bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A limited set of transcriptional programs define major histological types and provide the molecular basis for a cellular taxonomy of the human body

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]

[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]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, Castanet Tolosan, 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

*A. Breschi and M. Muñoz-Aguirre contributed equally to and are co-first authors of this article.

†Correspondence should be addressed to E-mail: roderic.guigo@crg.cat (Roderic Guigó) and gingeras@cshl.edu (Thomas R. Gingeras)

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The ENCODE project has produced a collection of RNA sequencing experiments from many cell lines and bulk tissues that constitutes an extensive catalogue of the expression programs utilized in the human body. However, the relationship between the transcriptomes of tissues and those of the constituent primary cells, and how these impact tissue phenotypes has not been well established. Here we have produced RNA sequencing data for a number of primary cells from ten human body locations. The analysis of this data, together with additional epigenetic data from a total of 146 primary cells, indicates that many cells in the human body belong to five major cell types of similar transcriptional complexity: three, epithelial, endothelial, and mesenchymal, are broadly distributed across the human body acting as components for many tissues and organs, and two, neural and blood cells, are more anatomically localized. Based on gene expression, these redefine the basic histological types by which tissues have been traditionally classified. We have identified genes whose expression is specific to these cell types, and have estimated the relative proportion of the major cell types in human tissues using the transcriptional profiles produced by the GTEx project. The inferred cellular composition is a characteristic signature of tissues and reflects 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. This transcriptionally based classification of human cells provide a new view of human biology and disease.

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 the expression profiles of 14,475 protein coding, 1,618 long non-coding RNAs (lncRNAs) and 1,347 pseudogenes revealed a number of well defined clusters (Fig. 1b-c, Fig. S1, 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. The clustering is supported by the silhouette analysis and the elbow method[4,5] (Fig. S2a-b). 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 reflect 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[6] (Fig. 1d, Fig. S3a,b, Table S6), 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[7], Table S5), 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

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. S3c), neural cells profiled by DNAse-seq exhibit a gene expression pattern similar to mesenchymal cells (Fig. S3a). 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 major types (Supplementary Information, Fig. S5-S7).

These results, all together, suggest the existence of a limited number of core transcriptional programs encoded in the human genome. 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). Three of these major cell types, epithelial, endothelial and mesenchymal, have a broad anatomical distribution and are present in almost any human organ. The other two, neural and blood cells, are more anatomically localized. They all show similar transcriptional heterogeneity, with blood being the most transcriptionally diverse (Fig. S8). These transcriptionally defined major cell types match broadly, but not exactly, the basic histological types in which tissues are usually classified (see for example[8–10]): epithelial, of which endothelial is often considered a subtype, muscular, connective, which includes blood, and neural. However, from the transcriptional standpoint, endothelial and blood constitute separate cell types, and are not subtypes of epithelial and connective types, respectively, while the connective (but not blood) and 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, suggesting that body location may actually play a role in subtype specialization. Remarkably, the epithelial cluster includes primary cells of all embryonic origin (ectoderm, endoderm and mesoderm), suggesting that the transcriptional programs of cells may not be majoritarily inherited through development, but partially adopted through function.

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 in our analyses, these specialized cells are likely to have their unique transcriptional program.

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 the epithelial, endothelial, mesenchymal or melanocyte cell types (Fig. 2a, Fig. S9, Table S7). These cell type specific genes include nearly all genes that we identified as the major drivers of the clustering (Supplementary Information, Fig. S10). 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. S11a). 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. S11b). Cell type specific genes show consistent restricted expression in the FANTOM CAGE data (Fig. S12), and they are enriched for encyclopedia cREs[11] specifically in the primary cells of that type (Fig. S13). Using ChIP-seq histone modification data obtained in a number of primary cells[12] (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. S14a). 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. S14b). 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, Rach et al., 2011[13], Pervouchine et al. 2015[14]).

Among cell type specific genes, we identified 167 Transcription Factors (TFs) from a total of 1,544 TFs annotated in the human genome[15]. We focused on 56 that showed the strongest co-expression patterns (Pearson's correlation coefficient $\geq$ 0.85, Fig. 2c, Fig. S15). They include previously annotated cell type-specific transcriptional regulators, such as ERG, which has been shown to regulate endothelial cell differentiation[16], and TP63, which is an established regulator of epithelial cell fate and is often altered in tumor cells[17]. 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. S16).

We found that transcriptional regulation appears to play the major role compared to post-transcriptional

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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[18] 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 post-transcriptional regulation 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. S17, 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 RAMPAGE data, we identified a number of cell-type specific TSSs (Fig. S18, Table S10, Supplementary Information). Figure 3c shows the case of the gene coding for the S100 Calcium Binding Protein A16 (*S100A16*) which is selectively transcribed from a proximal TSS in endothelial cells, whereas in the other cell types transcription starts from a more distal TSS. Second, examination of splicing isoforms revealed 230 cell-type specific alternative splicing events (Table S11, Supplementary Information), independently of the tissue of origin, consistent with earlier reports[19]. As an example, exon 6 of the *MYL6* gene, coding for the myosin light chain 6 protein has been previously reported to be more often included in muscle cells[20]; however, we found that it is also often included in other mesenchymal cell types, including fibroblasts and stem cells, but not in adipocytes (Fig. 3d, Fig. S19a, Fig. S19b). Interestingly, the exon is translated as part of the MYL6 N-terminal EF-hand domain, calcium-binding domain which mediates the interactions between actin and myosin (Fig. S19c).

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. 4a, Fig. S20a), but they are overrepresented in vertebrates, as early as in tetrapoda. One exception are epithelial genes, which are overrepresented only in mammals (Fig. 4b, Fig. S20b). Within the set of orthologous genes across tetrapoda[21], the expression of cell type specific genes is less conserved than that of protein coding genes overall, especially at larger evolutionary distances (Fig. 4c, Fig. S20c, Fig. S21). 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.

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[22] (version 6, 8,555 samples, 31 tissues,

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

544 individuals). We employed xCell[23], 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[24] as signatures, and computed the enrichments of these cell types in each GTEx tissue sample. We cannot obtain proper estimates of the cell type proportions, because in general we ignore the expression levels of the marker genes for the cell types specific to a given tissue. Taking as a proxy the expression of GTEx derived tissue specific (TS) genes, however, we obtained lower bound estimates using two different methods. Across all tissues, excluding blood and brain, we found the average tissue composition explained by the major cell types to be at least 58% using lsqlin[25] and 40% using CIBERSORT[26] (Fig. S22-S28, Tables S12, S13, Supplementary Information).

The xCell enrichments (Fig. 5a) and the estimated proportions by lsqlin and CIBERSORT (Fig. S24 and S25) 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[27], is also enriched in neural cells, a reflection of the similarity of the expression programs of these two organs[28]. Maybe unexpectedly, there is some 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. S29), 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). To address the issue whether skeletal muscle cells can indeed be included within the mesenchymal type, we analyzed single cell RNA-seq data produced during skeletal myoblast differentiation[29], 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. S30a-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. S30d). All together these results reveal that cells belonging to epithelial, endothelial, and mesenchymal types are broadly anatomically distributed, being presented in almost all tissues and organs, and that together with neural and blood cells, they are likely to constitute the major cell types in the human body.

To independently assess the xCell enrichments, we analyzed the histological images of the few tissues

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. 5b). 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 (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 decomposition (Fig. 5c), explaining the bimodality in the distribution of cell type enrichments observed specifically in the stomach samples (Fig. 5a). 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. 5d). Next, we used the classification of stomach images to train an SVM model (Fig. S31a,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. 5a) is again related to the unbalanced presence of the two tissue layers across samples (Fig. S31c). Considering that stomach and colon were not represented in our primary cell collection, this constitutes a strong validation of our estimates of the cellular composition of tissues.

Finally, we analyzed a large compendium of single cell RNASeq datasets comprising 300 human samples from 20 different organs, totaling about one million cells in PanglaoDB[30]. These have been clustered in 178 primary cell types, for 68 of which accurate sets of maker genes (sensitivity > 0, Methods) have been derived. We have intersected the marker genes for these primary cell types with the signature genes for our major cell types. About two thirds of the single cell types neatly cluster within the five major cell types (Fig. S32). Moreover, about one third of the remaining cells (including pulmonary alveolar cells, Goblet cells, enterocytes, hepatocytes and others) cluster together, sharing part of the epithelial signature. We also observed a cluster of unclassified cells that includes, in addition to enteric glial cells, different types of pancreatic endocrine cells, and that share part of the neural signature, consistent with the strong morphological, and physiological similarities between these two types of cells[31]. This likely reflects that our signatures for the major cell types are still incomplete, and that they could be refined, as the transcriptomes of additional primary cells are characterized.

Alterations of cellular composition in pathological states

We projected the GTEx tissue samples on a 3-dimensional space according to the enrichments of epithelial, endothelial and mesenchymal cell types in each sample (Fig. 6a, S33). The spatial arrangement of the samples recapitulates tissue type as strongly as the clustering based on gene expression (Fig. S34). This suggests that the basic cell type composition is a characteristic signature of tissues, and that departures

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from this composition may reflect pathological or diseased states. To assess this hypothesis, we analyzed the histological reports associated to 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 S14). 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 further identified six conditions with significant (FDR < 0.01) altered proportions of cell types when comparing affected and normal tissue (Fig. 6b-e)). Atherosclerosis in the tibial artery, which is more prevalent in older donors (Fig. S35a) is associated to an increase in endothelial cells (Fig. 6b); this might be attributed to endothelial proliferation stimulated in peripheral artery occlusion[32]. Atrophic skeletal muscle, a phenotype which is also correlated with age (Fig. S35b), is associated to an increase in mesenchymal cells, which is consistent with the reported increase of connective tissue[33] and intermuscular fat[34,35] in atrophy (Fig. 6c). 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 proportions of mesenchymal cells are also observed in liver congestion (Fig. S36a), a condition that often precedes fibrosis, which is characterized by an activation of matrix-producing cells, including fibroblasts, fibrocytes and myofibroblasts[36]. 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. S36c). In lung pneumonia, we also observe alteration of all cell types (Fig. S36b). The sixth condition is gynecomastia, a pathology which is characterized by ductal epithelial hyperplasia[37]. 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. 6d). 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. 6e), which is consistent with the impaired re-cellularization of lung epithelium that has been observed in decellularized lungs of aged mice[38]. Consistently, a similar pattern can be observed in the lungs of the individuals that died of respiratory related causes (Fig. S36e-f). In ovarian samples of women older than 48, a lower bound for menopause occurrence, we observe a decrease in endothelial cells (Fig. S36d), potentially related to an age-dependent decline in ovarian follicle vascularity[39].

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[40] (PCAWG) for 19 cancers affecting tissues also profiled in the GTEx collection, and estimated the cellular enrichments of the major cell types (Fig. S37). For same 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. 7a,b), likely reflecting

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 to the cancer (Fig. 7c-d). The exception are renal papillary cell carcinomas, which present, instead, reduced vascularity[41]. 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. 7e-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 cellular transcription and cellular composition shapes tissue histology, and ultimately impacts, human phenotypes. Our analyses suggest that a large fraction of human cells 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.

Methods

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).

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine* **9,** 75 (2017).
2. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome research* **25,** 1491–1498 (2015).
3. Regev, A. *et al. The Human Cell Atlas* 2016.
4. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20,** 53–65 (1987).
5. Thorndike, R. L. Who belongs in the family? *Psychometrika* **18,** 267–276 (1953).
6. The FANTOM Consortium *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).
7. The ENCODE Consortium. The ENCODE Encyclopedia for Human and Mouse. *in preparation.*
8. Young, B., Woodford, P. & O'Dowd, G. *Wheater's functional histology: a text and colour atlas* (Elsevier Health Sciences, 2013).
9. Mescher, A. L. *Junqueira's basic histology: text and atlas* (Mcgraw-hill, 2013).
10. Eroschenko, V. P. & Di Fiore, M. S. *DiFiore's atlas of histology with functional correlations* (Lippincott Williams & Wilkins, 2013).
11. Sheffield, N. C. *et al.* 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).
12. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
13. Rach, E. A. *et al.* Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS genetics* **7,** e1001274 (2011).
14. Pervouchine, D. *et al.* Enhanced Transcriptome Maps from Multiple Mouse Tissues Reveal Evolutionary Constraint in Gene Expression for Thousands of Genes. *Nat Commun* **6** (2015).
15. Zhang, H.-M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic acids research* **40,** D144–D149 (2012).
16. McLaughlin, F. *et al.* Combined genomic and antisense analysis reveals that the transcription factor Erg is implicated in endothelial cell differentiation. *Blood* **98,** 3332–3339 (2001).
17. Yoh, K. & Prywes, R. Pathway regulation of p63, a director of epithelial cell fate. *Frontiers in endocrinology* **6,** 51 (2015).
18. Gonzàlez-Porta, M., Calvo, M., Sammeth, M. & Guigó, R. Estimation of alternative splicing variability in human populations. *Genome research* **22,** 528–538 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

19. Mallinjoud, P. *et al.* Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. *Genome research* **24,** 511–521 (2014).

20. Lenz, S, Lohse, P, Seidel, U & Arnold, H. 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,** 9009–9015 (1989).

21. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338,** 1587–1593 (2012).

22. Consortium, G. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204 (2017).

23. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology* **18,** 220 (2017).

24. Yang, R. Y. *et al.* Deep profiling of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. *under revision.*

25. Gong, T. *et al.* Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PloS one* **6,** e27156 (2011).

26. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12,** 453–457 (2015).

27. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell reports* **3,** 2179–2190 (2013).

28. Guo, J., Huang, Q, Studholme, D., Wu, C. & Zhao, Z. 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).

29. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32,** 381 (2014).

30. Franzén, O., Gan, L.-M. & Björkegren, J. L. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019** (2019).

31. Arntfield, M. E. & van der Kooy, D. β-Cell evolution: How the pancreas borrowed from the brain: The shared toolbox of genes expressed by neural and pancreatic endocrine cells may reflect their evolutionary relationship. *Bioessays* **33,** 582–587 (2011).

32. Ziegler, M. A. *et al.* Marvels, mysteries, and misconceptions of vascular compensation to peripheral artery occlusion. *Microcirculation* **17,** 3–20 (2010).

33. Appell, H.-J. Muscular atrophy following immobilisation. *Sports Medicine* **10,** 42–58 (1990).

34. Manini, T. M. *et al.* Reduced physical activity increases intermuscular adipose tissue in healthy young adults–. *The American journal of clinical nutrition* **85,** 377–384 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. Addison, O., Marcus, R. L., LaStayo, P. C. & Ryan, A. S. Intermuscular fat: a review of the consequences and causes. *International journal of endocrinology* **2014** (2014).

36. Elpek, G. Ö. Cellular and molecular mechanisms in the pathogenesis of liver fibrosis: An update. *World journal of gastroenterology: WJG* **20,** 7260 (2014).

37. Cuhaci, N., Polat, S. B., Evranos, B., Ersoy, R., Cakir, B., *et al.* Gynecomastia: Clinical evaluation and management. *Indian journal of endocrinology and metabolism* **18,** 150 (2014).

38. Sokocevic, D. *et al.* The effect of age and emphysematous and fibrotic injury on the re-cellularization of de-cellularized lungs. *Biomaterials* **34,** 3256–3269 (2013).

39. Tatone, C. *et al.* Cellular and molecular aspects of ovarian follicle ageing. *Human reproduction update* **14,** 131–142 (2008).

40. ZHANG, K. & Hong, W. Cancer Genome Atlas Pan-cancer Analysis Project. *Chinese Journal of Lung Cancer* **18** (2015).

41. Aziz, S. A. *et al.* Vascularity of primary and metastatic renal cell carcinoma specimens. *Journal of translational medicine* **11,** 15 (2013).

42. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22,** 1760–1774 (2012).

43. Garrido-Martín, D., Palumbo, E., Guigó, R. & Breschi, A. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLoS computational biology* **14,** e1006360 (2018).

44. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution* (2017).

45. Breschi, A. *et al.* Gene-specific patterns of expression variation across organs and species. *Genome Biology* **17,** 151 (2016).

Acknowledgements

We thank Kristin Ardlie and Detlev Arendt for useful discussions. 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 and Ministerio de Educación, Cultura y Deporte, under the FPU programme (Formación de Profesorado Universitario) with pre-doctoral fellowship FPU15/03635, 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. 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. We acknowledge and thank the donors and their families

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Author Information

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. analysed 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.

Competing interests

The authors declare no competing financial interests.

Figure legends

Figure 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. 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.

Figure 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 (see Figure). 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

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(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 harbour 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 harbour 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 S5 for a complete description of the acronyms. Enrichment adjusted p-values: "*" < 0.05, "**" < 0.01, "***" < 0.001.

Figure 3

Transcriptional complexity of human primary cells. (**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. S13. (**c**) Expression signal for the *S100A16* gene, which shows the preferential usage of the proximal TSS in endothelial cells, compared to preferential use of the distal TSS in cells from the other types. The individual signal tracks are shown for endothelial cells, whereas overlaid tracks are shown for the other major cell types. The signal is scaled to the maximum of the track height to show the relative difference in TSS usage. (**d**) Sashimi plot depicting differential inclusion of exon 6 of the MYL6 gene. The exon is more included in mesenchymal cells, that comprise muscle cells, compared to the other major cell types. The signal is the average read count for each major cell type (y axis). The average number of reads supporting each splice junctions is reported for each splice junction within each major cell type. The bottom panel show the exonic structure of annotated transcripts in GENCODE[42]. The Sashimi plot was generated using *ggsashimi*[43].

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4

Evolutionary conservation of cell type specific genes. (**A**) 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. S15. (**b**) 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[44]. The black line is given as a reference and it indicates the proportion of 6-way orthologs (chimpanzee, rhesus, mouse, opossum, platypus and chicken) from[45] that are present in each species. The proportion is not 100% in these species because different versions of the GENCODE[42] 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. S15b. (**c**) 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 Figs. S16 and S17.

Figure 5

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.

Figure 6

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

bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

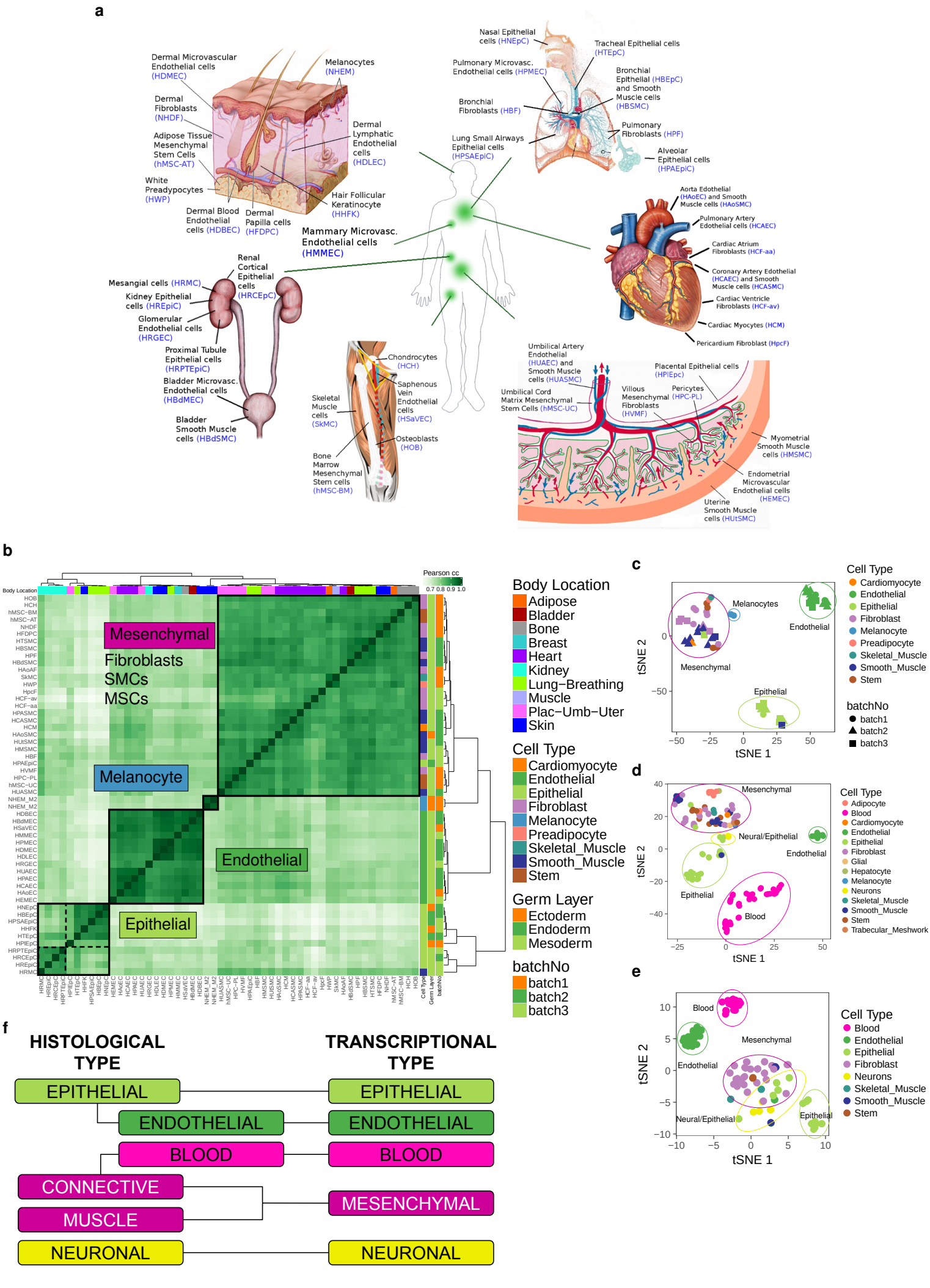
p-values as FDR) between affected and normal states. Histological images of affected and normal tissues are displayed (see text for details). Nrml: normal, Athr: atherosclerosis and Atrp: atrophy. **d)** Major cell type xCell enrichments in female (Fml) breast samples, and male breast samples with (MlGy) 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 correlation coefficient, adjusted p-values as FDR).

Figure 7

Alterations of the contributions of the major cell types to tissues in cancer. xCell enrichments in epithelial cells in lung cancers and matched normal controls from the PCAWG project separated by cancer project (**a**). LUAD-US: Lung Adenocarcinoma, TCGA, USA; LUSC-US: Lung Squamous Cell Carcinoma, TCGA, USA. Enrichment in matched normal and cancer lung samples by donor, pooled across the cancer projects (**b**). 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. xCell enrichment in endothelial cells in kidney cancers and matched normal controls from the PCAWG project separated by cancer project (**c**). 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. xCell Enrichments in matched normal and cancer kidney samples by donor (**d**). 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. xCell enrichments in epithelial cells in ovarian cancers from the PCAWG project separated by cancer project (**e**) or by donor for matched primary and recurrent samples (**f**). 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.

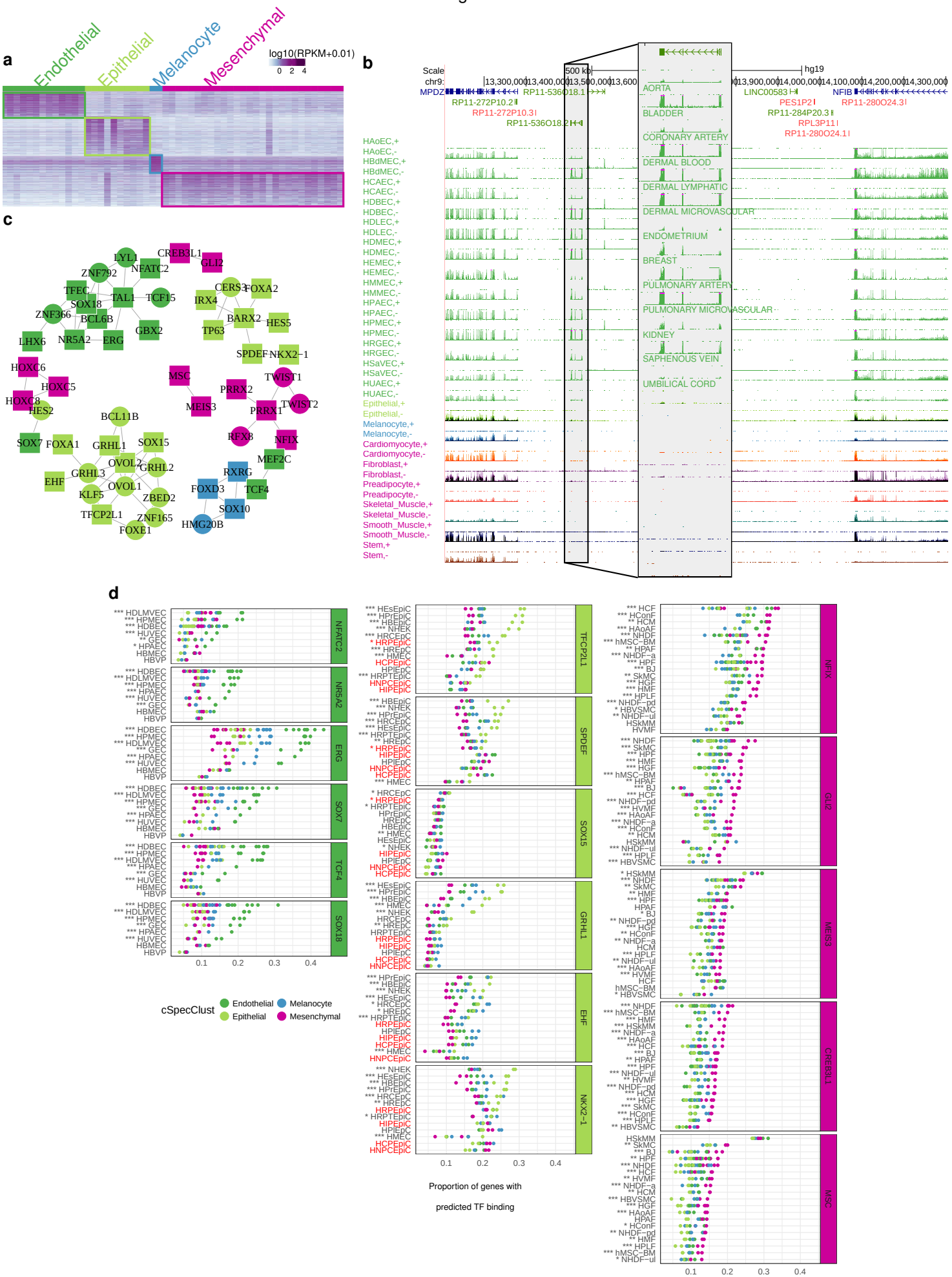
bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1



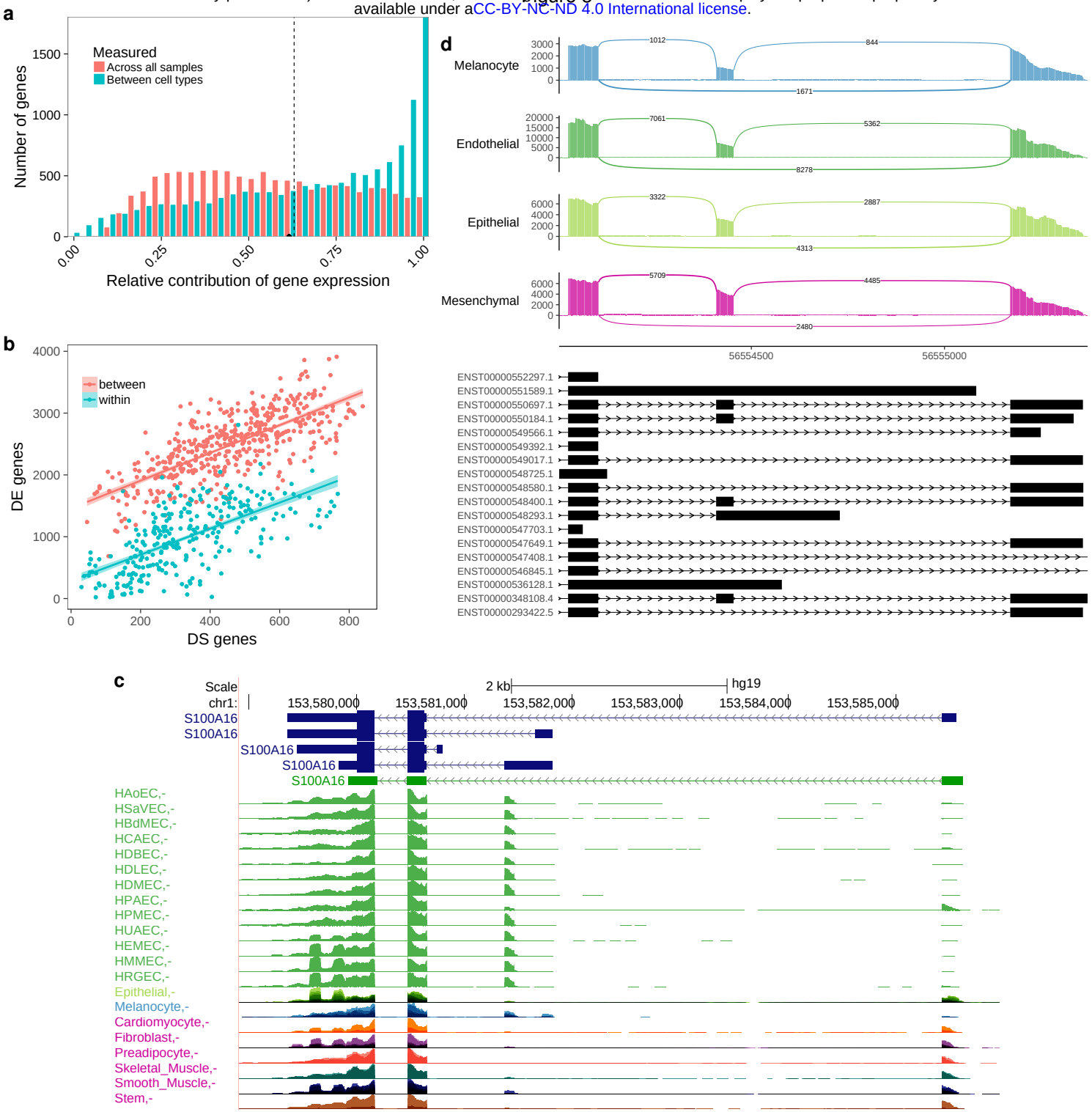
bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2



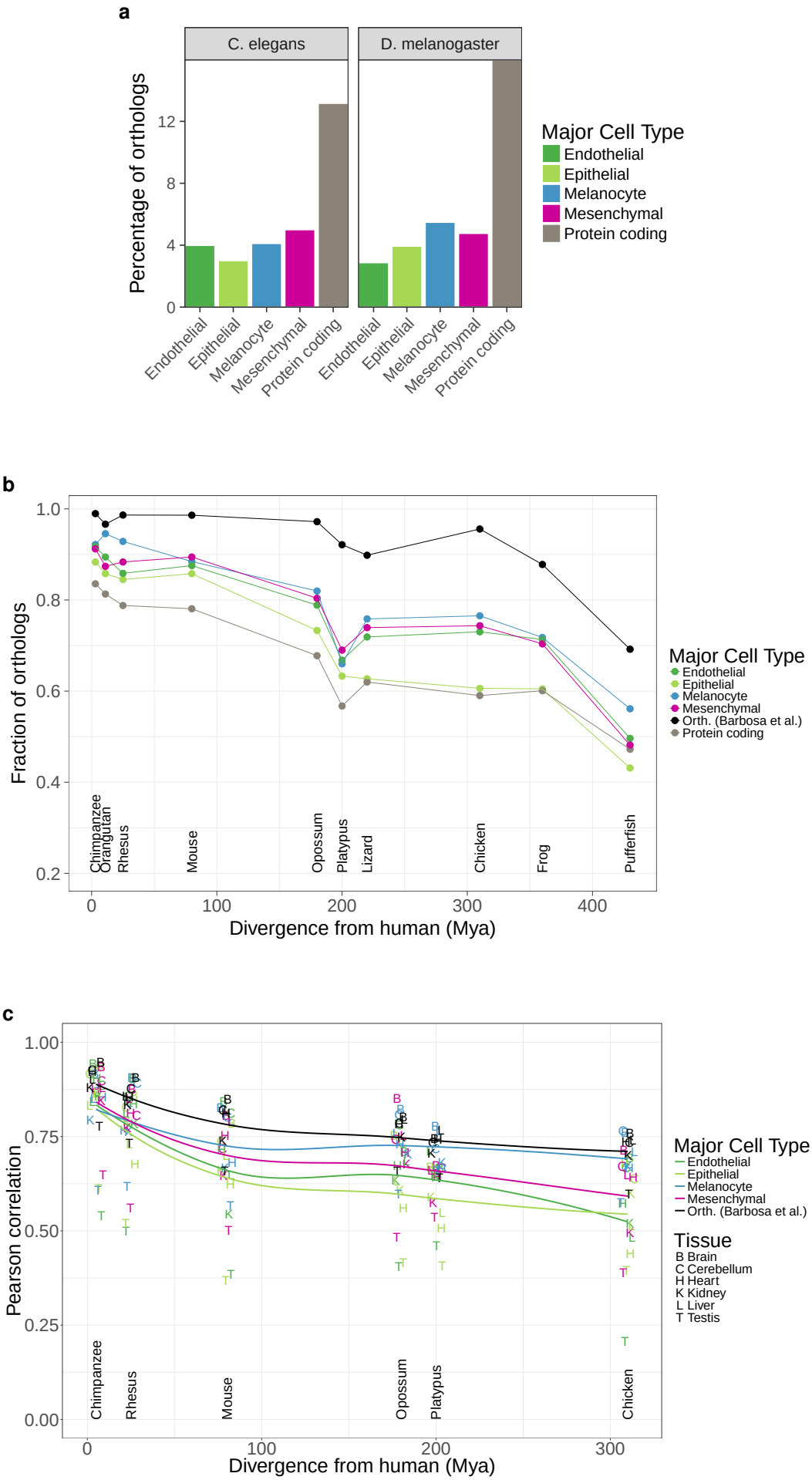
bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3



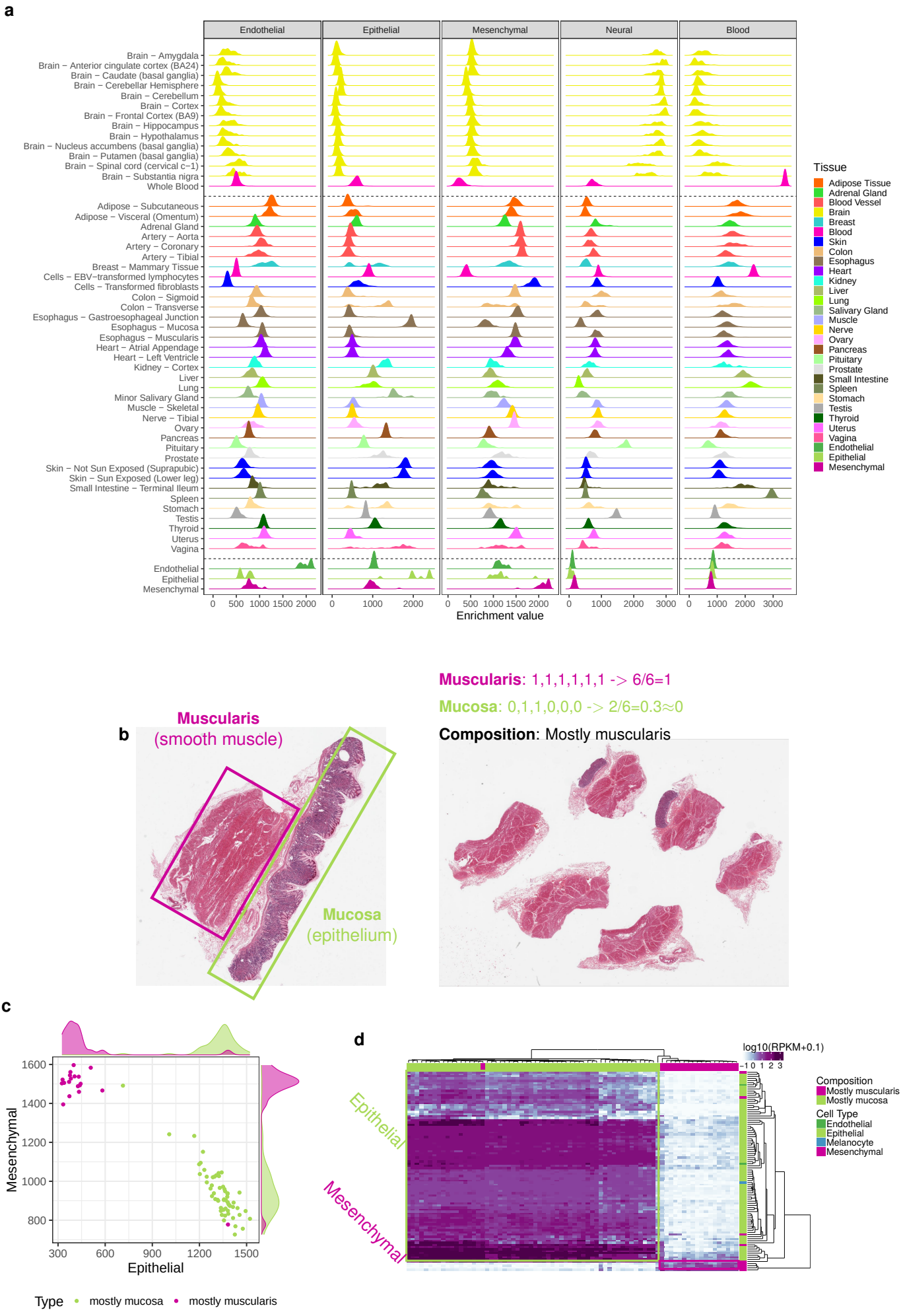
bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4



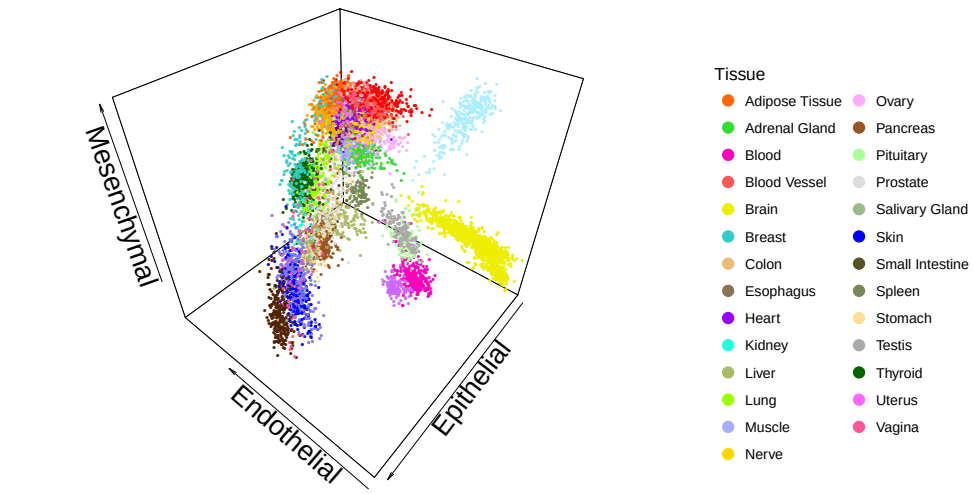
bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

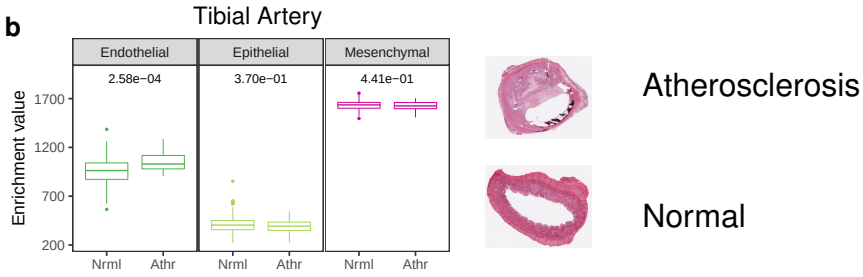
Figure 6

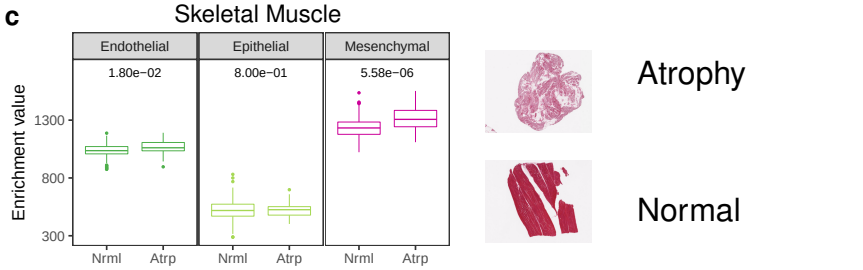


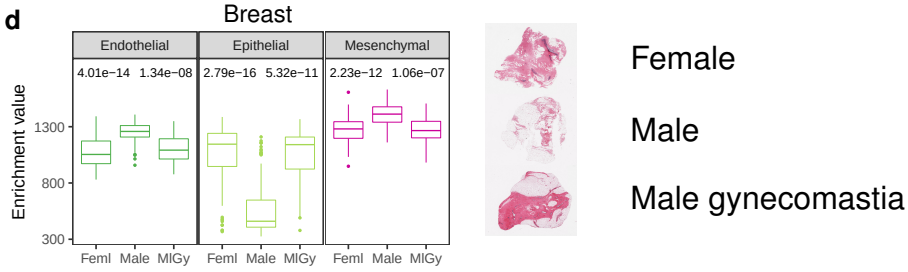
bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

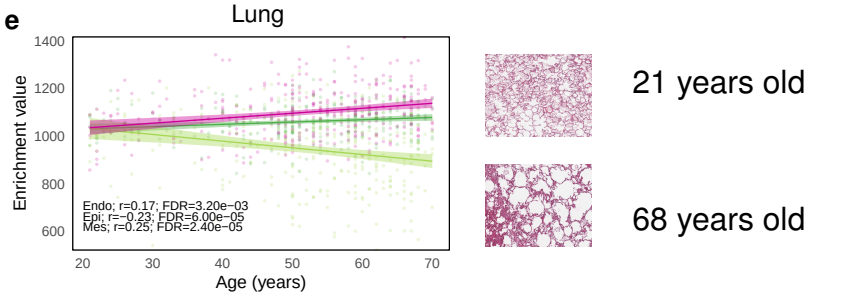
Figure 6





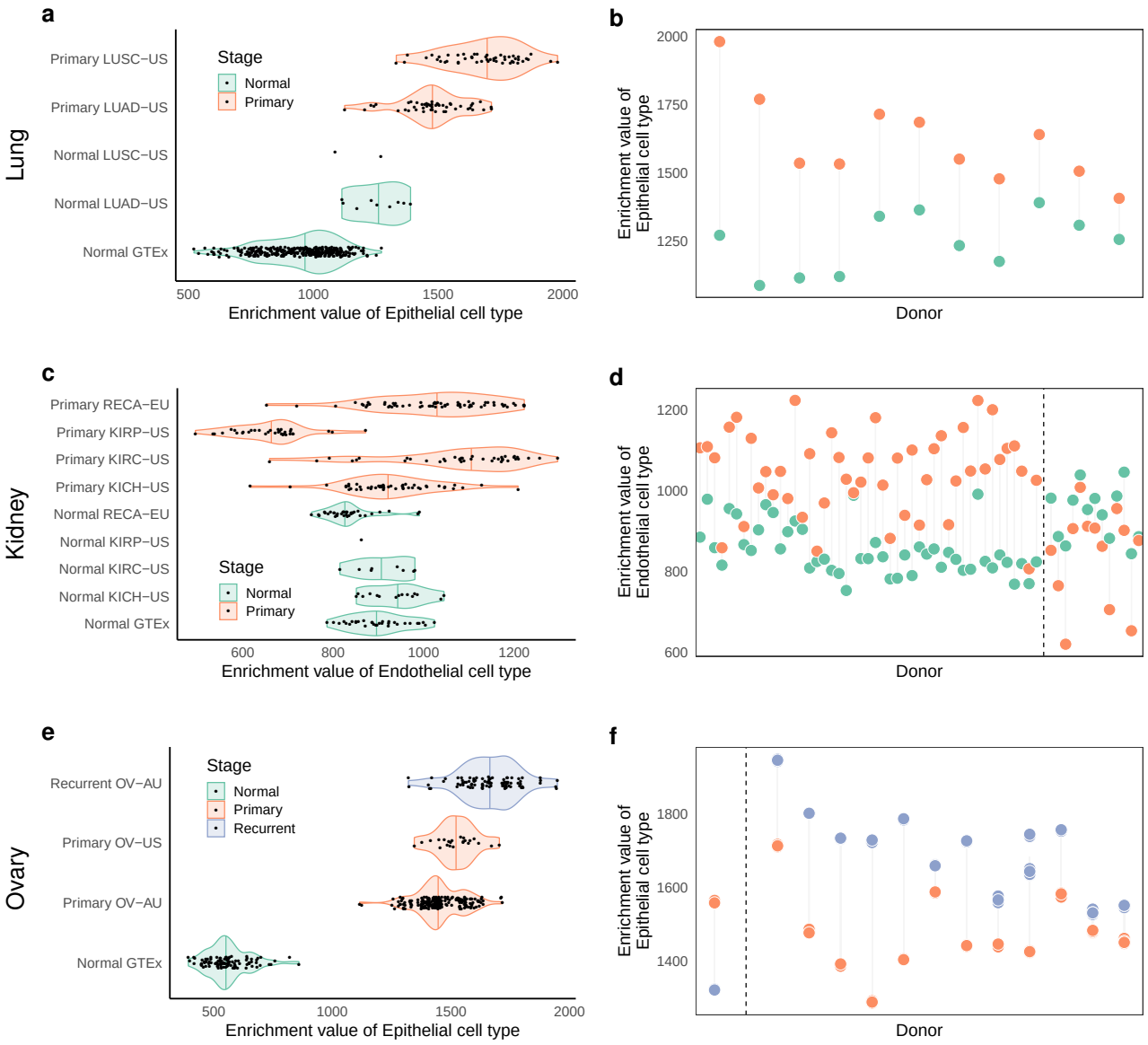






bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 7



bioRxiv preprint doi: https://doi.org/10.1101/857169; this version posted November 27, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.