# Cell Types of Origin in the Cell Free Transcriptome in Human Health and Disease

Sevahn K. Vorperian[1,2], Mira N. Moufarrej[3], Tabula Sapiens Consortium[4], Stephen R. Quake[3,5,6*]
[1]Department of Chemical Engineering, Stanford University, Stanford CA, USA
[2]ChEM-H, Stanford University, Stanford CA, USA
[3]Department of Bioengineering, Stanford University, Stanford CA, USA
[4]Members listed at the end
[5]Department of Applied Physics, Stanford University, Stanford CA, USA
[6]Chan-Zuckerberg Biohub, Stanford CA, USA
*steve@quake-lab.org

**Abstract:**
Cell-free RNA (cfRNA) can be used to noninvasively measure dynamic and longitudinal physiological changes throughout the body. While there is considerable effort in the liquid biopsy field to determine disease tissue-of-origin, pathophysiology occurs at the cellular level. Here, we describe two approaches to identify cell type contributions to cfRNA. First we used *Tabula Sapiens*, a transcriptomic cell atlas of the human body to computationally deconvolve the cell-free transcriptome into a sum of cell type specific transcriptomes, thus revealing the spectrum of cell types readily detectable in the blood. Second, we used individual tissue transcriptomic cell atlases in combination with the Human Protein Atlas RNA consensus dataset to create cell type signature scores which can be used to infer the implicated cell types from cfRNA for a variety of diseases. Taken together, these results demonstrate that cfRNA reflects cellular contributions in health and disease from diverse cell types, potentially enabling determination of pathophysiological changes of many cell types from a single blood test.

**Main Text**

Cell-free RNA (cfRNA) represents a mixture of transcripts reflecting the health status of multiple tissues, thereby affording broad clinical utility. Existing applications span oncology and bone marrow transplantation[12], obstetrics[3,45], neurodegeneration[6], and liver disease[7]. However, several aspects about the physiologic origins of cfRNA including the contributing cell types-of-origin remain unknown, and current assays focus on tissue level contributions at-best[2,4,6–8]. Incorporating knowledge from cellular pathophysiology, which often forms the basis of disease[9], into a liquid biopsy would more closely match the resolution afforded by invasive procedures.

We first characterized the landscape of cell type specific signal from healthy donor plasma (Fig. 1A) using published exome-enriched cf-transcriptome data[6] (Fig. 1A). After removing low-quality samples (Fig. S1, Methods), we intersected the set of genes detected in healthy individuals (n = 75) with a database of cell-type specific markers defined in context of the whole body[10]. Marker genes for blood, brain, and liver cell types were readily detected, as previously observed at tissue level[1,4,6,7 2]as well as the kidney, gastrointestinal track, and pancreas (Fig. 1B).

We then sought to deconvolve the fractions of cell-type specific RNA using support vector regression, a deconvolution method previously applied to decompose bulk tissue transcriptomes into fractional cell type contributions[11,12]. We used *Tabula Sapiens* v1.0 (TSP)[13], a multiple donor whole body cell atlas spanning 24 tissues and organs, to define a basis matrix whose gene set accurately and simultaneously resolved the distinct cell types in TSP. The basis matrix was defined using the gene space that maximized linear independence of the cell types; and does not include the whole transcriptome but

45 rather the minimum discriminatory gene set to distinguish between the cell types in Tabula Sapiens. To
46 reduce multicollinearity, transcriptionally similar cell types were grouped (fig. S2). We observed that the
47 basis matrix defined by this gene set appropriately described cell types as most similar to others from the
48 same organ compartment-and correspond to the highest off-diagonal similarity (Fig. 1C). We also
49 confirmed that the basis matrix accurately deconvolved cell type specific RNA fractional contributions
50 from several bulk tissue samples (fig. S3, S4, Supplementary Note 1).
51 We used this matrix to deconvolve the cell types-of-origin in the plasma cf-transcriptome (Fig.
52 1D; fig. S5, S6). Platelets, erythrocyte/erythroid progenitors, and leukocytes comprised the majority of
53 observed signal, whose respective proportions were generally consistent with recent estimates from serum
54 cfRNA[1] and plasma cfDNA[14]. Within this set of cell types, we suspect that the observation of platelets as
55 a majority cell type, rather than megakaryocytes[1], likely reflects annotation differences in reference data.
56 We observed distinct transcriptional contributions from solid tissue-specific cell types from the intestine,
57 liver, lungs, pancreas, heart, and kidney (Fig. 1D, fig. S5). Altogether, the observation of contributions
58 from many non-hematopoietic cell types underscores the ability to simultaneously noninvasively resolve
59 contributions to cfRNA from disparate cell types across the body.
60 Some cell types likely present in the plasma cf-transcriptome were missing in this decomposition
61 because the source tissues were not represented in TSP. Though ideally reference gene profiles for all cell
62 types would be simultaneously considered in this decomposition, a complete reference dataset spanning
63 the entire cell type space of the human body does not yet exist. To identify cell type contributions
64 possibly absent from this analysis, we intersected the genes measured in cfRNA absent from the basis
65 matrix with tissue specific genes from the human protein atlas RNA consensus dataset (HPA)[15]. This
66 identified both the brain and testes as tissues whose cell types were not found during systems-level
67 deconvolution and additional genes specific to the liver, skeletal muscle, and lymphoid tissues that were
68 not used by the basis matrix (Fig. 1E, Methods).
69 As an example of how to analyze cell type contributions from tissues which were not present in
70 TSP, we utilized an independent brain single cell atlas along with HPA to define cell type gene profiles
71 and examined their expression in cfRNA (Fig 2A, fig. S7-9). There was a strong signature score from
72 excitatory neurons and a reduced signature score from inhibitory neurons. We observed strong signals
73 from astrocytes, oligodendrocytes, and oligodendrocyte precursor cells. These glial cells facilitate brain
74 homeostasis, form myelin, and provide neuronal structure and support[9], consistent with evidence of RNA
75 transport across and the permeability of the blood brain barrier[16,17] and that some brain regions are in
76 direct contact with the blood [18]. Similarly, we utilized published cell atlases for the placenta[19,20], kidney[21],
77 and liver[22] to define cell type specific gene profiles (fig. S8, S10) from which to perform signature
78 scoring. These observations augment the resolution of previously observed tissue-specific genes reported
79 to date in cfRNA[4,6,7 21]and formed a baseline from which to measure aberrations in disease.
80 Cell type specific changes drive disease etiology[9], and we asked whether cfRNA reflected cellular
81 pathophysiology. We considered trophoblasts in preeclampsia[23,24], proximal tubules in chronic kidney
82 disease (CKD)[25,26], hepatocytes in non-alcoholic steatohepatitis (NASH)/non-alcoholic fatty liver disease
83 (NAFLD)[27], and multiple brain cell types in Alzheimer's Disease (AD)[28,29]. As an example of why whole
84 body cell type characterization is important, we observed that a prior attempt to infer trophoblast cell
85 types from cfRNA in preeclampsia[24] used genes that are not specific or readily measurable within their
86 asserted cell type (fig. S11, Supplementary Note 2). However, we found several other cases where
87 cellular pathophysiology can be measured in cfRNA.

88 The proximal tubule is a highly metabolic, predominant kidney cell type and is a major source for
89 injury and disease progression in CKD[25,26]. Tubular atrophy is a hallmark of CKD nearly independent of
90 disease etiology[30] and is superior to clinical gold standard as a predictor of CKD progression[31]. Using data
91 from Ibarra et al, we discovered a striking decrease in the proximal tubule cell signature score of CKD
92 patients (ages 67-91, CKD stage 3-5 or peritoneal dialysis) compared to healthy controls (Fig. 2B, fjg.
93 S12). These results demonstrate noninvasive resolution of proximal tubule deterioration observed in CKD
94 histology[31] and are consistent with findings from invasive biopsy.
95 Hepatocyte steatosis is a histologic hallmark of NASH and NAFLD phenotypes, whereby the
96 accumulation of cellular stressors results in hepatocyte death[27]. We found that several genes differentially
97 expressed in NAFLD serum cfRNA[7] were specific to the hepatocyte cell type profile derived above. ($P <$
98 $10^{-10}$, hypergeometric test). Notable hepatocyte specific DEG include cytochrome P450 enzymes
99 (including *CYP1A2, CYP2E1*, *CYP3A4*), lipid secretion (*MTTP)*, and hepatokine-coding genes (*AHSG,*
100 *LECT2)*[32]. We further observed striking differences in the hepatocyte signature score between healthy and
101 both NAFLD and NASH (Fig. 2C, fig. S13), and no difference between the NASH and NAFLD cohorts.
102 AD pathogenesis results in neuronal death and synaptic loss[29]. We used brain single cell data[28] to
103 define brain cell type gene profiles in both the AD and non-cognitively impaired (NCI) brain. Many of the
104 differentially expressed genes found in cfRNA analysis of AD plasma are brain cell type specific ($P < 10^{-}$
105 $^{5}$, hypergeometric test). Astrocyte specific genes included filament protein (*GFAP*[33]) and ion channels
106 (*GRIN2C* [28]). Excitatory neuron specific genes include solute carrier proteins (*SLC17A7*[28] & *SLC8A2*[34] ),
107 cadherin proteins (*CDH8*[35] & *CDH22*[36]), and a glutamate receptor (*GRM1*[29,37]). Oligodendrocyte-specific
108 genes encode proteins for myelin sheath stabilization (*MOBP*[29]) and a synaptic/axonal membrane protein
109 (*CNTN2*[29]*)*. Oligodendrocyte precursor cell-specific genes included transcription factors (*OLIG2*[38] &
110 *MYT1*[39]), neural growth and differentiation factor (*CSPG5*[40]), and a protein putatively involved in brain
111 extracellular matrix formation (*BCAN*[41]).
112 We then inferred neuronal death between AD and NCI and also observed differences in
113 oligodendrocyte, oligodendrocyte progenitor, and astrocyte signature scores (Fig. 2D, fig. S14). The
114 oligodendrocyte and oligodendrocyte progenitor cells signature score directionality agrees with reports of
115 their death and inhibited proliferation in AD respectively[42]. The observed astrocyte signature score
116 directionality is consistent with the cell type specificity of a subset of reported downregulated DEG[6] and
117 reflects that astrocyte-specific changes, which are known to in AD pathology[42], are noninvasively
118 measurable.
119 Taken together, this work demonstrates consistent noninvasive detection of cell-type specific
120 changes in human health and disease using cfRNA. Our findings uphold and further augment the scope of
121 prior work identifying immune cell types[1] and hematopoietic tissues[1,4] as primary contributors to the cf-
122 transcriptome cell type landscape. Our approach is complementary to prior work using cell free
123 nucleosomes[14] which depends on a more limited set of reference ChIP-seq data which is largely at the
124 tissue level[43]. Readily measurable cell types include those specific to the brain, lung, intestine, liver, and
125 kidney, whose pathophysiology affords broad prognostic and clinical importance. Consistent detection of
126 cell types responsible for drug metabolism (e.g. liver and renal cell types) as well as cell types that are
127 drug targets, such as neurons or oligodendrocytes for Alzheimer's-protective drugs, could provide strong
128 clinical trial end-point data when evaluating drug toxicity and efficacy. We anticipate that the ability to
129 noninvasively resolve cell type signatures in plasma cfRNA will both enhance existing clinical
130 knowledge and enable increased resolution in monitoring disease progression and drug response.

**Acknowledgements:** We thank M. Chen for single cell analysis input, feedback and helpful discussions. We thank E. Sattely and G.E. Marti for helpful discussions. We thank G. Loeb for kidney discussions. The human body in figure 1A was created using Biorender.com.

**Funding:** This work is supported by the Chan Zuckerberg Biohub. S.K.V. is supported by a NSF Graduate Research Fellowship (Grant # DGE 1656518), the Benchmark Stanford Graduate Fellowship, and the Stanford ChEM-H Chemistry Biology Interface (CBI) training program.

**Author contributions:** S.K.V. and S.R.Q. conceptualized the study. S.K.V. and S.R.Q. designed the study in collaboration with M.N.M. S.K.V. performed all analyses; M.N.M. wrote the bioinformatic preprocessing pipeline to map reads to the human genome and cell free sample QC. S.K.V, M.N.M, S.R.Q wrote the manuscript. All authors revised the manuscript and approved it for publication.

**Competing interests:** S.R.Q is a founder and shareholder of Molecular Stethoscope and Mirvie. M.N.M. is also a shareholder of Mirvie. S.K.V, M.N.M, and S.R.Q are inventors on a patent application covering the methods and compositions to detect specific cell types using cfRNA submitted by the Chan Zuckerberg Biohub and Stanford University.

**Data availability:** All datasets used for this work were publicly available, downloaded with permission, or directly requested from authors. Tissue gene lists and NX counts were downloaded from HPA (www.proteinatlas.org, v19). GTEx raw expression was downloaded from the GTEx portal (https://www.gtexportal.org/home/datasets, GTEx analysis V8). Tabula Sapiens single cell data were received from the CZ-Biohub (https://tabula-sapiens-portal.ds.czbiohub.org, version 1.0). The brain single cell data were downloaded with permission from Synapse (https://www.synapse.org/#!Synapse:syn18485175) and associated ROSMAP metadata were downloaded with permission from Synapse (https://www.synapse.org/#!Synapse:syn3157322). The liver Seurat object was requested from Aizarani et al. For the placenta atlases, a Seurat object was requested from Suryawanshi et al and AnnData requested from Vento-Tormo et al. Kidney AnnData was downloaded (https://www.kidneycellatlas.org, Mature Full dataset).

**Code availability:** Code for the work in this manuscript will be available on Github at www.github.com/sevahn/deconvolution.

## Materials & Methods

### *Data Processing*

**Data acquisition**

Cell free RNA: For samples from Ibarra et al, raw sequencing data was obtained from the SRA (PRJNA517339). For samples from Munchel et al, processed counts tables were directly downloaded.

For all individual tissue single cell atlases, Seurat objects or AnnData objects were downloaded or directly received from authors. Data from Mathys et al. were downloaded with permission from Synapse.

HPA v19 transcriptomic data, GTEx v8 raw counts, and Tabula Sapiens v1.0 were downloaded directly.

**Bioinformatic processing**

For each sample for which raw sequencing data were downloaded, we trimmed reads using trimmomatic (v 0.36) and then mapped them to the human reference genome (hg38) with STAR (v 2.7.3a). Duplicate reads were then marked and removed by GATK's (v 4.1.1) MarkDuplicates tool. Finally, mapped reads were quantified using htseq-count (v 0.11.1), and read statistics were estimated using FastQC (v 0.11.8).

The bioinformatic pipeline was managed using snakemake (v 5.8.1). Read and tool performance statistics were aggregated using MultiQC (v 1.7).

**Sample quality filtering**

For every sample for which raw sequencing data was available, we estimated three quality parameters as previously described[44,45]: RNA degradation, ribosomal read fraction, and DNA contamination.
RNA degredation was estimated by calculating a 3' bias ratio. Specifically, we first counted the number of reads per exon and then annotated each exon with its corresponding gene ID and exon number using htseq-count. Using these annotations, we measured the frequency of genes for which all reads mapped exclusively to the 3' most exon as compared to the total number of genes detected. We approximate RNA degradation for a given sample as the fraction of genes where all reads mapped to the 3' most exon.
To estimate ribosomal read fraction, we compared the number of reads that mapped to the ribosome (Region GL00220.1:105424-118780, hg38) relative to the total number of reads (Samtools view).
To estimate DNA contamination, we used an intron to exon ratio and quantified the number of reads that mapped to intronic as compared to exonic regions of the genome.

We applied the following thresholds as previously reported[44]:

- Ribosomal: > 0.2
- 3' Degradation: > 0.4
- Intron/Exon: > 3

We considered any given sample as low quality if its value for any metric was greater than any of these thresholds and excluded the sample from subsequent analysis.

**Data Normalization**

All gene counts were adjusted to counts per million reads and per milliliter of plasma used. For a given sample ($i$ denotes gene index and $j$ denotes sample index):

$$\eta_{ij} = \frac{Gene_{ij}}{(Library\ Size_j)*(mL\ Plasma_j)} \quad \text{where } Library\ Size_j = \sum_i G_{ij} \qquad (1)$$

For subjects who had samples with multiple technical replicates, these plasma volume CPM counts were averaged prior to nu-SVR deconvolution.

For all analyses except nu-SVR (e.g. all work except Fig. 1D and 1E), we next applied trimmed mean of M values (TMM) normalization as previously described[46]:

$$\frac{\eta_{ij}}{TMM_j} \qquad (2)$$

CPM TMM normalized gene counts across technical replicates for a given biological replicate were averaged for the count tables used in all analyses performed.

Sequencing batches and plasma volumes were obtained from the authors in Toden et al and Chalasani et al for per-sample normalization. For samples from Ibarra et al., plasma volume was assumed to be constant at 1 mL as we were unable to attain this information from the authors; sequencing batches were confirmed with authors (personal communication). All samples from Munchel et al were used to compute TMM scaling factors and 4.5 mL plasma[5] was used to normalize all samples within a given dataset (both PEARL-PEC and iPEC).

***Cell Type Marker Identification using PanglaoDB***

The PanglaoDB cell type marker database was downloaded on March 27, 2020. Markers were filtered for human ("Hs") only and for PanglaoDB's defined specificity (how often marker was not expressed in a given cell type) and sensitivity (how frequently marker is expressed in cells of this type). Gene synonyms from Panglao were determined using MyGene version 3.1.0 to ensure full gene space.

We then intersected this gene space with a cohort of healthy cfRNA samples (n = 74, non-cognitively impaired (NCI) individuals from Toden et al). A given cell type marker was counted in a given healthy cfRNA sample if its gene expression was greater than zero in log + 1 transformed CPM-TMM gene count space.

Cell types with markers filtered by sensitivity = 0.9 and specificity = 0.2 and samples with ≥ 5 cell type markers are shown in Fig 1B.

***Basis Matrix Formation***

Scanpy[47] (version 1.6.0) was used. Only cells from droplet sequencing ("10X") were used in analysis given that a more comprehensive set of unique cell types across the tissues in *Tabula Sapiens* were available[13]. Disassociation genes as reported[13] were eliminated from the gene space prior to subsequent analysis.

Given the non-specificity of the following annotations (e.g. other cell type annotations at finer resolution existed), cells with these annotations were excluded from subsequent analysis:

- 'epithelial cell'
- 'ocular surface cell'
- 'radial glial cell'
- 'lacrimal gland functional unit cell'
- 'connective tissue cell'
- 'corneal keratocyte'
- 'ciliary body'
- 'bronchial smooth muscle cell'
- 'fast muscle cell'
- 'muscle cell'
- 'myometrial cell'
- 'skeletal muscle satellite stem cell'
- 'slow muscle cell'
- 'tongue muscle cell'
- 'vascular associated smooth muscle cell'
- 'alveolar fibroblast'
- 'fibroblast of breast'
- 'fibroblast of cardiac tissue'
- 'myofibroblast cell'

All additional cells belonging to the ‘Eye’ tissue were excluded from subsequent analysis given discrepancies in compartment and cell type annotations and the unlikelihood of detecting eye-specific cell types. The resulting cell type space still possessed several transcriptionally similar cell types (e.g. various intestinal enterocytes, t cells, or dendritic cells), that left unaddressed, would reduce the linear independence of the basis matrix column space and hence would impact nu-SVR deconvolution.

Cells were therefore assigned broader annotations on a per-compartment basis as follows:
Epithelial, Stromal, Endothelial: using counts from the ‘decontXcounts’ layer of the adata object, cells were CPM normalized (sc.pp.normalize_total(target_sum = 1e6) and log-transformed (sc.pp.log1p). Hierarchical clustering with complete linkage (sc.tl.dendrogram) was performed per compartment on the feature space comprising the first 50 principal components (sc.pp.pca). Epithelial and stromal compartment dendrograms were then cut (scipy.cluster.hierarchy.cut_tree) at 20% and 10% of the height of the highest node respectively, such that cell types with high transcriptional similarity were grouped together but overall granularity of the cell type labels was preserved. This work is available in the script “cuttree.ipynb” on Github.

The endothelial compartment dendrogram revealed high transcriptional similarity across all cell types (maximum node height = 0.851) compared to epithelial (maximum node height = 3.78) and stromal (maximum node height = 2.34) compartments (fig. S2). To this end, only the ‘endothelial cell’ annotation was utilized for the ‘endothelial’ compartment.

Immune: given the high transcriptional similarity and the varying degree of annotation granularity across tissues and cell types, cell types were grouped on the basis of annotation. The following immune annotations were kept:

- 'b cell'
- 'basophil'
- 'erythrocyte'
- 'erythroid progenitor'
- 'hematopoietic stem cell'
- 'innate lymphoid cell'
- 'macrophage'
- 'mast cell'
- 'mature conventional dendritic cell'
- 'microglial cell'
- 'monocyte'
- 'myeloid progenitor'
- 'neutrophil'
- 'nk cell'
- 'plasma cell'
- 'plasmablast'
- 'platelet'
- 't cell'
- 'thymocyte'

All other immune compartment cell type annotations were excluded for being too broad when more detailed annotations existed (i.e. ‘granulocyte’, ‘leucocyte’, ‘immune cell’) or present in only one tissue (i.e. ‘erythroid lineage cell’; eye, ‘myeloid cell’; pancreas/prostate). The ‘erythrocyte’ and ‘erythroid progenitor’ annotations were further grouped to minimize multicollinearity.

Using the entire cell type space spanning all four organ compartments, either 30 observations (e.g. measured cells) were randomly sampled or the maximum number of available observations if less than 30 were subsampled, whichever was greater.

Cell type annotations were then reassigned based on the “broader” categories from hierarchical clustering (‘coarsegrain.py’). Raw count values from the Decont-X adjusted layer were used to minimize signal spread contamination impacting differentially expressed genes[13].

This subsampled counts matrix was then passed to the 'Create Signature Matrix' analysis module at cibersortx.stanford.edu, with the following parameters:

- Disable quantile normalization = True
- Min. expression = 0.25
- Replicates = 5
- Sampling = 0.5
- Kappa = 999
- q-value = 0.01
- No. barcode genes = 3000 - 5000
- Filter non-hematopoietic genes = False

The resulting basis matrix was used in our nu-SVR deconvolution code, available on Github, under the name "tsp_v1_basisMatrix.txt".

Abbreviations (left) of grouped cell types (right) in Figs. 1D as follows:

- gland cell: "acinar cell of salivary gland/myoepithelial cell"
- respiratory ciliated cell: "ciliated cell/lung ciliated cell"
- prostate epithelia: "club cell of prostate epithelium/hillock cell of prostate epithelium/hillock-club cell of prostate epithelium"
- salivary/bronchial secretory cell: "duct epithelial cell/serous cell of epithelium of bronchus"
- intestinal enterocyte: "enterocyte of epithelium of large intestine/enterocyte of epithelium of small intestine/intestinal crypt stem cell of large intestine/large intestine goblet cell/mature enterocyte/paneth cell of epithelium of large intestine/small intestine goblet cell"
- erythrocyte/erythroid progenitor: "erythrocyte/erythroid progenitor"
- fibroblast/mesenchymal stem cell: "fibroblast/mesenchymal stem cell"
- intestinal secretory cell: "intestinal enteroendocrine cell/paneth cell of epithelium of small intestine/transit amplifying cell of small intestine"
- ionocyte/luminal epithelial cell of mammary gland: "ionocyte/luminal epithelial cell of mammary gland"
- secretory cell: "mucus secreting cell/secretory cell/tracheal goblet cell"
- pancreatic alpha/beta cell: "pancreatic alpha cell/pancreatic beta cell"
- respiratory secretory cell: "respiratory goblet cell/respiratory mucous cell/serous cell of epithelium of trachea"
- basal prostate cell: "basal cell of prostate epithelia"

***Nu-SVR deconvolution***

We formulated the cell free transcriptome as a linear summation of the cell types from which it originates[4,48]. With this formulation, we adapted existing deconvolution methods developed with the objective of decomposing a bulk tissue sample into its single cell constituents[11,12], where the deconvolution problem is formulated as:

$$A\theta = b \quad (4)$$

Here, $A$ is the representative basis matrix (g x c) of **g** genes for **c** cell types, which represent the gene expression profiles of the **c** cell types. $\boldsymbol{\theta}$ is a vector (c x 1) of the contributions of each of the cell types and **b** is the measured expression of the genes observed in from blood plasma (g x 1). The goal here is to learn $\boldsymbol{\theta}$ such that the matrix product $A\theta$ predicts the measured signal **b**. The derivation of the basis matrix $A$ is described in the section 'Basis Matrix Formation'.

393 We performed nu-SVR using a linear kernel to learn $\boldsymbol{\theta}$ from a subset of genes from the signature matrix
394 to best recapitulate the observed signal **b**, where nu denotes the lower bound on the fraction of support
395 vectors and the upper bound on the fraction of errors at the margin[49]. Here, the support vectors are the
396 genes used from the basis matrix from which to learn $\boldsymbol{\theta}$; $\boldsymbol{\theta}$ reflects the weights of the cell types in the
397 basis matrix column space. For each sample, we learned coefficients for six values of nu, $\nu \in$
398 $\{0.05, 0.1, 0.15, 0.25, 0.5, 0.75\}$ and five values of C, $C \in \{0.1, 0.5, 0.75, 1, 10\}$ and estimated the
399 resulting deconvolution error using the root mean square error (RMSE). We determined the product of the
400 basis matrix with the learned coefficients ($A\theta$), which reflected some predicted expression value for each
401 of the genes in a given cfRNA mixture used in the deconvolution. The RMSE was then computed using
402 the predicted expression values and the measured values of these genes.
403
404 Only CPM counts ≥ 1 were considered in the mixture. The values in the basis matrix were also CPM-
405 normalized. Prior to deconvolution, the mixture and basis matrix were scaled to zero mean and unit
406 variance for improved runtime performance. We emphasize that we did not log-transform counts in $b$ or
407 in $A$, as this would destroy the requisite linearity assumption in equation 4. Specifically, the concavity of
408 the log function would result in the consistent underestimation of $\boldsymbol{\theta}$ during deconvolution[50].
409
410 Using the $\boldsymbol{\theta}$ resulting from the value of $\nu$ whose coefficients yielded the smallest RMSE was transformed
411 to. Specifically, the relative fractional contributions of cell type specific RNA from $\boldsymbol{\theta}$, we repeat what was
412 previously described [11,12]:
413 $$\forall \theta_j < 0 \in \{\theta_1, \dots, \theta_c\} \rightarrow 0 \quad (5)$$
414 All non-zero coefficients were then normalized by their sum to result in the relative fractions to determine
415 the relative fractional contributions of cell type specific RNA.
416
417 We used the function nuSVR from scikitlearn[51] version 0.23.2.
418
419 The samples used for nu-SVR deconvolution were 75 non-cognitively impaired (NCI) patients from
420 Toden et al spanning four sample collection centers. Given center-specific batch effects reported by
421 Toden et al, we report our results on a per-center basis (Fig 1D, fig S5 & S6). There was good pairwise
422 similarity of the learned coefficients amongst biological replicates within and across sample centers (fig.
423 S6A, S6B). Deconvolution performance yielded RMSE and pearson r consistent with deconvolved GTEx
424 tissues (fig. S3) whose distinct cell types were in the basis matrix column space (fig. S6C, S6D). In
425 interpreting the resulting cell type fractions, a limitation of nu-SVR is that it uses highly expressed genes
426 as support vectors, and consequently assigns a reduced fractional contribution to cell types expressing
427 genes at lower levels or that are smaller in cell volume. Comparison of nu-SVR to quadratic
428 programming[4] and non-negative linear least squares[52] yielded decreased deconvolution RMSE and
429 comparable Pearson correlation. Furthermore, the cell type contributors identified with nu-SVR were
430 consistent with the cell type markers detected using PanglaoDB in contrast to the other methods.
431
432 ***Evaluating Basis Matrix on GTEx samples***
433 Bulk RNA-seq samples from GTEx v8 were deconvolved with the derived basis matrix from tissues that
434 were present (i.e. kidney cortex, whole blood, lung, spleen) or absent (e.g. kidney medulla and brain)
435 from the basis matrix derived using Tabula Sapiens version 1. For each tissue type, the maximum number
436 available samples or thirty samples, whichever was smaller, was deconvolved. Please see Supplementary
437 Note 1 for additional discussion.
438
439 ***Identifying tissue specific genes in cfRNA absent from basis matrix***
440 To identify cell type specific genes in cfRNA that were distinct to a given tissue, we considered the set
441 difference of the non-zero genes measured in a given cfRNA sample with the row space of the basis
442 matrix and intersected this with HPA tissue specific genes:

$$(G_j - R) \cap HPA \qquad (6)$$

Where $G_j$ is the gene set in the j[th] deconvolved sample, where a given gene in the set's expression was ≥ 5 TMM-CPM. $R$ is the set of genes in the row space of the basis matrix used for nu-SVR deconvolution. $HPA$ denotes the total set of tissue specific genes from HPA.

The HPA tissue specific gene set ($HPA$) were genes across all tissues with Tissue Specificity assignments 'Group Enriched', 'Tissue Enhanced', 'Tissue Enriched' and NX expression ≥ 10. This approach yielded tissues with several distinct genes present in cfRNA which could then be subsequently interrogated using single cell data.

***Derivation of cell type specific gene profiles in context of the whole body using single cell data***

For this analysis, only cell types unique to a given tissue (i.e. hepatocytes unique to the liver, or excitatory neurons unique to the brain) were considered so that bulk transcriptomic data could be used to ensure specificity in context of the whole body. A gene was asserted to be cell type specific if it was (i) differentially expressed within a given single cell tissue atlas (ii) possessed a Gini coefficient ≥ 0.6 and was listed as specific to the native tissue for the cell type of interest, indicating comprehensive tissue specificity in context of the whole body (fig. S8, S10).

**(1) Single cell differential expression**

For data received as a Seurat object, conversion to AnnData was performed by saving as an intermediate loom objects (Seurat version 3.1.5) and converting to AnnData (loompy version 3.0.6). Scanpy (version 1.6.0) was used for all other single cell analysis. Reads per cell were normalized for library size (scanpy normalize_total, target_sum = 1e4), then logged (scanpy log1p). Differential expression was performed using the Wilcoxon rank sum test in Scanpy's filter_rank_genes_groups with the following arguments: min_fold_change = 1.5, min_in_group_fraction = 0.2, max_out_group_fraction = 0.5, corr_method = "benjamini-hochberg". For differentially expressed genes (DEG) with Benjamini Hochberg adjusted p-values < 0.01, the ratio of the highest out_group percent expressed to in-group percent expressed < 0.5 to ensure high specific expression in the cell type of interest within a given cell type atlas.

**(2) Quantifying comprehensive whole-body tissue specificity using the Gini coefficient**

The distribution of all the Gini coefficients and Tau values across all genes belonging to cell type gene profiles for cell types native to a given tissue were compared using the HPA gene expression Tissue Specificity and Tissue Distribution assignments[15] (fig. S9). The Gini coefficient better reflected the underlying distribution of gene expression tissue-specificity than Tau (fig. S9) and was hence used for subsequent analysis. As the Gini coefficient approaches unity, this indicates extreme gene expression inequality, or equivalently high specificity. A single threshold (Gini coefficient ≥ 0.6) was applied across all atlases to facilitate a generalizable framework from which to define tissue specific cell type gene profiles in context of the whole body in a principled fashion for signature scoring in cfRNA.

For the following definitions, $n$ denotes the total number of tissues and $x_i$ is the expression of a given gene in the i[th] tissue.

The Gini coefficient was computed as defined in [53]:

$$\text{Gini} = \frac{n+1}{n} - \frac{2\sum_{i=1}^{n}(n+1-i)x_i}{n\sum_{i=1}^{n}x_i}\text{; } x_i \text{ is ordered from least to greatest.} \quad (7)$$

Tau, as defined in [53]:

$$\tau = \frac{\sum_{i=1}^{n} 1-\bar{x}}{n-1} \text{ where } \bar{x} = \frac{x_i}{\max(x_i)\ \forall\ i \in \{1 \dots n\}} \quad (8)$$

HPA NX Counts from the HPA object entitled 'rna_tissue_consensus.tsv' accessed on July 1, 2019 were used for computing Gini coefficients and Tau.

Note for brain cell type gene profiles: given that there are multiple sub brain-regions in the HPA data, the determined Gini coefficients are lower (e.g. not as close to unity compared to other cell type gene profiles) since there are multiple regions of the brain with high expression, which would result in reduced count inequality.

***Gene Expression in GTEx***

We confirmed the specificity of a given gene profile to its corresponding cell type by comparing the aggregate expression of a given cell type signature in its native tissue compared to that of the average across remaining GTEx tissues (fig. S8C and fig. S10F, 10G). We uniformly observed a median fold change greater than one in the signature score of a cell type gene profile in its native tissue relative to the mean expression in other tissues, confirming high specificity.

Raw GTEx data v8 (accessed August 26 2019) was converted to log(counts-per-ten-thousand + 1) counts. The signature score was determined by summing the expression of the genes in a given bulk RNA sample for a given cell type gene profile. Since only gene profiles were derived for cell types that correspond to a given tissue, the mean signature score of a cell type profile across the non-native tissues was then computed and used to determine the log fold change.

***Cell Type specificity of differentially expressed genes in AD and NAFLD cfRNA***

After observing a significant intersection between the DEG in AD/NCI[6] or NAFLD[7] in cfRNA with corresponding cell type specific genes (fig. S13A, S14A), we then assessed the cell type specificity of DEG using a permutation test. To assess whether DEGs that intersected with a cell type gene profile were more specific to a given cell type than DEGs that were generally tissue specific, we performed a permutation test. Specifically, we compared the Gini coefficient for genes in these two groups, computed using the mean expression of a given gene across brain cell types from healthy brain[28] or liver[22] single cell data. We considered the cell type gene profiles as defined for signature scoring in Fig. 2.

The starting set of tissue specific genes were defined using in the HPA tissue transcriptional data annotated as either 'Tissue enriched', 'Group enriched', or 'Tissue enhanced' (brain, accessed January 13, 2021; liver, accessed November 28, 2020). These requirements ensured the specificity of a given brain/liver gene in context of the whole body. This formed the initial set of tissue specific genes $B$.

The union of all brain cell type specific genes is the set $C$. All genes in $C$ ('cell type specific') were a subset of the respective initial set of tissue specific genes:

$$B - C = 0 \quad (9)$$

Genes in $B$ that that did not intersect with $C$ (e.g. any cell type gene profile ('bcell type specific')) and intersected with DEG-up ($U$) or DEG-down genes ($D$)[6] were then defined as 'tissue specific'.

$$T = (B \cap U) + (B \cap D) \quad (10)$$

The Gini coefficients reflecting the gene expression inequality across the cell types within corresponding tissue single cell atlas were computed for the gene sets labelled as 'cell type specific' and 'tissue specific'. Brain reference data to compute Gini coefficients was the single cell brain atlas with diagnosis as 'Normal'[28]. Liver tissue data was used as-is. All Gini coefficients were computed using the mean log transformed CPTT (counts per ten thousand) gene expression per cell type.

A permutation test was then performed on the union of the Gini coefficients for the genes labeled as 'brain cell type specific' and 'brain tissue specific'. The purpose of this test was to assess probability that the observed mean difference in Gini coefficient for these two groups yielded no difference in specificity (e.g. $H_0$: $\mu_{cell\ type\ Gini\ Coefficient} = \mu_{tissue\ Gini\ coeffcient}$).

Gini coefficients were permuted and reassigned to the list of 'brain tissue' or 'brain cell type' genes, then the difference in mean of the two groups was computed. This procedure was repeated 10,000 times. The p-value was determined as follows:

$$p = \frac{\#\ trials\ with\ permuted\ (\mu_{cell\ type} - \mu_{tissue}) \geq \mu_{observed}}{10{,}000+1} \quad (11)$$

Where $\mu_{observed} \coloneqq (\mu_{cell\ type\ Gini\ Coefficient} - \mu_{brain\ tissue\ Gini\ coeffieint})$.

The additional 1 in the denominator reflects the original test between the true difference in means (e.g. the true comparison yielding $\mu_{observed}$)

NAFLD: We considered the space of reported NAFLD differentially expressed genes in serum[7]. Here $C$ = hepatocyte gene profile and $B$ = the liver specific genes.

AD: First, we intersected a given cell type gene profile in AD with the equivalent NCI profile for comparative analysis. Genes defined as 'brain cell type specific' for signature scoring in Fig. 2D were used in this comparison. Of note, no DEG-up genes intersected with any of the brain cell type signatures in Fig 2D. Microglia, though often implicated in AD pathogenesis were excluded given their high overlapping transcriptional profile with non-central nervous system macrophages[54]. Inhibitory neurons were also excluded given the low number of cell type specific genes intersecting between AD and NCI phenotypes.

***Estimating signature scores for each cell type***

The signature score is defined as the sum of the log-transformed CPM-TMM normalized counts per gene asserted to be cell type specific, where $i$ denotes the index of the gene in a cell type signature gene profile in the $j^{th}$ patient sample:

$$Signature\ Score_j = \sum_i G_{ij} \quad (9)$$

<u>Preeclampsia:</u> for signature scoring of syncytiotrophoblast and extravillous trophoblast gene profiles in PEARL-PEC and iPEC[5], a respective cell type gene profile used for signature scoring was derived as described in 'Derivation of cell type specific gene profiles in context of the whole body using single cell data' independently using two different placental single cell datasets[19,20]. Only the intersection of the cell type specific gene profiles for a given trophoblast cell type between the two datasets was included in the respective trophoblast gene profile for signature scoring.

<u>Chronic Kidney Disease:</u> We compared the signature score of the proximal tubule in CKD (9 patients; 51 samples) and healthy controls (3 patients; 9 samples). Given that all patients samples were longitudinally sampled over ~30 days (e.g. individual samples were taken on different days), we treated the samples as biological replicates and included all time points since the timescale over which renal cell type changes typically occur is longer than the collection period. The sequencing depth was comparable between the CKD and healthy cohorts though it was reduced in comparison to the other cfRNA datasets used in this

591 work. To account for gene measurement dropout, we required that the expression of a given gene in the
592 proximal tubule gene profile was non-zero in at least one sample in both cohorts. Given that all samples
593 were sequenced together, no batch correction was necessary, facilitating a representative comparison
594 between CKD and healthy.
595
596 Alzheimer's Disease:
597 Microglia, though often implicated in AD pathogenesis were excluded given their high overlapping
598 transcriptional profile with non-central nervous system macrophages[54]. Inhibitory neurons were also
599 excluded given the low number of cell type specific genes intersecting between AD and NCI phenotypes.
600 Brain gene profiles as defined in the AD section of '*Cell Type specificity of differentially expressed genes*
601 *in AD and NAFLD cfRNA'* were utilized.
602
603 ***Assessing p-value calibration for a given signature score***
604 Cell type signature scores were tested between control and sick samples with a Mann-Whitney U test. The
605 resulting p-values were calibrated with a permutation test. Here, the labels compared in a given test (i.e.
606 CKD vs. CTRL, or AD vs. NCI, NAFLD vs. CTRL, etc.) were randomly shuffled 10,000 times. We
607 observed a well-calibrated, uniform p-value distribution (fig. S12-S14), validating the experimentally
608 observed test statistics.
609
610 **Supplementary Note 1: Deconvolution of bulk GTEx tissues using the *Tabula Sapiens*-derived basis**
611 **matrix**
612 To assess the ability of the basis matrix to deconvolve tissues whose cell types were wholly present in the
613 cell type column space, we deconvolved a subset of bulk RNA-seq GTEx samples. The determined
614 fractions of cell type specific RNA recapitulated the predominant cell types within a given tissue (fig. S4).
615 Organs with increased cell type heterogeneity (lung, bladder, kidney, intestine, colon) in contrast to
616 tissues with reduced spatial heterogeneity (liver, spleen, whole blood)[13], exhibited greater variance in
617 deconvolved fractions (fig. S4) and deconvolution performance (fig. S3). Tissues with reduced spatial
618 heterogeneity whose cell types were wholly in the basis matrix column space include predominantly b
619 cells/plasma cells and erythrocytes in spleen (fig. S4F); hepatocytes, liver (fig. S4C); erythrocytes and
620 leukocytes, whole blood (fig. S4D). Cell types belonging to tissues with increased spatial heterogeneity
621 exhibited greater variance in deconvolved fractions: kidney cortex majority fractions were from kidney
622 epithelia and lymphocytes (fig. 4G); small intestine, intestinal enterocytes and lymphocytes (fig. S4I);
623 lung, pneumonocytes and immune cells (fig. S4J), colon, intestinal enterocytes, lymphocytes, and muscle
624 cells (fig. S4E). Cells with larger volume yielded larger deconvolved fractions across all tissues (fig. S4).
625 Variance in the relative cell type fractional contributions across the deconvolved bulk samples within a
626 given tissue reflects the underlying cell type heterogeneity, particularly in these complex samples. GTEx
627 kidney medulla samples recorded to be contaminated with renal cortex reflect the presence of the kidney
628 epithelia, the majority cell type in the renal cortex (fig. S4G). Given that the kidney medulla is not part of
629 TSP v1.0, we did not expect high deconvolution performance since its cell types are absent from the basis
630 matrix column space. The brain, whose cell types were wholly absent from the cell type column space
631 exhibited poor deconvolution performance, as expected (fig. S3, S4B). However, the majority cell type
632 fraction assigned was to the cell type belonging to the peripheral nervous system, the schwann cell,
633 underscoring the ability of our deconvolution method to assign fractional contributions to similar cell
634 types from those that are absent from the basis matrix column space.
635

636 **Supplementary Note 2: Noninvasive measurement of trophoblast cell type signatures in**
637 **preeclampsia**
638 In pregnancy, extravillous trophoblast (EVT) invasion is a stage in uteroplacental arterial remodeling[20,23].
639 Arterial remodeling occurs to ensure adequate maternal blood flow to the growing fetus[20,23] and is
640 sometimes reduced in preeclampsia[23]. Previously, EVT was reported by Tsang et al to be noninvasively
641 resolvable and elevated in early onset preeclampsia (gestational age at diagnosis < 34 weeks) as compared
642 to healthy pregnancy[24]. However, examination of the trophoblast gene profiles used by Tsang et al. using
643 two independent placental single-cell atlases[19,20] revealed several genes that were not cell type specific or
644 exhibited very low trophoblast expression (fig. S11C, 11D), thereby adversely impacting signature score
645 interpretation.
646 *CERCAM*, *IL18BP*, and *PYCR1* are not extravillous trophoblast specific, exhibiting higher
647 expression in fibroblast cell types in both atlases, despite Tsang's inclusion in their EVT gene profile (fig.
648 S11C, 11D). Furthermore, EVT genes in Tsang's gene profile, *RRAD*, *SLC6A2*, and *UPK1B* all exhibit
649 very low EVT expression across both placental atlases. Numerous PSG genes (*PSG11*, *PSG1/PSG2*,
650 *PSG3*, *PSG4*, *PSG6*, *PSG9*) do not exhibit high syncytiotrophoblast (SCT) expression, despite their
651 inclusion in Tsang's SCT gene profile. GH2 either exhibits no expression or comparable non-SCT
652 specific expression across cell types in both atlases (fig. S11C, 11D).
653 The presence of these non-cell type specific genes in a cell type gene profile consequently
654 impacted the interpretation of Tsang et al's signature scores. Using our criteria for deriving a given cell
655 type gene profile (Methods), we derived gene profiles for the same two cell types, EVT and SCT (fig.
656 S10), and then quantified their respective signature scores in two previously published preeclampsia
657 cohorts[5] (fig. S11). In contrast to Tsang et al, we observed no significant difference in either trophoblast
658 signature score in cfRNA samples collected at diagnosis for mothers with early-onset preeclampsia (fig.
659 S11) ($p = 0.703$ and $U = 1396$, 0.794 and $U = 1416$ respectively, two-sided Mann Whitney U) and for
660 mothers with either early- or late-onset preeclampsia ($p = 0.24$ and $H = 4.18$ , 0.54 and $H = 2.15$
661 respectively, Kruskal Wallace, fig. S10B) as compared to samples from mothers with no complications at
662 a matched gestational age.
663 In our work deriving cell type gene profiles for signature scoring in cfRNA, we only considered
664 genes with high log fold change in a given cell type population and low expression in any other measured
665 cell type (Methods). We acknowledge that this method may miss some genes for a given cell type
666 population with low uniform expression (i.e. low expression in a large fraction of cells of a given type) or
667 with heterogeneous expression (i.e. high expression in a small fraction of cells of a given type). However,
668 since this work is the first comprehensive examination of cell type specific origins in the cell free
669 transcriptome, we sought to be conservative in what we asserted to be cell type specific so that we could
670 be confident in measuring a cell type signature score noninvasively; this approach boded well for all
671 diseases presented in our work.
672 Taken together with validation in two independent placental cell atlases, we conclude that the
673 EVT and SCT cell type gene profiles by Tsang et al. do not enable estimation of trophoblast pathology
674 from cfRNA in preeclampsia. The role of extravillous trophoblast invasion and the ubiquity of its cellular
675 pathophysiology in preeclampsia thus remains an open question.
676
677
678
679
680

681 **References**


682 1. Ibarra, A. *et al.* Non-invasive characterization of human bone marrow stimulation and
683 reconstitution by cell-free messenger RNA sequencing. *Nat. Commun.* **11**, 400 (2020).

684 2. Larson, M. H. *et al.* A comprehensive characterization of the cell-free transcriptome reveals
685 tissue- and subtype-specific biomarkers for cancer detection. *Nat. Commun.* **12**, 2357 (2021).

686 3. Ngo, T. T. M. *et al.* Noninvasive blood tests for fetal development predict gestational age and
687 preterm delivery. *Science* **360**, 1133–1136 (2018).

688 4. Koh, W. *et al.* Noninvasive in vivo monitoring of tissue-specific global gene expression in
689 humans. *Proc Natl Acad Sci USA* **111**, 7361–7366 (2014).

690 5. Munchel, S. *et al.* Circulating transcripts in maternal blood reflect a molecular signature of early-
691 onset preeclampsia. *Sci. Transl. Med.* **12**, (2020).

692 6. Toden, S. *et al.* Noninvasive characterization of Alzheimer's disease by circulating, cell-free
693 messenger RNA next-generation sequencing. *Sci. Adv.* **6**, (2020).

694 7. Chalasani, N. *et al.* Noninvasive stratification of nonalcoholic fatty liver disease by whole
695 transcriptome cell-free mRNA characterization. *Am. J. Physiol. Gastrointest. Liver Physiol.* **320**,
696 G439–G449 (2021).

697 8. Ngo, T. T. M., Moufarrej, M. N. & Rasmussen, M. L. H. Noninvasive blood tests for fetal
698 development predict gestational age and preterm delivery. *science.sciencemag.org*.

699 9. Klatt, E. C. *Robbins & Cotran Atlas of Pathology*. (Elsevier, 2021).

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

702 11. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat.*
703 *Methods* **12**, 453–457 (2015).

704 12. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with
705 digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

706 13. The Tabula Sapiens Consortium & Quake, S. R. The Tabula Sapiens: a single cell transcriptomic
707 atlas of multiple organs from individual human donors. *BioRxiv* (2021)
708 doi:10.1101/2021.07.19.452956.

709 14. Sadeh, R. *et al.* ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of
710 the cells of origin. *Nat. Biotechnol.* **39**, 586–598 (2021).

711 15. Uhlen, M. *et al.* A genome-wide transcriptomic analysis of protein-coding genes in human blood
712 cells. *Science* **366**, (2019).

713 16. András, I. E. & Toborek, M. Extracellular vesicles of the blood-brain barrier. *Tissue Barriers* **4**,
714 e1131804 (2016).

715 17. Abbott, N. J. Inflammatory mediators and modulation of blood-brain barrier permeability. *Cell.*
716 *Mol. Neurobiol.* **20**, 131–147 (2000).

717 18. Ganong, W. F. Circumventricular organs: definition and role in the regulation of endocrine and
718 autonomic function. *Clin. Exp. Pharmacol. Physiol.* **27**, 422–427 (2000).

719 19. Suryawanshi, H. *et al.* A single-cell survey of the human first-trimester placenta and decidua. *Sci.*
720 *Adv.* **4**, eaau4788 (2018).

721 20. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans.
722 *Nature* **563**, 347–353 (2018).

723 21. Stewart, B. J. *et al.* Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–
724 1466 (2019).

725 22. Aizarani, N. *et al.* A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*
726 **572**, 199–204 (2019).

727 23. Kaufmann, P., Black, S. & Huppertz, B. Endovascular trophoblast invasion: implications for the
728 pathogenesis of intrauterine growth retardation and preeclampsia. *Biol. Reprod.* **69**, 1–7 (2003).

729 24. Tsang, J. C. H. *et al.* Integrative single-cell and cell-free plasma RNA transcriptomics elucidates
730 placental cellular dynamics. *Proc Natl Acad Sci USA* **114**, E7786–E7795 (2017).

731 25. Nakhoul, N. & Batuman, V. Role of proximal tubules in the pathogenesis of kidney disease.
732 *Contrib. Nephrol.* **169**, 37–50 (2011).

733 26. Chevalier, R. L. The proximal tubule is the primary target of injury and progression of kidney
734 disease: role of the glomerulotubular junction. *Am. J. Physiol. Renal Physiol.* **311**, F145-61
735 (2016).

736 27. Feldstein, A. E. & Gores, G. J. Apoptosis in alcoholic and nonalcoholic steatohepatitis. *Front.*
737 *Biosci.* **10**, 3093–3099 (2005).

738 28. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337
739 (2019).

740 29. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's
741 disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097
742 (2019).

743 30. Dhillon, P. *et al.* The Nuclear Receptor ESRRA Protects from Kidney Disease by Coupling
744 Metabolism and Differentiation. *Cell Metab.* **33**, 379-394.e8 (2021).

745 31. Schelling, J. R. Tubular atrophy in the pathogenesis of chronic kidney disease progression.
746 *Pediatr. Nephrol.* **31**, 693–706 (2016).

747 32. Meex, R. C. R. & Watt, M. J. Hepatokines: linking nonalcoholic fatty liver disease and insulin
748 resistance. *Nat. Rev. Endocrinol.* **13**, 509–520 (2017).

749 33. McCall, M. A. *et al.* Targeted deletion in astrocyte intermediate filament (Gfap) alters neuronal
750 physiology. *Proc Natl Acad Sci USA* **93**, 6361–6366 (1996).

751 34. Lytton, J. Na+/Ca2+ exchangers: three mammalian gene families control Ca2+ transport.
752 *Biochem. J.* **406**, 365–382 (2007).

753 35. Friedman, L. G. *et al.* Cadherin-8 expression, synaptic localization, and molecular control of
754 neuronal form in prefrontal corticostriatal circuits. *J. Comp. Neurol.* **523**, 75–92 (2015).

755 36. Arlotta, P. *et al.* Neuronal subtype-specific genes that control corticospinal motor neuron
756 development in vivo. *Neuron* **45**, 207–221 (2005).

757 37. Shigemoto, R., Nakanishi, S. & Mizuno, N. Distribution of the mRNA for a metabotropic
758 glutamate receptor (mGluR1) in the central nervous system: an in situ hybridization study in adult
759 and developing rat. *J. Comp. Neurol.* **322**, 121–135 (1992).

760 38. Zhou, Q., Choi, G. & Anderson, D. J. The bHLH transcription factor Olig2 promotes
761 oligodendrocyte differentiation in collaboration with Nkx2.2. *Neuron* **31**, 791–807 (2001).

762 39. Nielsen, J. A., Berndt, J. A., Hudson, L. D. & Armstrong, R. C. Myelin transcription factor 1
763 (Myt1) modulates the proliferation and differentiation of oligodendrocyte lineage cells. *Mol. Cell.*
764 *Neurosci.* **25**, 111–123 (2004).

765 40. Ichihara-Tanaka, K., Oohira, A., Rumsby, M. & Muramatsu, T. Neuroglycan C is a novel midkine
766 receptor involved in process elongation of oligodendroglial precursor-like cells. *J. Biol. Chem.*
767 **281**, 30857–30864 (2006).

768 41. Levine, J. M., Reynolds, R. & Fawcett, J. W. The oligodendrocyte precursor cell in health and
769 disease. *Trends Neurosci.* **24**, 39–47 (2001).

770 42. Liddelow, S. A. *et al.* Neurotoxic reactive astrocytes are induced by activated microglia. *Nature*
771 **541**, 481–487 (2017).

43. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

44. Moufarrej, M. N., Wong, R. J., Shaw, G. M., Stevenson, D. K. & Quake, S. R. Investigating Pregnancy and Its Complications Using Circulating Cell-Free RNA in Women's Blood During Gestation. *Front. Pediatr.* **8**, 605219 (2020).

45. Pan, W. Development of diagnostic methods using cell-free nucleic acids . (Stanford University, 2016).

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

47. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

48. Shen-Orr, S. S., Tibshirani, R. & Butte, A. J. Gene expression deconvolution in linear space. *Nat. Methods* **9**, 9–9 (2011).

49. Chang, C.-C. & Lin, C.-J. Training nu-support vector regression: theory and algorithms. *Neural Comput.* **14**, 1959–1977 (2002).

50. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat. Methods* **9**, 8–9; author reply 9 (2012).

51. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

52. Qiao, W. *et al.* PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.* **8**, e1002838 (2012).

53. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinformatics* **18**, 205–214 (2017).

54. van Rossum, D. & Hanisch, U.-K. Microglia. *Metab. Brain Dis.* **19**, 393–411 (2004).

**Tabula Sapiens Consortium**

**Overall Project Direction and Coordination**
Robert C Jones, Jim Karkanias, Mark Krasnow, Angela Oliveira Pisco, Stephen Quake, Julia Salzman, Nir Yosef

**Donor Recruitment**
Bryan Bulthaup, Phillip Brown, Will Harper, Marisa Hemenez, Ravikumar Ponnusamy, Ahmad Salehi, Bhavani Sanagavarapu, Eileen Spallino

**Surgeons**
Ksenia A. Aaron, Waldo Concepcion, James Gardner, Burnett Kelly, Nikole Neidlinger, Zifa Wang

**Logistical coordination**
Sheela Crasta, Saroja Kolluru, Maurizio Morri, Angela Oliveira Pisco, Serena Y. Tan, Kyle J. Travaglini, Chenling Xu

**Organ Processing**
Marcela Alcántara-Hernández, Nicole Almanzar, Jane Antony, Benjamin Beyersdorf, Deviana Burhan, Lauren Byrnes, Kruti Calcuttawala, Mathew Carter, Charles K. F. Chan, Charles A. Chang, Alex Colville, Sheela Crasta, Rebecca Culver, Ivana Cvijović, Jessica D'Addabbo, Gaetano D'Amato, Camille Ezran, Francisco Galdos, Astrid Gillich, William R. Goodyer, Yan Hang, Alyssa Hayashi, Sahar Houshdaran, Xianxi Huang, Juan Irwin, SoRi Jang, Julia Vallve Juanico, Aaron M. Kershner, Soochi Kim, Bernhard Kiss, Saroja Kolluru, William Kong, Maya Kumar, Rebecca Leylek, Baoxiang Li, Shixuan Liu, Gabriel Loeb, Wan-Jin Lu, Shruti Mantri, Maxim Markovic, Patrick L. McAlpine, Ross Metzger, Antoine de Morree, Maurizio Morri, Karim Mrouj, Shravani Mukherjee, Tyler Muser, Patrick Neuhöfer, Thi Nguyen, Kimberly Perez, Ragini Phansalkar, Angela Oliveira Pisco, Nazan Puluca, Zhen Qi, Poorvi Rao, Hayley Raquer, Koki Sasagawa, Nicholas Schaum, Bronwyn Lane Scott, Bobak Seddighzadeh, Joe Segal, Sushmita Sen, Sean Spencer, Lea Steffes, Varun R. Subramaniam, Aditi Swarup, Michael Swift, Kyle J Travaglini, Will Van Treuren, Emily Trimm, Maggie Tsui, Sivakamasundari Vijayakumar, Kim Chi Vo, Sevahn K. Vorperian, Hannah Weinstein, Juliane Winkler, Timothy T.H. Wu, Jamie Xie, Andrea R.Yung, Yue Zhang

**Sequencing**
Angela M. Detweiler, Honey Mekonen, Norma Neff, Rene V. Sit, Michelle Tan, Jia Yan

**Histology**
Gregory R. Bean, Gerald J. Berry, Vivek Charu, Erna Forgó, Brock A. Martin, Michael G. Ozawa, Oscar Silva, Serena Y. Tan, Pranathi Vemuri

**Computational Data Analysis**
Shaked Afik, Rob Bierman, Olga Botvinnik, Ashley Byrne, Michelle Chen, Roozbeh Dehghannasiri, Angela Detweiler, Adam Gayoso, Qiqing Li, Gita Mahmoudabadi, Aaron McGeever, Antoine de Morree, Julia Olivieri, Madeline Park, Angela Oliveira Pisco, Neha Ravikumar, Julia Salzman, Geoff Stanley, Michael Swift, Michelle Tan, Weilun Tan, Sevahn K. Vorperian, Sheng Wang, Galen Xing, Chenling Xu, Nir Yosef

**Expert Cell Type Annotation**
Marcela Alcántara-Hernández, Jane Antony, Charles A. Chang, Alex Colville, Sheela Crasta, Rebecca

Culver, Camille Ezran, Astrid Gillich, Yan Hang, Juan Irwin, SoRi Jang, Aaron M. Kershner, William Kong, Rebecca Leylek, Gabriel Loeb, Ross Metzger, Antoine de Morree, Patrick Neuhöfer, Kimberly Perez, Ragini Phansalkar, Zhen Qi, Hayley Raquer, Bronwyn Lane Scott, Rahul Sinha, Hanbing Song, Sean Spencer, Aditi Swarup, Michael Swift, Kyle J. Travaglini, Jamie Xie

**Tissue Expert Principal Investigators**

Steven E. Artandi, Philip Beachy, Michael F. Clarke, Linda Giudice, Franklin Huang, KC Huang, Juliana Idoyaga, Seung K Kim, Mark Krasnow, Christin Kuo, Patricia Nguyen, Stephen Quake, Thomas A. Rando, Kristy Red-Horse, Jeremy Reiter, Justin Sonnenburg, Bruce Wang, Albert Wu, Sean Wu, Tony Wyss-Coray

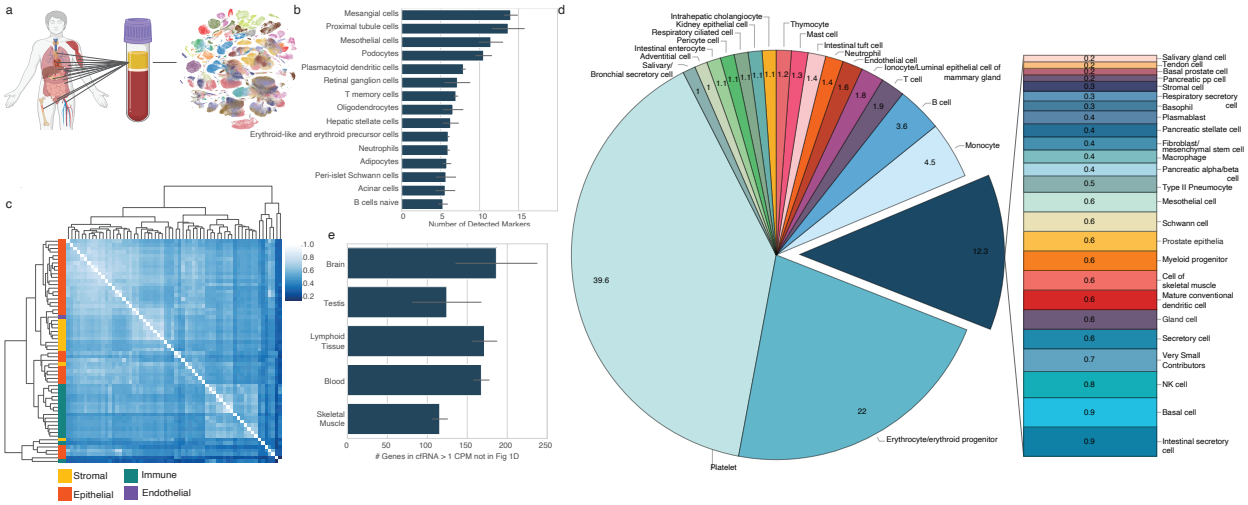
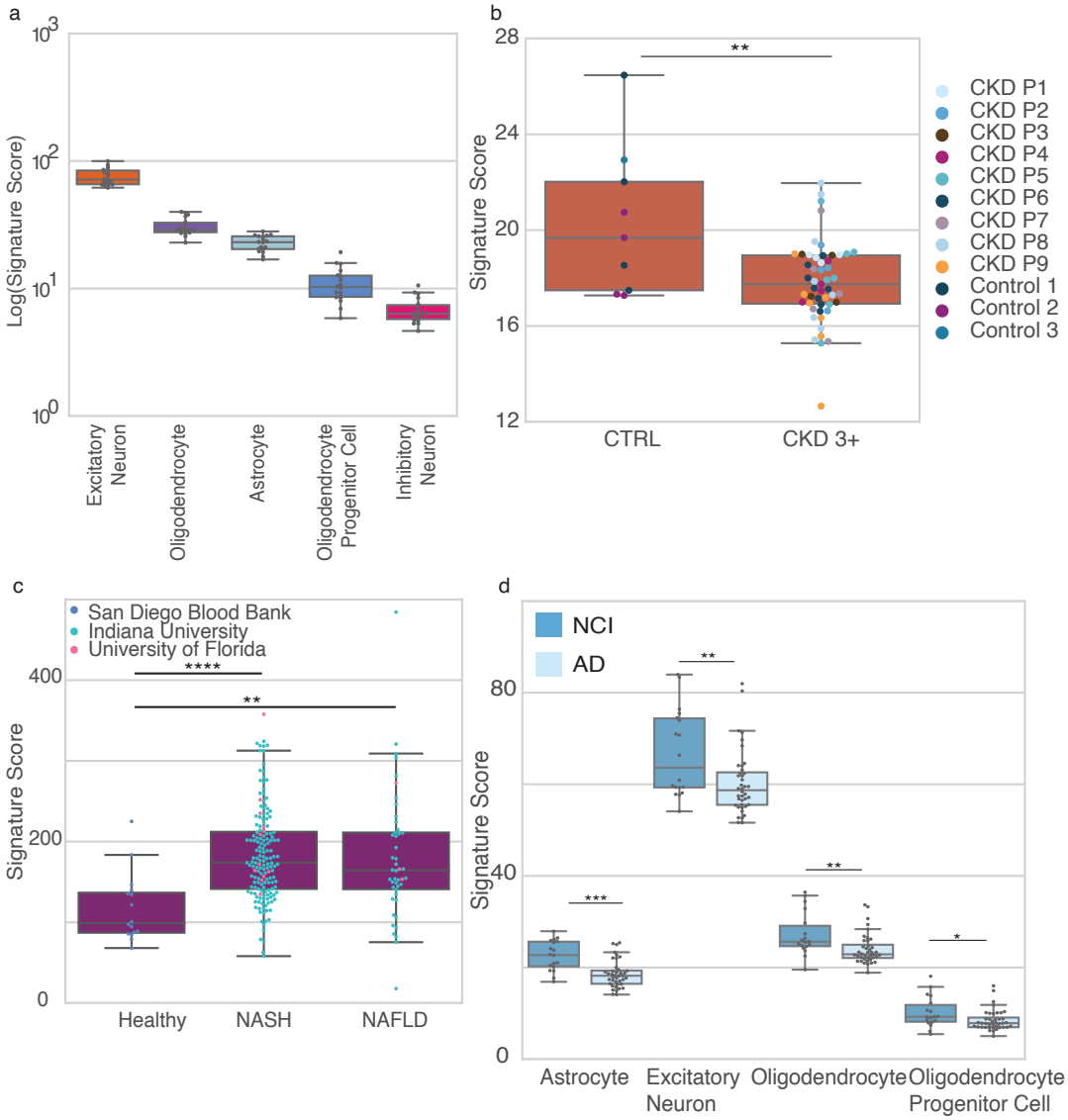



**Fig. 1: Cell type decomposition of the plasma cell free transcriptome using Tabula Sapiens reveals comprehensive cell type-of-origin landscape**

a. Schematic depicting integration of tissue-of-origin and single cell transcriptomics to identify cell types-of-origin in plasma cfRNA.
b. Number of cell type specific markers defined in context of the human body that are identified in cfRNA. Bars and error bars respectively denote the mean and s.d. of number of cell type specific markers across patients (n = 75). CPM-TMM counts for a given gene across technical replicates were averaged prior to intersection.
c. Cluster heatmap of Spearman correlations of cell type basis matrix column space derived from Tabula Sapiens.
d. Mean fractional contributions of cell type specific RNA in plasma cf-transcriptome (n = 18).
e. Top tissues in cfRNA not captured by basis matrix (e.g. the set difference of all genes detected in a given cfRNA sample and the row space of the basis matrix intersection with HPA tissue specific genes). Bars and error bars respectively denote the mean and s.d. of number of HPA tissue specific genes with NX counts > 10 and cell free CPM-TMM normalized expression ≥ 1 (n

895 = 18).




896

897 **Fig. 2: Cellular pathophysiology is noninvasively resolvable in cfRNA**

898 For a given boxplot, any cell type signature score is the sum of log-transformed CPM-TMM normalized counts. Box
899 plot: horizontal line, median; lower hinge, 25th percentile; upper hinge, 75th percentile; whiskers,1.5 interquartile
900 range; points outside whiskers indicate outliers. All P values were determined by a Mann Whitney U test; sidedness
901 specified in subplot caption. $*P < 0.05$, $**P < 10^{-2}$, $***P < 10^{-4}$, $****P < 10^{-5}$.

902 a. Neuronal and glial cell type signature scores in healthy cfRNA plasma (n = 18) on a logarithmic scale.
903 b. Comparison of dimensionless signature scores CKD stages 3+ (n = 51 samples; 9 patients) and healthy
904 controls (n = 9 samples; 3 patients) using data from Ibarra et al. Dot color denotes each patient ($P$ =
905 $9.66*10^{-3}$, U = 116, one-sided)
906 c. Hepatocyte signature scure between healthy (n = 16) and both NAFLD (n = 46) ($P$ =$3.15 \times 10^{-4}$, U =
907 155, one-sided) and NASH (n = 163) ($P = 4.68 \times 10^{-6}$, U = 427, one-sided); NASH vs. NAFLD
908 ($P$ = 0.464, U = 3483, two-sided). Color reflects sample collection center.
909 d. Neuronal and glial signature scores in AD (n = 40) and NCI (n = 18) cohorts. Excitatory neuron
910 ($P = 4.94 \times 10^{-3}$, U = 206, one-sided), oligodendrocyte ($P = 2.28 \times 10^{-3}$, U = 178, two-sided),
911 oligodendrocyte progenitor ($P = 2.27 \times 10^{-2}$, U = 224, two-sided), and astrocyte ($P = 6.1 \times 10^{-5}$, U
912 = 121, two-sided).

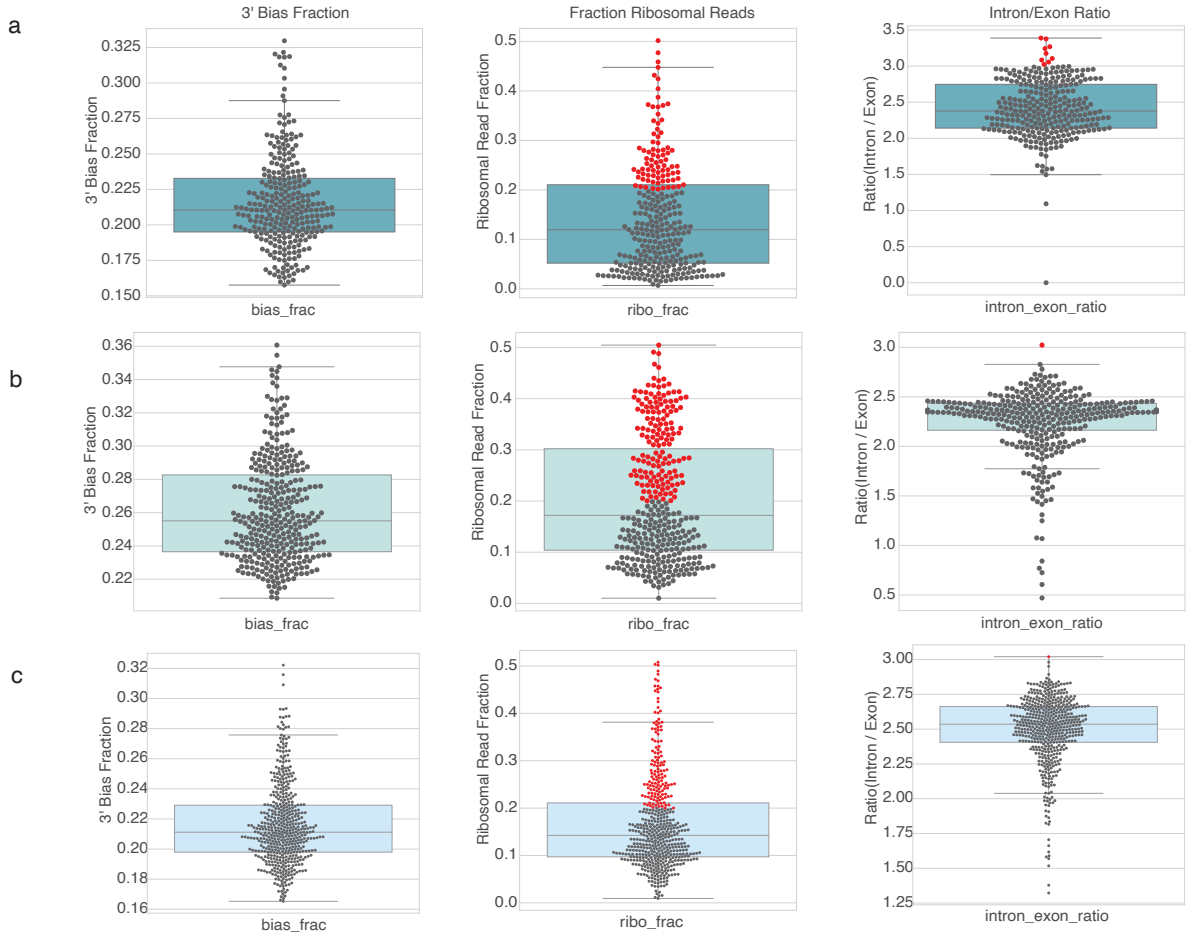



**Fig. S1 Identification of samples with outlier values for at least one quality control metric** including a measure of RNA degradation, ribosomal fraction, and DNA contamination from Ibarra et al. Samples with outlier values are highlighted in red. (See Methods section 'Sample quality filtering' for details). Box plot: horizonal line, median; lower hinge, 25$^{th}$ percentile; upper hinge, 75$^{th}$ percentile; whiskers span the 1.5 interquartile range; points outside the whiskers indicate outliers. Panels represent all published cfRNA samples from (a) Ibarra et al (b) Toden et al (c) Chalasani et al.

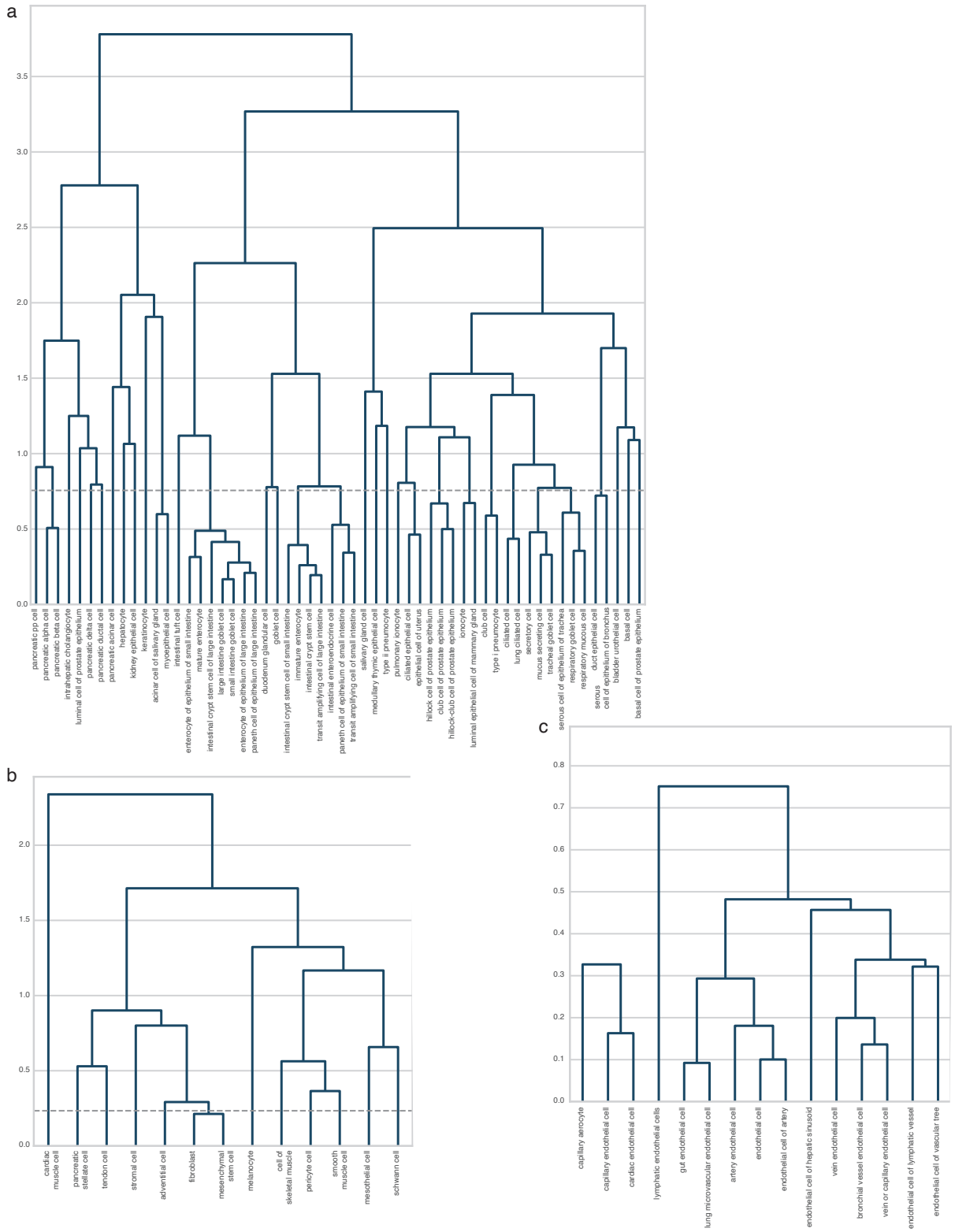



**Fig. S2 Hierarchical clustering on non-immune Tabula Sapiens organ compartments** for annotation coarsegraining in basis matrix generation. Dashed line indicates height at which tree was cut. Dendrograms correspond with the cell type annotations belonging to the (a) epithelial compartment, (b) the stromal compartment, (c) the endothelial compartment.

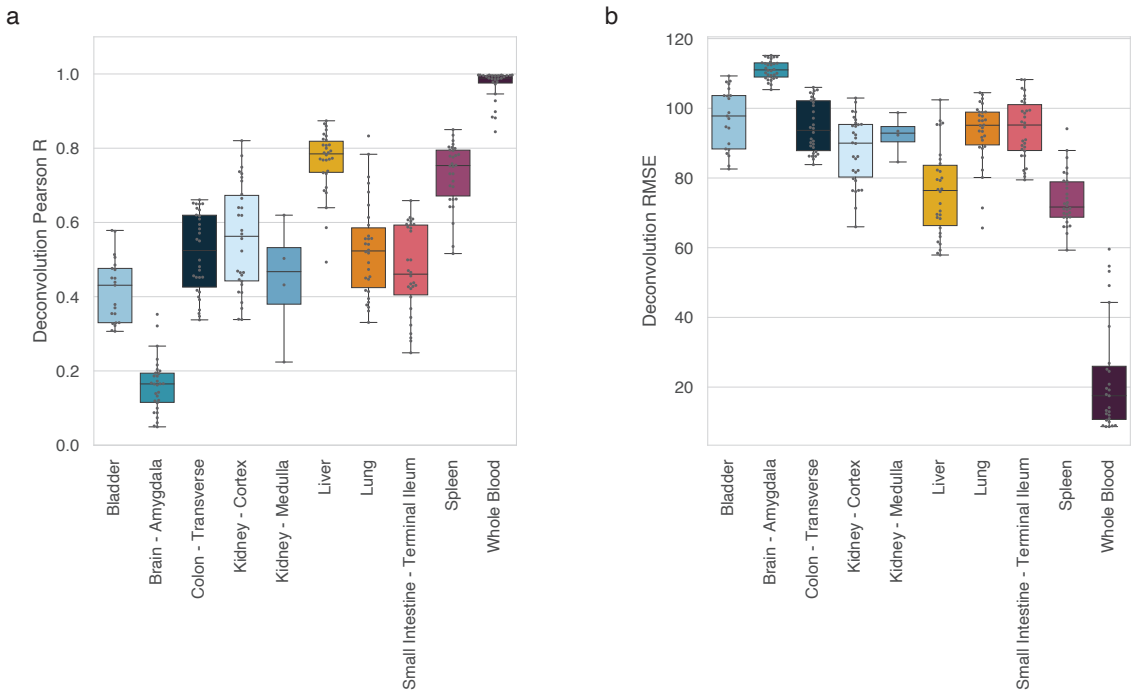
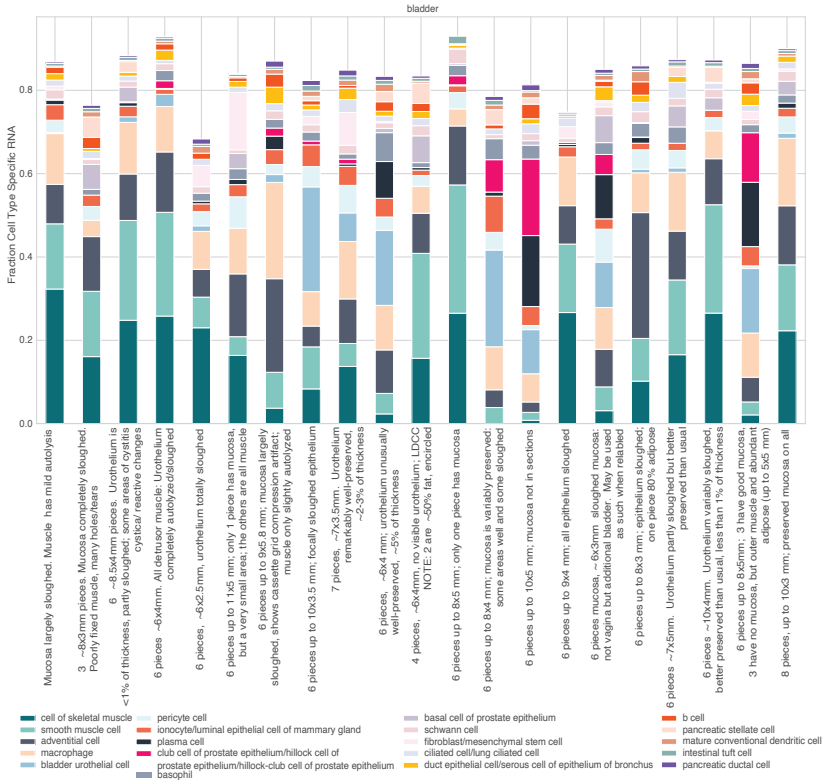


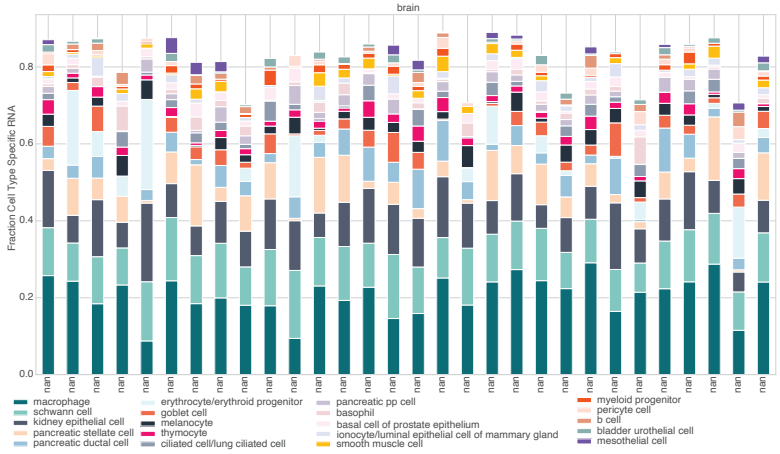

**Fig. S3. Basis matrix performance on GTEx bulk RNA samples using nu-SVR**. GTEx tissue samples possessing cell types wholly present (Kidney – Cortex, Spleen, Small Intestine – Terminal Ileum, Lung, Whole Blood) and absent from the basis matrix column space (Kidney – Medulla, Liver) were selected. For box plots: horizonal line, median; lower hinge, 25$^{th}$ percentile; upper hinge, 75$^{th}$ percentile; whiskers,1.5 interquartile range; points outside the whiskers indicate outliers.

(a) Pearson correlation between predicted expression and actual expression in cfRNA.
(b) Root Mean Square Error between predicted expression and actual expression in cfRNA. Units are zero-mean unit variance scaled CPM counts; tissues present in TSP have reduced RMSE compared to those that are absent (e.g. Kidney – Medulla and Brain). Tissues with high cellular heterogeneity (Lung, Bladder, Small Intestine, Kidney) exhibit reduced deconvolution performance compared to simpler tissues (Whole Blood, Spleen, Liver)

a

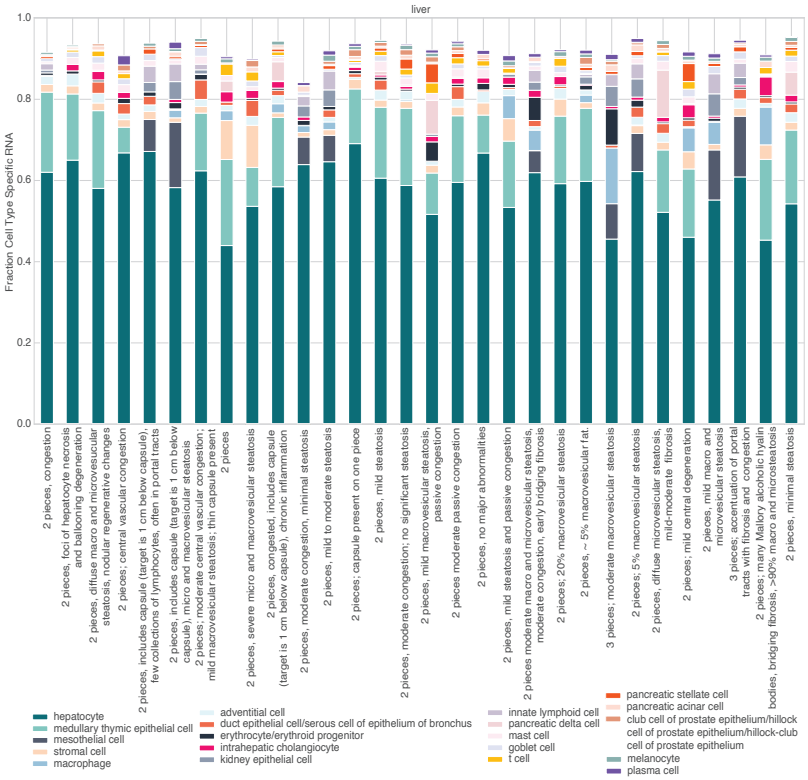

bladder
Fraction Cell Type Specific RNA
0.0
0.2
0.4
0.6
0.8
Mucosa largely sloughed. Muscle has mild autolysis
3, ~8x3 mm pieces. Mucosa completely sloughed. Poorly fixed muscle, many holes/tears
6 ~8.5x4 mm pieces. Urothelium is <1% of thickness, partly sloughed; some areas of cystitis cystica/ reactive changes
6 pieces ~6x4 mm. All detrusor muscle; Urothelium completely autolyzed/sloughed
6 pieces, ~6x2.5 mm, urothelium totally sloughed
6 pieces up to 11x5 mm; only 1 piece has mucosa, but a very small area; the others are all muscle
6 pieces up to 9x5.8 mm; mucosa largely sloughed, shows cassette grid compression artifact; muscle only slightly autolyzed
6 pieces up to 10x3.5 mm; focally sloughed epithelium
7 pieces, ~7x3.5 mm. Urothelium remarkably well preserved, ~2-3% of thickness
6 pieces, ~6x4 mm; urothelium unusually well-preserved, ~5% of thickness
4 pieces, ~6x4 mm, no visible urothelium; LDCC NOTE: 2 are ~50% fat, encircled
6 pieces up to 8x5 mm; only one piece has mucosa
6 pieces up to 8x4 mm; mucosa is variably preserved; some areas well and some sloughed
6 pieces up to 10x5 mm; mucosa not in sections
6 pieces up to 9x4 mm; all epithelium sloughed
6 pieces mucosa, ~6x3 mm sloughed mucosa; not vagina but additional bladder. May be used as such when relabled
6 pieces up to 8x3 mm; epithelium sloughed; one piece 80% adipose
6 pieces ~7x5 mm. Urothelium partly sloughed but better preserved than usual
6 pieces ~10x4 mm. Urothelium variably sloughed, better preserved than usual, less than 1% of thickness
6 pieces up to 8x5 mm; 3 have good mucosa, 3 have no mucosa, but outer muscle and abundant adipose (up to 5x5 mm)
8 pieces, up to 10x3 mm; preserved mucosa on all
cell of skeletal muscle
smooth muscle cell
adventitial cell
macrophage
bladder urothelial cell
pericyte cell
ionocyte/luminal epithelial cell of mammary gland
plasma cell
club cell of prostate epithelium/hillock cell of prostate epithelium
basophil
basal cell of prostate epithelium
schwann cell
fibroblast/mesenchymal stem cell
ciliated cell/lung ciliated cell
duct epithelial cell/serous cell of epithelium of bronchus
b cell
pancreatic stellate cell
mature conventional dendritic cell
intestinal tuft cell
pancreatic ductal cell

b

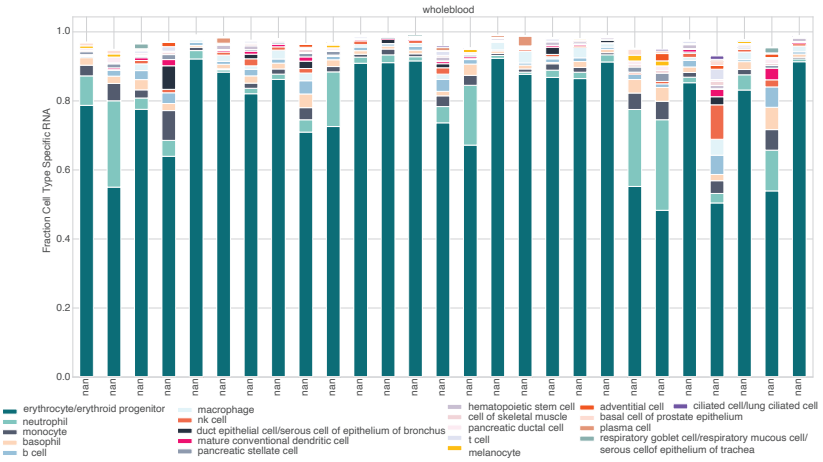

brain
Fraction Cell Type Specific RNA
0.0
0.2
0.4
0.6
0.8
nan
macrophage
schwann cell
kidney epithelial cell
pancreatic stellate cell
pancreatic ductal cell
erythrocyte/erythroid progenitor
goblet cell
melanocyte
thymocyte
ciliated cell/lung ciliated cell
pancreatic pp cell
basophil
basal cell of prostate epithelium
ionocyte/luminal epithelial cell of mammary gland
smooth muscle cell
myeloid progenitor
pericyte cell
b cell
bladder urothelial cell
mesothelial cell

946

c

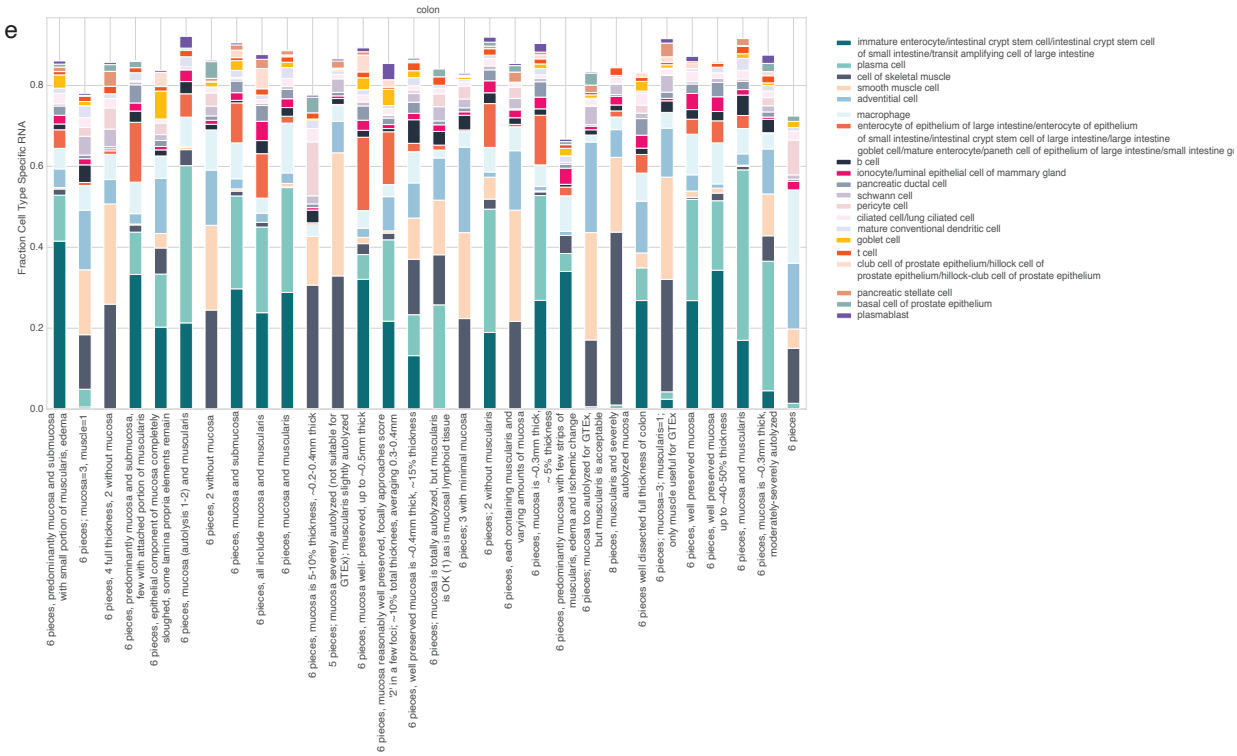

liver
Fraction Cell Type Specific RNA
1.0
0.8
0.6
0.4
0.2
0.0
2 pieces, congestion
2 pieces, foci of hepatocyte necrosis and ballooning degeneration
2 pieces, diffuse macro and microvesicular steatosis, nodular regenerative changes
2 pieces; central vascular congestion
2 pieces, includes capsule (target is 1 cm below capsule); few collections of lymphocytes, often in portal tracts
2 pieces, includes capsule (target is 1 cm below capsule), micro and macrovesicular steatosis
2 pieces; moderate central vascular congestion; mild macrovesicular steatosis; thin capsule present
2 pieces
2 pieces, severe micro and macrovesicular steatosis
2 pieces, congested, includes capsule (target is 1 cm below capsule), chronic inflammation
2 pieces, moderate congestion, minimal steatosis
2 pieces, mild to moderate steatosis
2 pieces; capsule present on one piece
2 pieces, mild steatosis
2 pieces, moderate congestion; no significant steatosis
2 pieces, mild macrovesicular steatosis, passive congestion
2 pieces moderate passive congestion
2 pieces, no major abnormalities
2 pieces, mild steatosis and passive congestion
2 pieces moderate macro and microvesicular steatosis, moderate congestion, early bridging fibrosis
2 pieces; 20% macrovesicular steatosis
2 pieces, ~ 5% macrovesicular fat.
3 pieces; moderate macrovesicular steatosis
2 pieces; 5% macrovesicular steatosis
2 pieces, diffuse microvesicular steatosis, mild-moderate fibrosis
2 pieces; mild central degeneration
2 pieces, mild macro and microvesicular steatosis
3 pieces; accentuation of portal tracts with fibrosis and congestion
2 pieces; many Mallory alcoholic hyalin bodies, bridging fibrosis, >80% macro and microsteatosis
2 pieces, minimal steatosis
hepatocyte
medullary thymic epithelial cell
mesothelial cell
stromal cell
macrophage
adventitial cell
duct epithelial cell/serous cell of epithelium of bronchus
erythrocyte/erythroid progenitor
intrahepatic cholangiocyte
kidney epithelial cell
innate lymphoid cell
pancreatic delta cell
mast cell
goblet cell
t cell
pancreatic stellate cell
pancreatic acinar cell
club cell of prostate epithelium/hillock cell of prostate epithelium/hillock-club cell of prostate epithelium
melanocyte
plasma cell

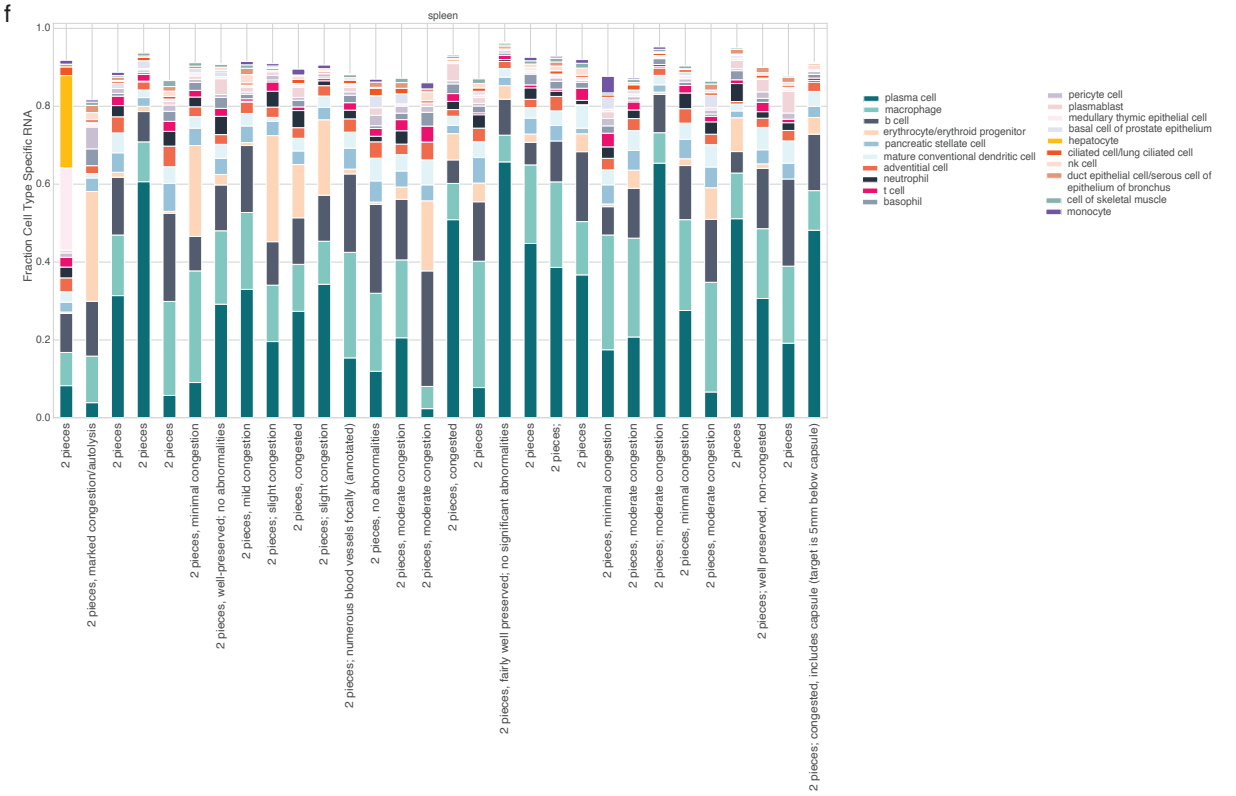
d


wholeblood
Fraction Cell Type Specific RNA
1.0
0.8
0.6
0.4
0.2
0.0
nan
erythrocyte/erythroid progenitor
neutrophil
monocyte
basophil
b cell
macrophage
nk cell
duct epithelial cell/serous cell of epithelium of bronchus
mature conventional dendritic cell
pancreatic stellate cell
hematopoietic stem cell
cell of skeletal muscle
pancreatic ductal cell
t cell
melanocyte
adventitial cell
basal cell of prostate epithelium
plasma cell
respiratory goblet cell/respiratory mucous cell/serous cellof epithelium of trachea
ciliated cell/lung ciliated cell

947

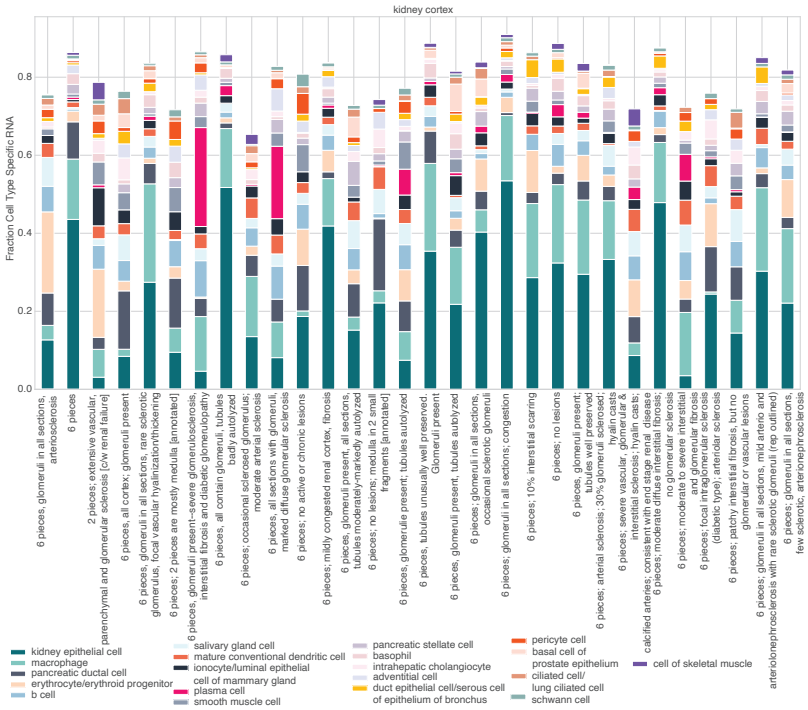
948


e
colon
Fraction Cell Type Specific RNA
0.0
0.2
0.4
0.6
0.8
6 pieces, predominantly mucosa and submucosa with small portion of muscularis, edema
6 pieces; mucosa=3, muscle=1
6 pieces, 4 full thickness, 2 without mucosa
6 pieces, predominantly mucosa and submucosa, few with attached portion of muscularis
6 pieces, epithelial component of mucosa completely sloughed, some lamina propria elements remain
6 pieces, mucosa (autolysis 1-2) and muscularis
6 pieces, 2 without mucosa
6 pieces, mucosa and submucosa
6 pieces, all include mucosa and muscularis
6 pieces, mucosa and muscularis
6 pieces, mucosa is 5-10% thickness, ~0.2-0.4mm thick
5 pieces; mucosa severely autolyzed (not suitable for GTEx); muscularis slightly autolyzed
6 pieces, mucosa well- preserved, up to ~0.5mm thick
6 pieces, mucosa reasonably well preserved, focally approaches score "2" in a few foci; ~10% total thickness, averaging 0.3-0.4mm
6 pieces, well preserved mucosa is ~0.4mm thick, ~15% thickness
6 pieces; mucosa is totally autolyzed, but muscularis is OK (1) as is mucosal lymphoid tissue
6 pieces; 3 with minimal mucosa
6 pieces; 2 without muscularis
6 pieces, each containing muscularis and varying amounts of mucosa
6 pieces, mucosa is ~0.3mm thick, 5% thickness
6 pieces, predominantly mucosa with few strips of muscularis, edema and ischemic change
6 pieces; mucosa too autolyzed for GTEx, but muscularis is acceptable
8 pieces, muscularis and severely autolyzed mucosa
6 pieces well dissected full thickness of colon
6 pieces; mucosa=3; muscularis=1; only muscle useful for GTEx
6 pieces, well preserved mucosa
6 pieces, well preserved mucosa up to ~40-50% thickness
6 pieces, mucosa and muscularis
6 pieces, mucosa is ~0.3mm thick, moderately-severely autolyzed
6 pieces
immature enterocyte/intestinal crypt stem cell/intestinal crypt stem cell of small intestine/transit amplifying cell of large intestine
plasma cell
cell of skeletal muscle
smooth muscle cell
adventitial cell
macrophage
enterocyte of epithelium of large intestine/enterocyte of epithelium of small intestine/intestinal crypt stem cell of large intestine/large intestine goblet cell/mature enterocyte/paneth cell of epithelium of large intestine/small intestine g
b cell
ionocyte/luminal epithelial cell of mammary gland
pancreatic ductal cell
schwann cell
pericyte cell
ciliated cell/lung ciliated cell
mature conventional dendritic cell
goblet cell
t cell
club cell of prostate epithelium/hillock cell of prostate epithelium/hillock-club cell of prostate epithelium
pancreatic stellate cell
basal cell of prostate epithelium
plasmablast

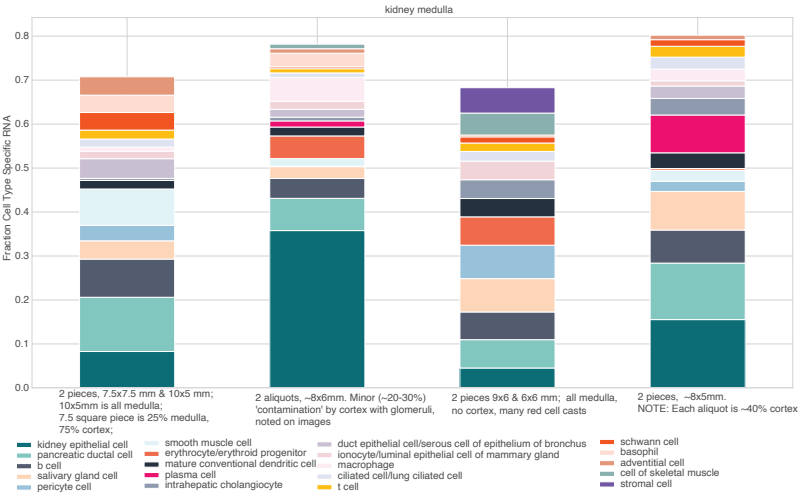

f
spleen
Fraction Cell Type Specific RNA
0.0
0.2
0.4
0.6
0.8
1.0
2 pieces
2 pieces, marked congestion/autolysis
2 pieces
2 pieces
2 pieces
2 pieces, minimal congestion
2 pieces, well-preserved; no abnormalities
2 pieces, mild congestion
2 pieces; slight congestion
2 pieces, congested
2 pieces; slight congestion
2 pieces; numerous blood vessels focally (annotated)
2 pieces, no abnormalities
2 pieces, moderate congestion
2 pieces, moderate congestion
2 pieces, congested
2 pieces
2 pieces, fairly well preserved; no significant abnormalities
2 pieces
2 pieces;
2 pieces
2 pieces, minimal congestion
2 pieces, moderate congestion
2 pieces; moderate congestion
2 pieces, minmal congestion
2 pieces, moderate congestion
2 pieces
2 pieces; well preserved, non-congested
2 pieces
2 pieces; congested, includes capsule (target is 5mm below capsule)
plasma cell
macrophage
b cell
erythrocyte/erythroid progenitor
pancreatic stellate cell
mature conventional dendritic cell
adventitial cell
neutrophil
t cell
basophil
pericyte cell
plasmablast
medullary thymic epithelial cell
basal cell of prostate epithelium
hepatocyte
ciliated cell/lung ciliated cell
nk cell
duct epithelial cell/serous cell of epithelium of bronchus
cell of skeletal muscle
monocyte

949

g

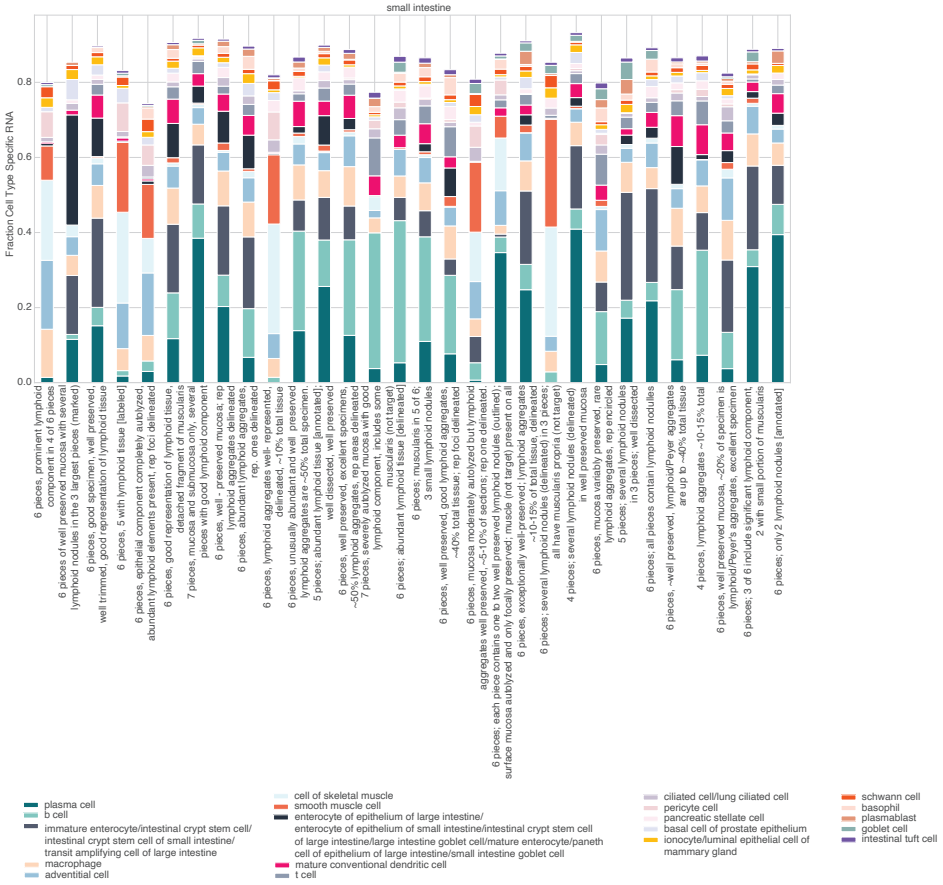

kidney cortex
Fraction Cell Type Specific RNA
0.0
0.2
0.4
0.6
0.8
6 pieces, glomeruli in all sections, arteriosclerosis
6 pieces
2 pieces; extensive vascular, parenchymal and glomerular sclerosis (c/w renal failure)
6 pieces, all cortex; glomeruli present
6 pieces, glomeruli in all sections, rare sclerotic glomerulus, focal vascular hyalinization/thickening
6 pieces; 2 pieces are mostly medulla [annotated]
6 pieces, glomeruli present–severe glomerulosclerosis, interstitial fibrosis and diabetic glomerulopathy
6 pieces, all contain glomeruli, tubules badly autolyzed
6 pieces; occasional sclerosed glomerulus; moderate arterial sclerosis
6 pieces, all sections with glomeruli, marked diffuse glomerular sclerosis
6 pieces; no active or chronic lesions
6 pieces; mildly congested renal cortex, fibrosis
6 pieces, glomeruli present, all sections, tubules moderately-markedly autolyzed
6 pieces; no lesions; medulla in 2 small fragments [annotated]
6 pieces, glomerulie present; tubules autolyzed
6 pieces, tubules unusually well preserved, Glomeruli present
6 pieces, glomeruli present, tubules autolyzed
6 pieces, glomeruli in all sections, occasional sclerotic glomeruli
6 pieces; glomeruli in all sections; congestion
6 pieces; 10% interstitial scarring
6 pieces; no lesions
6 pieces, glomeruli present; tubules well preserved
6 pieces; arterial sclerosis; 30% glomeruli sclerosed; hyalin casts
6 pieces; severe vascular, glomerular & interstitial sclerosis; renal disease
6 pieces; moderate diffuse interstitial fibrosis; calcified arteries; consistent with end stage renal disease
6 pieces; moderate to severe interstitial and glomerular fibrosis no glomerular sclerosis
6 pieces; focal intraglomerular sclerosis (diabetic type); arteriolar sclerosis
6 pieces; patchy interstitial fibrosis, but no glomerular or vascular lesions
6 pieces; glomeruli in all sections; mild arteriosclerosis with rare sclerotic glomeruli (rep outlined)
6 pieces; glomeruli in all sections; few sclerotic, arteriolonephrosclerosis
kidney epithelial cell
macrophage
pancreatic ductal cell
erythrocyte/erythroid progenitor
b cell
salivary gland cell
mature conventional dendritic cell
ionocyte/luminal epithelial cell of mammary gland
plasma cell
smooth muscle cell
pancreatic stellate cell
basophil
intrahepatic cholangiocyte
adventitial cell
duct epithelial cell/serous cell of epithelium of bronchus
pericyte cell
basal cell of prostate epithelium
ciliated cell/ lung ciliated cell
schwann cell
cell of skeletal muscle

h


kidney medulla
Fraction Cell Type Specific RNA
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2 pieces, 7.5x7.5 mm & 10x5 mm; 10x5mm is all medulla; 7.5 square piece is 25% medulla, 75% cortex;
2 aliquots, ~8x6mm. Minor (~20-30%) 'contamination' by cortex with glomeruli, noted on images
2 pieces 9x6 & 6x6 mm; all medulla, no cortex, many red cell casts
2 pieces, ~8x5mm. NOTE: Each aliquot is ~40% cortex
kidney epithelial cell
pancreatic ductal cell
b cell
salivary gland cell
pericyte cell
smooth muscle cell
erythrocyte/erythroid progenitor
mature conventional dendritic cell
plasma cell
intrahepatic cholangiocyte
duct epithelial cell/serous cell of epithelium of bronchus
ionocyte/luminal epithelial cell of mammary gland
macrophage
ciliated cell/lung ciliated cell
t cell
schwann cell
basophil
adventitial cell
cell of skeletal muscle
stromal cell

950

i


small intestine
Fraction Cell Type Specific RNA
0.0
0.2
0.4
0.6
0.8
6 pieces, prominent lymphoid component in 4 of 6 pieces
6 pieces of well preserved mucosa with several lymphoid nodules in the 3 largest pieces (marked)
6 pieces, good specimen, well preserved, well trimmed, good representation of lymphoid tissue
6 pieces, 5 with lymphoid tissue [labeled]
6 pieces, epithelial component completely autolyzed, abundant lymphoid elements present, rep foci delineated
6 pieces, good representation of lymphoid tissue, detached fragment of muscularis
7 pieces, mucosa and submucosa only, several pieces with good lymphoid component
6 pieces, well - preserved mucosa, rep lymphoid aggregates delineated
6 pieces, abundant lymphoid aggregates, rep. ones delineated
6 pieces, lymphoid aggregates well- represented, delineated, ~10% total tissue
6 pieces, unusually abundant and well preserved lymphoid aggregates are ~30% total specimen.
5 pieces; abundant lymphoid tissue [delineated]; well dissected, well preserved
6 pieces, well preserved, excellent specimens, ~50% lymphoid aggregates, rep areas delineated
7 pieces, severely autolyzed mucosa with good lymphoid component, includes some muscularis (not target)
6 pieces; abundant lymphoid tissue [delineated]
6 pieces; muscularis is 5 of 6; 3 small lymphoid nodules
6 pieces, well preserved, good lymphoid aggregates ~40% total tissue; rep foci delineated
6 pieces, mucosa moderately autolyzed but lymphoid aggregates well preserved, ~5-10% of sections; rep one delineated
6 pieces; each piece contains one to two well preserved lymphoid nodules (outlined); surface mucosa autolyzed and only focally preserved; muscle (not target) present on all
6 pieces, exceptionally well-preserved, lymphoid aggregates ~10-15% of total tissue, delineated
6 pieces; several lymphoid nodules (delineated) in 3 pieces; all have muscularis propria (not target)
4 pieces; several lymphoid nodules (delineated)
6 pieces, mucosa variably preserved, rare lymphoid aggregates, rep encircled
5 pieces; several lymphoid nodules in 3 pieces; well dissected
6 pieces; all pieces contain lymphoid nodules
6 pieces, ~well preserved, lymphoid/Peyer aggregates are up to ~40% total tissue
4 pieces, lymphoid aggregates ~10-15% total
6 pieces, well preserved mucosa, ~20% of specimen is lymphoid/Peyer's aggregates, excellent specimen
6 pieces; 3 of 6 include significant lymphoid component, 2 with small portion of muscularis
6 pieces; only 2 lymphoid nodules [annotated]
plasma cell
b cell
immature enterocyte/intestinal crypt stem cell/ intestinal crypt stem cell of small intestine/ transit amplifying cell of large intestine
macrophage
adventitial cell
cell of skeletal muscle
smooth muscle cell
enterocyte of epithelium of large intestine/ enterocyte of epithelium of small intestine/intestinal crypt stem cell of large intestine/large intestine goblet cell/mature enterocyte/paneth cell of epithelium of large intestine/small intestine goblet cell
mature conventional dendritic cell
t cell
ciliated cell/lung ciliated cell
pericyte cell
pancreatic stellate cell
basal cell of prostate epithelium
ionocyte/luminal epithelial cell of mammary gland
schwann cell
basophil
plasmablast
goblet cell
intestinal tuft cell

951

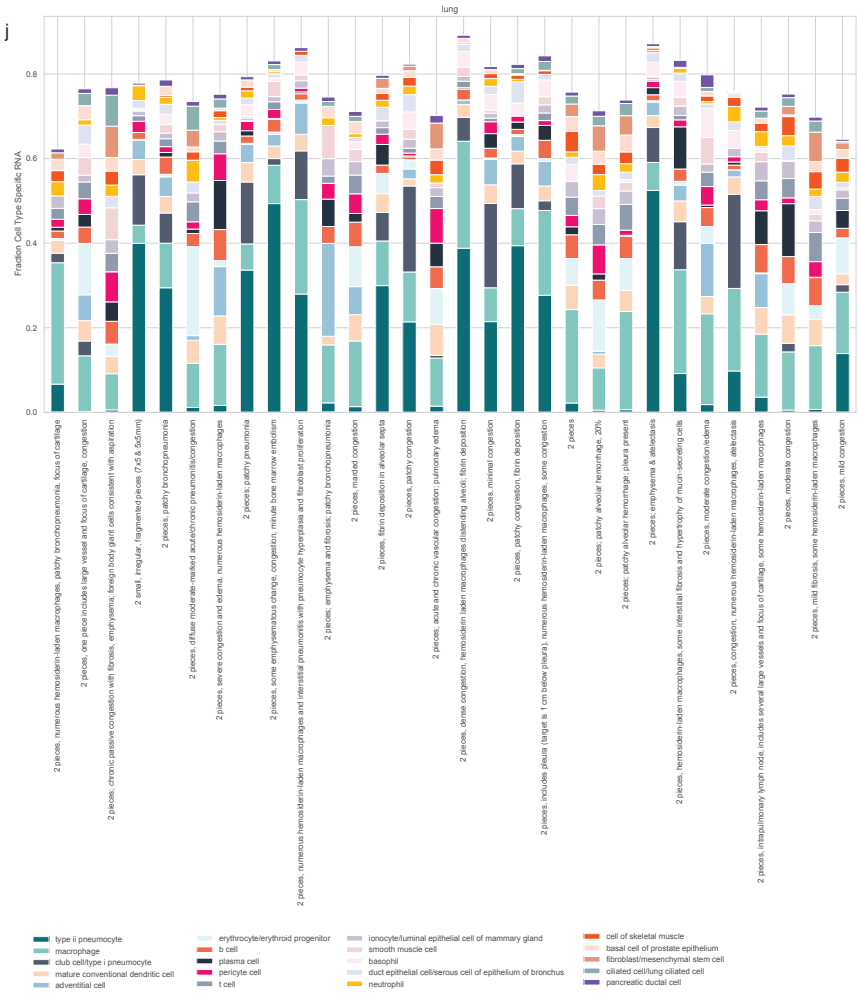



**Fig S4. Deconvolved fractions of cell type specific RNA from various GTEx tissues using nu-SVR to assess deconvolution performance of the *Tabula Sapiens*-derived basis matrix**. The two tissues whose cell types were absent from the basis matrix column space were Kidney – Medulla and Brain. Kidney medulla samples reported to be contaminated with cortex are reflected by deconvolved kidney epithelia fractions. The brain, which is absent from the TSP v1.0, has majority fractions of schwann cells, a peripheral nervous cell type. Majority cell types for a given tissue, such as lung pneumonocytes and immune cells in the lung or kidney epithelia for the kidney cortex underscore the ability for the signature matrix to capture representative fractions of cell type specific RNA and reflect underlying cell heterogeneity in bulk RNA-seq data. Additional discussion is in Supplementary Note 1.

(a) Bladder
(b) Brain
(c) Liver
(d) Whole Blood
(e) Colon – Transverse
(f) Spleen
(g) Kidney – Cortex
(h) Kidney – Medulla
(i) Small Intestine – Terminal Ileum
(j) Lung

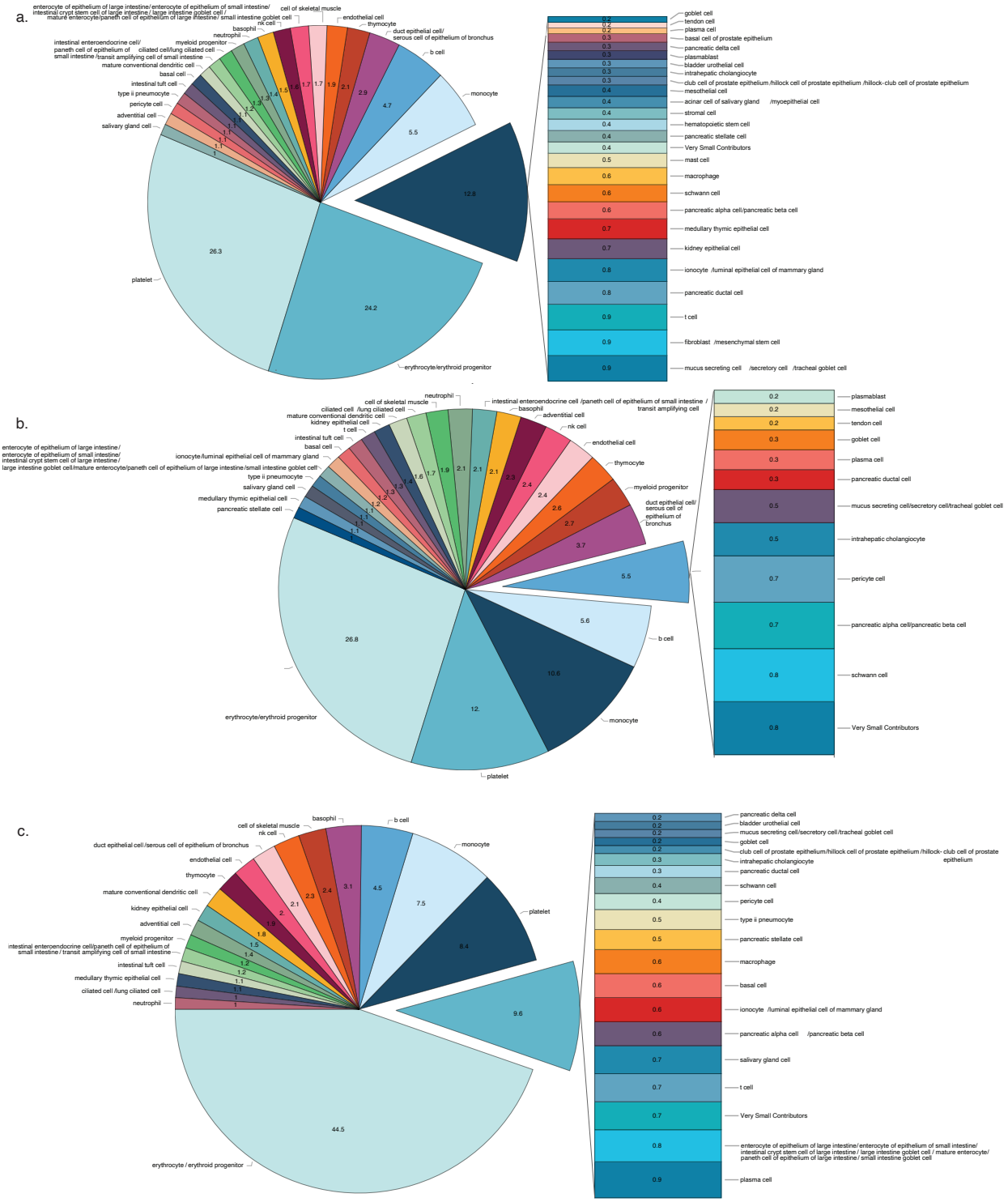



**Fig. S5. Deconvolution of healthy plasma samples from Toden et al using Tabula Sapiens.** Pie charts denote mean fractional cell type specific RNA contributions for (a) BioIVT (n = 18), (b) University of Indiana (n = 17), (c) Washington University in St. Louis (n = 22).

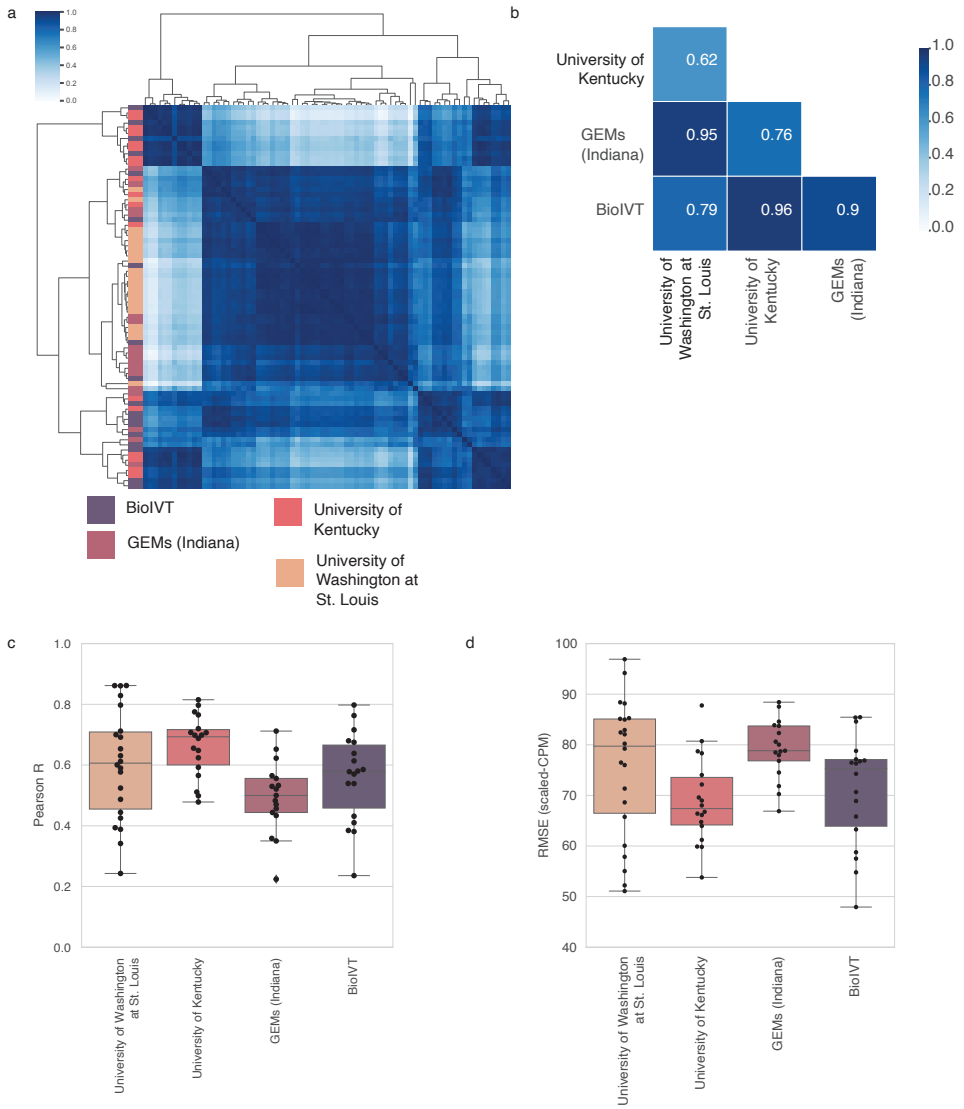



**Fig. S6. Deconvolving the plasma cell free transcriptome with Tabula Sapiens using nu-SVR (n = 75 samples)**. For boxplots, Horizonal line, median; lower hinge, 25th percentile; upper hinge, 75th percentile; whiskers span the 1.5 interquartile range; points outside the whiskers indicate outliers.

(a) Complete linkage clustermap of pairwise pearson correlation of deconvolved cell type fractions between patients from a given center; row color denotes a given center.
(b) Heatmap of pairwise pearson correlation of the mean cell type coefficients per center
(c) Deconvolution pearson correlation between predicted vs. measured expression for all biological replicates across all centers.
(d) Deconvolution RMSE between predicted vs. measured expression for all biological replicates across all centers.

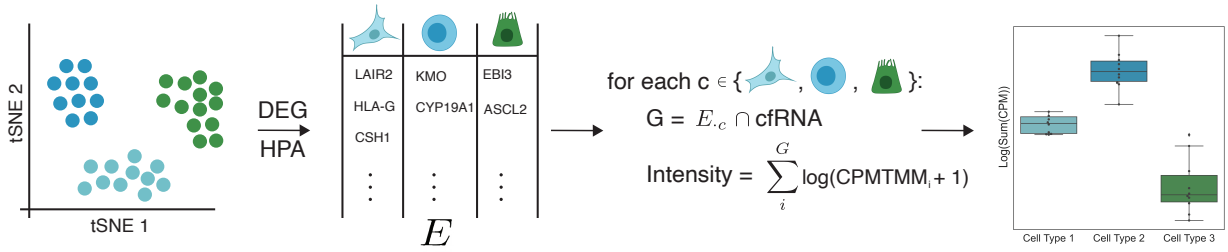



**Fig S7. Cell type signature scoring procedure**. See 'Signature Scoring' section of methods for derivation of cell type gene profiles.

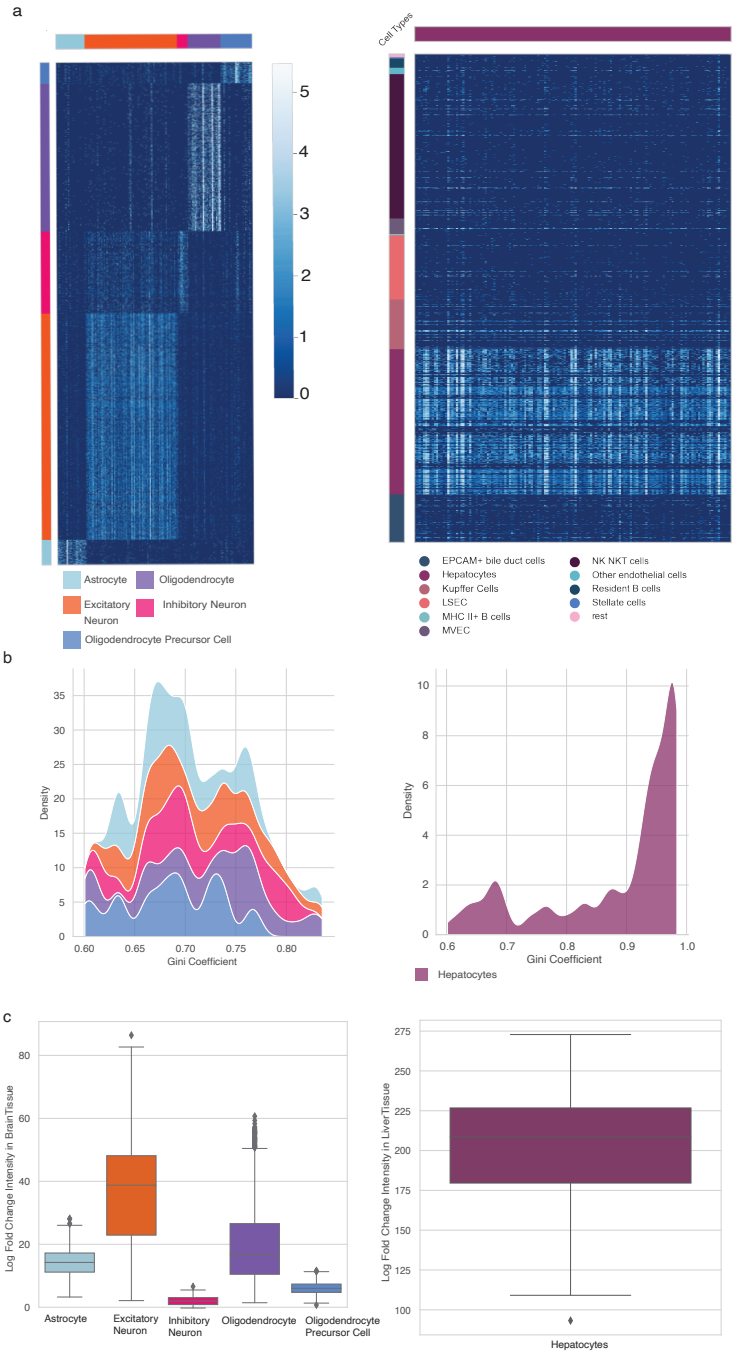



**Fig S8. Establishing gene profile cell type specificity in context of the whole human body using single cell and bulk RNA-seq data**

(a) Single cell heatmaps for gene cell type profiles within the corresponding tissue cell atlas, demonstrating that a cell type specific profile is unique to a given cell type across those within a given tissue. Columns denote marker genes for a given cell type; rows indicate individual cells. Heat mapped values are log-transformed counts-per-ten thousand.
(b) Gini coefficient density plot for genes in cell type profiles derived from brain and liver single cell atlases using HPA NX counts. The area under the curve for a given cell type sums to one.
(c) Log fold change in bulk RNA-seq data of a given cell type profile, demonstrating that the predominant expression of the cell type signature in its native tissue is highest relative to other non-native tissues. Values are the log-fold change of the signature score of a given cell type profile in the native tissue (indicated by the y-axis) to the mean expression in the remaining non-native tissues; this is the log ratio of the sum of log-transformed counts per ten-thousand balues. Box plot: horizontal line, median; lower hinge, 25th percentile; upper hinge, 75th percentile; whiskers span the 1.5 interquartile range; points outside the whiskers indicate outliers.

1000

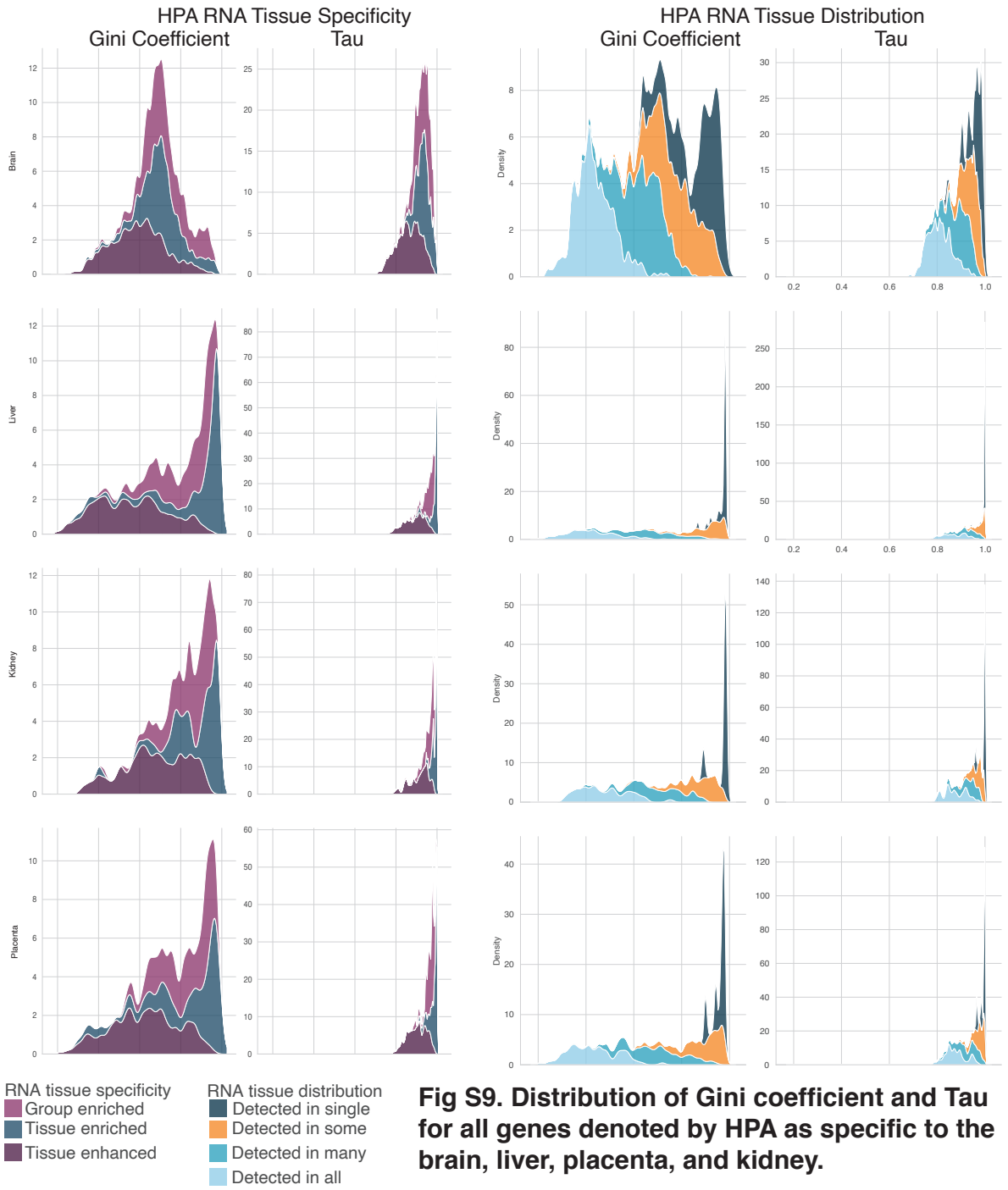



**Fig S9. Distribution of Gini coefficient and Tau for all genes denoted by HPA as specific to the brain, liver, placenta, and kidney.**

1001

**Fig. S10 Comprehensive placental and renal cell type gene profile specificity at single cell and whole body resolution.**
For box plots in f, g : horizontal line, median; lower hinge, 25th percentile; upper hinge, 75th percentile; whiskers span the 1.5 interquartile range; points outside whiskers indicate outliers.
(a) Violin plot of derived syncytiotrophoblast and extravillous trophoblast gene profiles from Vento-Tormo et al.
(b) Violin plot of derived syncytiotrophoblast and extravillous trophoblast gene profiles from Suryawanshi et al.
(c) Violin plot of derived proximal tubule gene profile
(d) Gini coefficient distribution for placental trophoblast cell types in (a) and (b)
(e) Gini coefficient distribution for renal cell type in (c)
(f) Distribution of placental trophoblast signature scores across all GTEx tissues. Note: given that the placenta is not in GTEx, the box plots correspond to the distribution of signature scores across non-placental tissues (sum of log-transformed counts-per-ten thousand).
(g) Log-fold change of renal cell type signature score in GTEx Kidney Cortex/Medulla samples relative to the mean non-kidney signature score, demonstrating that the predominant expression of the cell type signature in its native tissue is highest relative to other non-native tissues. Values are the log ratio of the sum of log-transformed counts per ten-thousand gene expression values.

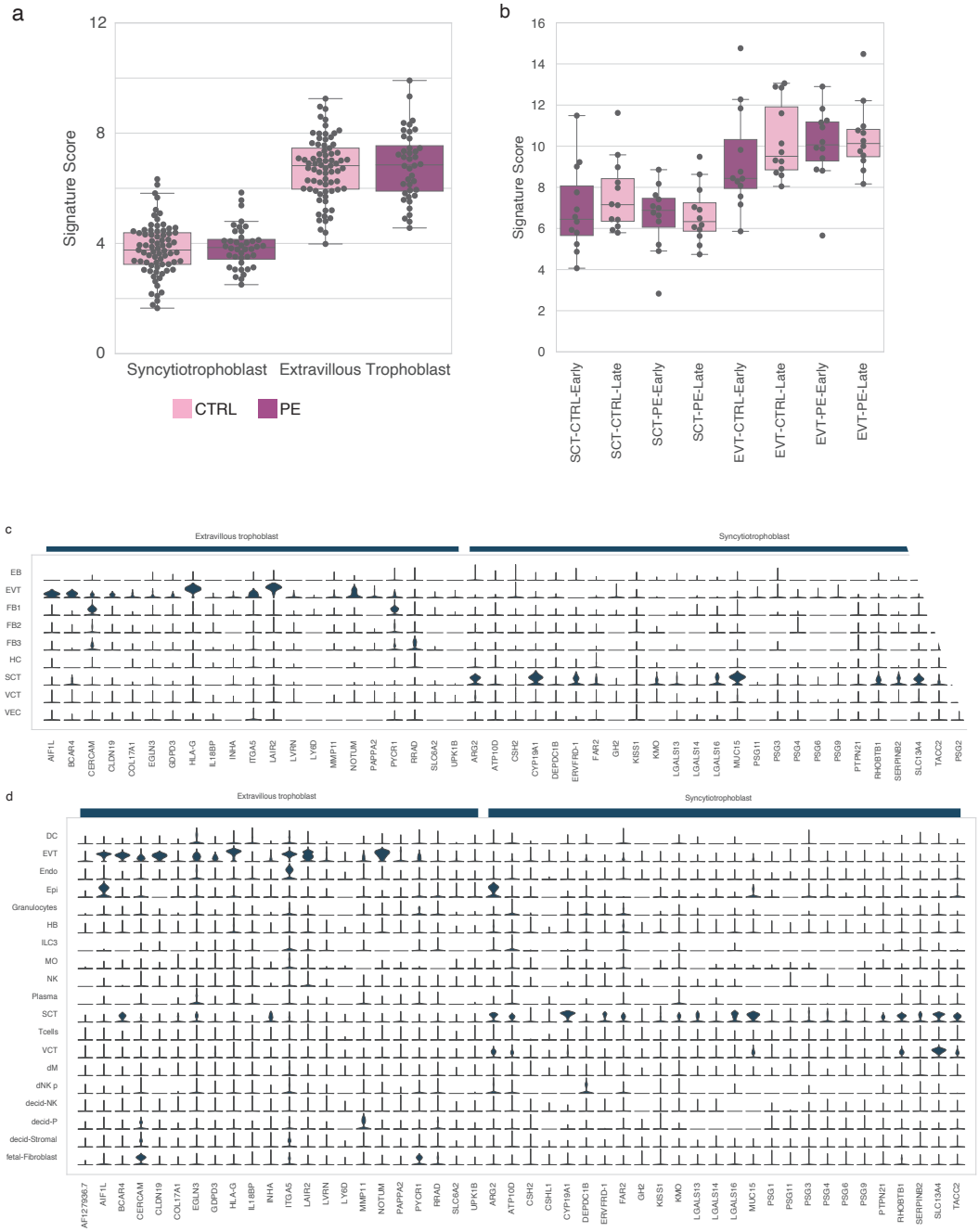
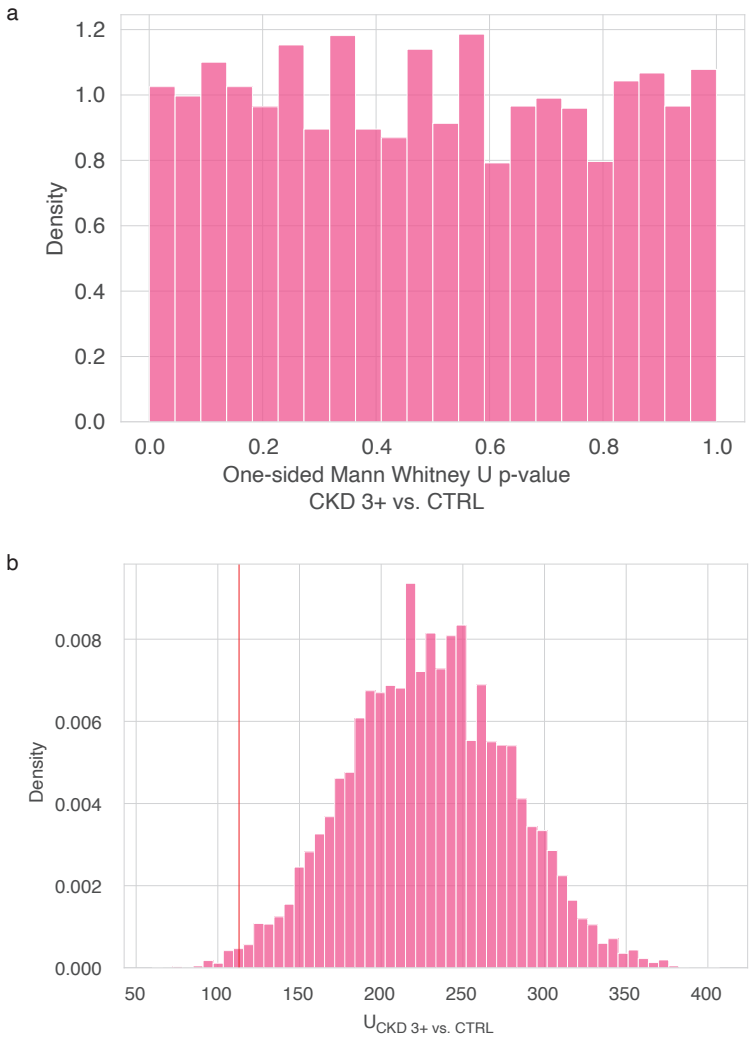



Fig. S11 **Expression distribution of Tsang et al trophoblast gene profiles and trophoblast signature scoring in preeclampsia cfRNA** from (a) iPEC (mothers with no complications, n = 73; mothers with preeclampsia, n = 40) and (b) PEARL-PEC (n = 12 for each early/late-onset PE cohorts and gestationally-age matched healthy controls). Box plot: horizontal line, median; lower hinge, 25$^{th}$ percentile; upper hinge, 75$^{th}$ percentile; whiskers span the 1.5 interquartile range; points outside the whiskers indicate outliers. Stacked violin plot of the genes comprising the extravillous trophoblast and syncytiotrophoblast gene profiles from Tsang et al. intersecting with the measured genes in (c) Suryawanshi et al and (d) Vento Tormo et al, reflecting the expression distribution across placental cell types.




1012
1013 **Fig. S12. Assessment of p-value calibration for the proximal tubule signature score** in CKD stage 3+
1014 vs healthy samples following one-sided Mann Whitney U in 10,000 trial permutation test with an
1015 alternative hypothesis that the PT signature in healthy is greater than CKD (a) Distribution of p-values;
1016 distribution is uniform, as expected under the null. (b) distribution of U values; vertical line denotes the
1017 actual comparison; we observe a test-statistic as or more extreme than what was experimentally observed
1018 0.63% of the time under the null. Taken together, this indicates the specificity of our signature score in
1019 discriminating between CKD and healthy patients.

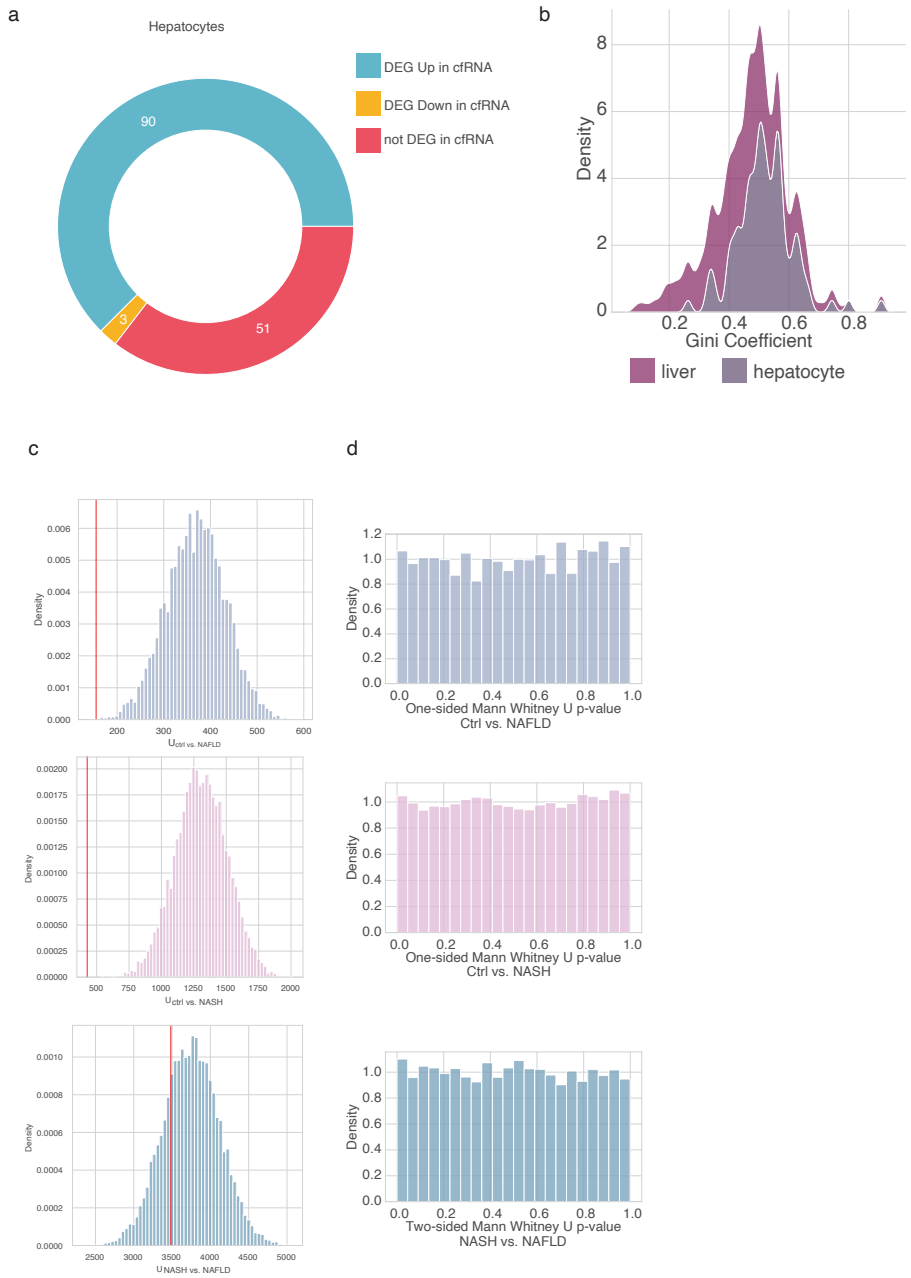



**Fig S13. NAFLD/NASH DEG intersection with the hepatocyte gene profile and assessing their discriminatory power in signature scoring**

(a) Donut plot reflecting the number of genes in the hepatocyte cell type gene profile that intersect with the reported NAFLD DEG[7]

(b) Density plot reflecting the Gini coefficient distribution corresponding to upregulated DEG in NAFLD[7] that are liver or hepatocyte specific. The Gini coefficient is computed using the mean expression per liver cell type in Aizarani et al (Methods). Area under each curve sums to one.

(c) Density of U values over 10,000 trial permutation test; red line indicates the U value corresponding to the experimental comparison reported in Fig. 2.

(d) Density of p-values over 10,000 trial permutation test.

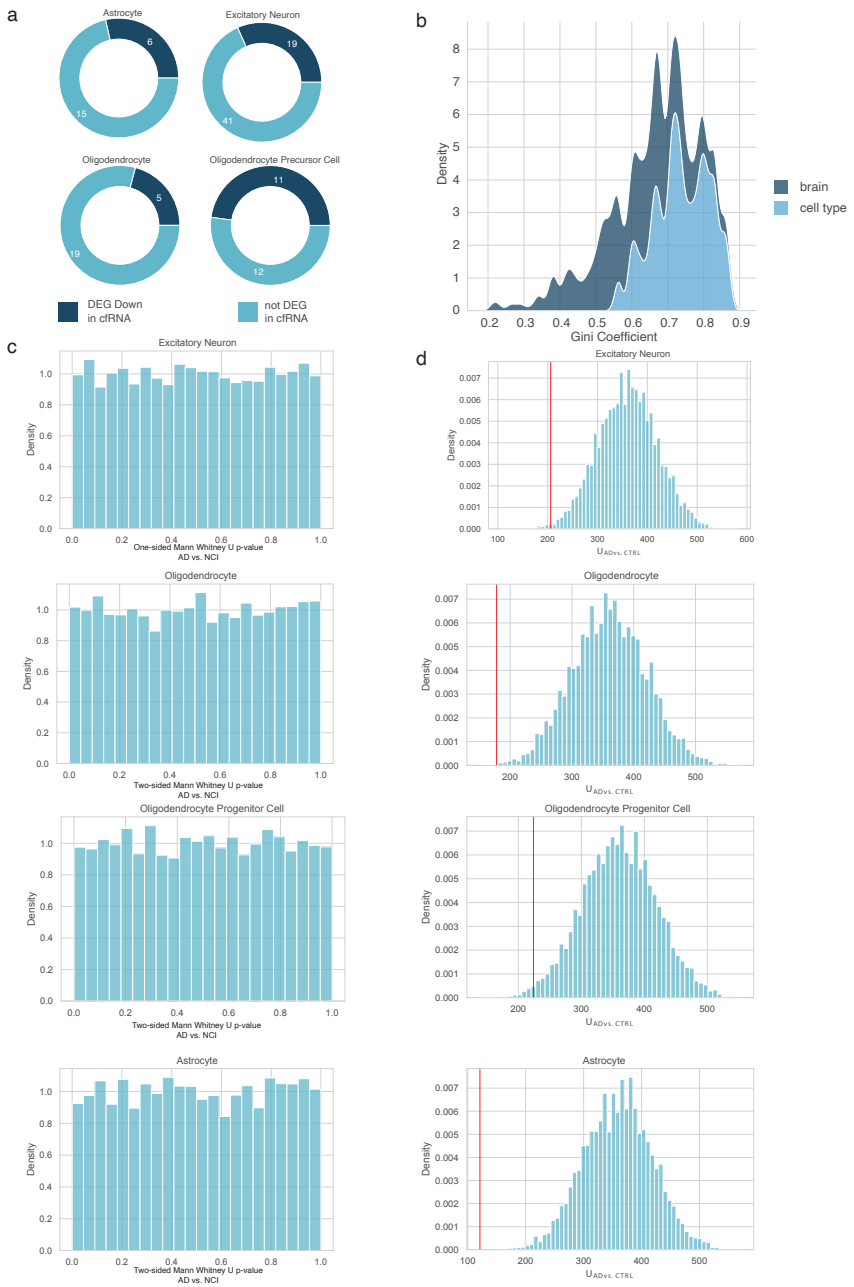



**Fig. S14. AD DEG intersection with neuronal/glial gene profiles and assessing their discriminatory power in signature scoring**

(a) Donut plots reflecting the number of genes in the hepatocyte cell type gene profile that intersect with the reported NAFLD DEG[7]

(b) Density plot reflecting the Gini coefficient distribution corresponding to DEG in AD[7] that are brain or brain cell type specific. The Gini coefficient is computed using the mean expression per brain cell type in the 'Normal' samples of Mathys et al (Methods). Area under each curve sums to one.

(c) Density of p-values over 10,000 trial permutation test

(d) Density of U values over 10,000 trial permutation test; red line indicates the U value corresponding to the experimental comparison reported in Fig. 2.