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

Liquid biopsies using cell-free RNA (cfRNA) can noninvasively measure dynamic physiological changes throughout the body. While there is much effort in the liquid biopsy field to determine disease tissue-of-origin, pathophysiology occurs at the cellular level. Here, we show that it is possible to determine cell type-of-origin from cfRNA by leveraging single cell transcriptomic atlases to perform computational deconvolution. We derived cell type gene signatures by combining the whole-body single cell atlas *Tabula Sapiens*, individual tissue single cell atlases, and bulk tissue atlases. Using deconvolution, we identified cell types-of-origin in the healthy human cell-free transcriptome, including contributions from multiple cell types in the brain, liver, lung, intestine, kidney, and pancreas in addition to hematopoietic cell types. We further showed that it is possible not only to detect cell types implicated in the pathology of chronic kidney disease (CKD) and Alzheimer's disease (AD), but also to measure changes in these cell types as a function of disease state. Altogether, our results show that cfRNA measurements reflect cellular contributions in health and disease from diverse tissue-specific cell types. These findings underscore the resolution at which one can monitor pathophysiological changes and the broad potential prognostic utility afforded by non-invasive transcriptomic measurement.

Introduction

Cell-free RNA (cfRNA) in blood plasma enables dynamic and longitudinal phenotypic insight into diverse physiological conditions, spanning oncology and bone marrow transplantation[1], obstetrics[2,3], neurodegeneration[4], and liver disease[5] . Liquid biopsies that measure cfRNA afford broad clinical utility since cfRNA represents a mixture of transcripts that reflects the health status of multiple tissues. However, several aspects about the physiologic origins of cfRNA including the contributing cell types-of-origin remain unknown, and most current assays focus on tissue level contributions[2–46,7]. Although information about tissue-of-origin can provide insight into transcriptional changes at a disease site, it would be even more powerful to incorporate knowledge from cellular pathophysiology which often forms the basis of disease[8]. This would also more closely match the resolution afforded by invasive biopsy.

Single cell transcriptomics (scRNA-seq) enable insight into the heterogeneous cellular transcriptional landscapes of tissues in health and disease[9]. Numerous scRNA-seq tissue atlases provide powerful reference data for defining cell type specific gene profiles in the context of an individual tissue. However, the starting set of cell types influences a differential expression analysis, which guides the assignment of a gene as cell type specific. cfRNA originates from cell types across the human body. Therefore, interpreting a measured gene in cfRNA as cell type specific relies on the completeness of relevant atlases. The *Tabula Sapiens* (TSP) cell atlas[10] from 14 tissues enables the most comprehensive

derivation of cell type specific gene profiles in the context of a single individual to date, all determined with uniform methods and sequencing, and we used this resource for our deconvolution process. For cell types originating from tissues absent from the draft TSP atlas, we derived specific gene profiles by combining a given single tissue cell atlas with comprehensive bulk transcriptomic datasets, including the Genotype-Tissue Expression (GTEx) project[11] and the Human Protein Atlas (HPA)[12].

In this work, we defined cell type specific gene profiles in the context of the whole body to identify the cell types comprising the cf-transcriptome. First, we computationally deconvolved the cell types-of-origin in the healthy human cf-transcriptome using the TSP cell atlas and individual scRNA-seq tissue atlases. Next, we measured striking cfRNA changes associated with cell types implicated in chronic kidney disease (CKD) and Alzheimer's disease (AD) that are consistent with observed clinical pathology. Altogether, we demonstrate that it is possible to decompose the cf-transcriptome into distinct cell type contributions even in the absence of a complete whole body single cell reference, and demonstrate that cell type specific changes in disease can be measured noninvasively using cfRNA.

Results

Deconvolution of cell type specific signals in the healthy cell free transcriptome

We used published exome-enriched cf-transcriptome data[1] to characterize the landscape of cell type specific signal in the plasma of healthy individuals (Fig. 1A). After eliminating low-quality cfRNA samples (Fig. S1, Methods), we intersected the set of genes detected in healthy individuals (n = 5) with a database of cell-type specific markers defined in context of the whole body[13] with stringent expression requirements (Fig. 1B, Methods). Marker genes for cell types originating from the blood, brain, and liver were readily detected, as previously observed at tissue level[1,3–5]. Kidney, GI track, and pancreas cell type markers were additionally detected (Fig. 1B).

Given the robust detection of several cell types contributing to the cf-transcriptome, we then deconvolved the fractions of cell-type specific RNA using TSP. We defined the cf-transcriptome as a linear combination of cell type specific RNA contributions using a deconvolution method, nu-SVR, originally developed to decompose bulk tissue transcriptomes into fractional cell type components[14,15] (fig. S2). This required specifying a basis matrix with a representative gene set (rows) that could accurately and simultaneously resolve the distinct cell types (columns). To reduce multicollinearity, we grouped transcriptionally similar cell types (Methods). We observed that the basis matrix appropriately described cell types as most similar to others from the same organ compartment, where cell types originating from the same compartment cluster together and correspond to the highest off-diagonal similarity (Fig. 1C). We also confirmed that the defined basis matrix can correctly deconvolve cell type specific RNA fractional contributions from several GTEx bulk tissue samples (fig. S3, S4, Supplementary Note).

We then deconvolved the cell types-of-origin contributing to the plasma cf-transcriptome (Fig. 1D). We observed a large signal from hematopoietic cell types, as well as smaller, distinct transcriptional contributions from tissue-specific cell types from the large and small intestine, lungs, and pancreas (Fig. 1D, fig. S5A, B). The highest cell type contributors were monocytes (18.6 ± 2.3%), platelets (13.6 ± 3.5 %), erythrocytes and erythroid progenitors (15.8 ± 9.1%), and lymphocytes (15.7 ± 2.7%). There was good pairwise similarity amongst all biological replicates ($r \geq 0.66$, fig. S5C). The predominant cell types and their respective proportions we observe are generally consistent with recently published estimates for serum cfRNA[1] and plasma cfDNA[16]. We also observed small fractional contributions from endothelial cells, pancreatic cells, intestinal enterocytes, kidney epithelia, club cells, goblet cells, pancreatic acinar

cells, and other pancreatic cells (Fig. 1D), underscoring the contributions of non-hematopoietic cell types to the cf-transcriptome.

Some cell types likely present in the plasma cf-transcriptome were not found in this decomposition because the source tissues were absent from the TSP version 0.9. Deconvolution performance yielded an elevated root mean square error (RMSE) (86 ± 5.2 CPM) and reduced Pearson correlation (0.43 ± 0.08) compared to deconvolved GTEx tissues whose cell types were completely in the basis matrix (fig. S3). To understand which cell type contributions might be absent from this present analysis, we intersected the genes measured in cfRNA but absent from the basis matrix with tissue specific genes annotated by the HPA transcriptomic atlas[12] (Methods). This identified brain, liver, testis, skeletal muscle, and cardiac muscle as tissues with several reliably measured genes in the plasma cfRNA (Fig. 1E, Methods) whose cell types were not found during systems-level deconvolution.

We then defined cell type specific gene profiles for these tissues in context of the whole body. To do so, we leveraged individual tissue cell atlases[17–19] but only considered cell types unique to a given tissue (fig. S6, Methods). This formulation allowed us to apply bulk GTEx and HPA transcriptomic data to ensure whole body specificity using stringent expression specificity constraints. First, we required a given gene to be differentially expressed in a given cell type against all others within an individual tissue cell atlas (Fig 2A, fig. S7A) (Methods). Second, we required high expression inequality across tissues measured by the Gini coefficient[20] (Fig 2B, fig. S7B & fig. S8) (Methods). We validated 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 2C, fig. S7C). 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.

Next, we estimated a signature score for each cell type in the cf-transcriptome using its specific gene profile by summing the measured level for all included genes (Fig. 2D), and observed contributions from multiple brain cell types, hepatocytes, and cardiomyocytes. Specifically, we measured a strong signature score from excitatory neurons and a reduced signature score from inhibitory neurons. We also 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]. Evidence of RNA transport across the blood brain barrier (BBB)[21], BBB permeability[22], and brain regions in direct contact with the blood[23] help rationalize brain cell type signature detection in the cf-transcriptome.

We additionally observed a strong hepatocyte signature score, which is consistent with their high turnover rate and cellular mass[24], a small signal for atrial cardiomyocytes, and negligible signal from ventricular cardiomyocytes, consistent with the low level of cardiomyocyte death in healthy adults[25] (Fig. 2D). These observations augment the resolution of previously observed brain-[3,4], liver-[5], and heart[26]-specific genes reported to date in cfRNA.

Plasma cfRNA measurement reflects cellular pathophysiology

Cell type specific changes drive disease etiology[8], and we asked whether cfRNA reflected changes in pathological cell types. We considered trophoblasts in preeclampsia[27,28], proximal tubules in CKD[29,30], and multiple brain cell types in AD[17,31].We utilized published cell atlases for the placenta[32,33], kidney[34], and brain[17] to define cell specific gene profiles in context of the whole body following the approach outlined above (fig. S7 and S9).

In pregnancy, extravillous trophoblast invasion is a stage in uteroplacental arterial remodeling[27,33]. Arterial remodeling occurs to ensure adequate maternal blood flow to the growing fetus[27,33] and is sometimes reduced in preeclampsia[27]. Previously, the extravillous trophoblast was reported by Tsang et al to be noninvasively resolvable and elevated in early onset preeclampsia (gestational age < 34 weeks) compared to healthy pregnancy[28]. The syncytiotrophoblast, involved in nutrient exchange[33], showed no difference[28]. We measured the respective signature scores in two previously published preeclampsia cohorts[35]. In contrast to Tsang et al, we observed no significant difference in either trophoblast signature score in cfRNA samples collected at diagnosis for mothers with early-onset preeclampsia (Fig. 3A) ($p = 0.703$, 0.794 respectively, two-sided Mann Whitney U) and for mothers with either early- or late-onset preeclampsia ($p = 0.24$, 0.54 respectively, Kruskal Wallace, fig. S10A) as compared to samples from mothers with no complications at a matched gestational age. Though we generally recapitulate the observed signature score directionality for both cell types[28], examination of the cell type gene profiles used by Tsang et al. in two independent placental atlases[32,33] revealed genes that were not cell type specific (fig. S10B,C). The presence of non-cell type specific genes in a cell type gene profile likely impaired Tsang et al's signature score interpretation. The role of extravillous trophoblast invasion and the ubiquity of its cellular pathophysiology in preeclampsia thus remains an open question.

As a second example, the proximal tubule is a highly metabolic, predominant kidney cell type and is a major source for injury and disease progression in CKD[29,30]. Tubular atrophy is a hallmark of CKD nearly independent of CKD etiology[36]. Atubular formation results from proximal tubule damage and death in renal pathologies[29,30]. We discovered a striking decrease in proximal tubule intensity of CKD ($n = 9$) patients (ages 67-91, CKD stage 3-5 or peritoneal dialysis) compared to healthy controls ($n = 7$) ($p = 7.45 * 10^{-4}$, one-sided Mann Whitney U) (Fig. 3B, Methods). Random sampling of a gene set of equivalent length as the proximal tubule gene profile in 10,000 trials yielded no discriminatory power between CKD and healthy 96.6% of the time (Benjamini-Hoschberg correction, FDR = 0.05, fig. S11) and an adjusted proximal tubule signature score of 0.038, validating the specificity of our proximal tubule gene profile. These findings demonstrate the ability to noninvasively resolve proximal tubule deterioration observed in CKD histology[37] and is consistent with findings from routine invasive biopsy.

As a third example, AD pathogenesis results in neuronal death and synaptic loss[31]. We derived brain cell type gene profiles for cell types in the AD brain in the same way as the NCI cell type profiles (Methods). We then intersected a given cell type gene profile in AD with the equivalent NCI profile for comparative analysis. Microglia, though often implicated in AD pathogenesis were excluded given their high overlapping transcriptional profile with non-central nervous system macrophages[38]. Inhibitory neurons were also excluded given the low number of cell type specific genes intersecting between AD and NCI phenotypes. Intersection of reportedly differentially downregulated AD genes (DEG) in plasma[4] with the derived cell type specific gene profiles, identified several genes as cell type specific (Fig. 3C). Astrocyte specific genes included filament protein (*GFAP*[39]) and ion channels (*GRIN2C* [17]). Excitatory neuron specific genes solute carrier proteins (*SLC17A7*[17] & *SLC8A2*[40]), cadherin proteins (*CDH8*[41] & *CDH22*[42]), and a glutamate receptor stimulated a major excitatory neuron neurotransmitter (*GRM1*[31,43]). Neuronal death in AD phenotypes[31], is likely the biochemical basis for the observed downregulation of these cell type specific markers. Oligodendrocyte-specific genes encode proteins for myelin sheath stabilization (*MOBP*[31]) and a synaptic/axonal membrane protein (*CNTN2*[31]*)*. Oligodendrocyte precursor cell-specific genes included transcription factors (*OLIG2*[44] & *MYT1*[45]), neural growth and differentiation factor (*CSPG5*[46]), and a protein putatively involved in brain extracellular matrix formation (*BCAN*[47]). A

177 permutation test using the Gini coefficients computed on the average single cell expression of the brain
178 specific and cell type specific DEGs corroborated that the DEGs assigned as cell type specific were more
179 specific to a given brain cell type than a brain specific DEG ($p < 1e-4$, 10000 trials, fig. S12) (Methods).
180 Taken together, these findings underscore the consistent detection cell-type specific changes in pathology
181 using noninvasive transcriptomic measurement of blood plasma and the resolution at which we can assert
182 the origins of cfRNA.
183
184 **Discussion**
185 We have shown that the cfRNA transcriptome can be reframed from a sum of tissue
186 transcriptomes to a sum of cell type transcriptomes. Using nu-SVR, we determined the fractional
187 contributions of cell type specific RNA relative to other cell types considered in the cell type column
188 space of the basis matrix[48]. Ideally reference gene profiles for all possible cell types in the human body
189 would be simultaneously considered in nu-SVR deconvolution. However, a complete reference dataset of
190 all cell types in the human body does not yet exist. Despite an incomplete cell atlas, we demonstrate the
191 ability to decompose the cf-transcriptome into distinct cell type contributions by further leveraging
192 individual cell atlases and bulk transcriptomic data from GTEx and HPA to define specificity in context
193 of the whole body.
194 The decomposition of the cf-transcriptome reveals platelets, lymphocytes, and monocytes as the
195 predominant cell type specific RNA contributors. This is consistent with what is known about the cf-
196 transcriptome, and may also reflect a bias in nu-SVR deconvolution, which uses highly expressed genes
197 as support vectors, and consequently assigns a reduced fractional contribution to cell types expressing
198 genes at lower levels or that are smaller in size, such as neutrophils. Furthermore, our finding that
199 platelets are a majority cell type, rather than megakaryocytes[1], likely reflects annotation differences in
200 reference data. Megakaryocytes are absent from the TSP v0.9 annotations; however, they are responsible
201 for platelet production[8]. Comparison of nu-SVR to quadratic programming[3] and non-negative linear least
202 squares[49] yielded similar deconvolution RMSE and slightly increased Pearson correlation. However, the
203 determined fractions with these methods excluded contributions from cell types with markers detected
204 using PanglaoDB, and so we chose nu-SVR for the comprehensive deconvolution in this work. Taken
205 together, these findings are consistent with prior work considering specific cell types[1] and the
206 hematopoietic tissues[1,3]. Shared features among the cell types contributing to the to the cf-transcriptome
207 are large volume and/or increased turnover[24], suggesting cell death as the possible predominant entry
208 mechanism of cfRNA to the bloodstream.
209 To identify contributions from cell types absent from the basis matrix, we derived individual cell
210 type gene profiles from individual tissue scRNA-seq atlases. By considering cell types unique to a given
211 tissue, we could leverage bulk RNA transcriptomic datasets from GTEx and HPA to ensure specificity in
212 context of the whole body. This directed approach could enable the application of many single tissue cell
213 atlases, whose meaningful integration into approaches like nu-SVR is otherwise limited by batch effects[50].
214 Defining cell type specific genes from single tissue atlases in the context of the whole body surmounts the
215 problem of missing cell types from TSP and more generally enables a means to address missing cell types
216 from the basis matrix column space required for deconvolution approaches like nu-SVR[14,15].
217 Previous work has demonstrated cell types-of-origin identification from cell free nucelosomes;[16]
218 in particular of enriched cardiomyocyte signal in patients with acute myocardial infarction, and liver cell
219 type specific signatures in patients with various hepatic disorders. However, reference ChIP-seq data
220 from pure cell types is limited, thereby reducing the scope of resolvable cell types with this approach[51]

and limiting interpretation of some aspects of the data to tissue resolution, and sensitivity appears to be limited. For example, in healthy (non-pathological) cases it was only possible to deconvolve blood cell types, and other than a generalized signal from liver it was not possible to detect cell types from solid tissues. Another interesting distinction which will potentially affect the overall applicability is that cfDNA measurement requires cell death, whereas cfRNA additionally incorporates information from living cells which secrete RNA by various mechanisms[1].

The results here reinforce the importance of reliable reference data annotation at both bulk tissue and single cell level; differences in either impact the ability to meaningfully integrate in cfRNA analysis. Cell type annotation differences across distinct cell atlases for the same tissue may impact the assignment of a gene as cell-type specific when considering a single dataset. Specifically, we observed that several genes reported as specific to a single trophoblast cell type[28] were not validated in two independent placental cell atlases[32,33]. Annotation discrepancies between atlases impacts the assignment of genes as cell type specific in context of the whole body, and consequently impact the interpretation of a cell type signature score in cfRNA.

This method for detecting cellular pathophysiology in the cf-transcriptome is most robust for diseases with cell type changes independent of disease etiology. CKD cfRNA samples reveal striking differences in proximal tubule signature score compared to healthy controls, in contrast with the minimal effect size of extravillous trophoblast signature score in preeclampsia. Multiple differentially downregulated genes in AD phenotypes are cell type specific, again reinforcing the ability to noninvasively resolve pathologically implicated cell types. Cellular pathology in CKD and AD are proximal tubule atrophy and neuronal death respectively, which occur irrespective of disease etiology, whereas preeclampsia etiology may be multifactorial and the extent of underlying cellular pathogenesis remains to be explored[52].

CKD impacts 9.1% of the global population[53] and invasive biopsy (often multiple) is the clinical gold-standard for diagnosis. The sensitivity and specificity of some glomerular filtration estimates for various patient populations and kidney function may result in clinical confusion [54]. Standard lab tests for serum creatinine and urine albumin levels merely indicate renal dysfunction and provide limited clinically actionable insight into the source(s) of renal pathophysiology. Tubular atrophy has been repeatedly shown to be superior to glomerular pathology as a predictor of CKD progression [37]. The ability to noninvasively measure proximal tubule pathophysiology and the general detection of multiple renal cell types could enable noninvasive classification of various renal disorders in future work and augment patient treatment plans. Given the small sample size used here (n = 9, CKD; n = 7, control), we emphasize that our findings, although striking, must still be validated in a larger follow-up study.

This work shows that one can apply cell atlases to measure disparate cell types that are disease-implicated in the blood, relevant to a myriad of questions impacting human health. Unlike model organisms which lack full translatability to human health, cf-transcriptomic measurement provides direct, immediate insights into patient health. Readily measurable cell types in cfRNA, including those specific to the brain, lung, intestine, liver, and kidney, have vast prognostic and clinical importance given the multitude of diseases in these tissues. Single cell RNA-seq reveals numerous cell type specific changes in pathologies within these tissues for investigation with cfRNA ranging from cancer to Crohn's disease, drug or vaccine response, and aging. Consistent detection of cell types responsible for drug metabolism (e.g. 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 powerful clinical trial end-point data in evaluating drug toxicity. Chemotherapy regimens are known to have severe systemic side-effects. We

anticipate that the ability to noninvasively resolve cell type signatures in plasma cfRNA will both enhance existing clinical knowledge in addition to enabling 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 for helpful discussions. We thank G. Loeb for kidney discussions.

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. 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 and materials availability: Code for the work in this manuscript will be made available on Github. All datasets used for this work were publicly available, downloaded with permission, or directly requested from authors.

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 appropriate approvals from Synapse.

HPA v19 transcriptomic data, GTEx v8 raw counts, and Tabula Sapiens v0.9 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 [55,56]. To estimate RNA degradation, we calculated 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 then identified outlier samples using the 95th percentile bound within a given cfRNA dataset. We considered any given sample a low quality sample if its value for any metric was greater than or equal to the 95th percentile bound.

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. 1b), we next applied trimmed mean of M values (TMM) normalization as previously described[57]:

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

CPM TMM normalized gene counts across technical replicates for a given biological replicate were averaged for the count tables used in the analyses performed in Figures 2 & 3.

Sequencing batches and plasma volumes were obtained from the authors in Toden 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 inferred based on the figure and confirmed with authors that sequencing strategy was figure-dependent (personal communication).

All samples from Munchel et al were used to compute TMM scaling factors and 4.5 mL plasma was used to normalize all samples within a given dataset (both PEARL-PEC and iPEC)

For the work in Figure 3B, longitudinal samples for a given CKD patient were averaged, given that the timescale over which renal cell type changes would occur were longer than the patient samples (~30 days). These samples were collected alongside two healthy patient biological replicates passing sample QC in this sequencing batch. We additionally used the five plasma biological replicates from above. This ultimately yielded n = 7 healthy biological replicates and n = 9 CKD patients.

Zero-Centered Batch Normalization

To account for center-specific effects that could impact meaningful comparison of data across centers in Figure 3B, we subtracted the mean normalized value across all samples measurements for given gene within a given batch from the measured normalized value for a given sample[58]:

$$\overline{G_{ij}} = G_{ij} - \mu_{ik} \quad (3)$$

Where the gene index is i, the sample is j, and k is the batch. The mean expression of the ith gene in the kth batch is denoted by μ_{ik}.

We defined a 'batch' of samples to reflect the experimental workflow for each of the corresponding analyses:

- For samples from Ibarra et al, given that only two control biological replicates were sequenced with the CKD samples and the other healthy controls came from another batch, we did not directly compare CKD vs. healthy samples. A difference in the raw median value of the proximal tubule signature score between the two sick and the nine healthy samples sequenced in the same batch was observed prior to grouping healthy plasma data from a different batch, consistent with the observed difference post-zero centered normalization. All CKD and healthy samples were treated as a single batch from which normalization was performed.
- For the datasets from Munchel et al., zero-centered batch normalization was not performed given that the data were compared within the same sequencing studies (e.g. iPEC and PEARL-PEC)

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. Specificity (how often marker was not expressed in a given cell type) and sensitivity (how frequently marker is expressed in cells of this type) thresholds determined the total gene space for intersection across the cfRNA samples. Gene synonyms from Panglao were determined using MyGene version 3.1.0 to ensure full gene space.

The intersection of this space with each cfRNA sample were then determined, where the error bars reflect the differences in number of markers detected across the samples for a given cell type across samples. A given cell type marker was counted in a given healthy cfRNA sample 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.
The samples used for nu-SVR deconvolution were the five healthy donor plasma samples as in Figure 1D of Ibarra et al.

Basis Matrix Formation

Only cells from droplet sequencing ("10X") were used in analysis. Disassociation genes as reported[10] and cell types too granular (i.e. fast muscle cell, smooth muscle cell, LYVE1 aortic macrophage, differentiated basal cell of epithelium of trachea, etc) or too broad (i.e. granulocyte, lymphocyte, monocyte subtypes, innate lymphoid cells, etc) in annotation were excluded from subsequent analysis.

Of the remaining cell types, either 30 observations were randomly sampled or the maximum number of available observations if less than 30 were subsampled, whichever was greater.

Cells were assigned broader labels to enable linear independence of the matrix column space, as several cell types that are very transcriptionally similar with few distinct gene would be challenging to resolve noninvasively and hence would impact nu-SVR deconvolution. In any sort of regression, multicollinearity between features will impact the learned coefficients.

Scanpy (version 1.6.0) was used to analyze all single cell data. Hierarchical clustering was performed on PCA-transformed (scanpy pca) CPM cell counts (scanpy normalize_total, target_sum = 1e6) log transformed (scanpy log1p) counts using the scanpy dendrogram function. Cell types that were close in clustering were grouped together, including:

- ‘pancreatic A/B/PP cell' comprised 'pancreatic A cell', 'pancreatic PP cell', 'type B pancreatic cell'
- Lung pneumocyte comprised 'type I pneumocyte' and 'type II pneumocyte'
- "intestinal crypt stem cell + transient amplifying cell" comprised 'intestinal crypt stem cell', 'intestinal transient amplifying cell', 'paneth cell'
- "vascular smooth muscle cell" comprised 'aortic smooth muscle cell' and 'vascular associated smooth muscle cell'

All cell types annotated as some type of ‘B cell’, ‘T cell’, ‘NK cell’, ‘dendritic cell’, ‘thymocyte’, etc were labelled respectively with the broader category.

All groupings are available on Github in the script entitled ‘coarsegraincells_forBMGeneration.py’.

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

- Disable quantile normalization = False
- Min. expression = 0.25
- Replicates = 5
- Sampling = 0.5
- Kappa = 999
- q-value = 0.01
- No. barcode genes = 2450 - 5000
- Filter non-hematopoietic genes = False

The resulting basis matrix was then saved as a .txt file and used in nu-SVR deconvolution

Nu-SVR deconvolution

We formulated the cell free transcriptome as a linear summation of the cell types from which it originates[3,59]. With this formulation, we adapted existing deconvolution methods developed with the objective of decomposing a bulk tissue sample into its single cell constituents[14,15], 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’.

We performed nu-SVR using a linear kernel to learn $\boldsymbol{\theta}$ from a subset of genes from the signature matrix to best recapitulate the observed signal **b**, where nu denotes the lower bound on the fraction of support vectors and the upper bound on the fraction of errors at the margin[60]. Here, the support vectors are the genes used from the basis matrix from which to learn $\boldsymbol{\theta}$; $\boldsymbol{\theta}$ reflects the weights of the cell types in the

basis matrix column space. For each sample, we learned coefficients for six values of nu, $\nu \in \{0.1, 0.15, 0.25, 0.5, 0.75, 0.9\}$ and estimated the resulting deconvolution error using the root mean square error (RMSE). We determined the product of the basis matrix with the learned coefficients ($A\theta$), which reflected some predicted expression value for each of the genes in a given cfRNA mixture. The RMSE was then computed using the predicted expression values and the measured values across all the non-zero CPM genes in a cfRNA mixture.

Only CPM counts > 0 were considered in the mixture. The values in the basis matrix were also in CPM space. Prior to deconvolution, the mixture and basis matrix were 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 4. Specifically, the concavity of the log function would result in the consistent underestimation of $\boldsymbol{\theta}$ during deconvolution[61].

Using the $\boldsymbol{\theta}$ resulting from the value of ν whose coefficients yielded the smallest RMSE was transformed to. Specifically, the relative fractional contributions of cell type specific RNA from $\boldsymbol{\theta}$, we repeat what was previously described [14,15]:

$$\forall \theta_j < 0 \in \{\theta_1, \ldots, \theta_c\} \rightarrow 0 \quad (5)$$

All non-zero coefficients were then normalized by their sum to result in the relative fractions to determine the relative fractional contributions of cell type specific RNA.

We used the function nuSVR from scikitlearn version 0.23.2.
The samples used for nu-SVR deconvolution were the five healthy donor plasma samples as in Figure 1D of Ibarra et al.

Evaluating Basis Matrix on GTEx samples

Bulk RNA-seq samples from GTEx v8 were deconvolved with the derived basis matrix from tissues that were present (kidney cortex, whole blood, small intestine – terminal ileum, lung, and spleen) or absent (kidney medulla and liver) from the basis matrix derived using Tabula Sapients version 0.9. For each tissue type, the maximum number available samples or ten samples, whichever was smaller, was deconvolved. Please 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:

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

Where G_j is the gene set in the jth deconvolved sample, where a given gene in the set's expression was $\geq$ 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 $\geq$ 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 tissue cell type atlas (ii) had a Gini coefficient ≥ 0.6, indicating comprehensive tissue specificity in context of the whole body.

(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[12] (fig S8). The Gini coefficient better reflected the underlying distribution of gene expression tissue-specificity than Tau (fig. S8) 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 [20]:

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

Tau, as defined in [20]:

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

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 used the raw GTEx data v8 (accessed August 26 2019) and converted to log(CPM + 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

Estimating signature scores for each cell type

The signature score is defined as the sum of genes 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.

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

For signature scoring of syncytiotrophoblast and extravillous trophoblast gene profiles in PEARL-PEC and iPEC[35]. The genes in a respective profile used for signature scoring were 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 [32,33]. Only the intersection of the cell type specific gene profiles was considered for signature scoring.

Comparison of proximal tubule signature score to random for discriminatory between CKD 3+ and Healthy

To assess the discriminatory power of a given cell type signature score with a statistically significant difference in Fig 3B, we randomly sampled an equivalent gene length as the proximal tubule gene profile in 10,000 trials and performed a one-sided Mann Whitney U with the alternative hypothesis that healthy would be greater than CKD 3+. For a given trial, the signature score of the random gene list was computed across all samples and tested. Benjamini-Hochberg correction at FDR = 0.05 was performed using 'multitest' function in statsmodels version 0.10.1 with the following arguments: alpha=0.05 method='fdr_bh'.

Cell Type specific differentially expressed neuronal and glial cell type specific genes in Alzheimer's plasma

To assess whether DEGs in AD/NCI plasma[4] that intersected with a brain cell type gene profile were more specific to a given brain cell type than DEGs in AD/NCI plasma that was generally brain 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 single cell data[17].

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

The union of all brain cell type specific genes is the set C.

Genes in B that that did not intersect with C (e.g. any brain cell type gene profile ('brain cell type specific')) and intersected with DEG-up (U) or DEG-down genes (D)[4] were then defined as 'brain tissue specific'.

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

All genes belonging to brain cell type gene profiles ('brain cell type specific') were a subset of the initial set of brain specific genes.

$$C - B = 0 \quad (11)$$

Genes defined as 'brain cell type specific' for signature scoring in Fig. 3C were intersected with differentially upregulated (DEG-up) and differentially downregulated genes (DEG-down) reported[4]. No DEG-up genes intersected with any of the brain signatures used in Fig 3C. Only DEG-down were considered in the subsequent analysis as 'brain cell type specific'.

The Gini coefficients reflecting the gene expression inequality across brain cell types were computed for the gene sets labelled as 'brain cell type specific' and 'brain tissue specific'. Brain reference data to compute Gini coefficients was the single cell brain atlas with diagnosis as 'Normal'[17]. 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: $\mu_{cell\ type\ Gini\ Coefficient} = \mu_{brain\ tissue\ Gini\ coeffieint}$).

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:

$$p = \frac{\#\ trials\ with\ permuted\ (\mu_{cell\ type} - \mu_{tissue}) \geq \mu_{observed}}{10{,}000+1} \quad (12)$$

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 $\mu_{observed}$)

Supplementary Note 1: Deconvolution of bulk GTEx tissues using the *Tabula Sapiens*-derived basis matrix

To assess the ability of the basis matrix to deconvolve tissues whose cell types were wholly present in the cell type column space, we deconvolved a subset of bulk RNA-seq GTEx samples. The determined fractions of cell type specific RNA generally recapitulated the predominant cell types within a given tissue (fig. S4). Kidney cortex majority fractions were from kidney epithelia and vascular endothelia (fig. 4A,B); small intestine, smooth muscle cells and intestinal enterocytes (fig. S4E); whole blood, erythrocytes (fig. S4G). Cells with larger volume yielded larger deconvolved fractions for all tissues (fig. S3). Variance in the relative cell type fractional contributions across the deconvolved bulk samples within a given tissue reflects the underlying cell type heterogeneity. GTEx kidney medulla samples recorded to be contaminated with renal cortex reflect the presence of the kidney epithelia, the majority cell type in the renal cortex (fig. S4A). Tissues absent from the cell type column space, such as liver, yielded cell types that are transcriptionally similar (kidney epithelia) (fig. S4C) and a higher deconvolution error (fig. S3) as expected.

References

1. 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).

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

3. 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).

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

5. Chalasani, N. *et al.* Non-invasive Stratification of Non-Alcoholic Fatty Liver Disease by Whole-transcriptome Cell-free mRNA Characterization. *Am. J. Physiol. Gastrointest. Liver Physiol.* (2021) doi:10.1152/ajpgi.00397.2020.

6. 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).

7. Basu, M., Wang, K., Ruppin, E. & Hannenhalli, S. Predicting tissue-specific gene expression from whole blood transcriptome. *Sci. Adv.* **7**, (2021).

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

9. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).

10. Tabula Sapiens: human transcriptome reference at single cell resolution. https://tabula-sapiens-portal.ds.czbiohub.org.

11. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

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

13. 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**, (2019).

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

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

16. Sadeh, R. *et al.* ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-020-00775-6.

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

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

19. Litviňuková, M. *et al.* Cells of the adult human heart. *Nature* **588**, 466–472 (2020).

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

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

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

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

693 24. Sender, R. & Milo, R. The distribution of cellular turnover in the human body. *Nat. Med.* **27**, 45–
694 48.

695 25. Zemmour, H. *et al.* Non-invasive detection of human cardiomyocyte death using methylation
696 patterns of circulating DNA. *Nat. Commun.* **9**, 1443 (2018).

697 26. Danielson, K. M. *et al.* Plasma Circulating Extracellular RNAs in Left Ventricular Remodeling
698 Post-Myocardial Infarction. *EBioMedicine* **32**, 172–181 (2018).

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

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

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

705 30. Chevalier, R. L. The proximal tubule is the primary target of injury and progression of kidney
706 disease: role of the glomerulotubular junction. *Am. J. Physiol. Renal Physiol.* **311**, F145-61
707 (2016).

708 31. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's
709 disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097
710 (2019).

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

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

715 34. Stewart, B. J. *et al.* Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–
716 1466 (2019).

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

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

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

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

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

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

41. 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).

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

43. 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).

44. 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).

45. 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).

46. 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).

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

48. Zaitsev, K., Bambouskova, M., Swain, A. & Artyomov, M. N. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.* **10**, 2209 (2019).

49. 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).

50. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

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

52. Huppertz, B. Placental origins of preeclampsia: challenging the current hypothesis. *Hypertension* **51**, 970–975 (2008).

53. GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **395**, 709–733 (2020).

54. Levin, A. & Stevens, P. E. Early detection of CKD: the benefits, limitations and effects on prognosis. *Nat. Rev. Nephrol.* **7**, 446–457 (2011).

55. 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).

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

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

58. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29–39 (2016).

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

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

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

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.

838 Rando, Kristy Red-Horse, Jeremy Reiter, Justin Sonnenburg, Bruce Wang, Albert Wu, Sean Wu, Tony
839 Wyss-Coray

840

841

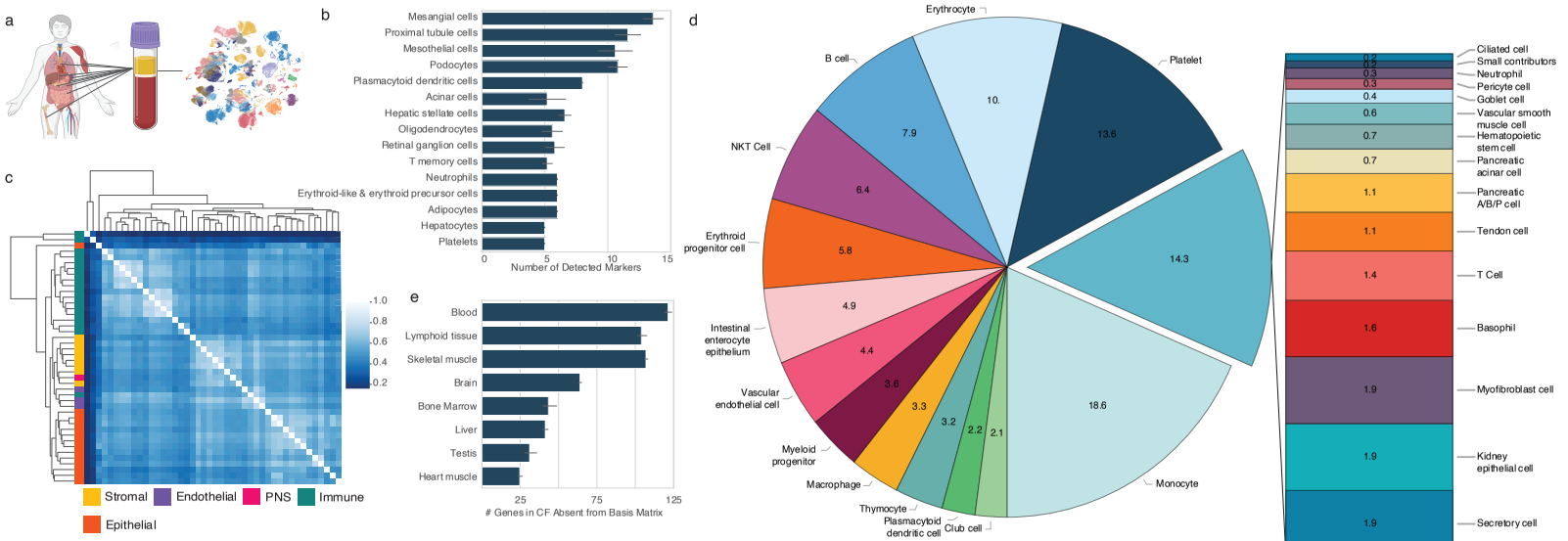


Fig. 1: Cell type decomposition of the plasma cell free transcriptome using Tabula Sapiens reveals the comprehensive cell type-of-origin landscape

a. Integration of tissue-of-origin and single cell transcriptomics to identify cell types-of-origin in plasma cfRNA.

b. Cell type specific markers defined in context of human body identified in cfRNA (n = 5). Error bar denotes variance among intersection size across biological replicates.

c. Cluster heatmap of Spearman correlations of cell types in basis matrix derived from Tabula Sapiens. PNS denotes 'peripheral nervous system'.

d. Mean fractional contributions of cell type specific RNA in plasma cf-transcriptome (n = 5). Full distributions of learned coefficients across biological samples are available in fig. S5.

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 intersected with HPA tissue specific genes).

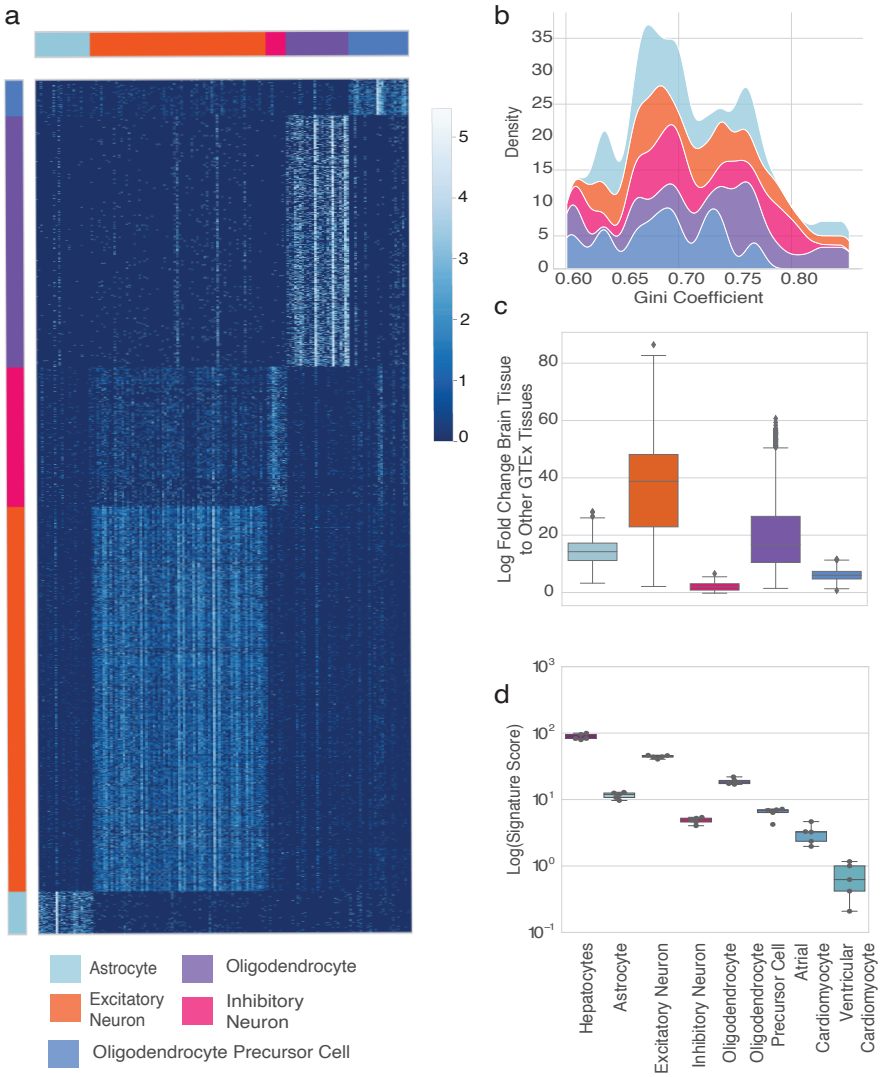


Fig. 2: Individual single cell atlases identify additional cell types-of-origin in the plasma cell free transcriptome

a. Heatmap of gene expression in NCI prefrontal cortex single cell data indicating cell type specificity across brain cell types. Rows are individual cells; columns are a given marker gene. Values are log-transformed counts per ten thousand.

b. Density plot of the Gini coefficients for the genes within a given cell type gene profile. The Gini coefficient for a given gene was computed using normalized counts across all HPA tissues. Area under curve for a given cell type sums to one.

c. Validation of specificity of cell type specific genes across the tissues of the human body. Log-fold change in total log-transformed counts-per-ten-thousand expression of genes for a given cell type signature across all GTEx tissues over the mean non-brain aggregate cell type signature score assess overall abundance in expression.

d. Hepatocyte, neuronal, glial, and cardiomyocyte cell type signature scores in healthy cfRNA plasma (n = 5) on a logarithmic scale.

843

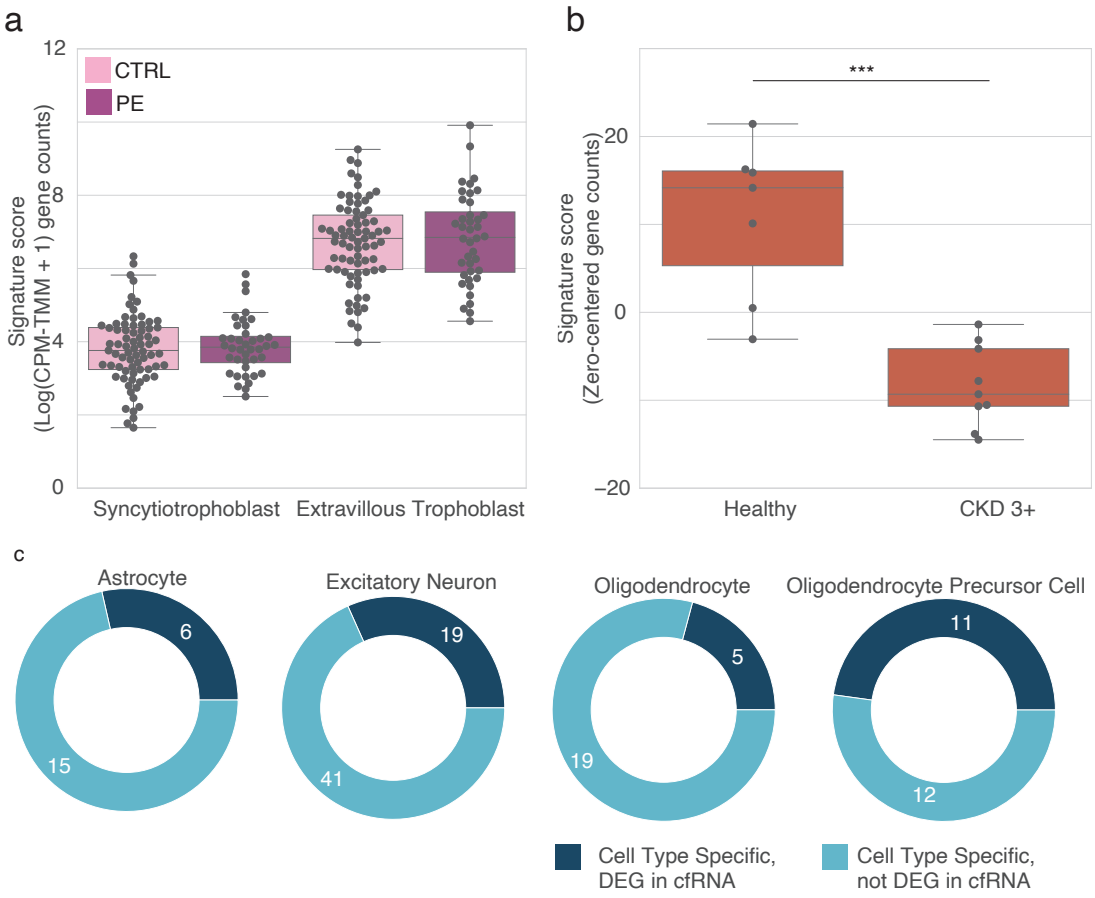


Fig. 3: Cell type signatures reveal implicated cell types in pathologic cfRNA samples

a. Comparison of signature scores for extravillous trophoblasts and syncytiotrophoblasts in cfRNA samples (Munchel et al. 2020, iPEC cohort) from mothers with no complications (n = 73) as compared to mothers with preeclampsia (n = 40). Gene counts were CPM-TMM normalized and then natural log transformed and summed to determine the signature score.

b. Comparison of signature scores for proximal tubule cells in cfRNA samples from CKD stages 3+ (n = 9) and healthy patients (n = 7) using data from Ibarra et al (*** denotes p < 10-3). Gene counts were CPM-TMM normalized and then zero centered across conditions and summed to determine the signature score.

c. Proportion of cell type specific genes from brain cell type single cell data intersecting with reported DEGs downregulated in AD-derived plasma[4].

844

845

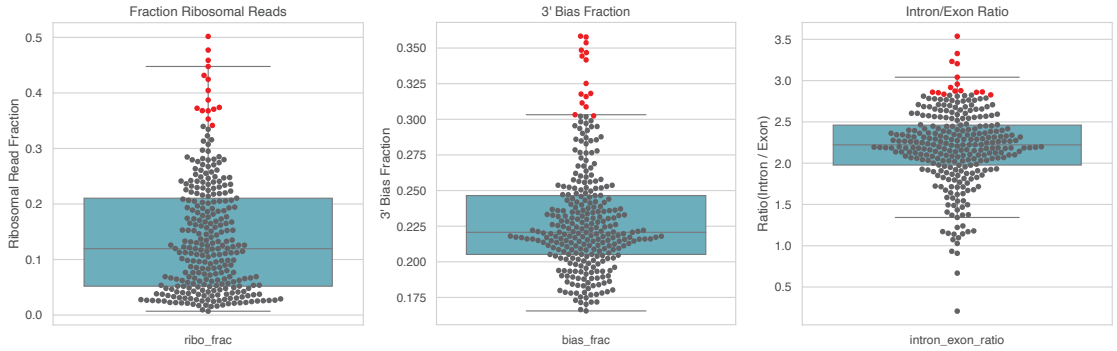


Fig S1 **Identification of samples with outlier values for at least one quality control metric** including a measure of RNA degradation, ribosomal fraction, and DNA contamination from Ibarra er al. Samples with outlier values are highlighted in red. (See Methods section 'Data Processing' for details)

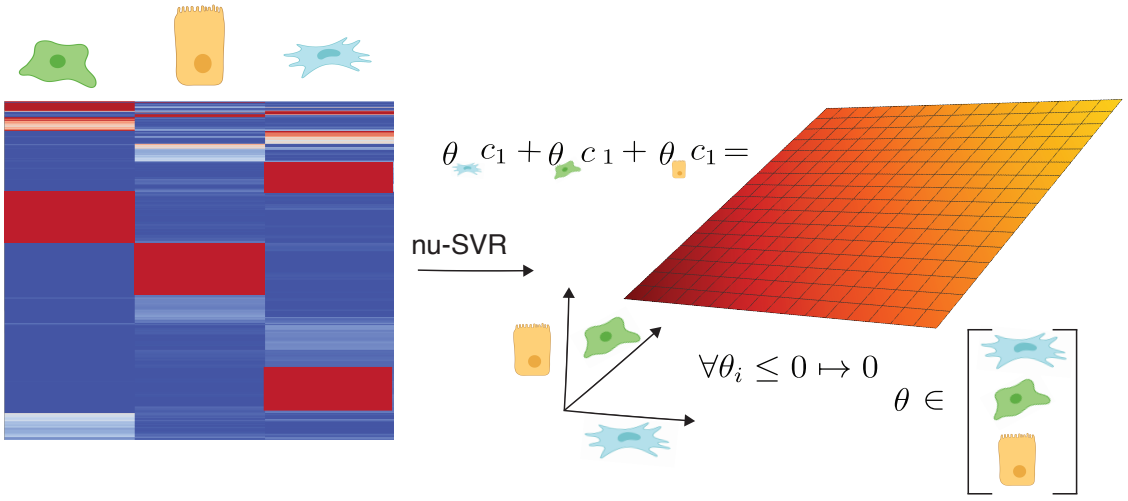


Fig. S2. **Schematic overview of nu-SVR deconvolution** in trivial 3-dimensional cell type dimensional space denoting the learning of a hyperplane in cell type dimensional space and subsequent normalization to infer relative fractions of cell type specific RNA.

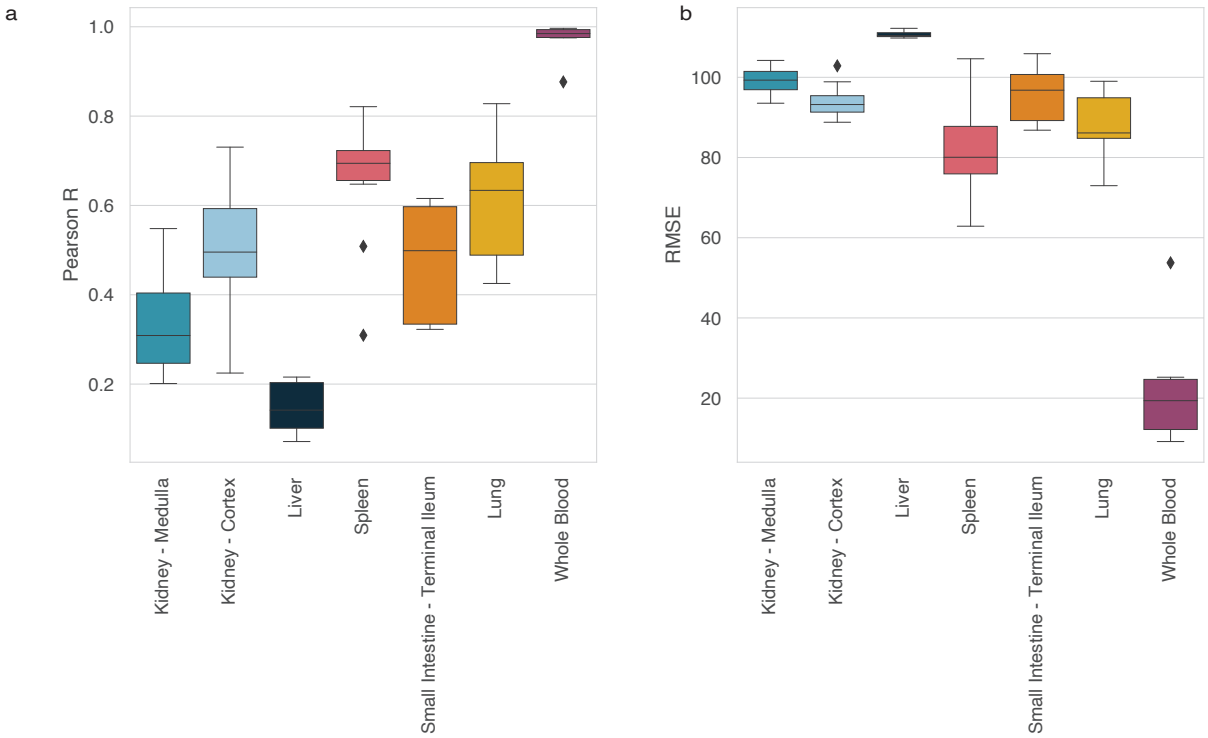
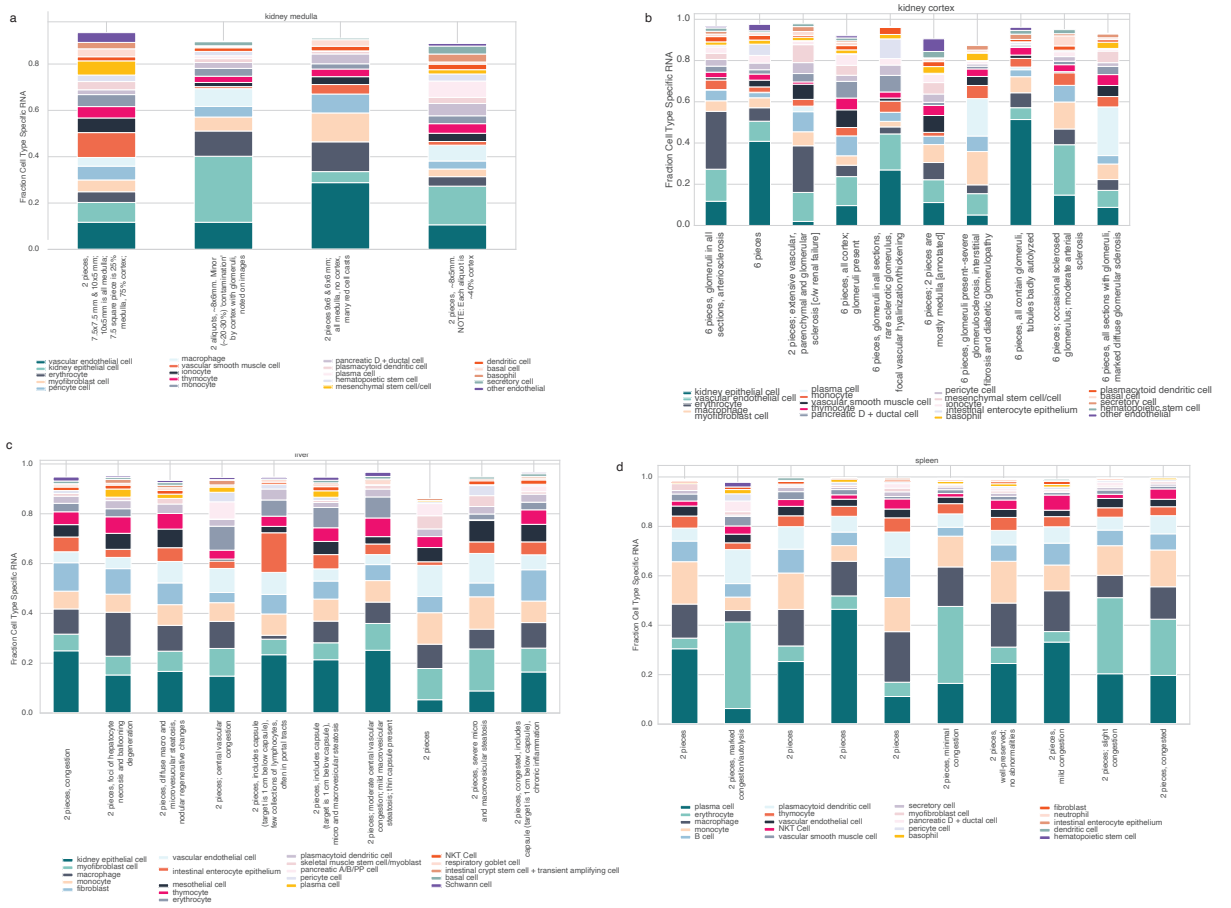


Fig. S3. **Basis matrix performance on GTEx bulk RNA samples using nu-SVR**. GTEx tissue samples possessing cell types wholly present (Kidney – Cortex, Spleen, Small Intestine – Terminal Illeum, Lung, Whole Blood) and absent from the basis matrix column space (Kidney – Medulla, Liver) were selected.

(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 Liver)



865

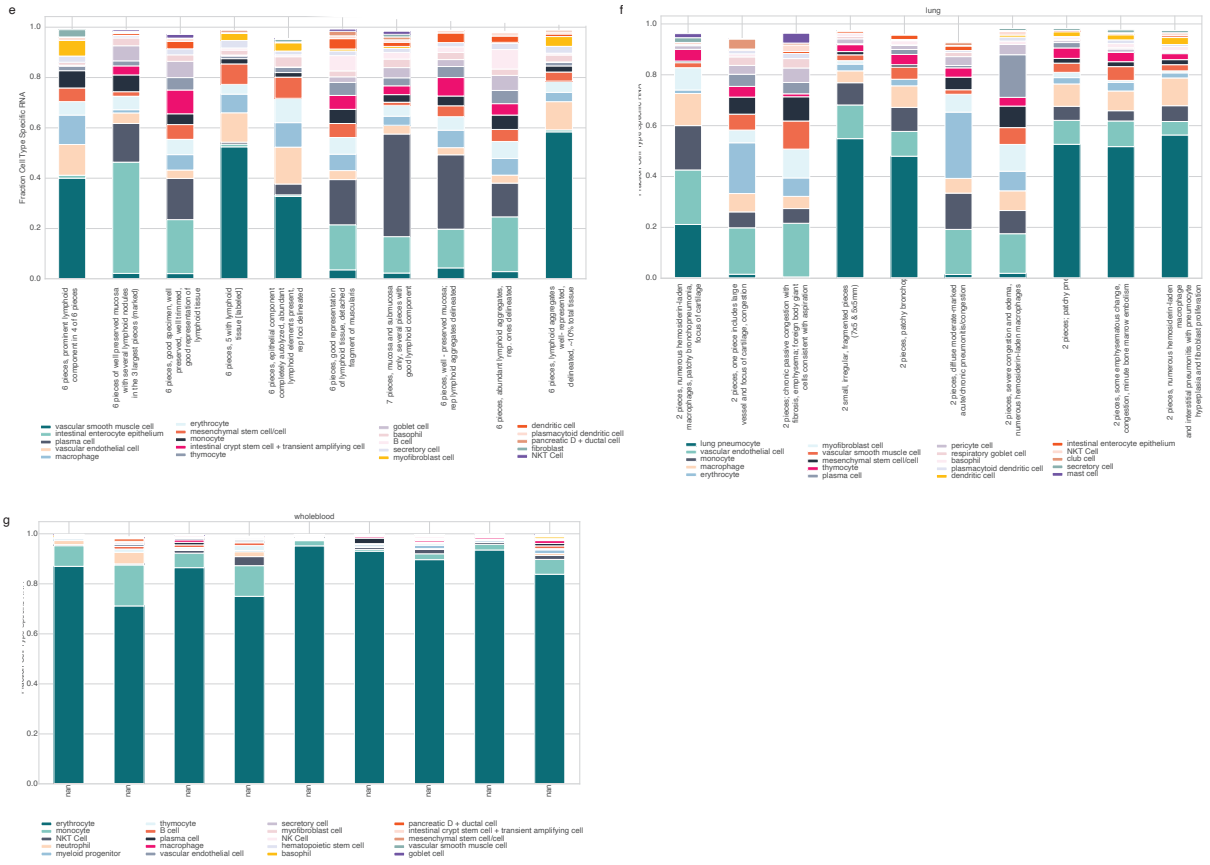
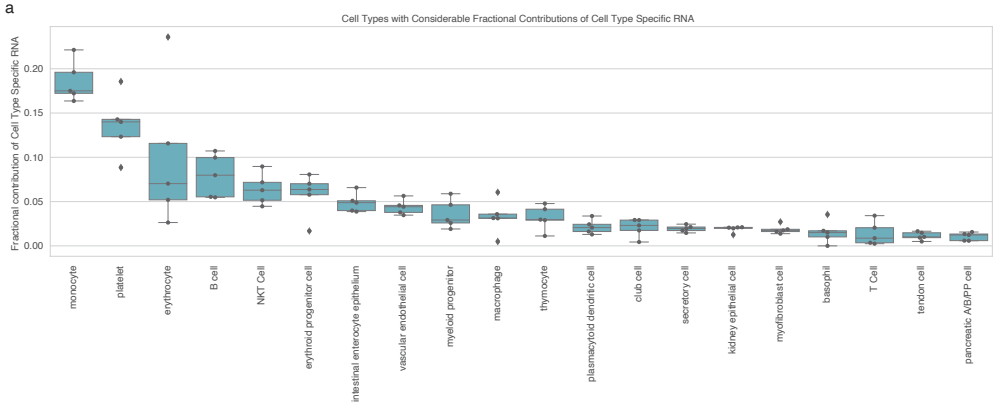


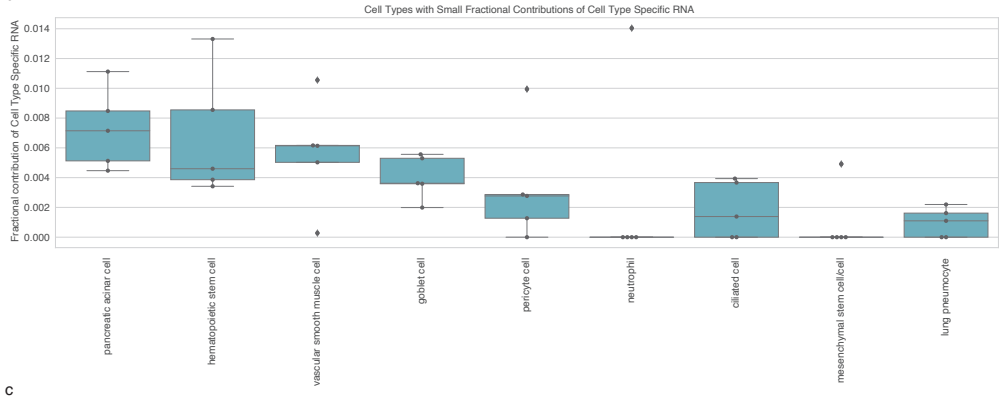
Fig S4. **Deconvolved fractions of cell type specific RNA from various GTEx tissues using nu-SVR to assess deconvolution performance of the *Tabula Sapiens*-derived basis matrix**. The two tissues whose cell types were absent from the basis matrix column space were Kidney – Medulla and Liver. Kidney medulla samples reported to be contaminated with cortex are reflected by deconvolved kidney epithelia fractions. Liver, which is absent from the TSP version used in this work, has majority fractions of kidney epithelia, which possesses several transcriptionally similar cell types to the liver.
Majority cell types for a given tissue, such as lung pneumonocytes and vascular endothelia 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 comments are in Supplementary Note 1.

(a) Kidney – Medulla
(b) Kidney – Cortex
(c) Liver
(d) Spleen
(e) Small Intestine – Terminal Ileum
(f) Lung
(g) Whole Blood

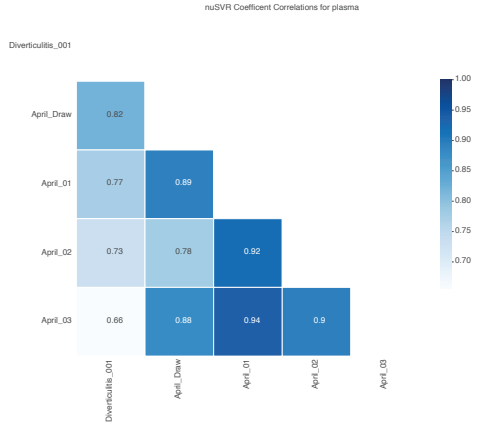
a



b



c



884

Fig. S5. **Deconvolving the plasma cell free transcriptome using nu-SVR**

(a) Distribution of learned relative fractional contributions with mean cell type proportions were > 1%

(b) Distribution of learned coefficients of small contributions using nu-SVR across samples (≤ 1%). 'Small contributors' slice in main text Fig 1D reflects cell types with cumulative fractions < 0.1% (mesenchymal stem cell/mesenchymal cell and lung pneumocyte).

(c) Pairwise Pearson correlations of nu-SVR learned coefficients between biological replicates

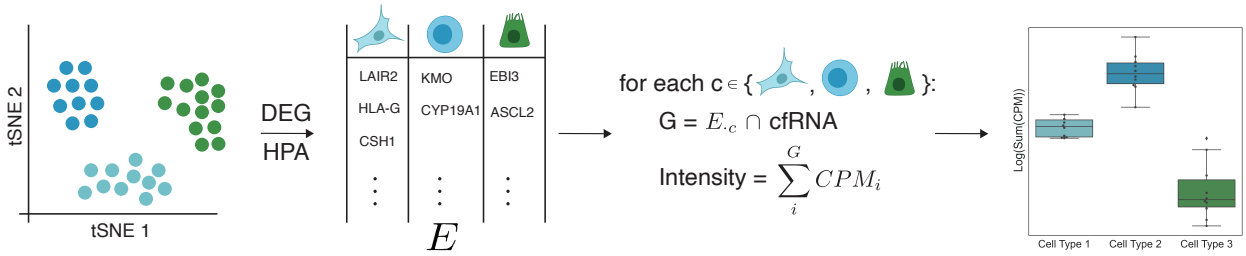


Fig S6. **Cell type signature score derivation overview**. See 'Signature Scoring' section of methods for filtering criteria and thresholds.

Fig S7. Establishing gene profile cell type specificity in context of the whole human body using single cell and bulk RNAseq data

(a) Single cell heatmaps for gene cell type profiles within the corresponding tissue cell atlas, demonstrating that predominant expression in bulk data is within the cell type of interest.

(b) Gini coefficient density plot for genes in cell type profiles derived from liver, testis, and heart single cell atlases using HPA NX counts.

(c) Log fold change in bulk RNA-seq data of the cell type profile

900

Fig S8. Distribution of Gini coefficient and Tau for all genes denoted by HPA as specific to the brain, liver, placenta, kidney, and heart

901

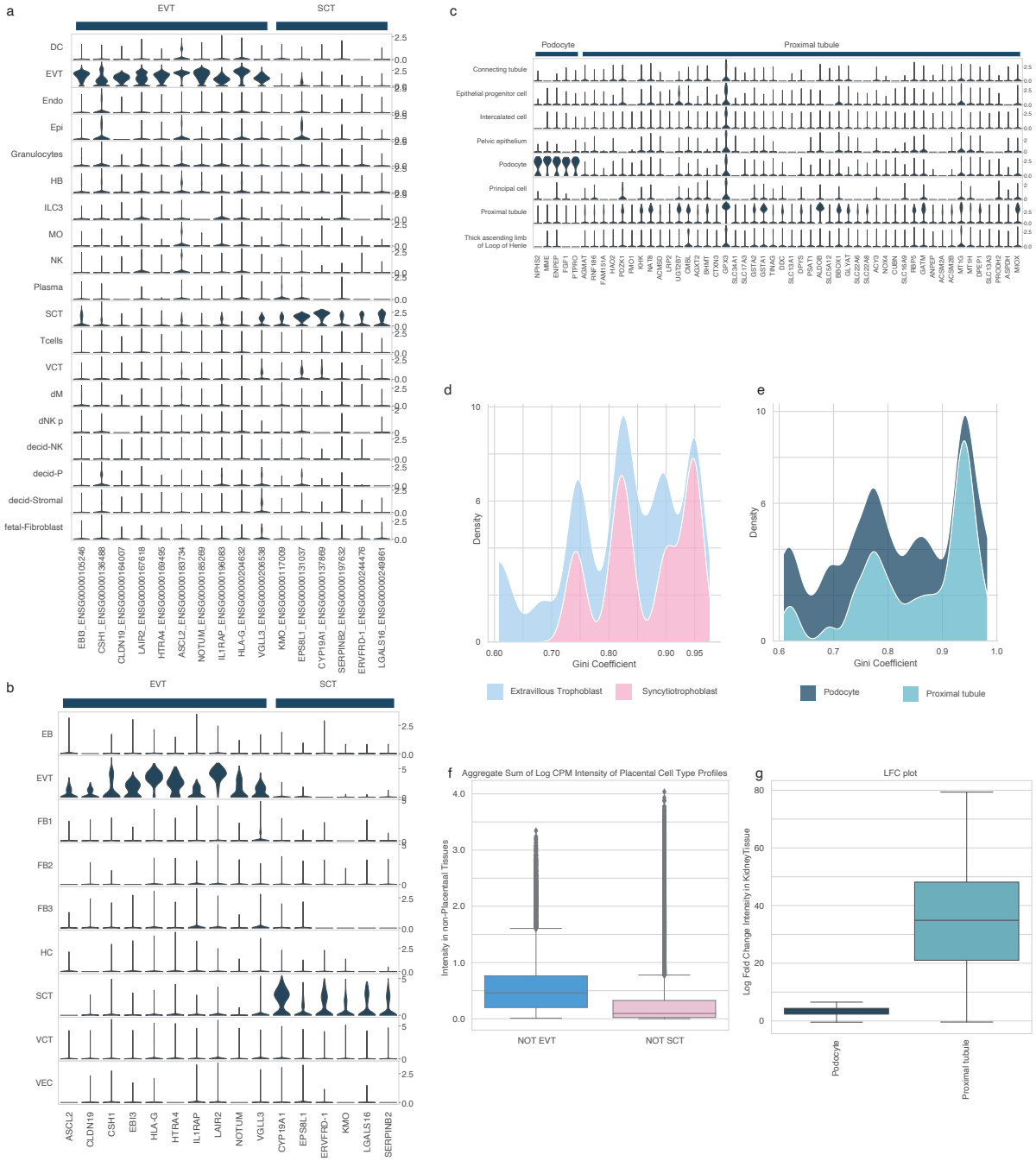


Fig. S9. **Comprehensive placental and renal cell type gene profile specificity** at single cell and whole body resolution

(a) Violin plot of derived syncytiotrophoblast and extravillous trophoblast from Vento-Tormo et al.

(b) Violin plot of derived syncytiotrophoblast and extravillous trophoblast from Suryawanshi et al.

(c) Violin plot of derived proximal tubule and podocyte markers

(d) Gini coefficient distribution for placental trophoblast cell types in (a) and (b)

(e) Gini coefficient distribution for renal cell types in (c)

(f) Distribution of placental trophoblast signature scores across all GTEx tissues, since the placenta is not in GTEx, so the values plotted are just the aggregate expression of genes in a given signature.

(g) Log fold change of renal cell type intensity in GTEx Kidney Cortex/Medulla samples (sum of log-transformed counts-per-ten thousand) relative to the mean non-kidney signature score intensity.

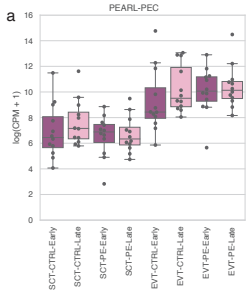
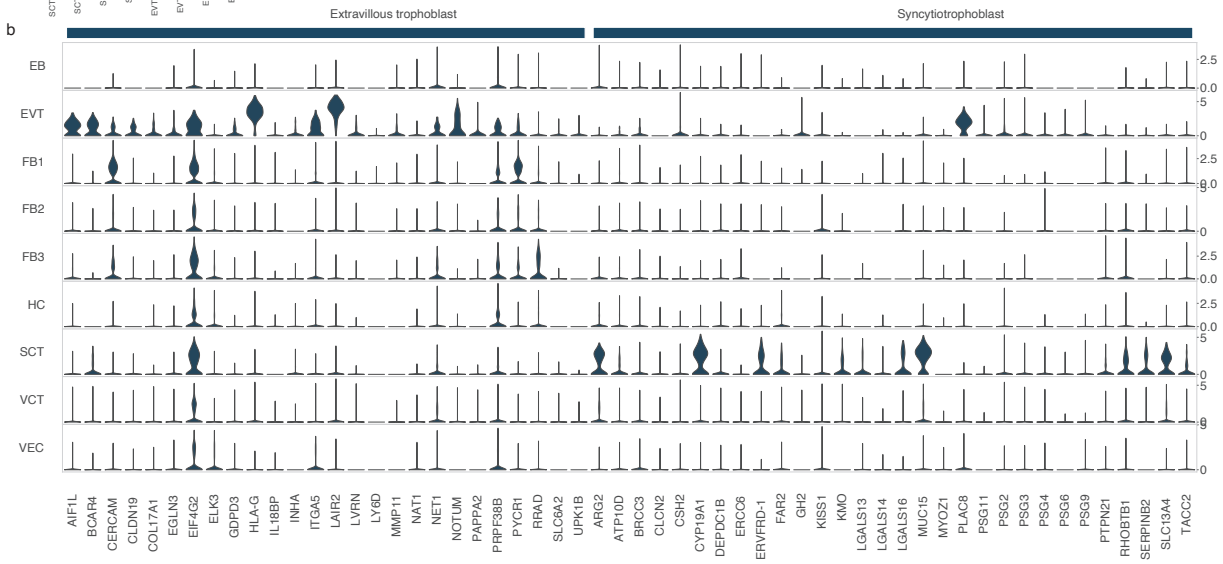
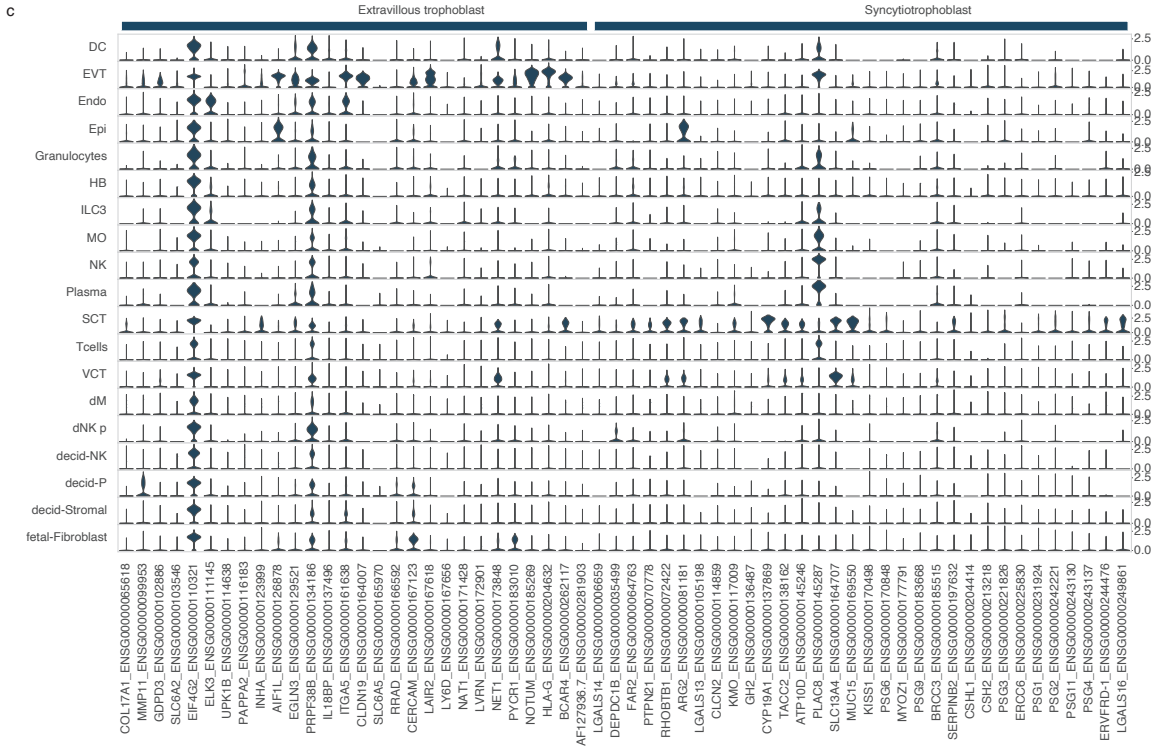


Fig. S10. **Trophoblast signature scoring in PEARL-PEC and expression distribution of two gene profiles** asserted by Tsang et al. in two independent placental cell atlases
(a) Placental trophoblast signature scores on PEARL-PEC validation dataset from Munchel et al (n = 12 each for early/late-onset PE cohorts and gestationally-aged matched healthy controls)
(b) Stacked violin plot of the genes comprising the extravillous trophoblast and syncytiotrophoblast signatures in Tsang et al. reflecting the distribution across placental cell types in Vento-Tormo et al
(c) Stacked violin plot of the genes comprising the extravillous trophoblast and syncytiotrophoblast signatures in Tsang et al. reflecting the distribution across placental cell types in Suryawanshi et al





906

907

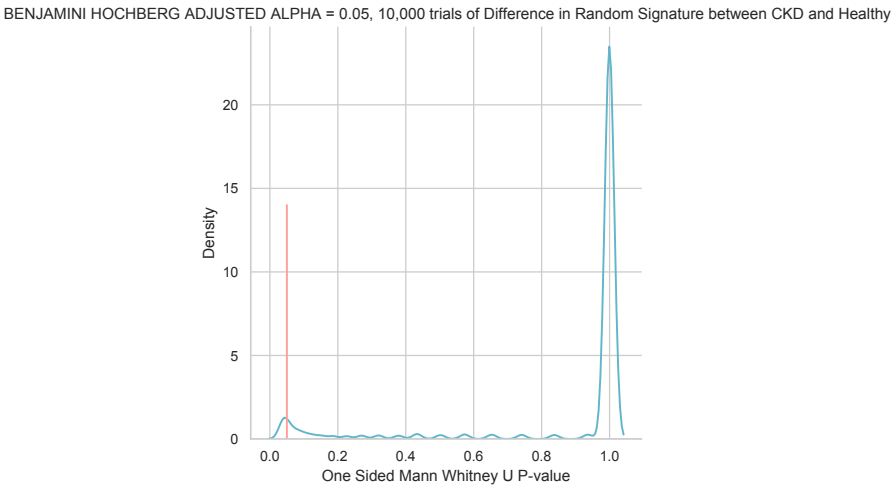


Fig. S11. **Validation of the discriminatory power in the proximal tubule signature** in CKD stage 3+ vs healthy samples
Distribution of one-sided Mann Whitney U in 10,000 trials of comparing a random signature score of an equivalent number of genes as the proximal tubule signature with an alternative hypothesis that the signature in healthy is greater than CKD. Red vertical line denotes the p = 0.05 threshold. Multiple hypothesis testing correction using Benjamini Hochberg with FDR = 0.05 was performed and this yielded an adjusted p-value of 0.038 between the CKD 3+ and healthy groups using the actual proximal tubule signature score. In 96.6% of the trials, no significance was observed in the random signature discriminating between sick and healthy, indicating the specificity of our signature score in discriminating between CKD and healthy patients.

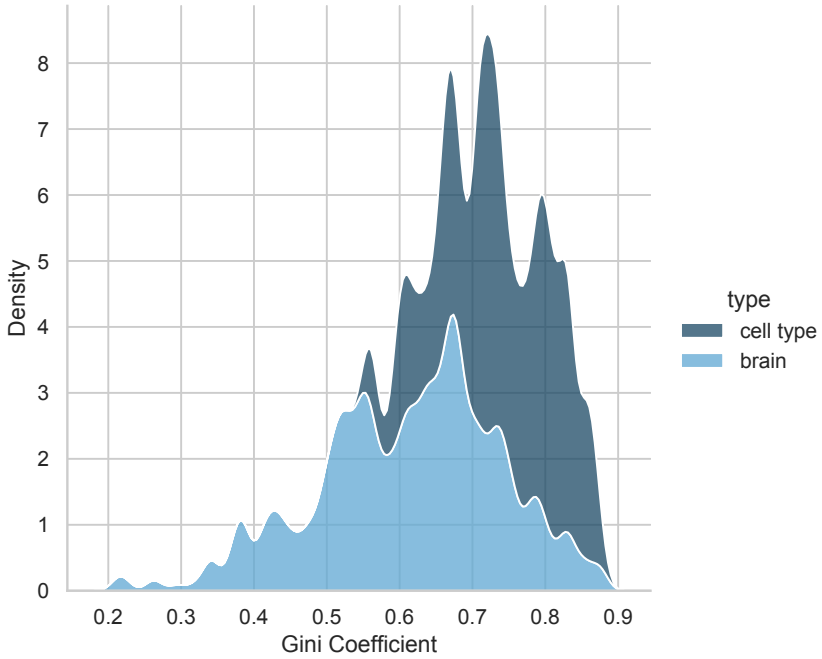


Fig. S12. **Comparison of brain specific DEG and cell type specific DEG.** Distribution in gini coefficients for AD downregulated DEG in Toden et al. that are brain-specific and cell type specific respectively. Area under curve for each group sums to 1.