UCLA
UCLA Previously Published Works

Title

Cell types of origin of the cell-free transcriptome

Permalink

https://escholarship.org/uc/item/97z3r1ww

Journal

Nature Biotechnology, 40(6)

ISSN

1087-0156

Authors

Jones, Robert C
Karkanias, Jim
Krasnow, Mark
et al.

Publication Date

2022-06-01

DOI

10.1038/s41587-021-01188-9

Peer reviewed

OPEN

Cell types of origin of the cell-free transcriptome

Sevahn K. Vorperian[1,2], Mira N. Moufarrej[3], Tabula Sapiens Consortium* and Stephen R. Quake[3,4,5] ✉

Cell-free RNA from liquid biopsies can be analyzed to determine disease tissue of origin. We extend this concept to identify cell types of origin using the Tabula Sapiens transcriptomic cell atlas as well as individual tissue transcriptomic cell atlases in combination with the Human Protein Atlas RNA consensus dataset. We define cell type signature scores, which allow the inference of cell types that contribute to cell-free RNA for a variety of diseases.

Cell-free RNA (cfRNA) represents a mixture of transcripts reflecting the health status of multiple tissues[1], thereby affording broad clinical utility. Existing applications span oncology and bone marrow transplantation[2,3], obstetrics[1,4,5], 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[1,3,4,5–7]. Incorporating knowledge from cellular pathophysiology, which often forms the basis of disease[8], 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 using published exome-enriched cell-free transcriptome data[6] (Fig. 1a). After removing low-quality samples (Extended Data Fig. 1 and 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[9]. Marker genes for blood, brain, and liver cell types were readily detected, as previously observed at tissue level[1,3,4,6,7], as well as the kidney, gastrointestinal tract, 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[10,11]. We used Tabula Sapiens version 1.0 (TSP)[12], 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 rather the minimum discriminatory gene set to distinguish between the cell types in TSP. To reduce multicollinearity, transcriptionally similar cell types were grouped (Extended Data Fig. 2). We observed that the basis matrix defined by this gene set appropriately described cell types as most similar to others from the same organ compartment and corresponded to the highest off-diagonal similarity (Fig. 1c). We also confirmed that the basis matrix accurately deconvolved cell-type-specific RNA fractional contributions from several bulk tissue samples[13] (Extended Data Fig. 3 and Supplementary Information).

We used this matrix to deconvolve the cell types of origin in the plasma cell-free transcriptome (Fig. 1d and Extended Data Figs. 4 and 5). Platelets, erythrocyte/erythroid progenitors and leukocytes comprised the majority of observed signal, whose respective proportions were generally consistent with recent estimates from serum cfRNA[2] and plasma cfDNA[14]. Within this set of cell types, we suspect that the observation of platelets as a majority cell type, rather than megakaryocytes[2], likely reflects annotation differences in reference data. We observed distinct transcriptional contributions from solid tissue-specific cell types from the intestine, liver, lungs, pancreas, heart, and kidney (Fig. 1d and Extended Data Fig. 4). Altogether, the observation of contributions from many non-hematopoietic cell types underscores the ability to simultaneously non-invasively resolve contributions to cfRNA from disparate cell types across the body.

Some cell types likely present in the plasma cell-free transcriptome were missing in this decomposition because the source tissues were not represented in TSP. Although, ideally, reference gene profiles for all cell types would be simultaneously considered in this decomposition, a complete reference dataset spanning the entire cell type space of the human body does not yet exist. To identify cell type contributions possibly absent from this analysis, we intersected the genes measured in cfRNA missing from the basis matrix with tissue-specific genes from the Human Protein Atlas (HPA) RNA consensus dataset[15]. This identified both the brain and the testis as tissues whose cell types were not found during systems-level deconvolution and additional genes specific to the blood, skeletal muscle and lymphoid tissues that were not used by the basis matrix (Fig. 1e and Methods).

As an example of how to analyze cell type contributions from tissues that were not present in TSP, we used an independent brain single-cell atlas along with HPA to define cell type gene profiles and examined their expression in cfRNA (Fig. 2a and Extended Data Figs. 6 and 7). There was a strong signature score from excitatory neurons and a reduced signature score from inhibitory neurons. We observed strong signals from astrocytes, oligodendrocytes and oligodendrocyte precursor cells. These glial cells facilitate brain homeostasis, form myelin and provide neuronal structure and support[8], consistent with evidence of RNA transport across and the permeability of the blood–brain barrier[16,17] and that some brain regions are in direct contact with the blood[18]. Similarly, we used published cell atlases for the placenta[19,20], kidney[21] and liver[22] to define cell-type-specific gene profiles (Extended Data Figs. 6 and 8) for signature scoring. These observations augment the resolution of previously observed tissue-specific genes reported to date in cfRNA[1–7] and formed a baseline from which to measure aberrations in disease.

Cell-type-specific changes drive disease etiology[8], and we asked whether cfRNA reflected cellular pathophysiology. We considered trophoblasts in preeclampsia[23,24], proximal tubules in chronic kidney
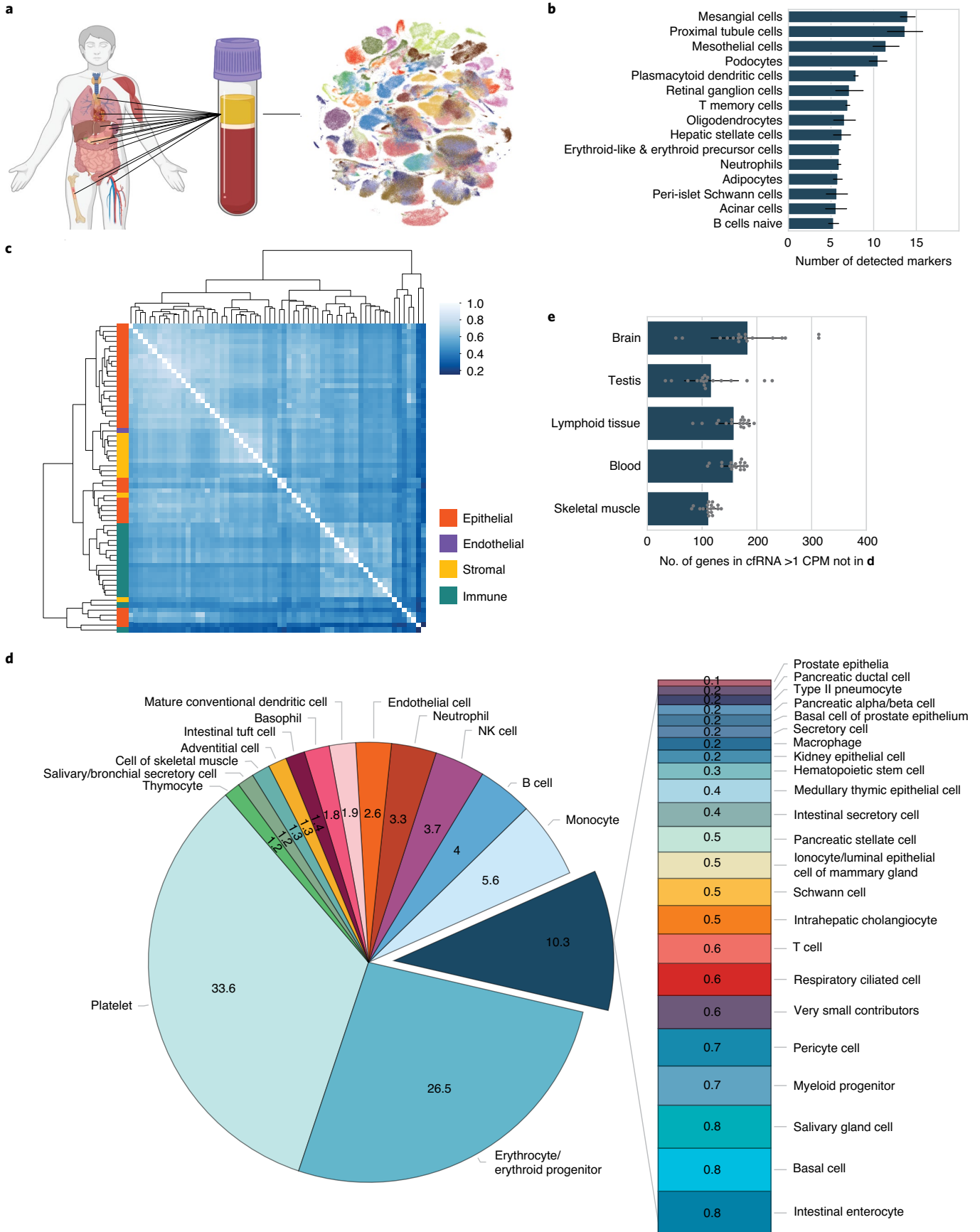
a

b

c

e

d

Fig. 1 | Cell type decomposition of the plasma cell-free transcriptome using Tabula Sapiens. a, Integration of tissue of origin and single-cell transcriptomics to identify cell types of origin in cfRNA. **b**, Cell-type-specific markers defined in context of the human body identified in plasma cfRNA. Error bars denote the s.d. of number of cell-type-specific markers ($n = 75$ patients); the measure of center is the mean. CPM-TMM counts for a given gene across technical replicates were averaged before intersection. **c**, Cluster heat map of Spearman correlations of the cell type basis matrix column space derived from Tabula Sapiens. Color bar denotes correlation value. **d**, Mean fractional contributions of cell-type-specific RNA in the plasma cell-free transcriptome ($n = 18$ patients). **e**, Top tissues in cfRNA not captured by basis matrix (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). Error bars denote the s.d. of number of HPA tissue-specific genes with NX counts >10 and cell-free CPM expression ≥ 1 ($n = 18$ patients); the measure of center is the mean.

disease (CKD)[25,26], hepatocytes in non-alcoholic steatohepatitis (NASH)/non-alcoholic fatty liver disease (NAFLD)[27] and multiple brain cell types in Alzheimer's disease (AD)[28,29]. As an example of why whole-body cell type characterization is relevant, we observed that a previous attempt to infer trophoblast cell types from cfRNA in preeclampsia[24] used genes that are not specific or readily measurable within their asserted cell type (Extended Data Fig. 9 and Supplementary Information). However, we found several other cases where cellular pathophysiology can be measured in cfRNA.

The proximal tubule is a highly metabolic, predominant kidney cell type and is a major source for injury and disease progression in CKD[25,26]. Tubular atrophy is a hallmark of CKD nearly independent of disease etiology[30] and is superior to clinical gold standard as a predictor of CKD progression[31]. Using data from Ibarra et al., we discovered a striking decrease in the proximal tubule cell signature score of patients with CKD (ages 67–91 years, CKD stage 3–5 or peritoneal dialysis) compared to healthy controls (Fig. 2b and Extended Data Fig. 10a,b). These results demonstrate non-invasive resolution of proximal tubule deterioration observed in CKD histology[31] and are consistent with findings from invasive biopsy.

Hepatocyte steatosis is a histologic hallmark of NASH and NAFLD phenotypes, whereby the accumulation of cellular stressors results in hepatocyte death[27]. We found that several genes differentially expressed in NAFLD serum cfRNA[7] were specific to the hepatocyte cell type profile derived above ($P < 10^{-10}$, hypergeometric test). Notable hepatocyte-specific differentially expressed genes (DEGs) include genes encoding cytochrome P450 enzymes (including *CYP1A2*, *CYP2E1* and *CYP3A4*), lipid secretion (*MTTP*) and hepatokines (*AHSG* and *LECT2*)[32]. We further observed striking differences in the hepatocyte signature score between healthy and both NAFLD and NASH cohorts and no difference between the NASH and NAFLD cohorts (Fig. 2c and Extended Data Fig. 10).

AD pathogenesis results in neuronal death and synaptic loss[29]. We used brain single-cell data[28] to define brain cell type gene profiles in both the AD and the normal brain. Several DEGs found in cfRNA analysis of AD plasma are brain cell type specific ($P < 10^{-5}$, hypergeometric test). Astrocyte-specific genes include those that encode filament protein (*GFAP*[33]) and ion channels (*GRIN2C*[28]). Excitatory neuron-specific genes encode solute carrier proteins (*SLC17A7*[28] and *SLC8A2*[34]), cadherin proteins (*CDH8*[35] and *CDH22*[36]) and a glutamate receptor (*GRM1*[29,37]). Oligodendrocyte-specific genes encode proteins for myelin sheath stabilization (*MOBP*[29]) and a synaptic/axonal membrane protein (*CNTN2*[29]). Oligodendrocyte-precursor-cell-specific genes encode transcription factors (*OLIG2*[38] and *MYT1*[39]), neural growth and differentiation factor (*CSPG5*[40]) and a protein putatively involved in brain extracellular matrix formation (*BCAN*[41]).

We then inferred neuronal death in plasma cfRNA between AD and healthy non-cognitive controls (NCIs) and also observed differences in oligodendrocyte, oligodendrocyte progenitor and astrocyte signature scores (Fig. 2d and Extended Data Fig. 10). The oligodendrocyte and oligodendrocyte progenitor cells signature score directionality agrees with reports of their death and inhibited proliferation in AD, respectively[42]. The observed astrocyte signature score directionality is consistent with the cell type specificity of a subset of
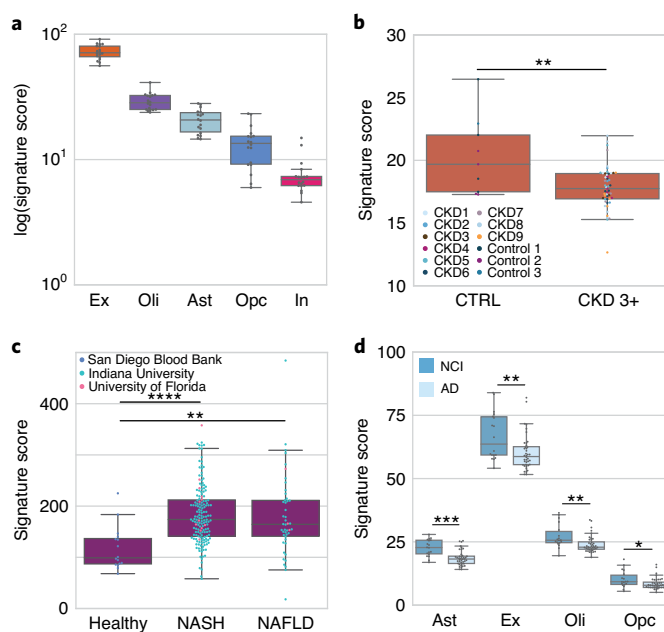

Fig. 2 | Cellular pathophysiology is non-invasively resolvable in cfRNA. For a given box plot, any cell type signature score is the sum of log-transformed CPM-TMM normalized counts. The horizontal line denotes the median; the lower hinge indicates the 25th percentile; the upper hinge indicates the 75th percentile; whiskers indicate the 1.5 interquartile range; and points outside the whiskers indicate outliers. All P values were determined by a Mann–Whitney U-test; sidedness is specified in the subplot caption. $*P < 0.05$, $**P < 10^{-2}$, $***P < 10^{-4}$, $****P < 10^{-5}$. **a**, Neuronal and glial cell type signature scores in healthy cfRNA plasma ($n = 18$) on a logarithmic scale. **b**, Comparison of the proximal tubule signature score in CKD stages 3+ ($n = 51$ samples; nine patients) and healthy controls ($n = 9$ samples; three patients) ($P = 9.66 \times 10^{-3}$, $U = 116$, one sided). Dot color denotes each patient. **c**, Hepatocyte signature score between healthy ($n = 16$) and both NAFLD ($n = 46$) ($P = 3.15 \times 10^{-4}$, $U = 155$, one sided) and NASH ($n = 163$) ($P = 4.68 \times 10^{-6}$, $U = 427$, one sided); NASH versus NAFLD ($P = 0.464$, $U = 3483$, two sided). Color reflects sample collection center. **d**, Neuronal and glial signature scores in AD ($n = 40$) and NCI ($n = 18$) cohorts. Excitatory neuron ($P = 4.94 \times 10^{-3}$, $U = 206$, one sided), oligodendrocyte ($P = 2.28 \times 10^{-3}$, $U = 178$, two sided), oligodendrocyte progenitor ($P = 2.27 \times 10^{-2}$, $U = 224$, two sided) and astrocyte ($P = 6.11 \times 10^{-5}$, $U = 121$, two sided). Ast, astrocyte; Ex, excitatory neuron; In, inhibitory neuron; Oli, oligodendrocyte; Opc, oligodendrocyte precursor cell.

reported downregulated DEGs[6] and reflects that astrocyte-specific changes, which are known in AD pathology[42], are non-invasively measurable.

Taken together, this work demonstrates consistent non-invasive detection of cell-type-specific changes in human health and disease using cfRNA. Our findings uphold and further augment the scope of previous work identifying immune cell types[2] and hematopoietic

tissues[1,2] as primary contributors to the cell-free transcriptome cell type landscape. Our approach is complementary to previous work using cell-free nucleosomes[14], which depends on a more limited set of reference chromatin immunoprecipitation sequencing data, which are largely at the tissue level[43]. Readily measurable cell types include those specific to the brain, lung, intestine, liver, and kidney, whose pathophysiology affords broad prognostic and clinical importance. Consistent detection of cell types responsible for drug metabolism (for example, liver and renal cell types) as well as cell types that are drug targets, such as neurons or oligodendrocytes for Alzheimer's-protective drugs, could provide strong clinical trial endpoint data when evaluating drug toxicity and efficacy. We anticipate that the ability to non-invasively resolve cell type signatures in plasma cfRNA will both enhance existing clinical knowledge and enable increased resolution in monitoring disease progression and drug response.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01188-9.

References

1. Koh, W. et al. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proc. Natl Acad. Sci. USA* **111**, 7361–7366 (2014).
2. Ibarra, A. et al. Non-invasive characterization of human bone marrow stimulation and reconstitution by cell-free messenger RNA sequencing. *Nat. Commun.* **11**, 400 (2020).
3. Larson, M. H. et al. A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat. Commun.* **12**, 2357 (2021).
4. Ngo, T. T. M., Moufarrej, M. N. & Rasmussen, M. L. H. Noninvasive blood tests for fetal development predict gestational age and preterm delivery. *Science* **360**, 1133–1136 (2018).
5. Munchel, S. et al. Circulating transcripts in maternal blood reflect a molecular signature of early-onset preeclampsia. *Sci. Transl. Med.* **12**, eaaz0131 (2020).
6. Toden, S. et al. Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing. *Sci. Adv.* **6**, eabb1654 (2020).
7. Chalasani, N. et al. Noninvasive stratification of nonalcoholic fatty liver disease by whole transcriptome cell-free mRNA characterization. *Am. J. Physiol. Gastrointest. Liver Physiol.* **320**, G439–G449 (2021).
8. Klatt, E. C. *Robbins & Cotran Atlas of Pathology* (Elsevier, 2021).
9. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* **2019**, baz046 (2019).
10. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
11. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
12. The Tabula Sapiens Consortium & Quake, S. R. The Tabula Sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors. Preprint at https://www.biorxiv.org/content/10.1101/2021.07.19.452956v1 (2021).
13. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
14. Sadeh, R. et al. ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nat. Biotechnol.* **39**, 586–598 (2021).
15. Uhlen, M. et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366**, eaax9198 (2019).
16. András, I. E. & Toborek, M. Extracellular vesicles of the blood–brain barrier. *Tissue Barriers* **4**, e1131804 (2016).
17. Abbott, N. J. Inflammatory mediators and modulation of blood–brain barrier permeability. *Cell. Mol. Neurobiol.* **20**, 131–147 (2000).
18. Ganong, W. F. Circumventricular organs: definition and role in the regulation of endocrine and autonomic function. *Clin. Exp. Pharmacol. Physiol.* **27**, 422–427 (2000).
19. Suryawanshi, H. et al. A single-cell survey of the human first-trimester placenta and decidua. *Sci. Adv.* **4**, eaau4788 (2018).
20. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
21. Stewart, B. J. et al. Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019).
22. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
23. Kaufmann, P., Black, S. & Huppertz, B. Endovascular trophoblast invasion: implications for the pathogenesis of intrauterine growth retardation and preeclampsia. *Biol. Reprod.* **69**, 1–7 (2003).
24. Tsang, J. C. H. et al. Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. *Proc. Natl Acad. Sci. USA* **114**, E7786–E7795 (2017).
25. Nakhoul, N. & Batuman, V. Role of proximal tubules in the pathogenesis of kidney disease. *Contrib. Nephrol.* **169**, 37–50 (2011).
26. Chevalier, R. L. The proximal tubule is the primary target of injury and progression of kidney disease: role of the glomerulotubular junction. *Am. J. Physiol. Renal Physiol.* **311**, F145–F161 (2016).
27. Feldstein, A. E. & Gores, G. J. Apoptosis in alcoholic and nonalcoholic steatohepatitis. *Front. Biosci.* **10**, 3093–3099 (2005).
28. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
29. Grubman, A. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).
30. Dhillon, P. et al. The nuclear receptor ESRRA protects from kidney disease by coupling metabolism and differentiation. *Cell Metab.* **33**, 379–394 (2021).
31. Schelling, J. R. Tubular atrophy in the pathogenesis of chronic kidney disease progression. *Pediatr. Nephrol.* **31**, 693–706 (2016).
32. Meex, R. C. R. & Watt, M. J. Hepatokines: linking nonalcoholic fatty liver disease and insulin resistance. *Nat. Rev. Endocrinol.* **13**, 509–520 (2017).
33. McCall, M. A. et al. Targeted deletion in astrocyte intermediate filament (Gfap) alters neuronal physiology. *Proc. Natl Acad. Sci. USA* **93**, 6361–6366 (1996).
34. Lytton, J. Na+/Ca2+ exchangers: three mammalian gene families control Ca2+ transport. *Biochem. J.* **406**, 365–382 (2007).
35. Friedman, L. G. et al. Cadherin-8 expression, synaptic localization, and molecular control of neuronal form in prefrontal corticostriatal circuits. *J. Comp. Neurol.* **523**, 75–92 (2015).
36. Arlotta, P. et al. Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo. *Neuron* **45**, 207–221 (2005).
37. Shigemoto, R., Nakanishi, S. & Mizuno, N. Distribution of the mRNA for a metabotropic glutamate receptor (mGluR1) in the central nervous system: an in situ hybridization study in adult and developing rat. *J. Comp. Neurol.* **322**, 121–135 (1992).
38. Zhou, Q., Choi, G. & Anderson, D. J. The bHLH transcription factor Olig2 promotes oligodendrocyte differentiation in collaboration with Nkx2.2. *Neuron* **31**, 791–807 (2001).
39. Nielsen, J. A., Berndt, J. A., Hudson, L. D. & Armstrong, R. C. Myelin transcription factor 1 (Myt1) modulates the proliferation and differentiation of oligodendrocyte lineage cells. *Mol. Cell. Neurosci.* **25**, 111–123 (2004).
40. Ichihara-Tanaka, K., Oohira, A., Rumsby, M. & Muramatsu, T. Neuroglycan C is a novel midkine receptor involved in process elongation of oligodendroglial precursor-like cells. *J. Biol. Chem.* **281**, 30857–30864 (2006).
41. Levine, J. M., Reynolds, R. & Fawcett, J. W. The oligodendrocyte precursor cell in health and disease. *Trends Neurosci.* **24**, 39–47 (2001).
42. Liddelow, S. A. et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).
43. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

Tabula Sapiens Consortium

Overall Project Direction and Coordination

Robert C. Jones[3], Jim Karkanias[6], Mark Krasnow[7,8], Angela Oliveira Pisco[6], Stephen R. Quake[3,4,6], Julia Salzman[7,9] and Nir Yosef[6,10,11,12]

Donor Recruitment

Bryan Bulthaup[13], Phillip Brown[13], William Harper[13], Marisa Hemenez[13], Ravikumar Ponnusamy[13], Ahmad Salehi[13], Bhavani A. Sanagavarapu[13] and Eileen Spallino[13]

Surgeons

Ksenia A. Aaron[14], Waldo Concepcion[13], James M. Gardner[15,16], Burnett Kelly[13,17], Nikole Neidlinger[13] and Zifa Wang[13]

Logistical coordination

Sheela Crasta[3,6], Saroja Kolluru[3,6], Maurizio Morri[6], Angela Oliveira Pisco[6], Serena Y. Tan[18], Kyle J. Travaglini[7] and Chenling Xu[10]

Organ Processing

Marcela Alcántara-Hernández[19], Nicole Almanzar[20], Jane Antony[21], Benjamin Beyersdorf[22], Deviana Burhan[23], Kruti Calcuttawala[24], Matthew M. Carter[19], Charles K. F. Chan[21,25], Charles A. Chang[26], Stephen Chang[7,22], Alex Colville[24,27], Sheela Crasta[3,6], Rebecca N. Culver[28], Ivana Cvijović[3,4], Gaetano D'Amato[29], Camille Ezran[7], Francisco X. Galdos[21], Astrid Gillich[7], William R. Goodyer[30], Yan Hang[26,31], Alyssa Hayashi[3], Sahar Houshdaran[32], Xianxi Huang[22,33], Juan C. Irwin[32], SoRi Jang[7], Julia Vallve Juanico[32], Aaron M. Kershner[21], Soochi Kim[24,26], Bernhard Kiss[21], Saroja Kolluru[3,6], William Kong[21], Maya E. Kumar[20], Angera H. Kuo[21], Rebecca Leylek[19], Baoxiang Li[34], Gabriel B. Loeb[35], Wan-Jin Lu[21], Sruthi Mantri[36], Maxim Markovic[3], Patrick L. McAlpine[14,37], Antoine de Morree[24,27], Maurizio Morri[6], Karim Mrouj[21], Shravani Mukherjee[34], Tyler Muser[20], Patrick Neuhöfer[7,38,39], Thi D. Nguyen[40], Kimberly Perez[19], Ragini Phansalkar[29], Angela Oliveira Pisco[6], Nazan Puluca[21], Zhen Qi[21], Poorvi Rao[23], Hayley Raquer-McKay[19], Nicholas Schaum[21,24], Bronwyn Scott[34], Bobak Seddighzadeh[41], Joe Segal[23], Sushmita Sen[32], Shaheen Sikandar[21], Sean P. Spencer[19], Lea Steffes[20], Varun R. Subramaniam[34], Aditi Swarup[34], Michael Swift[3], Kyle J. Travaglini[7], Will Van Treuren[19], Emily Trimm[29], Stefan Veizades[22,42], Sivakamasundari Vijayakumar[21], Kim Chi Vo[32], Sevahn K. Vorperian[1,3], Wanxin Wang[32], Hannah N. W. Weinstein[41], Juliane Winkler[44], Timothy T. H. Wu[7], Jamie Xie[41], Andrea R. Yung[7] and Yue Zhang[7]

Sequencing

Angela M. Detweiler[6], Honey Mekonen[6], Norma F. Neff[6], Rene V. Sit[6], Michelle Tan[6] and Jia Yan[6]

Histology

Gregory R. Bean[18], Vivek Charu[18], Erna Forgó[18], Brock A. Martin[18], Michael G. Ozawa[18], Oscar Silva[18], Serena Y. Tan[18], Angus Toland[18] and Venkata N. P. Vemuri[6]

Data Analysis

Shaked Afik[10], Kyle Awayan[6], Rob Bierman[7,9], Olga Borisovna Botvinnik[6], Ashley Byrne[6], Michelle Chen[3], Roozbeh Dehghannasiri[7,9], Angela M. Detweiler[6], Adam Gayoso[10], Alejandro A. Granados[6], Qiqing Li[6], Gita Mahmoudabadi[3], Aaron McGeever[6], Antoine de Morree[24,27], Julia Eve Olivieri[7,9,45], Madeline Park[6], Angela Oliveira Pisco[6], Neha Ravikumar[3], Julia Salzman[7,9], Geoff Stanley[3], Michael Swift[3], Michelle Tan[6], Weilun Tan[6], Alexander J. Tarashansky[6], Rohan Vanheusden[6], Sevahn K. Vorperian[1,3], Peter Wang[7,9], Sheng Wang[6], Galen Xing[6], Chenling Xu[9] and Nir Yosef[6,9,10,11]

Expert Cell Type Annotation

Marcela Alcántara-Hernández[19], Jane Antony[21], Charles K. F. Chan[21,25], Charles A. Chang[26], Alex Colville[24,27], Sheela Crasta[3,6], Rebecca Culver[28], Les Dethlefsen[46], Camille Ezran[7], Astrid Gillich[7], Yan Hang[26,31], Po-Yi Ho[19], Juan C. Irwin[32], SoRi Jang[7], Aaron M. Kershner[21], William Kong[21], Maya E. Kumar[20], Angera H. Kuo[21], Rebecca Leylek[19], Shixuan Liu[7,43], Gabriel B. Loeb[35], Wan-Jin Lu[21], Jonathan S. Maltzman[47,48], Ross J. Metzger[30,49], Antoine de Morree[24,27], Patrick Neuhöfer[7,38,39], Kimberly Perez[19], Ragini Phansalkar[29], Zhen Qi[21], Poorvi Rao[23], Hayley Raquer-McKay[19], Koki Sasagawa[22], Bronwyn Scott[34], Rahul Sinha[18,21,38], Hanbing Song[41], Sean P. Spencer[19], Aditi Swarup[34], Michael Swift[3], Kyle J. Travaglini[7], Emily Trimm[29], Stefan Veizades[22,42], Sivakamasundari Vijayakumar[21], Bruce Wang[23], Wanxin Wang[32], Juliane Winkler[44], Jamie Xie[41] and Andrea R. Yung[7]

Tissue Expert Principal Investigators

Steven E. Artandi[7,38,39], Philip A. Beachy[21,26,50], Michael F. Clarke[21], Linda C. Giudice[32], Franklin W. Huang[41,51], Kerwyn Casey Huang[3,19], Juliana Idoyaga[19], Seung K. Kim[26,31], Mark Krasnow[7,8], Christin S. Kuo[20], Patricia Nguyen[22,41,48], Stephen R. Quake[3,4,6], Thomas A. Rando[24,27], Kristy Red-Horse[29], Jeremy Reiter[52], David A. Relman[19,46,48], Justin L. Sonnenburg[19], Bruce Wang[23], Albert Wu[34], Sean M. Wu[22,42] and Tony Wyss-Coray[24,27]

[6]Chan Zuckerberg Biohub, San Francisco, CA, USA. [7]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA. [8]Howard Hughes Medical Institute, San Francisco CA, USA. [9]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [10]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. [11]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA. [12]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. [13]Donor Network West, San Ramon, CA, USA. [14]Department of Otolaryngology-Head and Neck Surgery, Stanford University School of Medicine, Stanford, CA, USA. [15]Department of Surgery, University of California, San Francisco, San Francisco, CA, USA. [16]Diabetes Center, University of California, San Francisco, San Francisco, CA, USA. [17]DCI Donor Services, Sacramento, CA, USA. [18]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [19]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. [20]Department of Pediatrics, Division of Pulmonary Medicine, Stanford University, Stanford, CA, USA. [21]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA. [22]Department of Medicine, Division of Cardiovascular Medicine, Stanford University, Stanford, CA, USA. [23]Department of Medicine and Liver Center, University of California, San Francisco, San Francisco, CA, USA. [24]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA. [25]Department of Surgery - Plastic and Reconstructive Surgery, Stanford University School of Medicine, Stanford, CA, USA. [26]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA, USA. [27]Division of Infectious Diseases & Geographic Medicine, Department

of Medicine, Stanford University, School of Medicine, Stanford, CA, USA. [28]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [29]Department of Biology, Stanford University, Stanford, CA, USA. [30]Department of Pediatrics, Division of Cardiology, Stanford University School of Medicine, Stanford, CA, USA. [31]Stanford Diabetes Research Center, Stanford University School of Medicine, Stanford, CA, USA. [32]Center for Gynecology and Reproductive Sciences, Department of Obstetrics, Gynecology and, Reproductive Sciences, University of California, San Francisco, San Francisco, CA, USA. [33]Department of Critical Care Medicine, The First Affiliated Hospital of Shantou University Medical College, Shantou, China. [34]Department of Ophthalmology, Stanford University School of Medicine, Stanford, CA, USA. [35]Division of Nephrology, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. [36]Stanford University School of Medicine, Stanford, CA, USA. [37]Mass Spectrometry Platform, Chan Zuckerberg Biohub, Stanford, CA, USA. [38]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. [39]Department of Medicine, Division of Hematology, Stanford University School of Medicine, Stanford, CA, USA. [40]Department of Biochemistry and Biophysics, Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA, USA. [41]Division of Hematology and Oncology, Department of Medicine, Bakar Computational Health Sciences Institute, Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. [42]Stanford Cardiovascular Institute, Stanford, CA, USA. [43]Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, CA, USA. [44]Department of Cell & Tissue Biology, University of California, San Francisco, San Francisco, CA, USA. [45]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. [46]Paul F. Glenn Center for the Biology of Aging, Stanford University School of Medicine, Stanford, CA, USA. [47]Division of Nephrology, Stanford University School of Medicine, Stanford, CA, USA. [48]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA. [49]Vera Moulton Wall Center for Pulmonary and Vascular Disease, Stanford University School of Medicine, Stanford, CA, USA. [50]Department of Urology, Stanford University School of Medicine, Stanford, CA, USA. [51]Division of Hematology/Oncology, Department of Medicine, San Francisco Veterans Affairs Health Care System, San Francisco, CA, USA. [52]Department of Biochemistry, University of California, San Francisco, San Francisco, CA, USA.

Methods

Data processing. *Data acquisition.* cfRNA: For samples from Ibarra et al. (PRJNA517339), Toden et al. (PRJNA574438) and Chalasani et al. (PRJNA701722), raw sequencing data were obtained from the Sequence Read Archive with the respective accession numbers. 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 the authors. Data from Mathys et al. were downloaded with permission from Synapse. The liver Seurat object was requested from Aizarani et al. For the placenta cell atlases, a Seurat object was requested from Suryawanshi et al., and AnnData was requested from Vento-Tormo et al. Kidney AnnData were downloaded (https://www.kidneycellatlas.org, Mature Full dataset).

HPA version 19 transcriptomic data, Genotype-Tissue Expression (GTEx) version 8 raw counts and Tabula Sapiens version 1.0 were downloaded directly.

Bioinformatic processing. All analyses were performed using Python (version 3.6.0) and R (version 3.6.1) For each sample for which raw sequencing data were downloaded, we trimmed reads using trimmomatic (version 0.36) and then mapped them to the human reference genome (hg38) with STAR (version 2.7.3a). Duplicate reads were then marked and removed by the MarkDuplicates tool in GATK (version 4.1.1). Finally, mapped reads were quantified using htseq-count (version 0.11.1), and read statistics were estimated using FastQC (version 0.11.8).

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

Sample quality filtering. For every sample for which raw sequencing data were available, we estimated three quality parameters as previously described[44,45]: RNA degradation, ribosomal read fraction and DNA contamination.

RNA degradation 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 approximated 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 GL00220.1:105,424–118,780, 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′ Bias Fraction: >0.4
- DNA Contamination: >3

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

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

$$\eta_{ij} = \frac{\text{Gene}_{ij}}{(\text{Library size}_j) \times (\text{mL plasma}_j)} \text{ where Library size}_j = \sum_i G_{ij} \quad (1)$$

For individuals who had samples with multiple technical replicates, these plasma volume CPM counts were averaged before nu support vector regression (nu-SVR) deconvolution.

For all analyses except nu-SVR (all work except Fig. 1d,e), we next applied trimmed mean of M values (TMM) normalization as previously described[46] using edgeR (version 3.28.1):

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

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 obtain this information from the authors; sequencing batches were confirmed with the authors (personal communication). All samples from Munchel et al. were used to compute TMM scaling factors, and 4.5 ml of 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 27 March 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 = 75, 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 on average 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 was available[12]. Disassociation genes as reported[12] were eliminated from the gene space before subsequent analysis.

Given the non-specificity of the following annotations (for example, 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 (for example, various intestinal enterocytes, T cells or dendritic cells), which, left unaddressed, would reduce the linear independence of the basis matrix column space and, hence, would affect 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 = 1×10^6)) 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 'treecutter.ipynb' on GitHub; the scipy version used is 1.5.1.

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 (Extended Data Fig. 2). To this end, only the 'endothelial cell' annotation was used 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 (that is, 'granulocyte', 'leucocyte' and 'immune cell') or present in only one tissue (that is, 'erythroid lineage cell'; eye, 'myeloid cell'; and 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 (for example, measured cells) were randomly sampled or the maximum number of available observations (if less than 30) was 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 DecontX adjusted layer were used to minimize signal spread contamination that could affect DEG analysis[12]. This subsampled counts matrix was then passed to the 'Create Signature Matrix' analysis module at https://cibersortx.stanford.edu/, with the following parameters:

- Disable quantile normalization = True
- Minimum expression = 0.25
- Replicates = 5
- Sampling = 0.5
- Kappa = 999
- q value = 0.01
- No. of barcode genes = 3,000–5,000
- 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 Fig. 1d and the Extended Data are 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'
- intestinal crypt stem cell: 'immature enterocyte/intestinal crypt stem cell/intestinal crypt stem cell of small intestine/transit amplifying cell of large intestine'
- 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[1,48]. With this formulation, we adapted existing deconvolution methods developed with the objective of decomposing a bulk tissue sample into its single-cell constituents[10,11], where the deconvolution problem is formulated as:

$$A\theta = b \quad (3)$$

Here, A is the representative basis matrix ($g \times c$) of g genes for c cell types, which represent the gene expression profiles of the c cell types. θ is a vector ($c \times 1$) of the contributions of each of the cell types, and b is the measured expression of the genes observed in blood plasma ($g \times 1$). The goal here is to learn θ 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'.

We performed nu-SVR using a linear kernel to learn θ from a subset of genes from the basis matrix to best recapitulate the observed signal b, where nu corresponds to a lower bound on the fraction of support vectors and an upper bound on the fraction of margin errors[49]. Here, the support vectors are the genes from the basis matrix used to learn θ; θ reflects the learned weights of the cell types in the basis matrix column space. For each sample, a set of θ was learned by performing a grid search on the two SVR hyperparameters: $\nu \in \{0.05, 0.1, 0.15, 0.25, 0.5, 0.75\}$ and $C \in \{0.1, 0.5, 0.75, 1, 10\}$.

For each sample, we next enforce two constraints: θ can contain only non-negative weights, and the weights in θ must sum to 1. Each θ corresponding to a hyperparameter combination was normalized as previously described in two steps[10,11]. First, only non-negative weights were kept:

$$\forall \theta_j < 0 \in \{\theta_1, ..., \theta_c\} \rightarrow 0 \quad (4)$$

Second, the remaining non-zero weights were then normalized by their sum to yield the relative proportions of cell-type-specific RNA.

We then determined the basis matrix dot product with the set of normalized weights for each sample. This dot product yields the predicted expression value for each gene in a given cfRNA mixture with imposed non-negativity on the normalized coefficient vector. The root mean square error (RMSE) was then computed using the predicted expression values and the measured values of these genes for each hyperparameter combination in a given cfRNA mixture. The model yielding the smallest RMSE in predicting expression for a given cfRNA sample was then chosen and assigned as the final deconvolution result for a given sample.

Only CPM counts ≥1 were considered in the mixture, b. The values in the basis matrix were also CPM normalized. Before deconvolution, the mixture and basis matrix were centered and scaled to zero mean and unit variance for improved runtime performance. We emphasize that we did not log-transform counts in b or in A, as this would destroy the requisite linearity assumption in equation (3). Specifically, the concavity of the log function would result in the consistent underestimation of θ during deconvolution[50].

We used the function nu-SVR from scikitlearn[51] version 0.23.2.

The samples used for nu-SVR deconvolution were 75 NCI patients from Toden et al. spanning four sample collection centers. Given center-specific batch effects reported by Toden et al., we report our results on a per-center basis (Fig. 1d and Extended Data Figs. 4 and 5). There was good pairwise similarity of the learned coefficients among biological replicates within and across sample centers (Extended Data Fig. 5a,b). Deconvolution performance yielded RMSE and Pearson r consistent with deconvolved GTEx tissues (Extended Data Fig. 3) whose distinct cell types were in the basis matrix column space (Extended Data Fig. 5c,d). In interpreting the resulting cell type fractions, a limitation of nu-SVR is that it uses highly expressed genes as support vectors and, consequently, assigns a reduced fractional contribution to cell types expressing genes at lower levels or that are smaller in cell volume. Comparison of nu-SVR to quadratic programming[1] and non-negative linear least squares[52] yielded similar deconvolution RMSE and Pearson correlation. In contrast to the other methods, nu-SVR cell type contributions were the most consistent with the cell type markers detected using PanglaoDB and was, hence, chosen as the deconvolution model for this work.

Evaluating basis matrix on GTEx samples. Bulk RNA sequencing samples from GTEx version 8 were deconvolved with the derived basis matrix from tissues that were present (that is, kidney cortex, whole blood, lung and spleen) or absent (for example, kidney medulla and brain) from the basis matrix derived using Tabula Sapiens version 1.0. For each tissue type, the maximum number of available samples or 30 samples, whichever was smaller, was deconvolved. See Supplementary Note 1 for additional discussion.

Identifying tissue-specific genes in cfRNA absent from basis matrix. To identify cell-type-specific genes in cfRNA that were distinct to a given tissue, we considered the set difference of the non-zero genes measured in a given cfRNA sample with the row space of the basis matrix and intersected this with HPA tissue-specific genes:

$$(G_j - R) \cap HPA \quad (5)$$

where G_j is the gene set in the jth deconvolved sample, where a given gene in the set's expression was ≥1 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) comprised 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 (that is, 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 (1) differentially expressed within a given single-cell tissue atlas, (2) 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 (Extended Data Figs. 6 and 8).

(1) **Single-cell differential expression**

For data received as a Seurat object, conversion to AnnData (version 0.7.4) was performed by saving as an intermediate loom object (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 = 1×10^4) and 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'. The set of resulting DEGs with Benjamini–Hochberg-adjusted P values <0.01 whose ratio of the highest out-group percent expressed to in-group percent expressed <0.5 was selected 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 coefficiets 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] (Extended Data Fig. 7). The Gini coefficient better reflected the underlying distribution of gene expression tissue specificity than Tau (Extended Data Fig. 7) and, hence, were 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_j is the expression of a given gene in the ith tissue.

The Gini coefficient was computed as defined[53]:

$$\text{Gini} = \frac{n+1}{n} - \frac{2\sum_{i=1}^{n}(n+1-i)\,x_i}{n\sum_{i=1}^{n}x_i} \; ; \; x_i \text{ is ordered from least to greatest.} \quad (6)$$

Tau, as defined in ref. [53]:

$$\tau = \frac{\sum_{i=1}^{n}1-\bar{x}}{n-1} \text{ where } \bar{x} = \frac{x_i}{\max(x_i)\,\forall i \in \{1...n\}} \quad (7)$$

HPA NX Counts from the HPA object titled 'rna_tissue_consensus.tsv' accessed on 1 July 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 (for example, not as close to unity compared to other cell type gene profiles) because 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 (Extended Data Figs. 6d and 8f,g). We uniformly observed a median fold change greater than 1 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 version 8 (accessed 26 August 2019) were 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. Because 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 DEGs in AD and NAFLD cfRNA. After observing a significant intersection between the DEGs in AD[6] or NAFLD[7] in cfRNA with corresponding cell-type-specific genes (Extended Data Fig. 10c,e), we then assessed the cell type specificity of DEGs 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 was defined using the HPA tissue transcriptional data annotated as 'Tissue enriched', 'Group enriched' or 'Tissue enhanced' (brain, accessed on 13 January 2021; liver, accessed on 28 November 2020).

These requirements ensured the specificity of a given brain/liver gene in context of the whole body. For a given tissue, this formed the initial set of tissue-specific genes B.

The union of all brain or liver 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:

$$C - B = 0 \quad (8)$$

Genes in B that did not intersect with C and intersected with DEG-up (U) or DEG-down (D) genes in a given disease[6,7] were then defined as 'tissue specific'.

$$T = (B \cap U) \cup (B \cap D) - C \quad (9)$$

The Gini coefficients reflecting the gene expression inequality across the cell types within corresponding tissue single-cell atlas were computed for the gene sets labeled as 'cell type specific' and 'tissue specific'. Brain reference data to compute Gini coefficients were from the single-cell brain atlas with diagnosis as 'Normal'[28]. Liver single cell data were used as-is[22]. 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 'cell type specific' and 'tissue specific'. The purpose of this test was to assess probability that observed mean difference in Gini coefficient for these two groups yielded no difference in specificity (that is, H_0: $\mu_{\text{cell type Gini coefficient}} = \mu_{\text{tissue Gini coefficient}}$).

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

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

where $\mu_{\text{observed}} := (\mu_{\text{cell type Gini coefficient}} - \mu_{\text{tissue Gini coefficient}})$.

The additional 1 in the denominator reflects the original test between the true difference in means (for example, the true comparison yielding μ_{observed}).

NAFLD: We considered the space of reported NAFLD DEGs 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 Normal 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, although 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 G in the jth patient sample:

$$\text{Signature score}_j = \sum_i G_{ij} \quad (11)$$

Preeclampsia. 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.

CKD. We compared the signature score of the proximal tubule in CKD (nine patients; 51 samples) and healthy controls (three patients; nine samples). Given that all patient samples were longitudinally sampled over ~30 d (individual samples were taken on different days), we treated the samples as biological replicates and included all time points because the time scale over which renal cell type changes typically occur is longer than the collection period. The sequencing depth was similar between the CKD and healthy cohorts, although it was reduced in comparison to the other cfRNA datasets used in this work. To account for gene measurement dropout, we required that the expression of a given gene in the proximal tubule gene profile was non-zero in at least one sample in both cohorts. Given that all samples were sequenced together, no batch correction was necessary, facilitating a representative comparison between CKD and healthy cohorts.

AD. Microglia, although 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. Brain gene profiles as defined in the AD section of 'Cell type specificity of DEGs in AD and NAFLD cfRNA' were used.

Assessing P value calibration for a given signature score. Cell type signature scores were tested between control and diseased samples with a Mann–Whitney U-test. The resulting P values were calibrated with a permutation test. Here, the labels compared in a given test (that is, CKD versus control, AD versus NCI, NAFLD versus control, etc.) were randomly shuffled 10,000 times. We observed a well-calibrated, uniform P-value distribution (Extended Data Fig. 10a), validating the experimentally observed test statistics.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article

Data availability

All datasets used for this work are publicly available, were downloaded with permission or were directly requested from the authors. Samples from Ibarra et al. (PRJNA517339), Toden et al. (PRJNA574438) and Chalasani et al. (PRJNA701722) were downloaded from the Sequence Read Archive with the respective accession numbers. Reads were mapped to the reference human genome (hg38). For data from Munchel et al., sample gene count tables were directly downloaded. Tissue gene lists and NX counts were downloaded from the Human Protein Atlas (www.proteinatlas.org, version 19). GTEx raw expression data were directly downloaded (https://www.gtexportal.org/home/datasets, GTEx analysis version 8). Tabula Sapiens was downloaded from the Chan Zuckerberg 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 were requested from Vento-Tormo et al. Kidney AnnData were downloaded (https://www.kidneycellatlas.org, Mature Full dataset). Source data are provided with this paper.

Code availability

Code for the work in this manuscript is available on GitHub at www.github.com/sevahn/deconvolution.

References

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. https://searchworks.stanford.edu/view/11686039 (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 (2012).
49. Chang, C.-C. & Lin, C.-J. Training ν-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 (2012).
51. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **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. Bioinform.* **18**, 205–214 (2017).
54. van Rossum, D. & Hanisch, U.-K. Microglia. *Metab. Brain Dis.* **19**, 393–411 (2004).

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 pre-processing pipeline to map reads to the human genome and cell-free sample quality control. S.K.V., M.N.M. and 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.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-021-01188-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01188-9.

Correspondence and requests for materials should be addressed to Stephen R. Quake.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
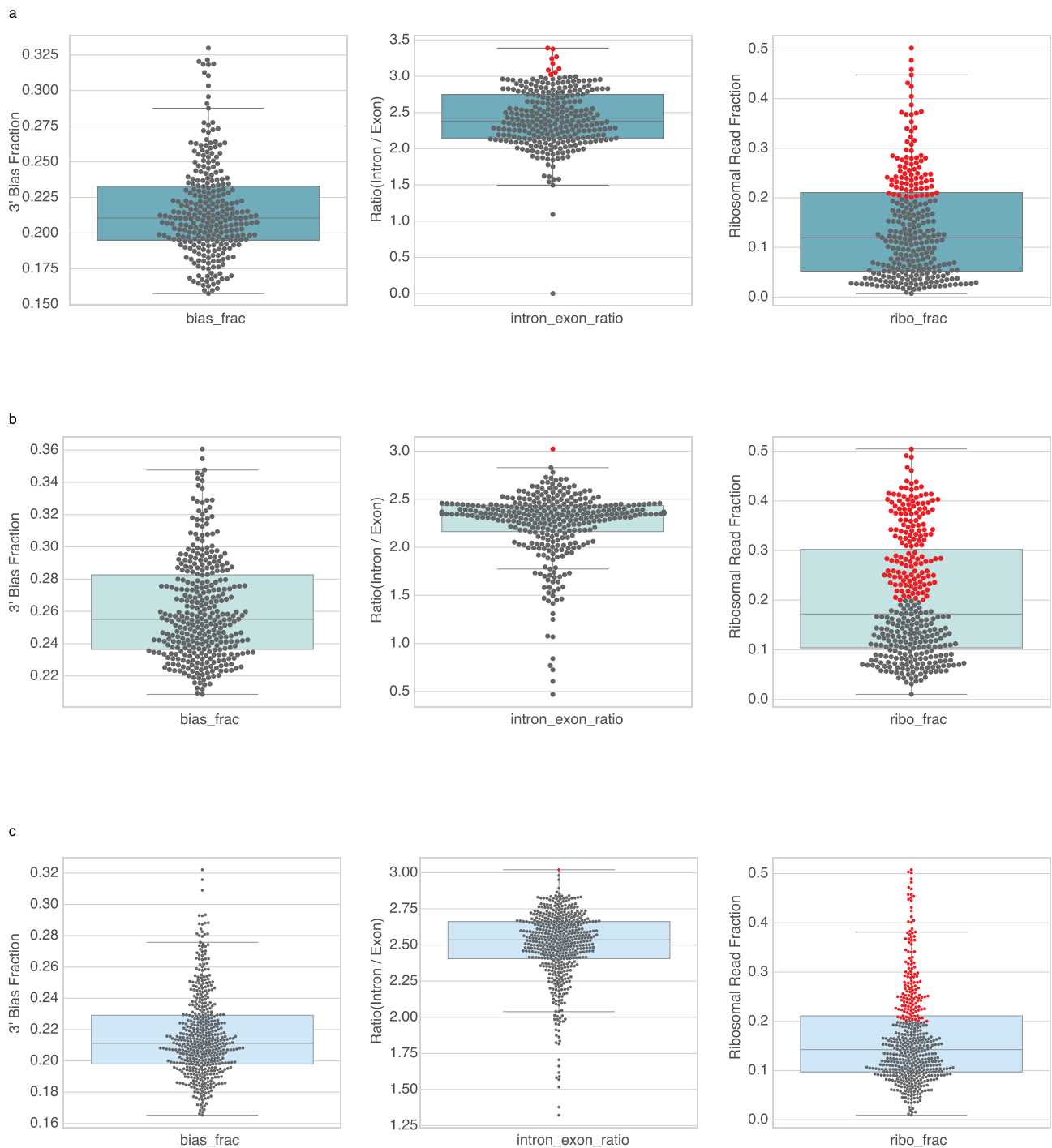
Extended Data Fig. 1 | Cell-free RNA Sample Quality Control. Quality control metrics (3' bias fraction, ribosomal fraction, and DNA contamination) were determined for each cfRNA sample downloaded from a given SRA accession number. Samples with outlier values are highlighted in red and were not considered in subsequent analyses (see Methods section 'Sample quality filtering'). **(a)** Ibarra et al (n = 285) **(b)** Toden et al (n = 339) **(c)** Chalasani et al (n = 500). Box plot: horizonal line, median; lower hinge, 25th percentile; upper hinge, 75th percentile; whiskers span the 1.5 interquartile range; points outside the whiskers indicate outliers. Each point corresponds to a downloaded cfRNA sample from the corresponding SRA accession number.

a

b

c
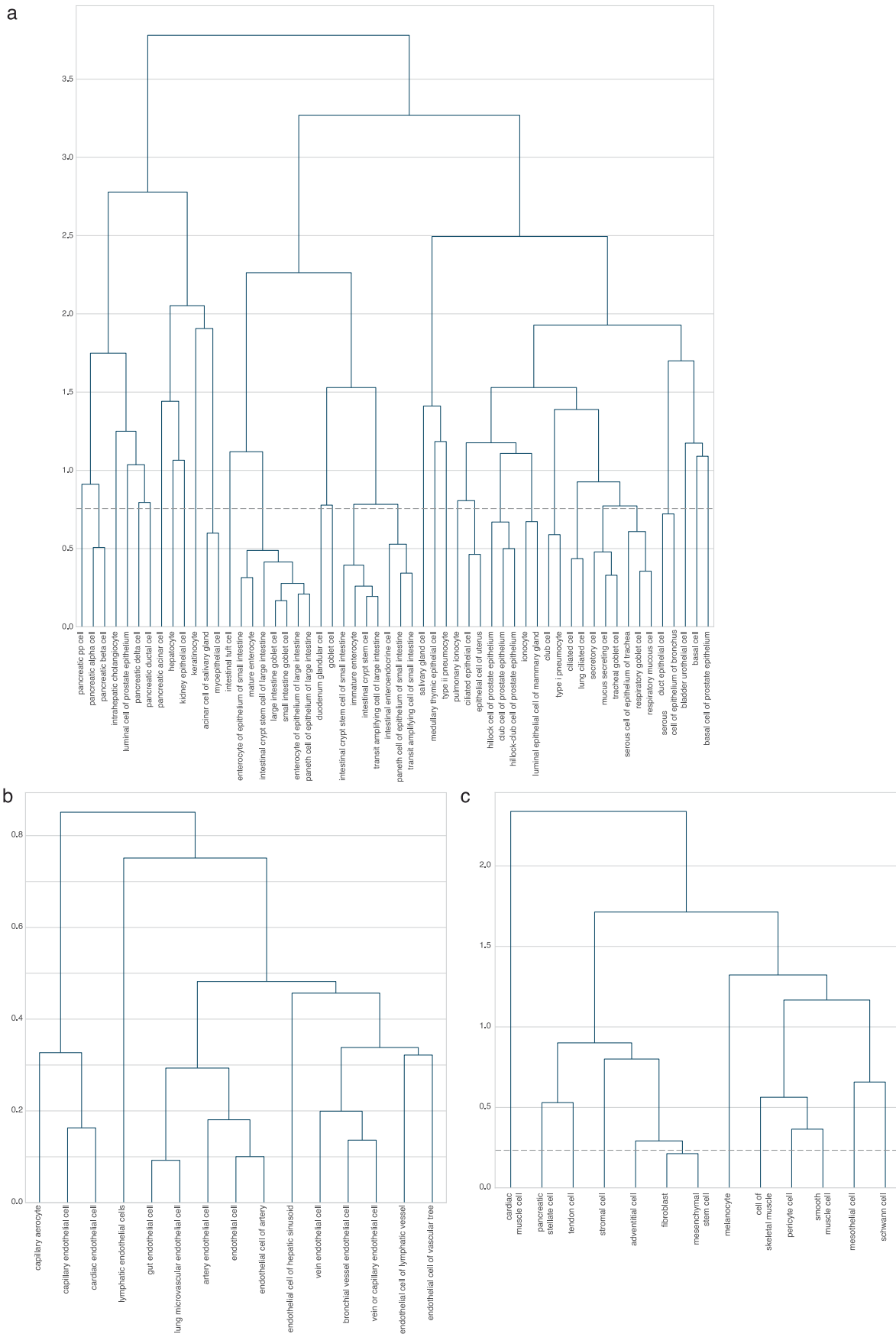
Extended Data Fig. 2 | Hierarchical clustering on non-immune Tabula Sapiens organ compartments. Dashed line indicates the height at which tree was cut. Dendrograms correspond with the cell type annotations belonging to **(a)** the epithelial compartment, **(b)** the endothelial compartment **(c)** the stromal compartment.
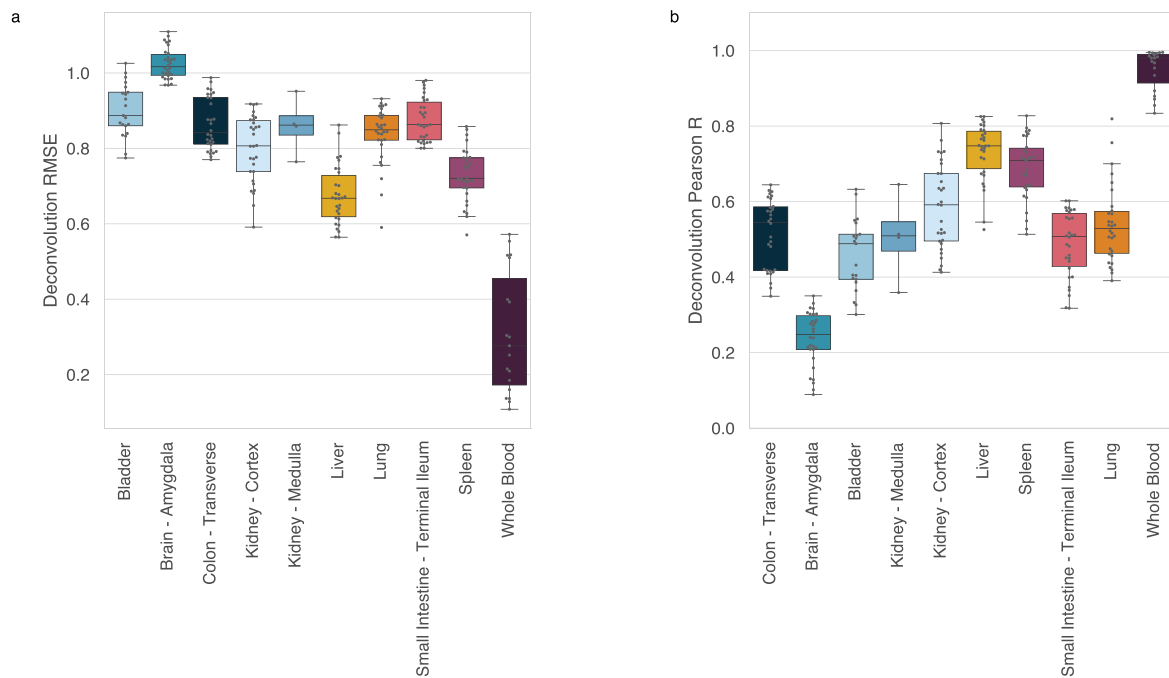
Extended Data Fig. 3 | Tabula Sapiens basis matrix performance on GTEx bulk RNA samples using nu-SVR. GTEx tissue samples possessing cell types wholly present and absent from the basis matrix column space were selected. For box plots: horizonal line, median; lower hinge, 25th percentile; upper hinge, 75th percentile; whiskers, 1.5 interquartile range; points outside the whiskers indicate outliers. There are 30 bulk RNA seq samples for a given tissue except for the Bladder (n = 21), Kidney – Medulla (n = 4), and Whole Blood (n = 19). **(a)** Root mean square error between predicted expression and measured expression in a given GTEx tissue. Units are zero-mean unit variance scaled CPM counts. Tissues present in TSP have reduced RMSE compared to those that are absent (Kidney – Medulla and Brain). Tissues with high cellular heterogeneity (for example Lung, Bladder, Small Intestine, Kidney) exhibit reduced deconvolution performance compared to less heterogeneous tissues (for example Whole Blood, Spleen, Liver). **(b)** Pearson correlation between predicted expression and measured expression in a given GTEx tissue.

a

Mature conventional dendritic cell
Basophil
Adventitial cell
Thymocyte
Intestinal tuft cell
Cell of skeletal muscle
Salivary gland cell
Myeloid progenitor
Salivary/Bronchial secretory cell
Endothelial cell
Neutrophil
B cell
NK cell
Monocyte
Platelet
Erythrocyte/Erythroid progenitor

2.1 | 2.2 | 2.3 | 2.4 | 2.8 | 3.7 | 4.2 | 5.7 | 5.9 | 9.3 | 10.4 | 16.2 | 27.5
1.6 | 1.3 | 1.1

0.2	Innate lymphoid cell
0.2	Type ii pneumocyte
0.2	Pancreatic/alpha beta cell
0.2	Respiratory secretory cell
0.2	Secretory cell
0.2	Prostate epithelia
0.2	Basal cell of prostate epithelium
0.4	Pericyte cell
0.4	Intestinal secretory cell
0.4	Tcell
0.5	Intestinal enterocyte
0.5	Very Small Contributors
0.6	Kidney epithelial cell
0.6	Medullary thymic epithelial cell
0.7	Ionocyte/luminal epithelial cell of mammary gland
0.7	Intrahepatic cholangiocyte
0.7	Pancreatic stellate cell
0.7	Schwann cell
0.8	Respiratory ciliated cell
0.9	Basal cell

b

Adventitial cell
Intestinal tuft cell
Endothelial cell
NK cell
B cell
Neutrophil
Monocyte
Platelet
Erythrocyte/erythroid progenitor

1. | 1.5 | 2.1 | 2.3 | 2.6 | 2.6 | 5. | 13.2 | 22.8 | 46.7

0.2	Myeloid progenitor
0.2	Basal cell of prostate epithelium
0.2	Intestinal secretory cell
0.2	Mast cell
0.2	Hematopoietic stem cell
0.3	Prostate epithelia
0.4	Secretory cell
0.4	Pancreatic stellate cell
0.4	Salivary gland cell
0.5	Pancreatic alpha/beta cell
0.5	Pericyte cell
0.5	Schwann cell
0.6	Cell of skeletal muscle
0.6	Respiratory ciliated cell
0.6	Ionocyte/luminal epithelial cell of mammary gland
0.6	Intrahepatic cholangiocyte
0.7	Intestinal enterocyte
0.8	Thymocyte
0.8	Basal cell
0.9	Mature conventional dendritic cell
0.9	Salivary/bronchial secretory cell
0.9	Basophil
0.9	T cell
0.9	Very Small Contributors

c

NK cell
B cell
Basophil
Mature conventional dendritic cell
Endothelial cell
Neutrophil
Cell of skeletal muscle
Adventitial cell
Intestinal tuft cell
Thymocyte
Salivary/Bronchial secretory cell
Monocyte
Platelet
Erythrocyte/erythroid progenitor

4.6 | 4.1 | 3.2 | 2.8 | 2.7 | 2. | 1.9 | 1.7 | 1.4 | 1.4 | 1.2 | 7.8 | 10.4 | 10.4 | 44.4

0.2	T cell
0.2	Type ii pneumocyte
0.3	Pancreatic alpha/beta cell
0.3	Secretory cell
0.3	Schwann cell
0.4	Pericyte cell
0.4	Ionocyte/luminal epithelial cell of mammary gland
0.5	Intestinal secretory cell
0.5	Intrahepatic cholangiocyte
0.6	Respiratory ciliated cell
0.6	Myeloid progenitor
0.6	Pancreatic stellate cell
0.7	Very Small Contributors
0.7	Basal cell
0.7	salivary gland cell
0.7	Intestinal enterocyte
0.9	Medullary thymic epithelial cell
0.9	Kidney epithelial cell
0.9	Macrophage

Extended Data Fig. 4 | Deconvolution of healthy plasma samples from Toden et al using Tabula Sapiens. Pie charts denote mean fractional cell type specific RNA contributions for **(a)** University of Indiana (n=17), **(b)** University of Kentucky (n=18), **(c)** Washington University in St. Louis (n=22).

Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | nuSVR decomposition of the plasma cell free transcriptome with Tabula Sapiens. For boxplots, horizonal line, median; lower hinge, 25th percentile; upper hinge, 75th percentile; whiskers span the 1.5 interquartile range; points outside the whiskers indi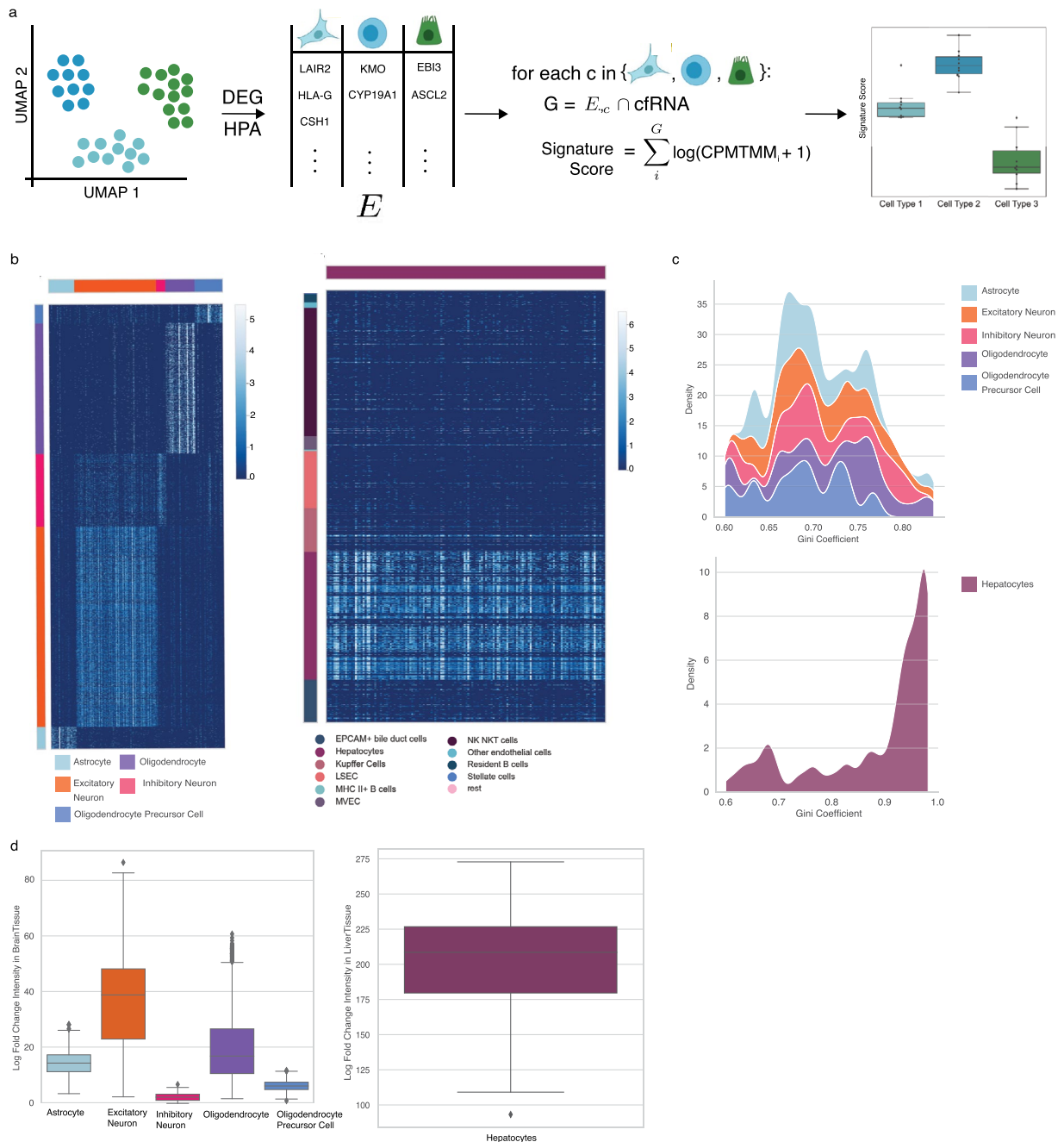cate outliers. Each point corresponds to a patient in a given cohort; University of Indiana (n=17), University of Kentucky (n=18), Washington University in St. Louis (n=22), and BioIVT (n=18). For heatmaps or clustermaps, the scale bar denotes the pearson correlation value. **(a)** Complete linkage clustermap of pairwise pearson correlation of deconvolved cell type fractions between patients from a given center; row color denotes a given center (n=75 patients). **(b)** Heatmap of pairwise pearson correlation of the mean cell type coefficients per center. **(c)** Deconvolution RMSE between predicted vs. measured expression for all biological replicates across all centers. **(d)** Deconvolution pearson correlation between predicted vs. measured expression for all biological replicates across all centers.

Extended Data Fig. 5 | nuSVR decomposition of the plasma cell free transcriptome with Tabula Sapiens.

Extended Data Fig. 6 | Establishing gene profile cell type specificity in context of the whole body using single cell and bulk RNA-seq data. (a) Cell type signature scoring procedure; please see the 'Signature Scoring' in the Methods for the full derivation procedure of a given cell type gene profile. **(b)** 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. The color bar scale corresponds to log-transformed counts-per-ten thousand. **(c)** 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. **(d)** 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. 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 (n = 2462 GTEx brain samples for box plot on left; n = 226 GTEx liver samples, right).
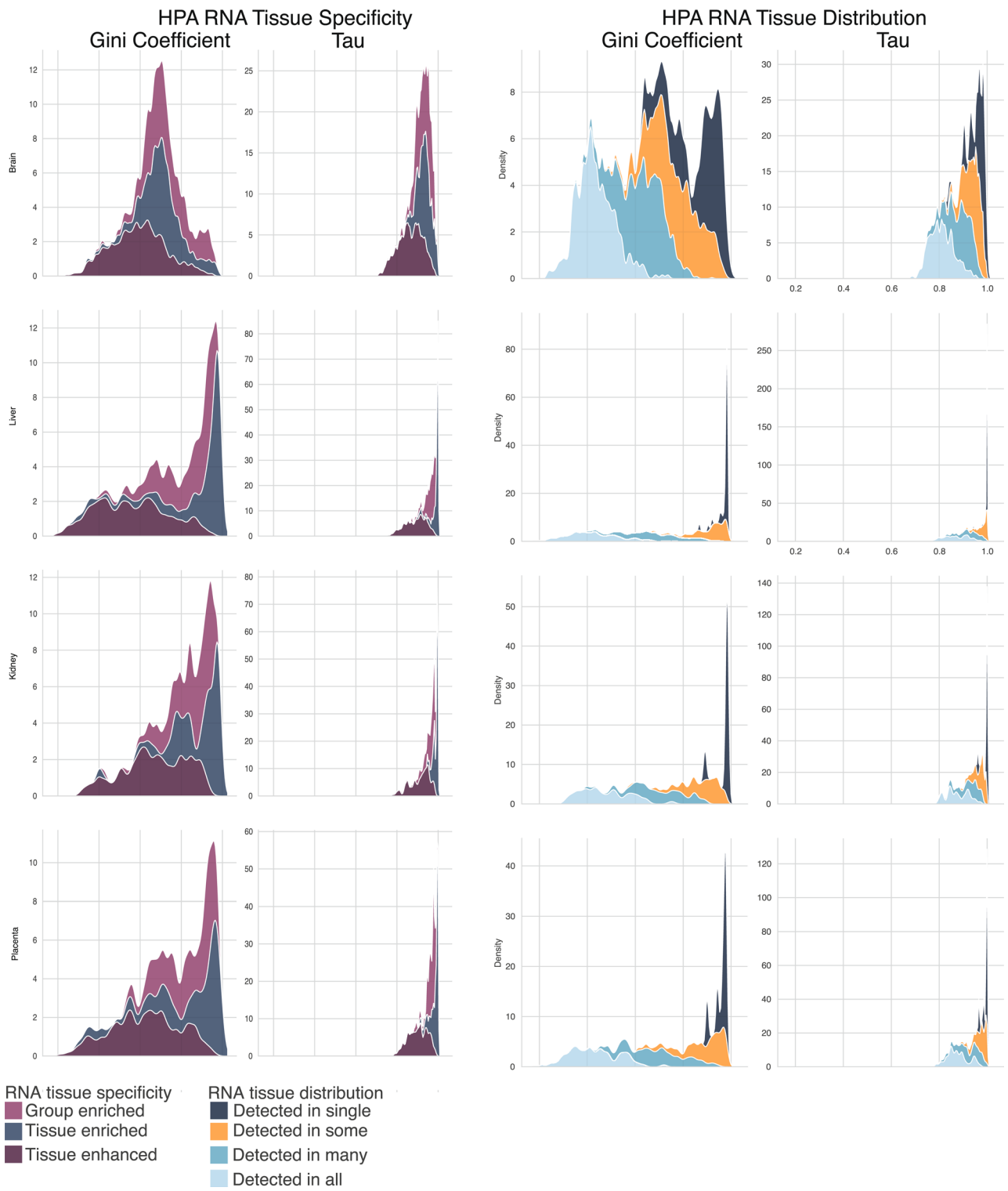
Extended Data Fig. 7 | Distribution of Gini coefficient and Tau for all genes denoted by HPA as specific to the brain, liver, placenta, and kidney.
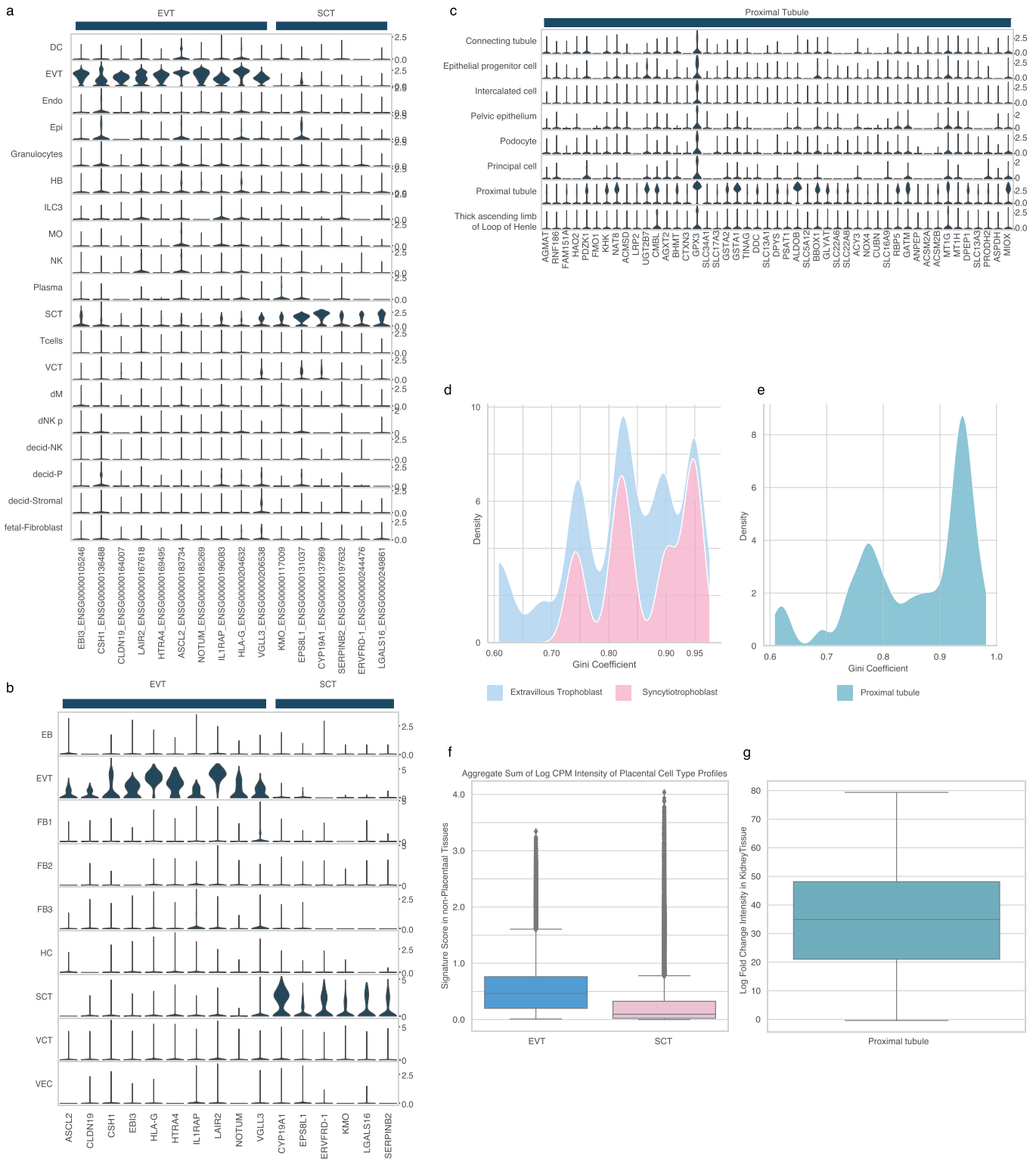
Extended Data Fig. 8 | Comprehensive placental and renal cell type gene profile specificity at single cell and whole body resolution. For box plots in f, g: horizontal line, median; lower hinge, 25th percentile; upper hinge, 75th percentile; whiskers span the 1.5 interquartile range; points outside whiskers 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) (n = 17382 non-placenta GTEx samples) **(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 signature score in the kidney to the mean signature score in the remaining non-kidney GTEx tissue samples (n = 89 GTEx renal cortex or medulla samples).
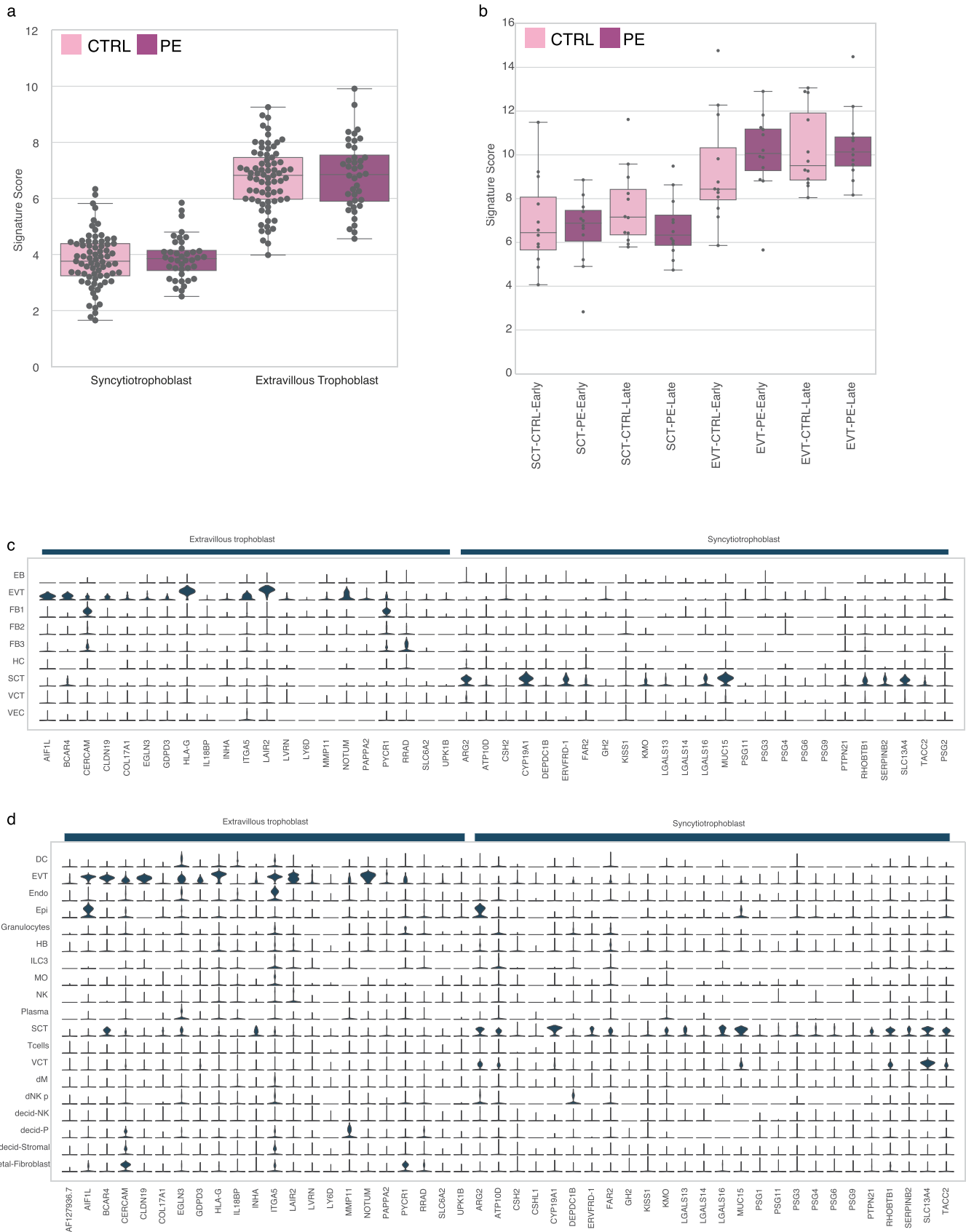
Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Expression distribution of Tsang et al trophoblast gene profiles in placenta scRNA atlases and in preeclampsia cfRNA. Derived trophoblast signature scores in the **(a)** iPEC dataset (mothers with no complications, n = 73 patients; mothers with preeclampsia, n = 40 patients) and **(b)** PEARL-PEC (n = 12 patients for each early/late-onset PE cohorts and gestationally- age matched healthy controls) datasets from Munchel et al. 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. 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 all observed placental cell types.
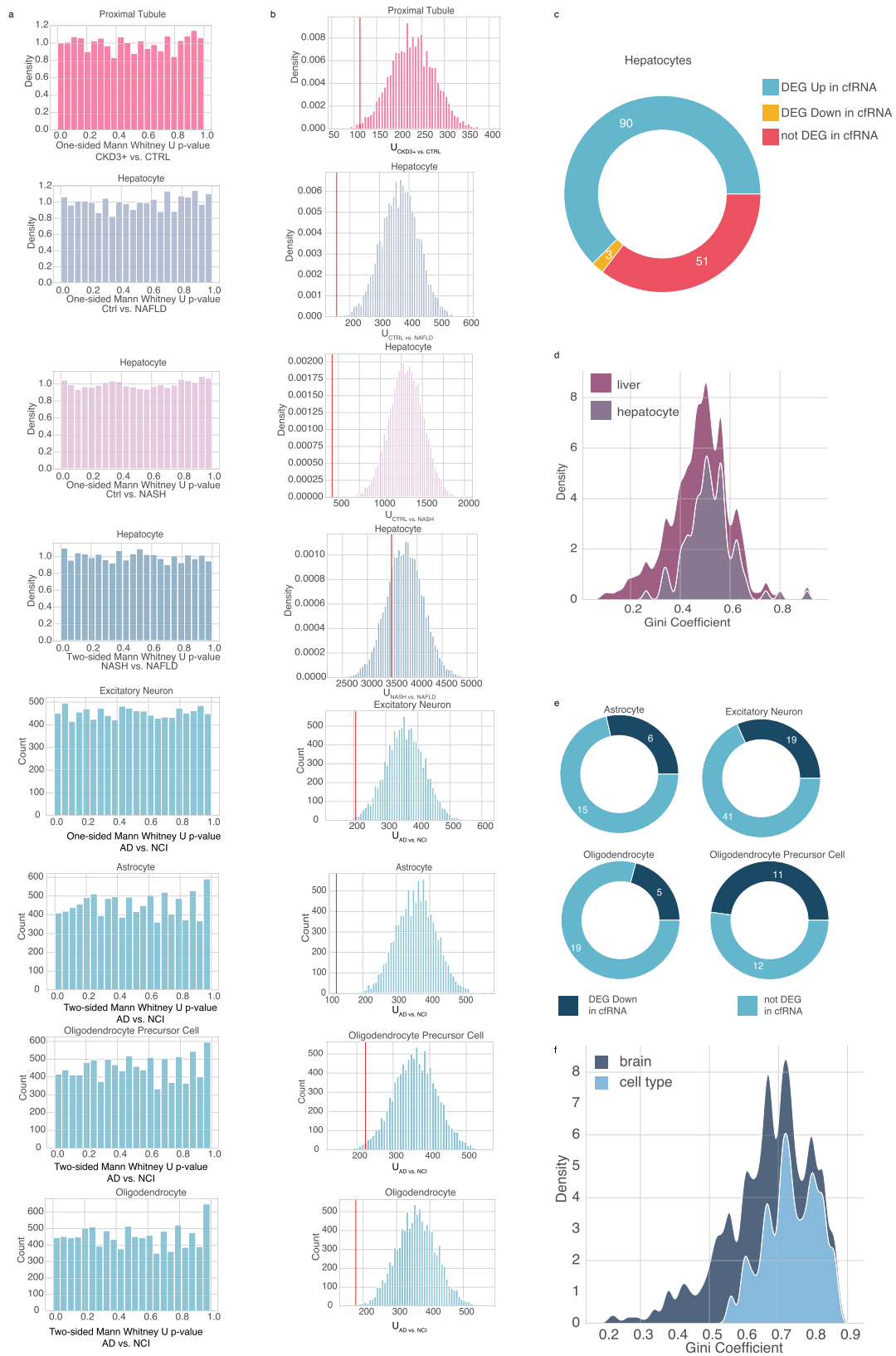
Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Assessment of cell type gene profile discriminatory power during signature scoring. (a) Density of p-values over 10,000 trial permutation test to assess p-value calibration for a given signature score. In all cases, the distribution is uniform, as expected under the null. **(b)** 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. **(c)** Donut plot reflecting the number of genes in the hepatocyte cell type gene profile that intersect with the reported NAFLD DEG in Chalasani et al. **(d)** Density plot reflecting the Gini coefficient distribution corresponding to DEG in NAFLD 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. **(e)** Donut plots reflecting the number of genes in brain cell type gene profiles that intersect with the reported AD DEG in Toden et al. **(f)** Density plot reflecting the Gini coefficient distribution corresponding to DEG in AD 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.

nature research

Corresponding author(s): Dr. Stephen R. Quake

Last updated by author(s): November 21, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data were collected in this study; all data used were from other studies.
Data analysis	All analyses were performed using Python (version 3.6) and R (version 3.6.1) Deconvolution was coded using scikitlearn (version 0.23.2) Bioinformatic processing: STAR (version 2.7.3a), GATK (version 4.1.1), htseq-count (version 0.11.1), FastQC (v ersion0.11.8), snakemake (version 5.8.1), MultiQC (version 1.7). Data structures: AnnData (version 0.7.4). Single cell objects received from authors as Seurat objects were converted to an intermediate loom file, loom files were read into python using loompy (version 3.0.6). Statistics: scipy (version 1.5.1) Single cell analysis: scanpy (version 1.6.0), Seurat (version 3.1.5) Basis matrix was generated using CIBERSORTx (cibersortx.stanford.edu) Normalization: in addition to built-in functions in scanpy, edgeR (version 3.28.1) for TMM normalization

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

All datasets used for this work were publicly available, downloaded with permission, or directly requested from authors.
Cell free RNA: For samples from Ibarra et al (PRJNA517339), Toden et al (PRJNA574438), and Chalasani et al (PRJNA701722), raw sequencing data was obtained from the SRA with the respective accession number. Reads were mapped to the reference human genome (hg38). For samples from Munchel et al, processed counts tables were directly downloaded.

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).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Cell free RNA samples: the datasets involved in this study were selected on the basis of availability, size, and high data quality. All datasets that we had access to meeting these three criteria were used in this work. No sample size calculation was performed; all samples used in this work were from published peer-reviewed studies. The entirety of the published samples passing QC were used in this work.
Data exclusions	Cell free RNA samples: we estimated the 3' bias ratio, ribosomal fraction, and the ratio of the number of reads that mapped to intronic as compared to exonic regions of the genome. A sample with a value greater than previously published thresholds for any of these three metrics was excluded from subsequent analysis. Single cell: a list of disassociation genes were eliminated prior to downstream analysis (e.g. differential expression) while working with the Tabula Sapiens data given that observed disassociation artifact in single cell data.
Replication	The cell free transcriptome in human health: We used several independent methods to assess the presence of cell-type specific signal, using cell type markers from PanglaoDB, systems-level deconvolution using Tabula Sapiens, and then individual cell type signatures scores derived from independent scRNA-seq tissue cell atlases. For systems level deconvolution on 75 healthy plasma samples, concordance was observed between the coefficients of cell type specific RNA between independent biological replicates between four different sample centers. For signature scoring and the cell type markers analyses, findings were again upheld over independent biological replicates. The cell free transcriptome in pathology: For the preeclampsia cell type signature scoring, we performed signature scoring using two independent datasets (PEARL-PEC and iPEC, from Munchel et al.). We validated our placental cell type signatures using two independent placental cell atlases (Munchel et al + Suryawanshi et al). All cell type signature scores were tested between control and sick samples with a Mann-Whitney U test. We ensured that the resulting p-values were calibrated with a permutation test. Here, the labels compared in a given test (i.e. CKD vs. CTRL, AD vs. NCI, NAFLD vs. CTRL, etc.) were randomly shuffled 10,000 times. We observed a well-calibrated, uniform p-value distribution, validating the experimentally observed test statistics. Of the differentially expressed genes that we observed to be cell type specific in AD/NAFLD, we performed a 10,000 trial permutation test on the Gini coefficients that are tissue-specific (e.g. brain/liver) vs. cell type specific. We found that the DEG that were identified as cell type specific possessed higher Gini than just tissue-specific. Together, this underscored that a subset of the DEG in cfRNA liquid biopsy for AD/NAFLD are associated with pathologically implicated cell types and are resolvable at cell type resolution. All attempts at replication were successful.
Randomization	Randomization was not relevant for this study. For the determination of the healthy cf-transcriptome landscape, we looked at the signal observed within a given sample independently, then compared the observed results between different patients. In disease, comparisons were made solely on the basis of patient disease status, no treatments were applied.

Blinding	All data used in this work were publicly available; no data collection was performed, hence no blinding was necessary. QC filtering of samples was performed blinded before revealing disease status of the samples. During data analysis, as purpose of this study was to determine the comprehensive landscape of cell type specific signal in cfRNA as a baseline from which to measure aberrations in Chronic Kidney Disease/Preeclampsia/Alzheimer's Disease/NAFLD and NASH, blinding was not performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	Antibodies
☒	Eukaryotic cell lines
☒	Palaeontology and archaeology
☒	Animals and other organisms
☒	Human research participants
☒	Clinical data
☒	Dual use research of concern

Methods

n/a	Involved in the study
☒	ChIP-seq
☒	Flow cytometry
☒	MRI-based neuroimaging