Title: A limited set of transcriptional programs define major cell types

Authors: Alessandra Breschi[†1,2,3], Manuel Muñoz-Aguirre[†1,4], Valentin Wucher[†1], Carrie A. Davis[5], Diego Garrido-Martín[1,2], Sarah Djebali[1,2,6], Jesse Gillis[3], Dmitri D. Pervouchine[1,7], Anna Vlasova[8], Alexander Dobin[5], Chris Zaleski[5], Jorg Drenkow[5], Cassidy Danyko[5], Alexandra Scavelli[5], Ferran Reverter[1,2], Michael P. Snyder[3], Thomas R. Gingeras*[5] and Roderic Guigó*[1,2].

Affiliations:

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Catalonia, E-08003.

[2] Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, E-08003.

[3] Stanford University, Department of Genetics, Stanford, 94305, USA.

[4] Universitat Politècnica de Catalunya. Departament d'Estadística i Investigació Operativa. 08034 Barcelona, Catalonia, E-08003.

[5] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742.

[6] IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, Toulouse, France.

[7] Skolkovo Institute for Science and Technology, 3 Nobel st., Moscow, Russia 143025.

[8] Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), 1030 Vienna, Austria.

*Correspondence to: roderic.guigo@crg.cat, gingeras@cshl.edu.

[†]A. Breschi, M. Muñoz-Aguirre and V. Wucher contributed equally to this work.

Abstract: We have produced RNA sequencing data for a number of primary cells from different locations in the human body. The clustering of these primary cells reveals that most cells in the human body share a few broad transcriptional programs, which define five major cell types: epithelial, endothelial, mesenchymal, neural and blood cells. These act as basic components of many tissues and organs. Based on gene expression, these cell types redefine the basic histological types by which tissues have been traditionally classified. We identified genes whose expression is specific to these cell types, and from these genes, we estimated the contribution of the major cell types to the composition of human tissues. We found this cellular composition to be a characteristic

signature of tissues, and to reflect tissue morphological heterogeneity and histology. We identified changes in cellular composition in different tissues associated with age and sex and found that departures from the normal cellular composition correlate with histological phenotypes associated to disease.

One Sentence Summary: A few broad transcriptional programs define the major cell types underlying the histology of human tissues and organs.

Main Text:

Transcriptional profiles reflect cell type, condition and function. In tissues and organs, they are monitored in RNA extracted from millions to billions of cells (11^6-10^9) *(1)* likely including multiple cell types. As a consequence, the transcriptional profiles obtained from tissue samples represent the average expression of genes across heterogeneous cellular collections, and gene expression differences measured in bulk tissue transcriptomes may thus reflect changes in cellular composition rather than changes in the expression of genes in individual cells. Single-cell RNA sequencing (scRNA-seq) has indeed revealed large cellular heterogeneity in many tissues and organs *(2)*, and the Human Cell Atlas (HCA) project *(3)* has been recently initiated with the aim of defining all human cell types and to infer the cellular taxonomy of the human body. As a step in that direction and to bridge the transcriptomes of tissues with the transcriptomes of the constituent primary cells, and to understand how these impact tissue phenotypes, we have generated bulk expression profiles of 53 primary cell lines isolated from ten different anatomical sites in the human body. These profiles include long and short strand-specific RNA-seq, and RAMPAGE data (Fig. 1a, Table S1-4).

Major cell types in the human body

Clustering of the primary cells based on gene expression profiles revealed a number of well-defined clusters (Fig. 1b-c, Fig. S1, S2a-b, Supplementary Information). One cluster was composed of endothelial cells, a second large cluster included a mixture of cell types: fibroblasts, stem cells and muscle cells, among others, which we collectively termed as mesenchymal, two smaller clusters, which clustered together, were composed of epithelial cells, and finally, the melanocytes clustered separately. Almost all of the individual primary cells are assigned to the proper major cell type. The exceptions are renal mesangial cells, which have contractile properties, but are classified as epithelial, and lung epithelial cells, that are classified as mesenchymal. These two cell types, however, are of embryonic origin — in contrast to the vast majority of primary cells in our study, which are adult (Table S1) — and their transcriptomes may not reflect the transcriptomes of fully differentiated cells.

The clustering of primary cells does not appear to be dominated by body location, or embryological origin. Body location actually contributes very little to the expression profile of primary cells, explaining only about 4% of the variance in gene expression (Fig. S2c). Variation of gene expression among organs is similar for the different clusters (Fig. S2d). Remarkably, the transcriptional diversity among cells within a given organ can be as high as that across the entire human body (Fig. S2e). A similar clustering is obtained using FANTOM CAGE-based transcriptomic data on 105 primary cells *(4)* (Fig. 1d, Fig. S3a-b, Table S5), which reveals, in addition, two clusters corresponding to blood and neural cells, which were not represented in our set of primary cells. The analysis of a different set of primary cells from the ENCODE encyclopedia Candidate Regulatory Elements (cREs *(5)*, Table S6), based on DNAse Hypersensitive Sites (DHSs), also recapitulates the clustering (Fig. 1e, Fig. S3c). The clustering remains in the set of 146 non-redundant primary cells, that results from merging the RNA-Seq, the CAGE and the DHS data. The clustering is thus conserved despite the heterogeneity of the underlying assays and experimental protocols used to generate these different data sets (Fig. S4). In the clustering, neural cells (mostly astrocytes from different brain regions and neurons) cluster together with a few neuroepithelial primary cells (we labelled them epithelial, but they are mostly ciliate cells from different sites in the eye). While the neural cells profiled by CAGE seem to have a distinct transcriptional signature (Fig. S3a), neural cells profiled by DNAse-seq exhibit a gene expression pattern similar to mesenchymal cells (Fig. S3c). However, the neural cells profiled by DNAse-seq are, in contrast to most primary cells investigated here, of embryonic origin, and thus

they are not likely to express the transcriptional program characteristic of adult neural cells. The analysis of publicly available transcriptomics data from nervous tissues including single-cell and bulk RNA-seq strongly support that the neural cell type is a proper major type clearly differentiated from the other types (Supplementary Information, Fig. S5-S7).

Comparable multi-tissue RNASeq data has become recently available at the single cell level for twenty mouse organs and tissues, through the "Tabula Muris" project *(6)*. Principal Component Analysis (PCA) of the individual cells and hierarchical clustering of the primary cell types show that most individual cells, and most cell types, clustered into the five major cell types above, irrespective of the organ of origin (Fig. S8, S9). As in the case of melanocytes above, we also found a few specialized cell types which do not properly belong to these types. Hepatocytes are a notable example (Fig. S8a, S9a). While closer to the epithelial cells than to cells of other types, they seem to have a quite specialized transcriptional program.

These results, all together, suggest the existence of a limited number of core transcriptional programs encoded in the human genome, and likely in mammalian genomes, in general. These programs underlie the morphology and function common to a few major cellular types, which are at the root of the hierarchy of the many cell types that exist in the human body (Table 1). They all show similar transcriptional heterogeneity, with blood, and epithelial within the solid tissues, being the most transcriptionally diverse (Fig. S10). These transcriptionally defined major cell types correspond broadly, but not exactly, the basic histological types in which tissues are usually classified (see for example *(7-9)*): epithelial, of which endothelial is often considered a subtype, muscular, connective, which includes blood, and neural. However, from the transcriptional standpoint, endothelial constitutes a separate type, closer, if any, to the mesenchymal than to the epithelial type. Blood is also a separate major cell type, while the connective (but not blood) and the muscular histological types cluster together into a single mesenchymal transcriptional type (Fig. 1f).

Within each of the major types, further hierarchical organization of cell types may exist. While we have not profiled enough diversity of primary cells to resolve the taxonomic substructure within each major cell type, hints of this substructure can be clearly seen in the epithelial type. Within the epithelial cluster, two well defined subclusters can be identified (Fig. 1b-e; see also Fig. S2a). One of the clusters is made mostly by renal cells, indicating that body location may actually play a role in subtype specialization. Remarkably, the epithelial cluster includes primary cells of all embryonic origins (ectoderm, endoderm and mesoderm), suggesting that the transcriptional

programs of cells may not be fully inherited through development, but partially adopted through function. The more heterogeneous composition of the epithelial type is also apparent in the mouse scRNASeq (Fig. S8, S9).

Our results also suggest that, while many cells are likely to adhere to these basic transcriptional programs, many other primary cells are likely highly specialized and very tissue specific. As with melanocytes and hepatocytes in our analyses, these specialized cells are likely to have their unique transcriptional program.

Table 1

Cell Type:	sets of cells with similar phenotype (morphology and functions). The similarity threshold induces a taxonomic hierarchy of cell types, by means of which similar cell types are recursively aggregated into higher order types.
Primary Cell Type:	cell types at the bottom of the taxonomic hierarchy. They denote specialized cells phenotypically identical (to some resolution); they cannot further be segregated into biologically meaningful subtypes; for example, pancreatic beta cells. In our work, we do not include here, **cell lines**, which are primary cells that have been transformed to proliferate indefinitely.
Major Cell Type:	cell types at the root of the taxonomic hierarchy. They cannot be further aggregated in biologically meaningful higher order types; for example, epithelial cells.
Tissue-Specific Cell Type:	cell type topologically restricted to a specific anatomical region (tissue, organ, body location); for instance, hepatocytes.
Transcriptional Program:	The pattern of gene expression characteristic of a given cell type.

Table 1: Cell types in the human body.

Figure 1

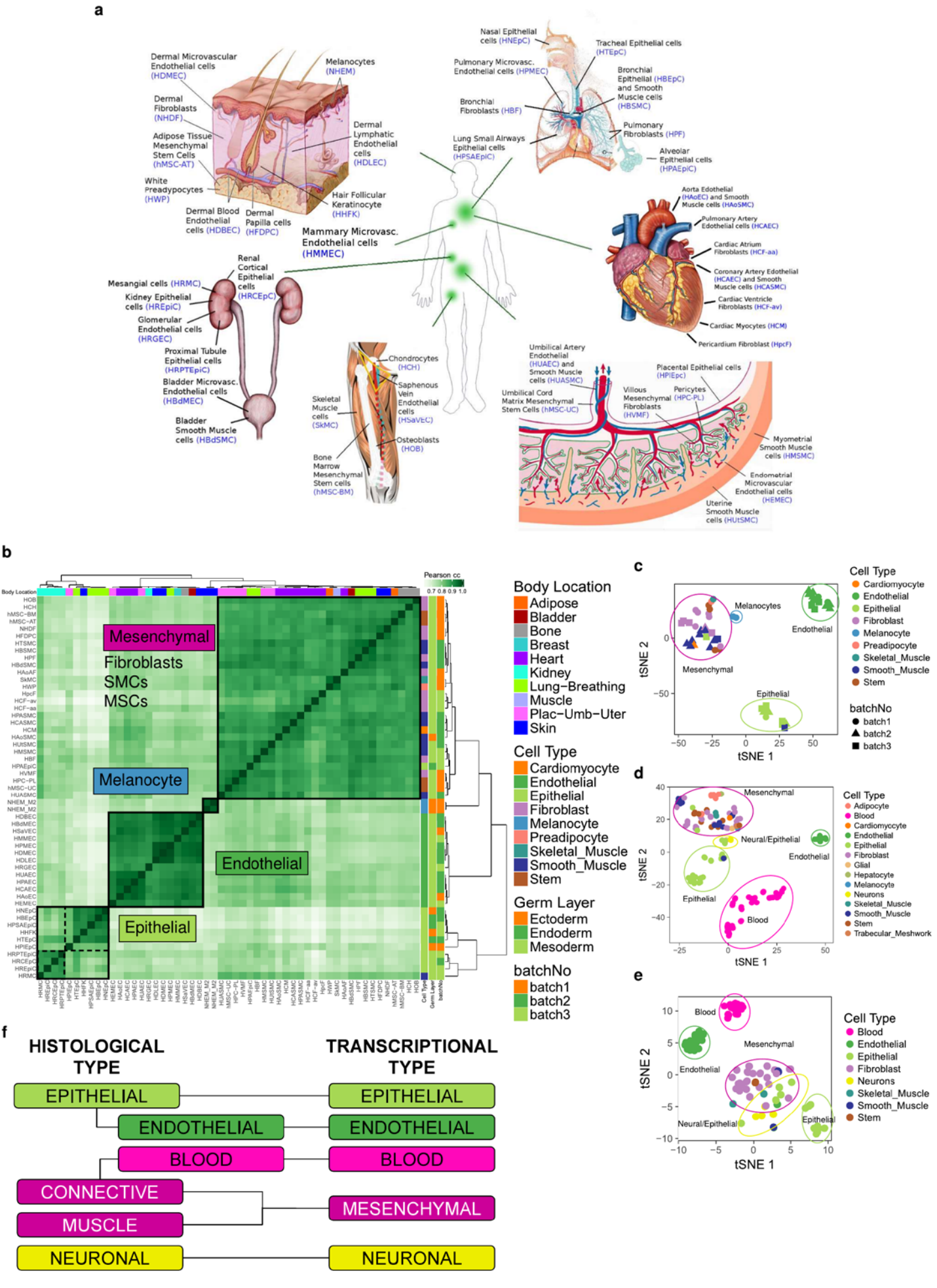

Fig. 1. Basic transcriptional programs of human primary cells.

(a) Overview of primary cells analyzed in this study and the body location they are extracted from. **(b)** Hierarchical clustering of human primary cells based on the correlation of gene expression. The clustering in four major clusters is supported by the silhouette analysis and the elbow method (Fig. S2a-b). tSNE of human primary cells based on gene expression measured here **(c)**, on gene expression measured by CAGE by the FANTOM consortium **(d)** and on Candidate Regulatory Elements (cREs) by the ENCODE encyclopedia scored DNAse hypersensitivity signal **(e)**. **(f)** Correspondence between transcriptionally derived major cell types and classical histological types.

Cell type specific genes

We identified a total of 2,871 genes (including 2,463 protein coding genes, 283 long non-coding RNAs and 125 pseudogenes), the expression of which is specific to epithelial, endothelial, mesenchymal or melanocyte cell types (Fig. 2a, Fig. S11, Table S7). These cell type specific genes include nearly all genes that we identified as the major drivers of the clustering (Supplementary Information, Fig. S12). Examples of these genes include collagen (*COL1/3/6*), expressed in mesenchymal cells, epithelial transcription factors genes *OVOL1/2*, *VWF* gene encoding for the endothelial marker von Willebrand Factor, and *TYR* gene encoding for the melanocyte-specific enzyme tyrosinase (see Table S8 for a list of manually curated driver genes). Figure 2b shows the expression pattern of *RP11-536O18.2*, an endothelial specific long non-coding RNA (lncRNA) of unknown function. The gene is expressed in nearly all endothelial cells analyzed here, but not in cells from other types, and its expression is correlated to protein coding genes with endothelial-related functions (Fig. S13a). The gene, however, is expressed in multiple tissues, and, therefore, it is not tissue specific.

The functions of annotated tissue-specific genes closely match the expected biology of the primary cells in each type (Fig. S13b). Cell type specific genes show consistent restricted expression in the FANTOM CAGE data (Fig. S14), and they are enriched for encyclopedia cREs *(10)* specifically in the primary cells of that type (Fig. S15). Using ChIP-seq histone modification data obtained in a number of primary cells *(11)* (Supplementary Information, Table S9), we found the promoters of genes specific to a given type to be enriched for activating chromatin marks in primary cells of that type compared with primary cells of different type (Fig. S16a). However, overall, except for H3K4me1, we found low levels of most activating marks in the promoters of cell type specific genes compared with all genes, even after controlling for differences in gene expression. In contrast, the promoters of cell type specific genes exhibit similar or higher levels of repressive

histone modifications compared to all genes (Fig. S16b). This is consistent with previous reports showing that genes under tighter regulation show lower levels of activating histone modifications than broadly expressed genes (see for example *(12-13)*).

Among cell type specific genes, we identified 167 Transcription Factors (TFs) from a total of 1,544 TFs annotated in the human genome *(14)*. We focused on 56 that showed the strongest co-expression patterns (Pearson's correlation coefficient ≥ 0.85, Fig. 2c, Fig. S17). They include previously annotated cell type-specific transcriptional regulators, such as ERG, which has been shown to regulate endothelial cell differentiation *(15)*, and TP63, which is an established regulator of epithelial cell fate and is often altered in tumor cells *(16)*. Consistent with the hypothesis that the cell type specific TFs might regulate cell type specificity, we found that genes specific to a given type are enriched for binding motifs for TFs specific to that type in most cell lines (Fig. 2d). The enrichment arises specifically when the motifs occur in open chromatin domains in primary cells of that type (e.g. in epithelial primary cells, epithelial specific genes are enriched, compared to genes specific to other types, in epithelial specific TF motifs occurring in open chromatin domains, Fig. 2d, Fig. S18).

Figure 2

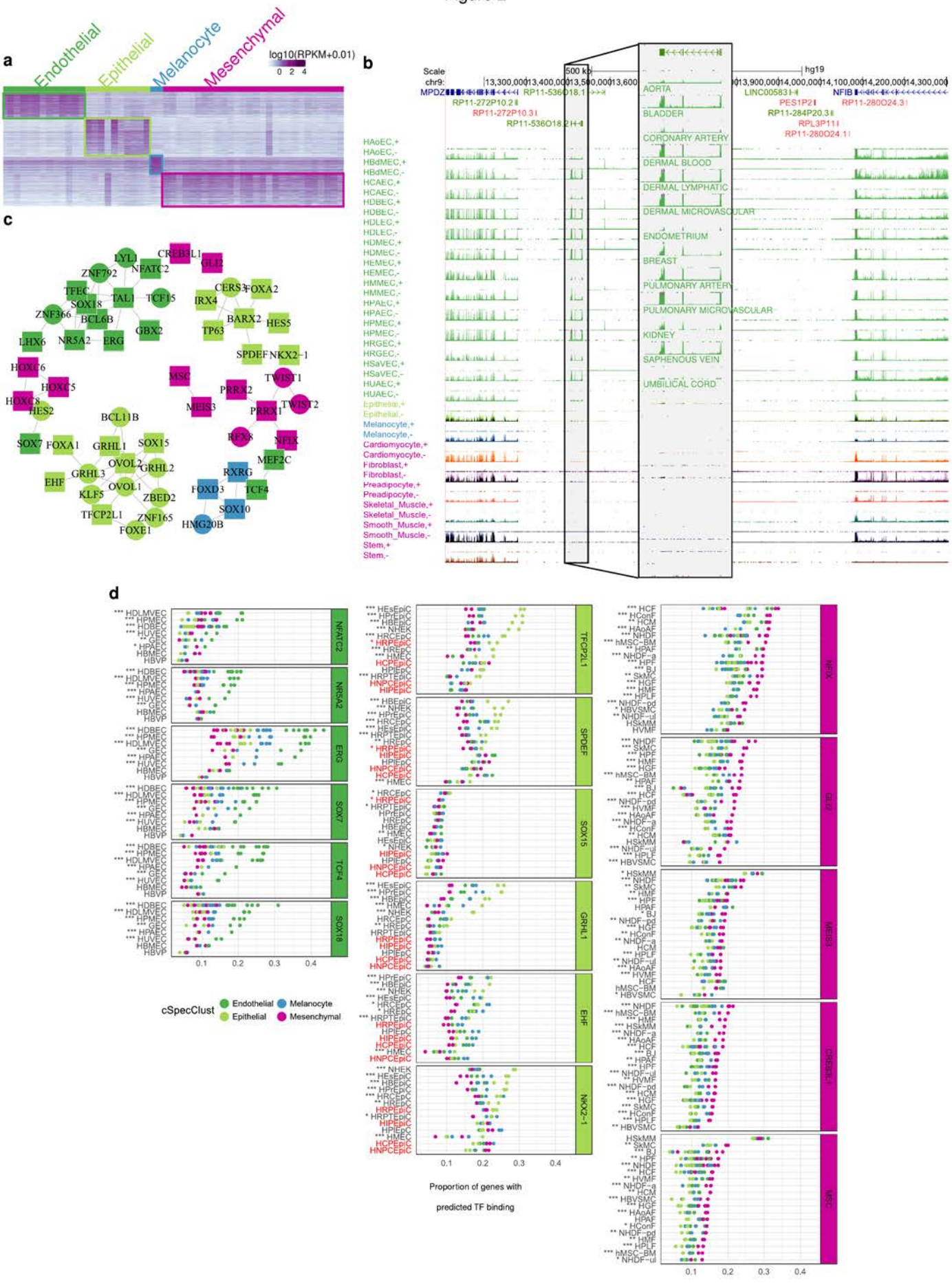

Fig. 2. Cell-cluster-specific genes.

(a) Expression of 2,871 genes specific to major cell types. **(b)** Expression of the endothelial-specific lncRNA RP11-536O18.1. Separate strand-specific signal tracks are shown for endothelial cells, while the other tracks contain overlaid signal for each cell type. The lncRNA has highly correlated (correlation coefficient > 0.9) expression with 72 protein coding genes across our set of primary cells. Nearly all these genes are endothelial specific, and they are functionally enriched for vessel development and angiogenesis (Fig. S13A). The gene appears to be under relatively strong regulation, since it has almost 1,500 eQTLs across multiple tissues in GTEx (v7) well above the average eQTLs for lncRNAs (about 450). **(c)** Network of the most strongly co-expressed (Pearson's correlation coefficient > 0.85) cell type specific transcription factors (TFs). Nodes are colored according to the cell-type-specificity of the TF, and shaped based on the availability of sequence motif (square: available, circle: not available). **(d)** Proportion of cell type specific genes with predicted TF binding over cell type specific genes that harbor a DHS around their TSS (-10kb/+5kb), individually for each cell type specific TF (with binding motif available) and cell line for which DNAse-seq data was available. In general, we found that genes specific to a given type are enriched for binding motifs for TFs specific to that type. For instance, the proportion of endothelial specific genes with DHS sites that harbor motifs for the endothelial specific TF ERG in dermal blood endothelial cells (HDBEC) is larger than the proportion of genes with DHS sites specific of other major cell types. Primary cells highlighted in red, although included within the epithelial major cell type, they have been labelled as neural/epithelial in Fig. 1d, and they are therefore not proper epithelial; consistently, they do not show the enrichment in binding motifs for epithelial specific transcription factors. Refer to Table S6 for a complete description of the acronyms. Enrichment adjusted p-values: "*" < 0.05, "**" < 0.01, "***" < 0.001.

We found that transcriptional regulation appears to play the major role compared to post-transcriptional (splicing) regulation, both in defining the major cell types as well the individual primary cells within the types. We estimated the fraction of the variation in isoform abundance explained by variation in gene expression *(17)* to be on average 67% across transcriptional types and 55% across primary cells (Fig. 3a). The lower proportion of variance explained across primary cells suggests that splicing plays comparatively a more important role in defining the transcriptomes of primary cells within a given type, than in setting the transcriptional programs of the major cell types. In additional support of this conclusion, we have found that while the number of differentially expressed genes in pairwise comparisons of primary cells is much larger between than within cell types, the number of differentially spliced genes is similar (Fig. 3b, Fig. S19, Supplementary Information).

While bulk gene expression is the main contributor to define cell type specificity, other transcriptional events are also cell type specific. First, using the RNA-seq data, we identified cell-

type specific splicing events, independent of the tissue of origin (Table S10, Fig. S20, Supplementary Information). Second, using the RAMPAGE data, we identified cell-type specific TSSs (Fig. S21, Table S11, Supplementary Information).

The basic human transcriptional programs seem to have been established early in vertebrate evolution: genes orthologous of cell type specific genes are underrepresented compared to orthologues of all genes in invertebrate genomes (Fig. 3c, Fig. S22a), but they are overrepresented in vertebrates, as early as in tetrapoda. One exception are epithelial genes, which are overrepresented only in mammals (Fig. 3d, Fig. S22b). Within the set of orthologous genes across tetrapoda *(18)*, the expression of cell type specific genes is less conserved than that of protein coding genes overall, especially at larger evolutionary distances (Fig. 3e, Fig. S22c, Fig. S23). This suggests an important role for the evolution of gene expression regulation in shaping the basic transcriptional programs in the human genome. Epithelial specific genes also show the lowest conservation of expression levels. The transcriptional program characteristic of the epithelium appears to be therefore the most dynamic evolutionarily — possibly reflecting a greater need for adaptation of the epithelial layer in constant interaction with the environment, and it is also consistent with the greater transcriptional heterogeneity of this major cell type.

Figure 3

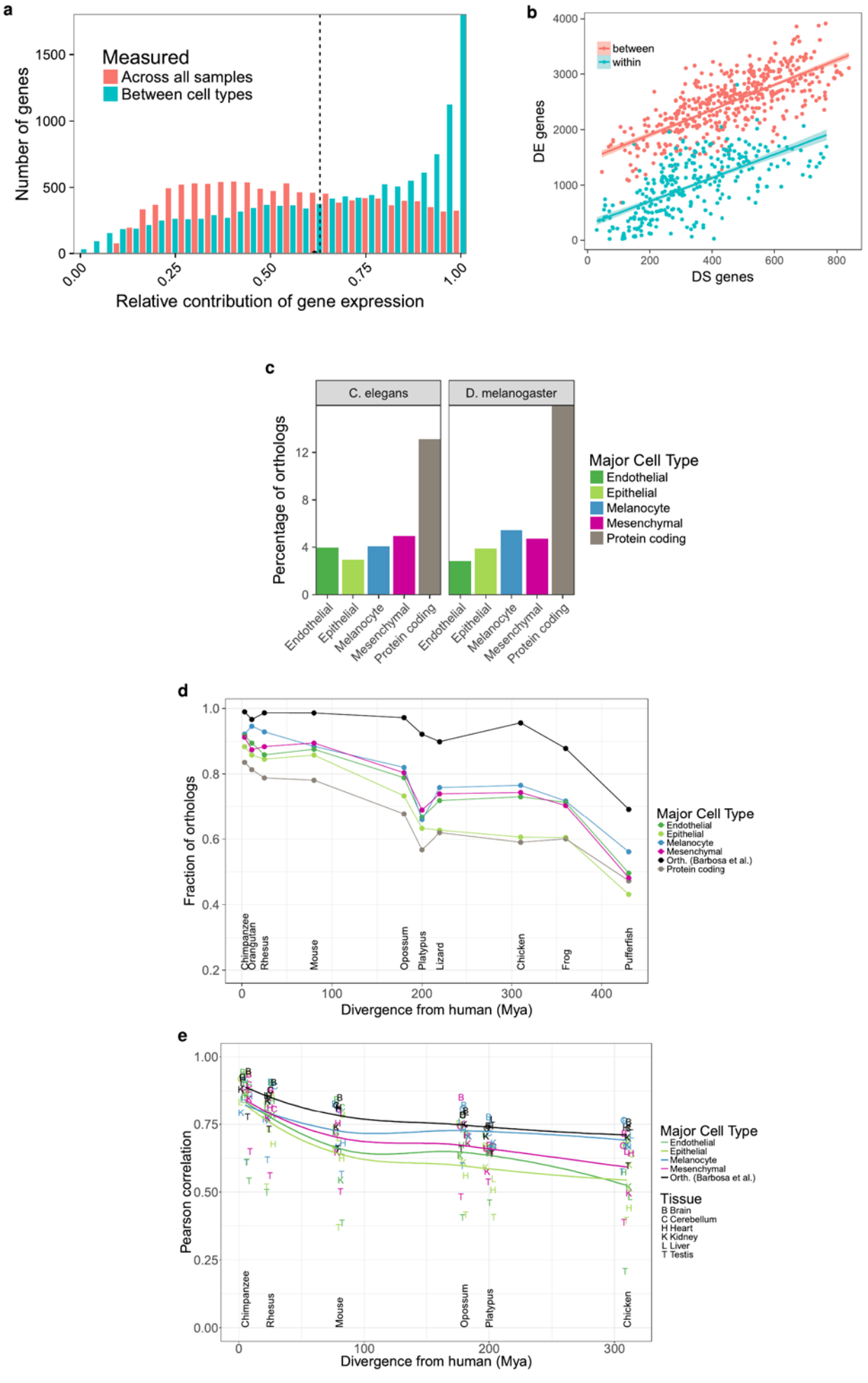

Fig. 3. Transcriptional complexity of human primary cells and evolutionary conservation of cell type specific genes.

(a) Distribution of the relative contribution of gene expression to the variation in isoform abundance between major cell types (blue) and between all primary cells. Large values of the contribution of gene expression indicate that changes in isoform abundance from one condition (primary cell, cell type) to another can be simply explained by changes in gene expression. Small values, by contrast, indicate that changes of isoform abundance are mostly independent of changes in gene expression, and can obey to changes in the relative abundance of the isoform. **(b)** Number of differentially expressed genes (DE, y axis) vs number of genes with differentially spliced exons (DS, x axis), between pairs of samples of the same cell type (within, blue) or different cell types (between, red). DS genes have been obtained using IPSA (https://github.com/pervouchine/ipsa-full). See also Fig. S19. **(c)** Percentage of cell type specific genes and protein coding genes with detected 1 to 1 orthologs in worm (*Caenorhabditis Elegans*) and fly (*Drosophila Melanogaster*). See also Fig. S22a. **(d)** Fraction of 1 to 1 orthologs between each species and human for major cell type specific genes and for protein coding genes overall. Species are sorted by increasing evolutionary distance from human. The black line is given as a reference and it indicates the proportion of 6-way orthologs (chimpanzee, rhesus, mouse, opossum, platypus and chicken) that are present in each species. The proportion is not 100% in these species because different versions of the GENCODE gene set reference were used. The genes in this set of 6-way orthologs are used for the comparison of gene expression in **(c)**. See also Fig. S22b. **(e)** Pearson's correlation coefficient between gene expression in each human organ and the corresponding one in every other species. The correlation is computed across all the genes in each major cell type separately. See also Fig. S23.

Estimation of the cellular composition of complex organs from the expression of cell type specific genes

We used the patterns of expression of cell type specific genes to estimate the cellular composition of human tissues and organs from GTEx bulk tissue transcriptome data *(19)* (version 6, 8,555 samples, 31 tissues, 544 individuals). We employed xCell *(20)*, using the sets of genes specific to epithelial, endothelial, and mesenchymal major cell types derived from ENCODE, and specific to brain (neural) and blood derived from GTEx *(21)* as signatures, and computed the enrichments of these cell types in each GTEx tissue sample (Supplementary Information).

The xCell enrichments (Fig. 4a, Table S12) are largely consistent with the histology of the tissues. For instance, esophagus mucosa is enriched for epithelial cells, while esophagus muscularis is enriched for mesenchymal cells. Skin (both exposed and unexposed) is enriched in epithelial cells; fibroblasts, in mesenchymal cells, etc. Blood and brain are only enriched in blood and neural cells, respectively. Most other tissues are not enriched in these two major cell types, with the expected

exceptions of spleen enriched in blood cells, and pituitary enriched in neural cells. Testis, which is widespread transcription *(22)*, is also enriched in neural cells, a reflection of the similarity of the expression programs of these two organs *(23)*. Consistent with previous observations *(24)*, we found enrichment of cells of endothelial type in adipose tissue. The analysis of the pathology reports of the subcutaneous adipose tissue shows that often is contaminated with other tissues, in particular blood vessels, which would explain the enrichment in cells of the endothelial type. We have further processed and analyzed the histopathology images available from the GTEx adipose samples (Supplementary Information), and estimated that on average about 84% of the adipose tissue does actually correspond to adipocytes (Fig. S24), which would explain the endothelial enrichment. In skeletal muscle we do not observe a particularly large enrichment in cells of the mesenchymal type, in apparent contradiction with our initial classification (Fig. 1b, f). The samples in GTEx, however, are all from differentiated skeletal muscle, while the ENCODE primary cells that we used to identify the mesenchymal specific genes are undifferentiated satellite cells (SkMC), and smooth muscle cells (Table S1). We analyzed single cell RNA-seq data produced during skeletal myoblast differentiation *(25)*, and found that differentiating skeletal muscle cells retain the mesenchymal signature through most of the differentiation pathway, acquiring only the GTEx muscle specific signature when fully differentiated (Fig. S25a-c). Further supporting that muscle is indeed of mesenchymal type, potentially forming a well-defined subtype, gene expression profiles cluster together myoblast differentiating single cells with ENCODE mesenchymal cells, rather than with epithelial or endothelial cells, or forming a separate cluster (Fig. S25d).

To independently assess the xCell enrichments, we analyzed the histological images of the few tissues in which samples were obtained from different subregions. This is most notable in the case of transverse colon and stomach. The GTEx stomach samples are all from the gastric body, whose walls consist of two broad layers: the mucosa, which is mostly epithelial, and the muscularis, which is smooth muscle (Fig. 4b). We processed the histological images, and identified a subset of samples that presented mostly the muscularis or the mucosa layer (Supplementary Information). This partition of the samples has been also observed by the GTEx consortium using laser capture microdissection (K. Ardlie, personal communication). The enrichment of epithelial cells in the samples from the muscularis layer is much lower than in the samples from the mucosa layer; conversely, the enrichment of mesenchymal cells is much higher in the muscularis than in the mucosa layer. The two sets of samples are almost perfectly separated by our cellular enrichments (Fig. 4c), explaining the bimodality in the distribution of cell type enrichments observed

specifically in the stomach samples (Fig. 4a). Consistently, we found that epithelial specific genes were exclusively expressed in the mucosa layer and mesenchymal specific genes were exclusively expressed in the muscularis layer (Fig. 4d). Next, we used the classification of stomach images to train an SVM model (Fig. S26a-b), and used this model to predict the presence of the two layers in 196 transverse colon samples — with histology similar to that of stomach (Supplementary Information). The SVM-predicted classification closely matches the differences observed at the transcriptional level, and confirms that the bimodality of cellular composition (Fig. 4a) is again related to the unbalanced presence of the two tissue layers across samples (Fig. S26c). Considering that stomach and colon were not represented in our primary cell collection, this constitutes a strong validation of our estimates of the cellular enrichments in tissues.

Figure 4

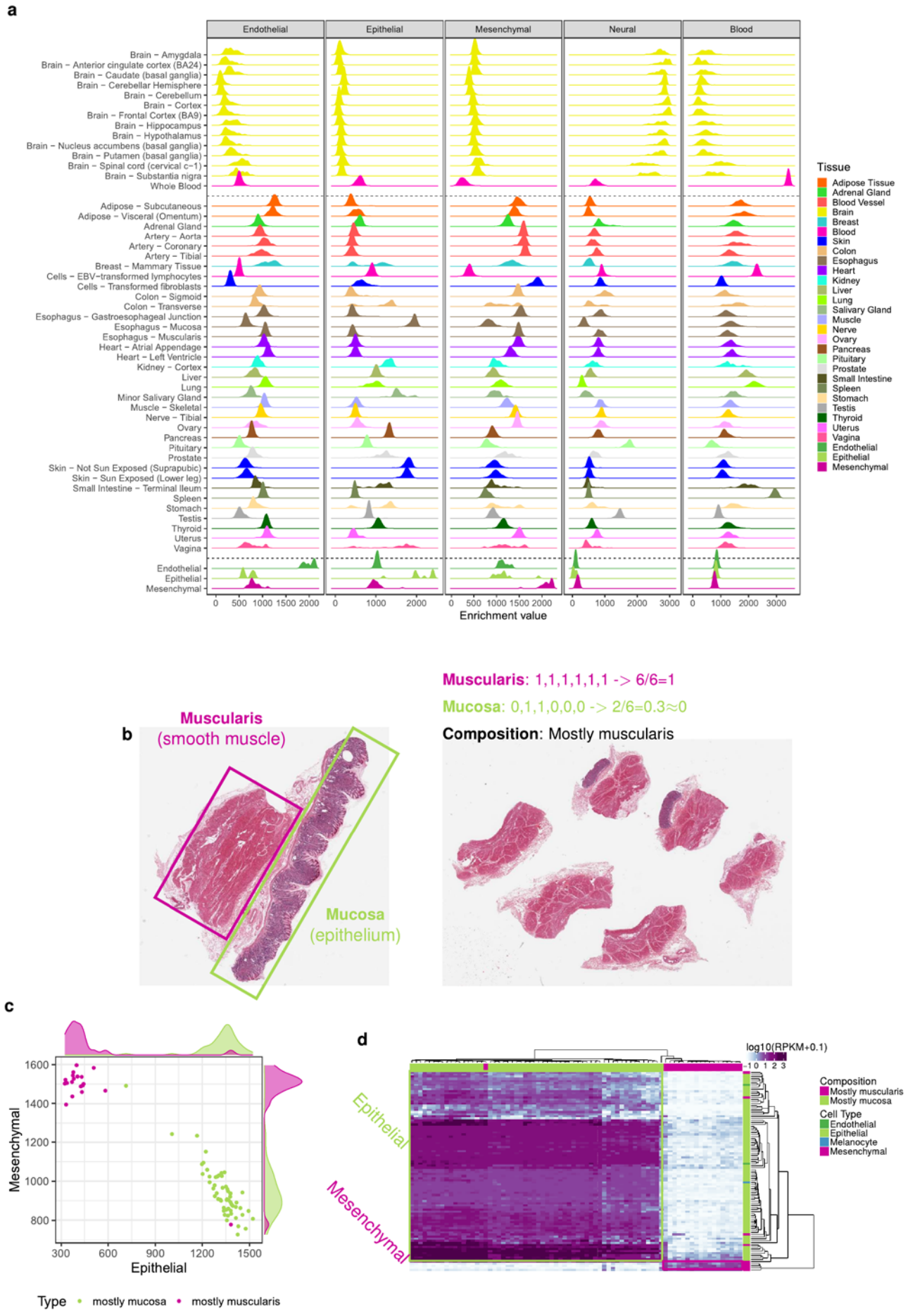

Fig. 4. Expression of cell type cluster-specific genes in GTEx organs.

(a) Enrichment of each major cell type in GTEx tissues, estimated from bulk tissue RNA-seq using the xCell method. As a control, we also include the enrichments in the primary cells monitored here. As expected, the highest enrichment for cells of a particular cell type occurs in cells of that cell type. **(b)** Example of stomach histological slides which represent the two main tissue layers and the procedure for the manual annotation of the images based on the presence of those layers. Each GTEx histological image displays up to six tissue slices. For the stomach samples, we scored each slice for the presence (1) or absence (0) of the muscularis and mucosa layers, summed up the values for each layer separately and divided by the number of slices. If the proportion of slices with mucosa layer, or muscularis layer, is more than 50% we classify the entire slide as mc1, or ms1, respectively. If the proportion is lower, we classify the slide as mc0 or ms0. A combined class, for example mc0ms1, is assigned to the slides. Thus, samples labeled mc0ms1 are mostly muscularis, while samples labelled mc1ms0 are mostly mucosa. **(c)** Enrichment of cells of epithelial and mesenchymal types in stomach samples containing mostly the mucosa (green) or mostly the muscularis (purple) layer. **(d)** Expression of the cell type-specific genes that drive the separation of stomach samples in mostly muscularis or mostly mucosa samples. Among discriminant cell type specific genes, mucosa only samples express almost exclusively epithelial specific genes, while muscularis only samples express exclusively mesenchymal specific genes.

Alterations of cellular composition in pathological states

We projected the solid non neural GTEx tissue samples on a 3-dimensional space according to the enrichments of epithelial, endothelial and mesenchymal cell types in each sample (Fig. 5a, Fig. S27). The spatial arrangement of the samples recapitulates tissue type as strongly as the clustering based on gene expression (Fig. S28). This suggests that the basic cell type composition is a characteristic signature of tissues, and that departures from this composition may reflect pathological or diseased states. To assess this hypothesis, we analyzed the histological reports associated with the GTEx images (7,911 reports). We employed fuzzy string search and parse trees to convert the natural language annotations produced by the pathologists to annotations in a controlled vocabulary that can be analyzed automatically (Supplementary Information, Table S13). In this way, we identified 19 histological phenotypes affecting one or more tissues for which there were at least 30 affected samples. From these, we identified six conditions with significant (FDR < 0.01) altered contributions of major cell types when comparing composition of affected and normal tissue (Fig. 5b-e)). Atherosclerosis in the tibial artery, which is more prevalent in older donors (Fig. S29a) is associated with an increase in endothelial cells (Fig. 5b); this might be attributed to endothelial proliferation stimulated in peripheral artery occlusion *(26)*. Atrophic skeletal muscle, a phenotype which is also correlated with age (Fig. S29b), is associated with an

increase in mesenchymal cells, which is consistent with the reported increase of connective tissue *(27)* and intermuscular fat *(28-29)* in atrophy (Fig. 5c). Indeed, analysis of the pathology reports of GTEx muscle histological images reveals that the proportion of fat is almost twice as high in atrophic than in non-atrophic muscle (24% vs 13%, Supplementary Information). Elevated enrichments of mesenchymal cells are also observed in liver congestion (Fig. S30a), a condition that often precedes fibrosis, which is characterized by an activation of matrix-producing cells, including fibroblasts, fibrocytes and myofibroblasts *(30)*. In spite of the low presence of cells of the major cell types in the testis, we found a further reduction of cells of all these types, mostly endothelial, in testis undergoing spermatogenesis (Fig. S30b). In lung pneumonia, we also observe alteration of all cell types (Fig. S30c). The sixth condition is gynecomastia, a pathology which is characterized by ductal epithelial hyperplasia *(31)*. We investigated differences in cellular composition between males and females, and found them significant only in mammary tissue, where female breasts exhibit much higher enrichment in epithelial cells than male breasts, possibly due to the presence of epithelial ducts and lobules (Fig. 5d). Remarkably, males diagnosed with gynecomastia show a cellular composition similar to that of females, mirroring tissue morphology.

We also observed specific age-related changes in cellular composition in lung and ovarian tissues. In lung samples we observe changes of all cell types, in particular, a significant reduction of epithelial cells in older donors (Fig. 5e), which is consistent with the impaired re-cellularization of lung epithelium that has been observed in decellularized lungs of aged mice *(32)*. Consistently, a similar pattern can be observed in the lungs of the individuals that died of respiratory related causes (Fig. S30d-e). In ovarian samples of women older than 48, a lower bound for menopause occurrence, we observe a decrease in endothelial cells (Fig. S30f), potentially related to an age-dependent decline in ovarian follicle vascularity *(33)*.

Figure 5

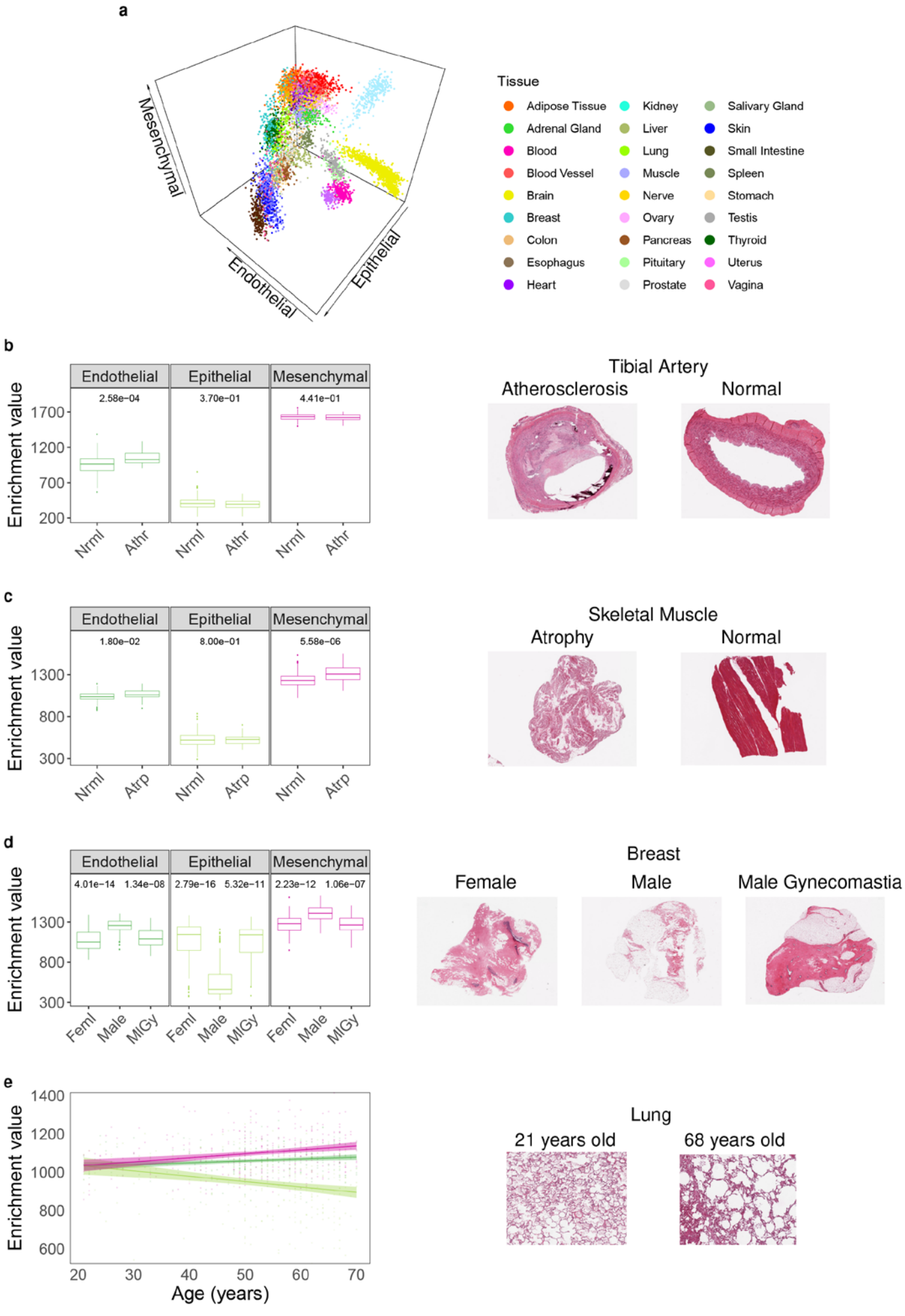

Fig. 5. Alterations of the contributions of the major cell types to tissues in histological phenotypes.

(a) GTEx samples represented in a 3D space where the axes are the enrichments of endothelial, epithelial and mesenchymal cells. (**b** and **c**) Differences in xCell enrichments of major cell types (Wilcoxon test, adjusted p-values as FDR) between affected and normal states. Histological images of affected and normal tissues are displayed (see text for details). Nrml: normal, Athr: atherosclerosis and Atrp: atrophy. **(d)** Major cell type xCell enrichments in female (Fml) breast samples, and male breast samples with (MlGy) or without gynecomastia (Male). Only significant FDR (≤ 0.05) are shown, all of them being between female and male without gynecomastia (left FDR) and between male without gynecomastia and male with gynecomastia (right FDR). **(e)** Changes in major cell type xCell enrichments in lung samples with age. Pearson's *r* and adjusted p-values as FDR for: Endothelial *r*=0.17 and FDR=3.20e-03, Epithelial *r*=-0.23 and FDR=6.00e-05, Mesenchymal *r*=0.25 and FDR=2.40e-05.

Altered cellular composition is likely to be particularly relevant in cancer. We analyzed, therefore, transcriptome data from the Cancer Genome Atlas Pan-Cancer analysis project (PCAWG) *(34)* for 19 cancers affecting tissues also profiled in the GTEx collection, and estimated the cellular enrichments of the major cell types (Fig. S31). For some cases there is also transcriptome data for normal samples from the same cancer project, which serve as a control for the highly different methodologies employed in GTEx and in the cancer projects. Thus, in lung cancer, there is an increase in epithelial cells (Fig. 6a-b), likely reflecting the epithelial origin of most lung cancers. In kidney primary tumors, in contrast, there is an overall increase of endothelial cells across most cancer subtypes, consistent with the increased vascularity associated with the cancer (Fig. 6c-d). The exceptions are renal papillary cell carcinomas, which present, instead, reduced vascularity *(35)*. In both cases, the cellular composition of GTEx samples and normal samples from the cancer projects are similar, supporting the robustness of our cellular characterization. Alterations in cellular composition can also reflect cancer progression. For ovary, even though we lack a comparable set of normal samples from the cancer projects, there is data on different stages of the disease, which serve as an internal control (Fig. 6e-f). Compared to GTEx normal data, there is markedly increase in epithelial cells in cancer, which is more evident as the severity of the cancer progresses, from primary to recurrent.

Overall, the data collected here on the transcriptomics of human primary cells constitute a unique resource, serving as an intermediate resolution of complexity between single cell and whole organ transcriptomics. This resource will contribute to the understanding of how the interplay between transcription and cellular composition shapes tissue histology, and ultimately impacts, human

phenotypes. Our analyses suggest that a large fraction of human cells and cell types in tissues belong to a few major cell types, providing a high level transcriptionally-based hierarchical classification of human cells. Extending the variety of profiled cell types, achieving single cell resolution and integrating expression data with epigenetics data, as proposed in the Human Cell Atlas project *(3)*, will enrich our understanding of the constitutive cell types in the human body and of their functional relationship.

Figure 6

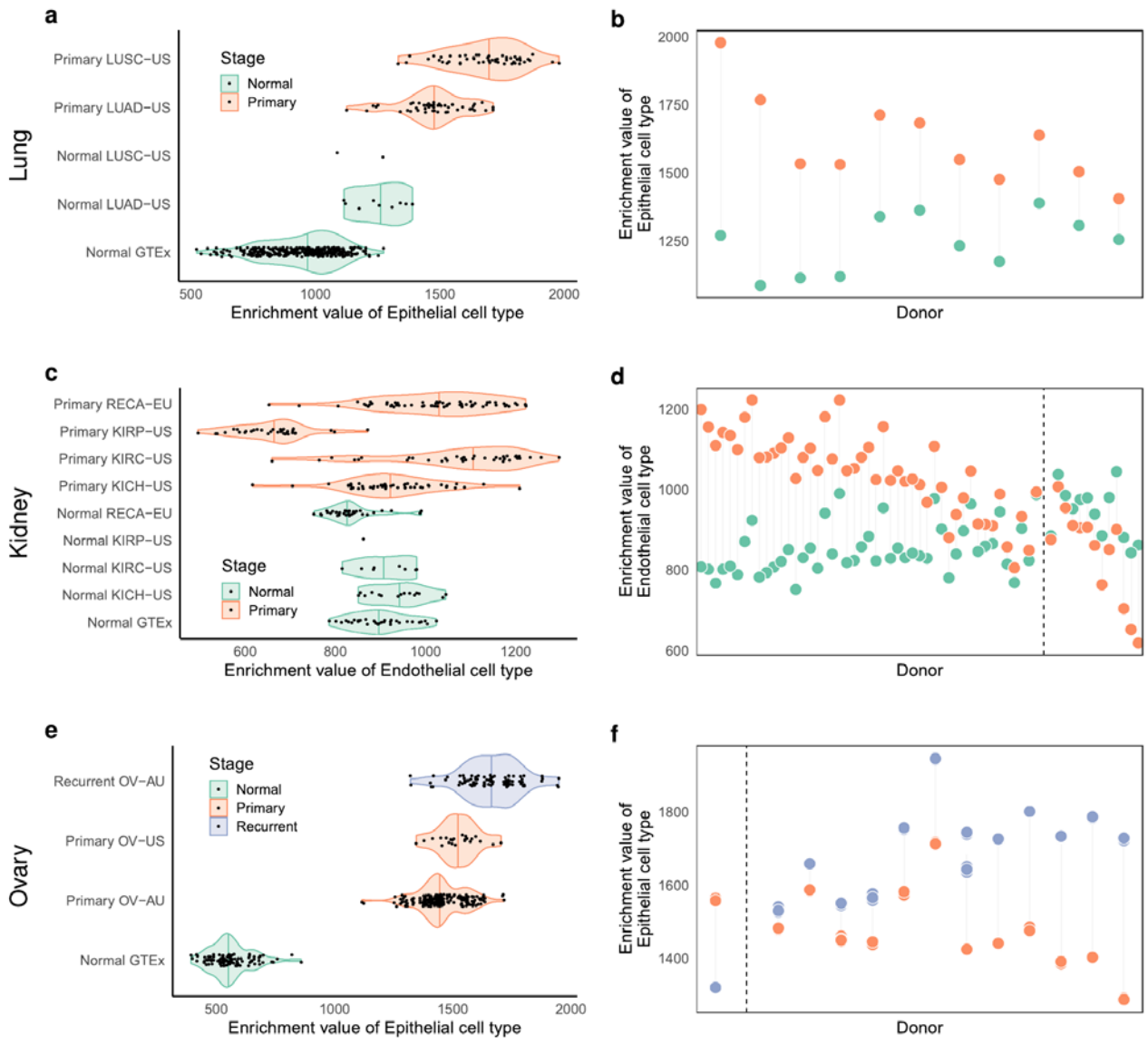

Fig. 6. Alterations of the contributions of the major cell types to tissues in cancer.

(a) xCell enrichments in epithelial cells in lung cancers and matched normal controls from the PCAWG project separated by cancer project. LUAD-US: Lung Adenocarcinoma, TCGA, USA; LUSC-US: Lung Squamous Cell Carcinoma, TCGA, USA. **(b)** Enrichment in matched normal and cancer lung samples by donor, pooled across the cancer projects. The p-value for the Wilcoxon test for the differences in epithelial contribution between normal and

cancer samples in the LUAD-US project is: 8.1e-06. **(c)** xCell enrichment in endothelial cells in kidney cancers and matched normal controls from the PCAWG project separated by cancer project. RECA-EU: Renal Cell Cancer, France, EU; KIRP-US: Kidney Renal Papillary Cell Carcinoma, TCGA, USA; KIRC-US: Kidney Renal Clear Cell Carcinoma, TCGA, USA; KICH-US: Kidney Chromophobe, TCGA, USA. **(d)** xCell Enrichments in matched normal and cancer kidney samples by donor. The adjusted p-values for the Wilcoxon tests for the differences in endothelial contribution between normal and cancer samples in the RECA-EU, KIRC-US, KICH-US projects are respectively: 3.8e-12, 0.0024, 0.65. **(e)** xCell enrichments in epithelial cells in ovarian cancers from the PCAWG project separated by cancer project or by **(f)** donor for matched primary and recurrent samples. OV-AU: Ovarian Cancer, Austria; OV-US: Ovarian Serous Cystadenocarcinoma, TCGA, USA. The p-value for the Wilcoxon test for the differences in endothelial contribution between primary and recurrent samples in the OV-AU project is: 3.6e-27. The donors in displays b, d, f are sorted based on the difference between the enrichments. The dashed lines in d, f separate the matched samples in which the enrichment of endothelial (epithelial) cells is larger in the cancer sample from those in which it is larger in the normal sample.

Materials and Methods

RNA Isolation, Library Construction and Sequencing

For each cell type to be made into a library we obtained cell pellets that were stored in RNAlater (Thermofisher) as catalogue items from PromoCell (http://www.promocell.com) and ScienCell (https://www.sciencellonline.com/) (see Table S1 for a list of primary cells). We rely on the providers' standards for quality assurance. Quality sheets are available through the ENCODE portal (see for example: https://www.encodeproject.org/search/?type=Biosample&organism.scientific_name=Homo+sapiens&biosample_ontology.classification=primary+cell&lab.title=Thomas+Gingeras%2C+CSHL&source.title=PromoCell&award.rfa=ENCODE3). We ordered 3 vials per cell type per donor for a total of 3 million cells. The 3 vials were combined together and we isolated Total RNA from them using the Ambion mirVana miRNA Isolation kit (cat #AM1561). The rRNA was removed using the RiboZero Gold Protocol (cat #RZG1224). The libraries are made using a homebrew "dUTP" protocol *(36)*, which generates stranded libraries. They were sequenced on the Illumina platform in mate-pair fashion and processed though the data processing pipeline at the ENCODE DCC. Additional, information about each of these steps, metadata and files can be found at: https://www.encodeproject.org/.

RAMPAGE sample preparation

Isolation of RNA is described in the above section. The RAMPAGE protocol *(37)* was used to make libraries. Each library was sequenced in mate-pair fashion on the Illumina platform. Detailed protocol and quality control images and metrics on a per library basis can be found in the "Production Documents" appended to each RAMPAGE assay at the ENCODE Data Coordination center: https://www.encodeproject.org/.

Small RNA Isolation, Library Construction and Sequencing

Isolation of RNA is described in the above section. The Illumina TruSeq protocol was used to make libraries. Each library was sequenced in single end fashion on the Illumina platform. Detailed protocol and quality control images and metrics on a per library basis can be found in the "Production Documents" appended to each Small RNA assay at the ENCODE Data Coordination center: https://www.encodeproject.org/.

RNA-seq processing pipeline

Raw reads from the 106 RNA-seq libraries (see Table S1 for a list of ENCODE library ids and https://www.encodeproject.org/ for submitted fastq files) were aligned with STAR v2.3.1z *(38)* to the human genome assembly hg19. Reads mapping to more than 20 multiple positions were discarded. Read counts for all long genes annotated in GENCODE v19 *(39)* were computed with RSEM 1.2.19 *(40)* (expected read counts).

Since for most of the analyses we average expression values for a given pair of replicates and sometimes the two biological replicates are from donors of opposite sex, we remove genes on chromosome Y. The lack of an enrichment step for polyadenylated transcripts preserves the presence of some short biotype genes, which are still longer than 200bp. Thus, we remove genes with at least one transcript annotated as short RNA in GENCODE. These genes are often of repetitive nature which makes the quantification of their expression problematic, this is why we decided to remove them.

Read counts which are not reproducible between two replicates (npIDR > 0.1) *(41)* are set to 0. The matrix of read counts after npIDR is provided as Table S2. After filtering for reproducibility, read counts are normalized to a slightly modified version of RPKM (reads per kilobase of exon

model per million mapped reads *(42)*). Specifically, read counts were first normalized to cpm (counts per million), where the library sizes are the TMM (trimmed mean of M values *(43)*) scaled sums of exonic reads, and then normalized by gene length. Finally, RPKM values from the two replicates were averaged, and genes with RPKM<1 in all samples were discarded, resulting in 16,265 genes, including 13,990 protein coding, 1,380 long non-coding RNAs and 895 pseudogenes.

As the samples were prepared and sequenced in three known distinct batches (see Table S1), we used the *removeBatchEffect()* function from R limma package *(44)* to build a linear model with the batch information and the cell types on log10-transformed RPKM (with a pseudocount of 0.01), and we regressed out the batch variable.

Data and materials availability:

All experimental protocols for the samples described here are available on the ENCODE portal www.encodeproject.org. Detailed information about data processing and analyses are available as Supplementary Information. All the data generated for this study are also publicly available on the ENCODE portal www.encodeproject.org. Additional data tables derived from the analyses are included in this published article (and its supplementary information files). GTEx gene expression is available in the GTEx portal at www.gtexportal.org.

Acknowledgments:

We thank Kristin Ardlie and Detlev Arendt for useful discussions. We acknowledge and thank the donors and their families for their generous gifts of organ donation for transplantation and tissue donations for the GTEx research study. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We acknowledge the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership.

Disclosure declaration:

The authors declare no competing financial interests.

Author contributions:

A.B., C.A.D., M.M., V.W., R.G. and T.R.G. conceived and designed the experiments and analyses. J.D., C.A.D., A.S. and C.D. performed the experiments. A.B., M.M., V.W., D.G. analyzed the data. J.G., D.D.P., A.V., A.D., C.Z., D.G., F.R., M.P.S. contributed with ideas and statistical advice. A.B., M.M., V.W., R.G. and T.R.G. wrote the manuscript.

Funding:

This project was supported by awards U54HG007004, U41HG007234 and R01MH101814 from the National Human Genome Research Institute of the National Institutes of Health, as well as from the Spanish Ministry of Economy and Competitiveness, Centro de Excelencia Severo Ochoa 2013–2017, SEV-2012-0208, Programa de Ayudas FPI del Ministerio de Economía y Competitividad BES-2012-055848 to A.B., and Ministerio de Educación, Cultura y Deporte, under the FPU programme (Formación de Profesorado Universitario) with pre-doctoral fellowship FPU15/03635 to M.M-A., as well as the support of the CERCA programme / Generalitat de Catalunya. D.G.-M. is supported by a “la Caixa”-Severo Ochoa pre-doctoral fellowship LCF/BQ/SO15/52260001. We would also like to acknowledge support from the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement 294653.

References

1. A. Haque, J. Engel, S. A. Teichmann, T. Lönnberg, A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine.* **9** (2017).

2. C. Trapnell, Defining cell types and states with single-cell genomics. *Genome Research.* **25**, 1491–1498 (2015).

3. A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, The Human Cell Atlas. eLife. **6** (2017).

4. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature.* **507**, 462–470 (2014).

5. The ENCODE Consortium. The ENCODE Encyclopedia for Human and Mouse. in preparation.

6. The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature.* **562**, 367–372 (2018).

7. B. Young, G. O'Dowd, P. Woodford. *Wheater's Functional Histology* (Elsevier Health Sciences, 2013).

8. A. L. Mescher. Junqueira's basic histology: text and atlas (McGraw-Hill Medical, ed. 13, 2013).

9. V. P. Eroschenko, M. S. H. di Fiore. *DiFiore's Atlas of Histology with Functional Correlations*. (Lippincott Williams & Wilkins, ed. 12, 2013).

10. N. C. Sheffield, R. E. Thurman, L. Song, A. Safi, J. A. Stamatoyannopoulos, B. Lenhard, G. E. Crawford, T. S. Furey, Patterns of regulatory activity across diverse human cell types

predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research*. **23**, 777–788 (2013).

11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

12. E. A. Rach, D. R. Winter, A. M. Benjamin, D. L. Corcoran, T. Ni, J. Zhu, U. Ohler, Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level. *PLoS Genetics*. **7**, e1001274 (2011).

13. D. D. Pervouchine, S. Djebali, A. Breschi, C. A. Davis, P. P. Barja, A. Dobin, A. Tanzer, J. Lagarde, C. Zaleski, L.-H. See, M. Fastuca, J. Drenkow, H. Wang, G. Bussotti, B. Pei, S. Balasubramanian, J. Monlong, A. Harmanci, M. Gerstein, M. A. Beer, C. Notredame, R. Guigó, T. R. Gingeras, Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nature Communications*. **6** (2015).

14. H.-M. Zhang, H. Chen, W. Liu, H. Liu, J. Gong, H. Wang, A.-Y. Guo, AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Research. 40, D144–D149 (2011).

15. F. McLaughlin, V. J. Ludbrook, J. Cox, I. von Carlowitz, S. Brown, A. M. Randi, Combined genomic and antisense analysis reveals that the transcription factor Erg is implicated in endothelial cell differentiation. *Blood*. **98**, 3332–3339 (2001).

16. K. Yoh, R. Prywes, Pathway Regulation of p63, a Director of Epithelial Cell Fate. *Frontiers in Endocrinology*. **6** (2015).

17. M. Gonzalez-Porta, M. Calvo, M. Sammeth, R. Guigo, Estimation of alternative splicing variability in human populations. *Genome Research*. **22**, 528–538 (2011).

18. N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, B. J. Blencowe, The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*. **338**, 1587–1593 (2012).

19. GTEx consortium. Genetic effects on gene expression across human tissues. *Nature*. **550**, 204–213 (2017).

20. D. Aran, Z. Hu, A. J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*. **18** (2017).

21. R. Y. Yang, *et al.*, https://www.biorxiv.org/content/10.1101/311563v1.article-info (2018).

22. M. Soumillon, A. Necsulea, M. Weier, D. Brawand, X. Zhang, H. Gu, P. Barthès, M. Kokkinaki, S. Nef, A. Gnirke, M. Dym, B. de Massy, T. S. Mikkelsen, H. Kaessmann, Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports*. **3**, 2179–2190 (2013).

23. J. H. Guo, Q. Huang, D. J. Studholme, C. Q. Wu, Z. Zhao, Transcriptomic analyses support the similarity of gene expression between brain and testis in human as well as mouse. *Cytogenetic and Genome Research*. **111**, 107–109 (2005).

24. A. Frontini, A. Giordano, S. Cinti, Endothelial cells of adipose tissues: A niche of adipogenesis. *Cell Cycle*. **11**, 2765–2766 (2012).

25. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*. **32**, 381–386 (2014).

26. M. A. Ziegler, M. R. Distasi, R. G. Bills, S. J. Miller, M. Alloosh, M. P. Murphy, A. G. Akingba, M. Sturek, M. C. Dalsing, J. L. Unthank, Marvels, Mysteries, and Misconceptions of Vascular Compensation to Peripheral Artery Occlusion. *Microcirculation*. **17**, 3–20 (2010).

27. H.-J. Appell, Muscular Atrophy Following Immobilisation. *Sports Medicine*. **10**, 42–58 (1990).

28. T. M. Manini, B. C. Clark, M. A. Nalls, B. H. Goodpaster, L. L. Ploutz-Snyder, T. B. Harris, Reduced physical activity increases intermuscular adipose tissue in healthy young adults. *The American Journal of Clinical Nutrition*. **85**, 377–384 (2007).

29. O. Addison, R. L. Marcus, P. C. LaStayo, A. S. Ryan, Intermuscular Fat: A Review of the Consequences and Causes. *International Journal of Endocrinology*. 2014, 1–11 (2014).

30. G. Ö. Elpek, Cellular and molecular mechanisms in the pathogenesis of liver fibrosis: An update. *World Journal of Gastroenterology*. **20**, 7260 (2014).

31. N. Cuhaci, S. Polat, B. Evranos, R. Ersoy, B. Cakir, Gynecomastia: Clinical evaluation and management. *Indian Journal of Endocrinology and Metabolism*. **18**, 150 (2014).

32. D. Sokocevic, N. R. Bonenfant, D. E. Wagner, Z. D. Borg, M. J. Lathrop, Y. W. Lam, B. Deng, M. J. DeSarno, T. Ashikaga, R. Loi, A. M. Hoffman, D. J. Weiss, The effect of age and emphysematous and fibrotic injury on the re-cellularization of de-cellularized lungs. *Biomaterials*. **34**, 3256–3269 (2013).

33. C. Tatone, F. Amicarelli, M. C. Carbone, P. Monteleone, D. Caserta, R. Marci, P. G. Artini, P. Piomboni, R. Focarelli, Cellular and molecular aspects of ovarian follicle ageing. *Human Reproduction Update*. **14**, 131–142 (2008).

34. The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*. **45**, 1113–1120 (2013).

35. S. A. Aziz, J. Sznol, A. Adeniran, J. W. Colberg, R. L. Camp, H. M. Kluger, Vascularity of primary and metastatic renal cell carcinoma specimens. *Journal of Translational Medicine*. **11**, 15 (2013).

36. D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, A. Soldatov, Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research*. **37**, e123–e123 (2009).

37. P. Batut, T. R. Gingeras, RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5′-Complete cDNAs. *Current Protocols in Molecular Biology*. **104** (2013).

38. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2012).

39. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, T. J. Hubbard, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*. **22**, 1760–1774 (2012).

40. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics.* **12** (2011).

41. S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, T. R. Gingeras, Landscape of transcription in human cells. *Nature.* **489**, 101–108 (2012).

42. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods.* **5**, 621–628 (2008).

43. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology.* **11**, R25 (2010).

44. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research.* **43**, e47–e47 (2015).

45. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics.* **20**, 53–65 (1987).

46. R. L. Thorndike, Who belongs in the family? *Psychometrika*. **18**, 267–276 (1953).

47. A. Breschi, S. Djebali, J. Gillis, D. D. Pervouchine, A. Dobin, C. A. Davis, T. R. Gingeras, R. Guigó, Gene-specific patterns of expression variation across organs and species. *Genome Biology.* **17** (2016).

48. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* **26**, 139–140 (2009).

49. S. Falcon, R. Gentleman, Using GOstats to test gene lists for GO term association. *Bioinformatics.* **23**, 257–258 (2006).

50. M. T. Nogalski, A. Solovyov, A. S. Kulkarni, N. Desai, A. Oberstein, A. J. Levine, D. T. Ting, T. Shenk, B. D. Greenbaum, A tumor-specific endogenous repetitive element is induced by herpesviruses. *Nature Communications.* **10** (2019).

51. S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. Hayden Gephart, B. A. Barres, S. R. Quake, A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences.* **112**, 7285–7290 (2015).

52. Y. Zhang, S. A. Sloan, L. E. Clarke, C. Caneda, C. A. Plaza, P. D. Blumenthal, H. Vogel, G. K. Steinberg, M. S. B. Edwards, G. Li, J. A. Duncan III, S. H. Cheshier, L. M. Shuer, E. F. Chang, G. A. Grant, M. G. H. Gephart, B. A. Barres, Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron.* **89**, 37–53 (2016).

53. B. S. Carvalho, R. A. Irizarry, A framework for oligonucleotide microarray preprocessing. *Bioinformatics.* **26**, 2363–2367 (2010).

54. B. S. Carvalho, pd.huex.1.0.st.v2 Platform Design Info for Affymetrix HuEx-1 0-st-v2. R package version 3.14.1. (2015), doi:10.18129/B9.BIOC.PD.HUEX.1.0.ST.V2.

55. P. Kheradpour, M. Kellis, Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research.* **42**, 2976–2987 (2013).

56. M. Mele, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segre, S. Djebali, A. Niarchou, T. G. Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, R. Guigo, The human transcriptome across tissues and individuals. *Science.* **348**, 660–665 (2015).

57. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution.* **34**, 1812–1819 (2017).

58. J. Herrero, M. Muffato, K. Beal, S. Fitzgerald, L. Gordon, M. Pignatelli, A. J. Vilella, S. M. J. Searle, R. Amode, S. Brent, W. Spooner, E. Kulesha, A. Yates, P. Flicek, Ensembl comparative genomics resources. *Database*, Volume 2016. doi:10.1093/database/bav096.

59. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics.* **27**, 1017–1018 (2011).

60. R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, J. K. Pritchard, Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research.* **21**, 447–455 (2010).

61. A. Pohl, M. Beato, bwtool: a tool for bigWig files. *Bioinformatics.* **30**, 1618–1619 (2014).

62. D. D. Pervouchine, D. G. Knowles, R. Guigo, Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics.* **29**, 273–274 (2012).

63. Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, J. K. Pritchard, Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*. **50**, 151–158 (2017).

64. M. Sammeth, S. Foissac, R. Guigó, A General Definition and Nomenclature for Alternative Splicing Events. *PLoS Computational Biology.* **4**, e1000147 (2008).

65. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, Missing value estimation methods for DNA microarrays. *Bioinformatics.* **17**, 520–525 (2001).

66. D. Garrido-Martín, E. Palumbo, R. Guigó, A. Breschi, ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLOS Computational Biology.* **14**, e1006360 (2018).

67. S. Lenz, P. Lohse, U. Seidel, H. H. Arnold, The alkali light chains of human smooth and nonmuscle myosins are encoded by a single gene. Tissue-specific expression by alternative splicing pathways. *Journal of Biological Chemistry.* **264** (15), 9000-9015 (1989).

68. J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive,

D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, H. F. Moore, The Genotype-Tissue Expression (GTEx) project. *Nature Genetics.* **45**, 580–585 (2013).

69. M. Fontes, C. Soneson, The projection score - an evaluation criterion for variable subset selection in PCA visualization. *BMC Bioinformatics.* **12** (2011), doi:10.1186/1471-2105-12-307.

70. J. H. Krijthe. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. R package version 0.13. 2015. URL: https://github.com/jkrijthe/Rtsne

71. M. E. Arntfield, D. van der Kooy, β-Cell evolution: How the pancreas borrowed from the brain. BioEssays. 33, 582–587 (2011).

72. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological).* **57**, 289–300 (1995).

Supplementary Materials:

Supplementary text

Figures S1-S31

Tables S1-S13

References (*45-72*)