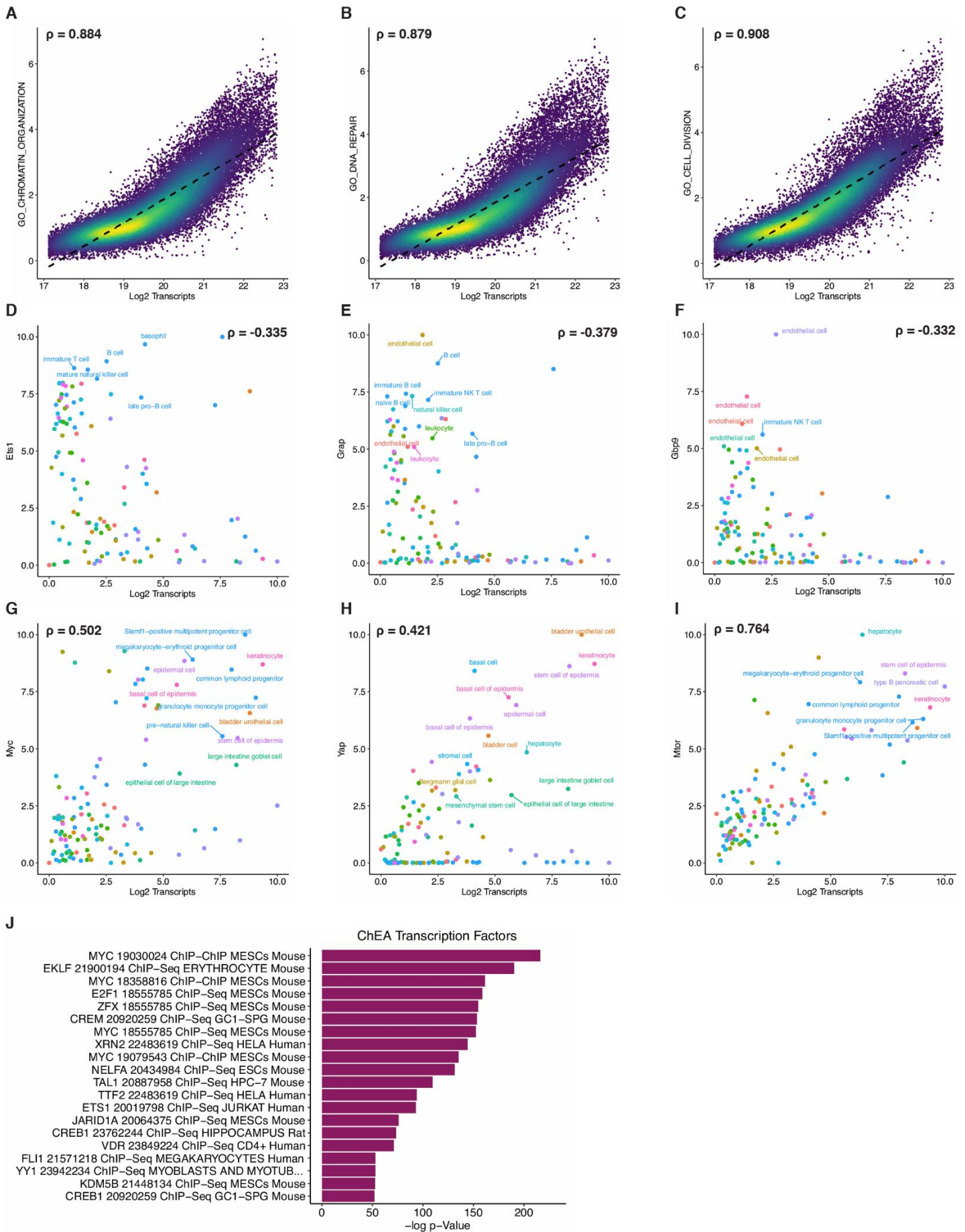
Figure S9

Supplementary Figure 9: Hypertranscription Hallmarks are Correlated with Transcript Abundance. Supplemental to Figure 3.
(A-C) Single-cell correlation of transcription signatures with log2 transcript abundance. Signature scores were determined using VISION with absolute-scaled FACS data. Each point represents a single cell, with color scale depicting plotting density.
(D-I) Positive and negative correlation of select gene expression with log2 transcript abundance. Points are displayed as averages within cell types, with colors matching organs in **(Fig 2)**. Correlation values represent Spearman coefficients.
**(J)** Enrichment analysis for ChEA sets using top 3000 highly correlated genes (to log2 transcript abundance).
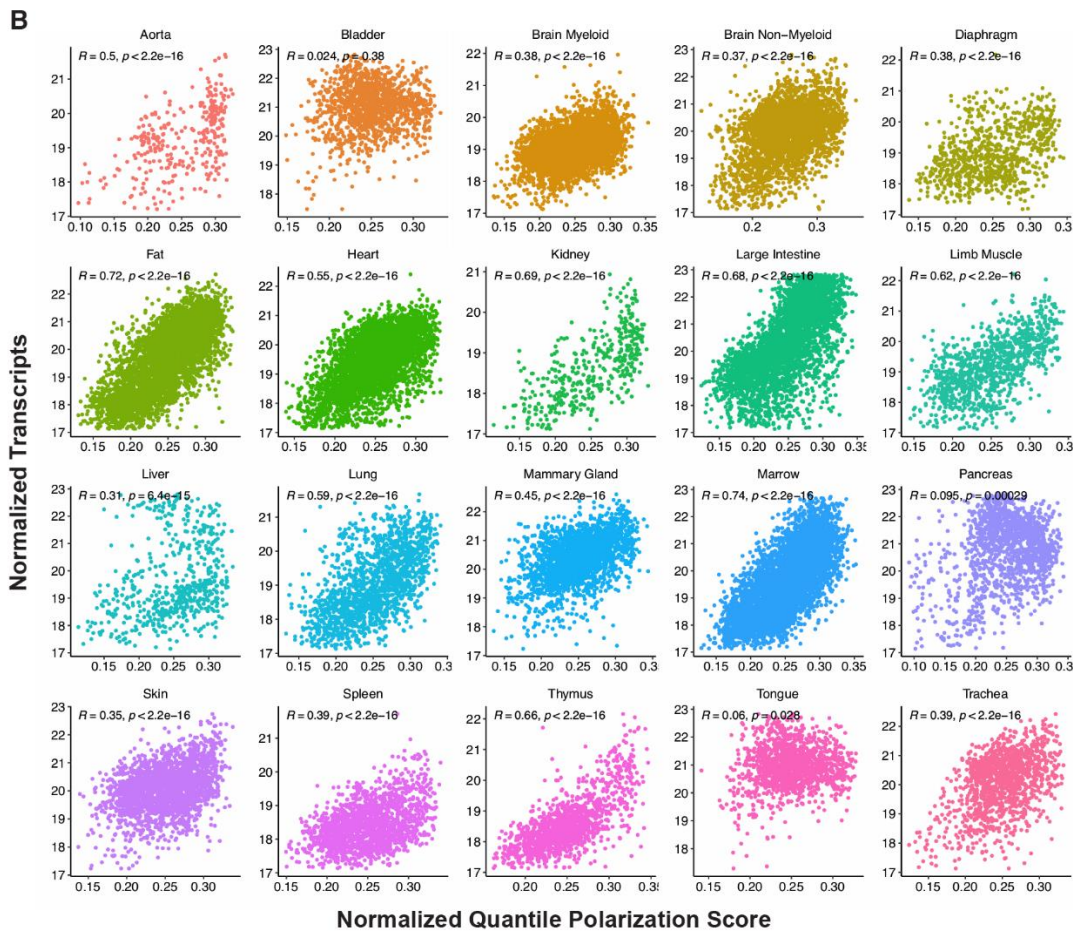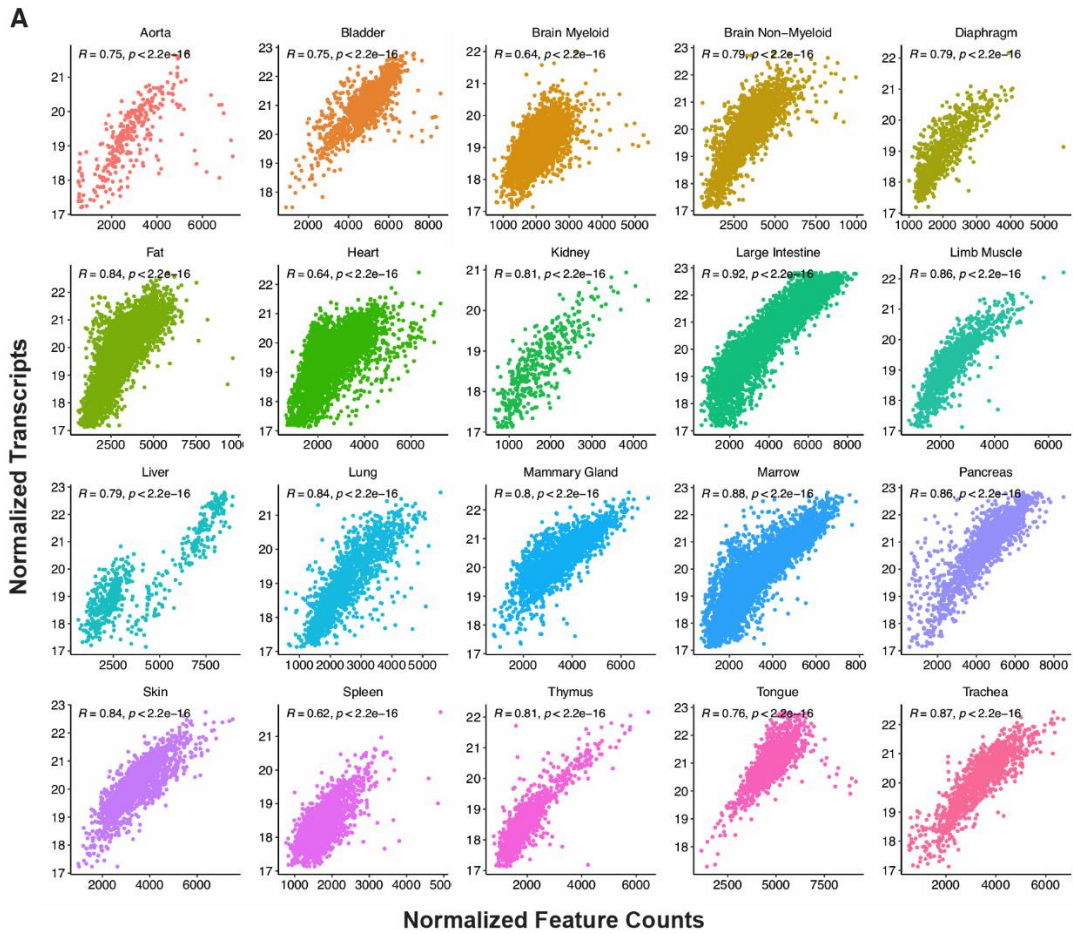
Figure S10

Supplementary Figure 10: Transcript Content is a Correlated to Markers of Developmental Progression. Supplemental to Figure 3.
(A) Comparison of log2 absolute-scaled transcript content with feature counts (number of expressed genes).
(B) Comparison of log2 absolute-scaled transcript content with quantile progression scores, calculated using globally-scaled FACS data. Correlation values in **(A-B)** represent Spearman coefficients.
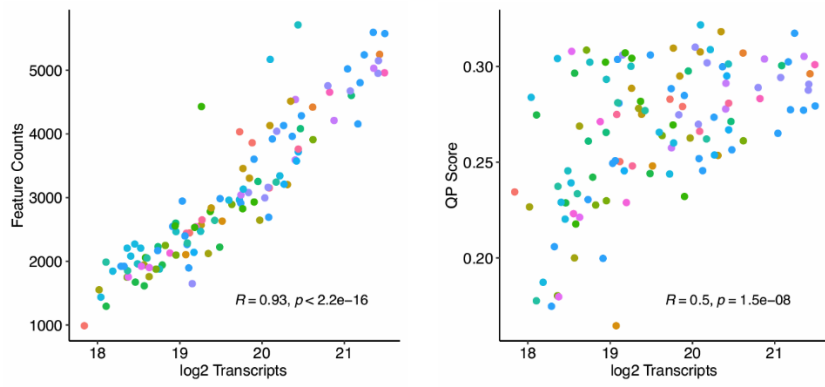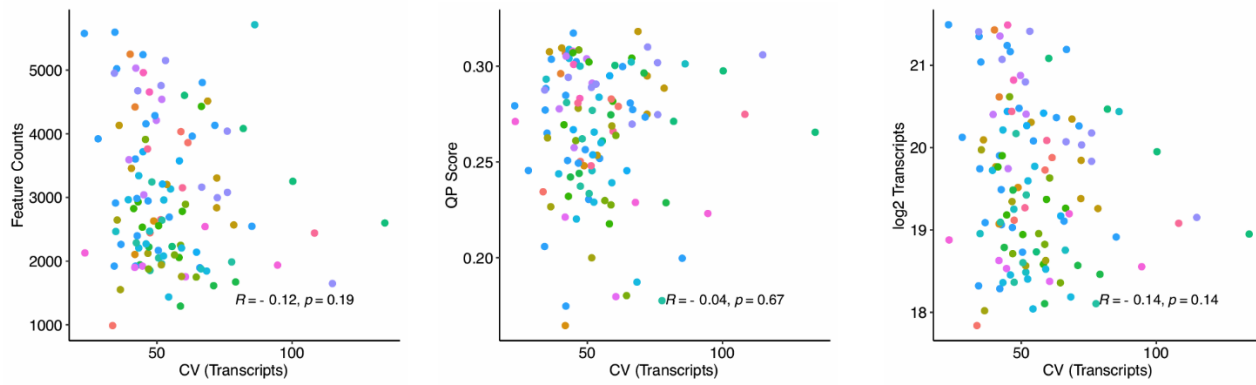
Figure S11

Supplementary Figure 11: The Cell-Type Variance of Transcript Content is Not Correlated to Markers of Developmental Progression. Supplemental to Figure 3.
(A) Single-cell correlation of log2 Transcripts with markers of developmental progression.
(B) Single-cell correlation of the coefficient of variation (CV) of transcripts with markers of developmental progression. Values in **(A-B)** were derived from cell-types within absolute-scaled FACS data. Points are displayed as averages within cell types, with colors matching organs in **(Fig 2)**. Correlation values represent Spearman coefficients.
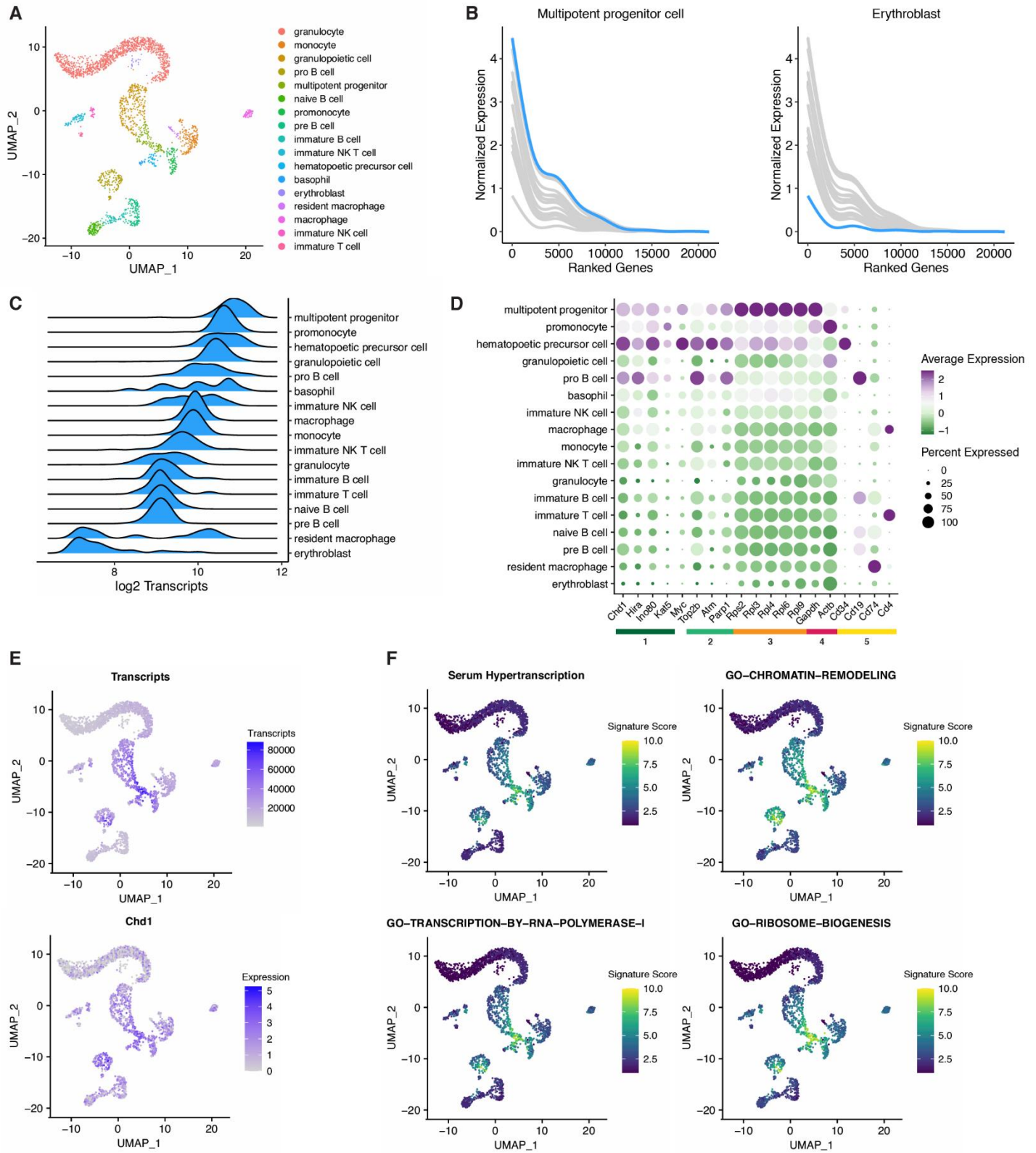
Figure S12

Supplementary Figure 12: Independent Single-Cell Analysis Reveals Hallmarks of Hypertranscription in Hematopoietic Progenitors. Supplemental to Figure 4.
(A) UMAP visualization of cell types within in-house generated 10X adult bone marrow dataset using dimensionality reduction performed under global scaling.
(B) Transcriptome curves depicting gene expression across top 20,000 genes in representative cell types of the bone marrow. Individual genes are ranked using combined log2 expression between all cell types. Indicated cell types correspond to the highlighted curve.
(C) Distribution of cellular transcript content between cell types of the adult bone marrow dataset. Transcript counts represent log2 UMIs.
(D) Expression of select genes relevant to hallmarks of hypertranscription, including (1) chromatin remodelers, (2) DNA repair factors, (3) ribosomal genes, (4) housekeeping genes, and (5) hematopoietic markers.
(E) UMAP visualization of total transcripts and *Chd1* expression under absolute scaling.
(F) UMAP visualization of signature scores generated by VISION using absolute-scaled expression data.
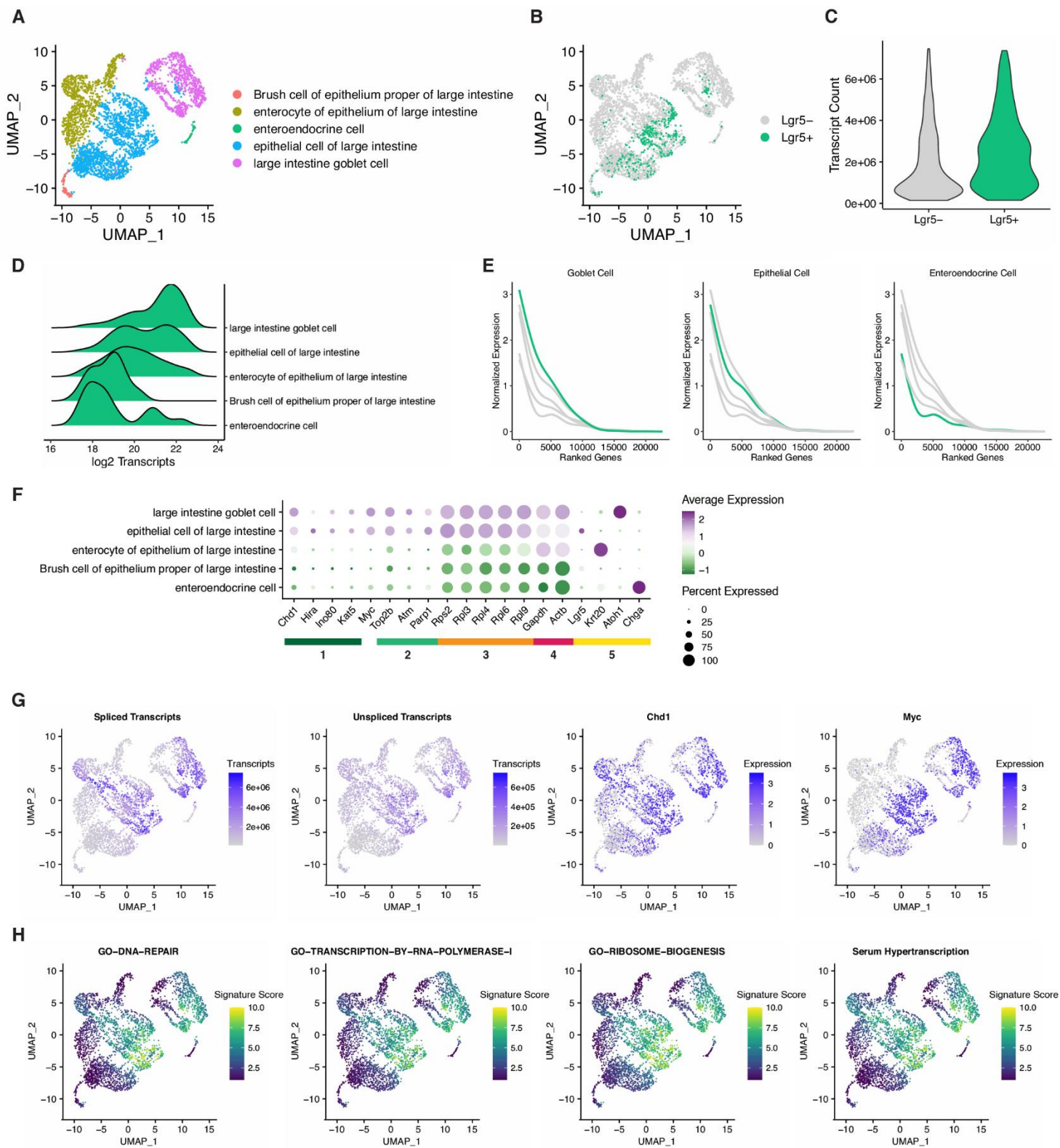
Figure S13

Supplementary Figure 13: Colonic Epithelium Progenitors Display Hallmarks of Hypertranscription. Supplemental to Figure 5.

(A) UMAP visualization of cell types within the large intestine FACS dataset.

(B) Visualization of cycling Lgr5+ intestinal stem cells within the dataset. Plots in **(A-B)** depict dimensionality reduction performed under absolute scaling.

(C) Transcript content between Lgr5+ and Lgr5- cells.

(D) Distribution of cellular transcript content between cell types of the colonic epithelium. Transcript counts represent log2 ERCC-normalized reads.

(E) Transcriptome curves depicting gene expression across top 20,000 genes in representative cell types of the colonic epithelium. Individual genes are ranked using combined log2 expression between all cell types. Indicated cell types correspond to the highlighted curve.

(F) Expression of select genes relevant to hallmarks of hypertranscription, including (1) chromatin remodelers, (2) DNA repair factors, (3) ribosomal genes, (4) housekeeping genes, and (5) crypt cell markers.

(G) UMAP visualization of total spliced/unspliced transcripts, Chd1 and Myc expression under absolute scaling.

(H) UMAP visualization of signature scores generated by VISION using absolute scaling.
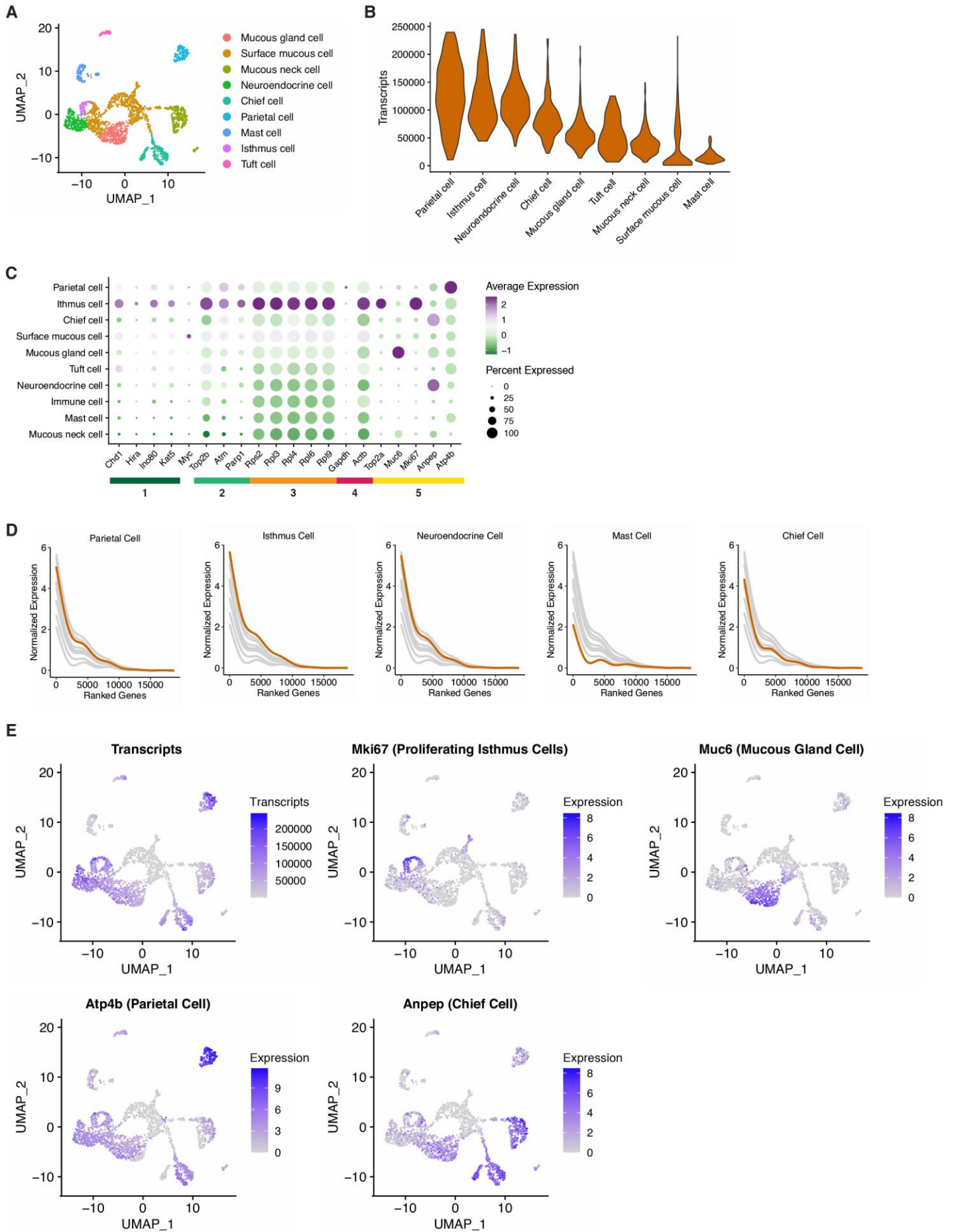
Figure S14

Supplementary Figure 14: Secretory Cells of the Gastric Pit Display Hallmarks of Hypertranscription. Supplemental to Figure 5.

(A) UMAP visualization of cell types within an adult mouse corpus and pylorus atlas. Plots depicts dimensionality reduction performed under absolute scaling.

(B) Distribution of cellular transcript content between cell types of the gastric pit. Transcript counts represent ERCC-normalized reads.

(C) Expression of select genes relevant to hallmarks of hypertranscription, including (1) chromatin remodelers, (2) DNA repair factors, (3) ribosomal genes, (4) housekeeping genes, and (5) gastric pit cell markers.

(D) Transcriptome curves depicting gene expression across top 20,000 genes in representative cell types of the gastric pit. Individual genes are ranked using combined log2 expression between all cell types. Indicated cell types correspond to the highlighted curve.

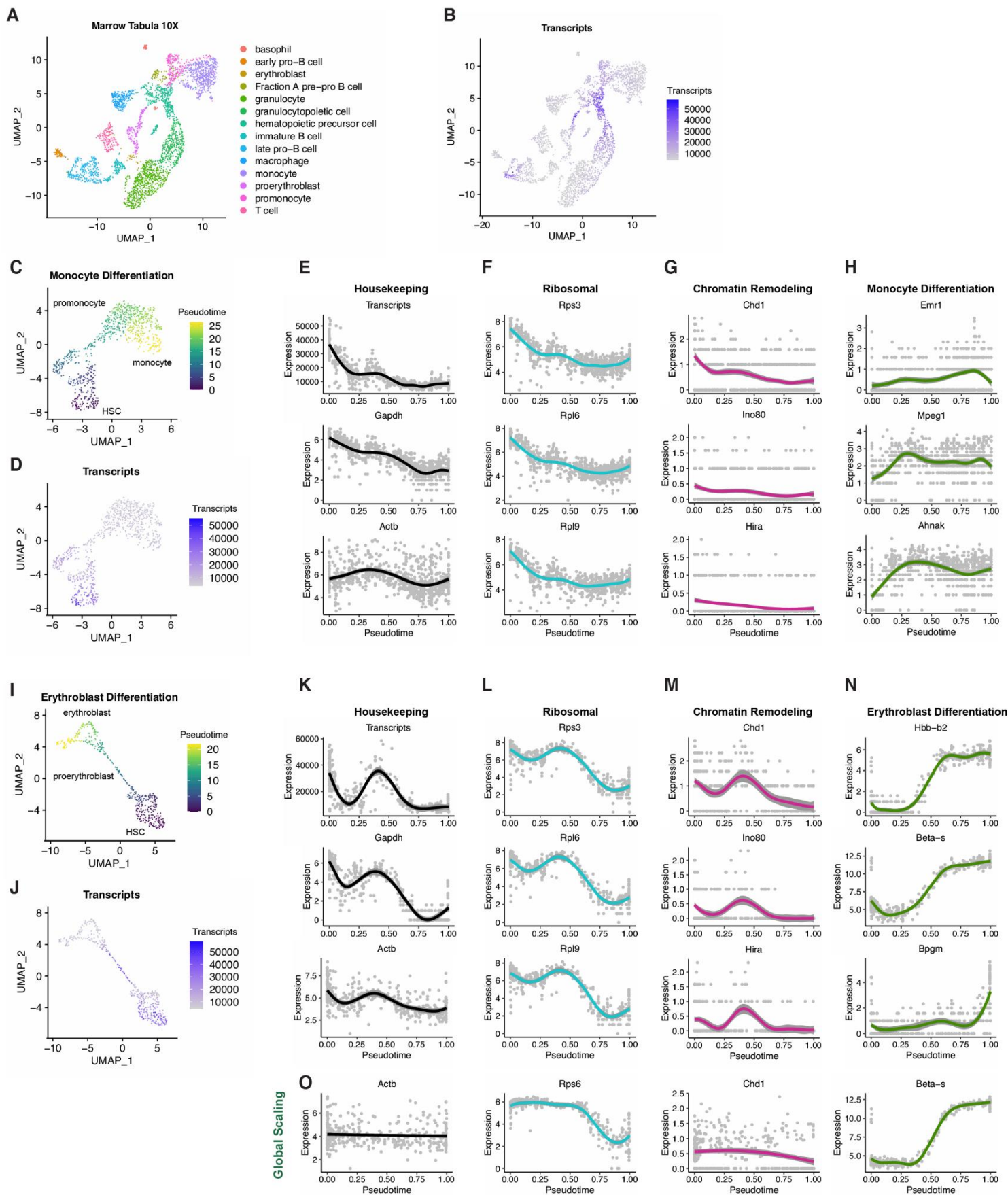**(E)** UMAP visualization of selected markers under absolute scaling.

Figure S15

Supplementary Figure 15: Hypertranscription Marks Developmental Progress Through Erythroblast and Monocyte Differentiation. Supplemental to Figure 6.
(A) UMAP representation of cell types within Tabula Muris 10X bone marrow dataset.
(B) Transcript abundance (raw UMIs) across all cells of the bone marrow.
(C-D, I-J) UMAP visualization of Monocle pseudotime scores and total transcript abundance under absolute scaling. Hematopoietic stem cells were defined as root cells for analysis. Plots in **(A-B), (C-D)**, and **(I-J)** depict dimensionality reduction performed under absolute scaling.
(E-H, K-N) Gene curves representing absolute-scaled expression of indicated genes through progression of pseudotime. Each point represents single cells ordered by pseudotime values.
(O) Gene curves representing globally-scaled expression of indicated genes through progression of pseudotime. Each point represents single cells ordered by pseudotime values. Expression values in **(E, K)** represent raw UMIs, values in **(F-H, L-N)** represent log2 raw UMIs.