RNA-Sieve: A likelihood-based deconvolution of bulk gene expression data using single-cell references

Dan D. Erdmann-Pham*[1], Jonathan Fischer*[2,3,4], Justin Hong[3], and Yun S. Song†[2,3,5]

[1]Department of Mathematics, University of California, Berkeley, California 94720
[2]Department of Statistics, University of California, Berkeley, California 94720
[3]Computer Science Division, University of California, Berkeley, California 94720
[4]Department of Biostatistics, University of Florida, Gainesville, Florida 32611
[5]Chan Zuckerberg Biohub, San Francisco, California 94158

October 1, 2020

Abstract

Direct comparison of bulk gene expression profiles is complicated by distinct cell type mixtures in each sample which obscure whether observed differences are actually due to changes in expression levels themselves or simply cell type compositions. Single-cell technology has made it possible to measure gene expression in individual cells, achieving higher resolution at the expense of increased noise. If carefully incorporated, such single-cell data can be used to deconvolve bulk samples to yield accurate estimates of the true cell type proportions, thus enabling one to disentangle the effects of differential expression and cell type mixtures. Here, we propose a generative model and a likelihood-based inference method that uses asymptotic statistical theory and a novel optimization procedure to perform deconvolution of bulk RNA-seq data to produce accurate cell type proportion estimates. We demonstrate the effectiveness of our method, called RNA-Sieve, across a diverse array of scenarios involving real data and discuss several extensions made uniquely possible by our probabilistic framework, including general hypotheses tests and confidence intervals.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: yss@berkeley.edu

Introduction

Bulk RNA sequencing (RNA-seq) has proven a useful tool to investigate transcriptomic variation across organs, individuals, and various other biological conditions (Melé et al., 2015; Sudmant et al., 2015). Despite many successes, this technology's full potential is inherently limited because each experiment measures the average gene expression among a large group of cells, the composition of which is unknown. Thus, despite the reduction in technical and biological variability attained by averaging, bulk experiments are potentially confounded by cell type proportions when considering heterogeneous cell mixtures (Lowe and Rakyan, 2014; Shiwa et al., 2016). Such confounding impedes the direct comparison of samples, possibly leading to the spurious or missed inference of biologically relevant genes when attempting to identify clinically important differences. Moreover, cell type compositions are often independently informative of biological processes including organ function (Cabrera et al., 2006; Hagenauer et al., 2018; Kalisky et al., 2013; Yu and He, 2017) and development (Hagenauer et al., 2018; Hu et al., 2017). For example, cell type infiltration has been found to correlate with disease progression, disease status, and complex phenomena such as aging (Funada et al., 2003; Bense et al., 2017; Zhou et al., 2019; Bremnes et al., 2016; Stout et al., 2017). Unlike bulk experiments, single-cell technologies allow us to query the transcriptome at the resolution of individual cells. Resulting analyses often seek to characterize the heterogeneity within, or the differences between, specified cell types (Saliba et al., 2014). By isolating the expression patterns of each measured cell type, single-cell gene expression data can provide a reference to aid the inference of the cell type compositions of bulk samples; this is known as deconvolution.

Computational rather than experimental estimation of cell type compositions is attractive for several reasons. Single-cell experiments are more expensive than their bulk counterparts and require heightened technical expertise to perform, often rendering the large-scale generation of single-cell gene expression data infeasible (Goldman et al., 2019). Furthermore, even when performed correctly, many protocols fail to capture cell types in an unbiased fashion, meaning empirical cell type proportions often are not reliable estimators of true organ compositions (Trapnell, 2015). Finally, deconvolution can be applied to the deep compendium of available bulk RNA-seq data to refine earlier analyses and probe previously unanswerable or heretofore unformulated questions. Consequently, the computational deconvolution problem has become a topic of intense methodological research (as detailed in Avila Cobos et al. (2018)). The problem may be represented mathematically as

$$M\boldsymbol{\alpha} = \boldsymbol{b}, \tag{1}$$

where M is a gene-by-cell type matrix of cell type-specific gene expression averages, $\boldsymbol{\alpha}$ a vector of cell type mixing proportions, and $\boldsymbol{b}$ a vector of gene expression values in a bulk RNA-seq experiment. Depending on which of M, $\boldsymbol{\alpha}$, and $\boldsymbol{b}$ are measured, different approaches are appropriate. We focus on the case in which both M and $\boldsymbol{b}$ have been observed, albeit noisily, and it remains to infer $\boldsymbol{\alpha}$; this is known as supervised deconvolution. Early approaches frequently utilized pre-defined marker genes for well-studied cell types, restricting their applicability. More recent methods formulate the problem as a regression task to be solved by variants of non-negative least squares (e.g., MuSiC (Wang et al., 2019), DWLS (Tsoucas et al., 2019), SCDC (Dong et al., 2020), and Bisque (Jew et al., 2020)) or with more sophisticated machine learning techniques (such as CIBERSORTx (Newman et al., 2019) and Scaden (Menden et al., 2020)). Though each paradigm presents its own strengths, both fail to replicate the benefits of explicit generative modeling. The result is algorithms which may perform well but lack the flexibility to extend beyond the estimation of cell type proportions.

In this paper, we propose a new method, RNA-Sieve, which employs asymptotic theory and a novel optimization procedure to solve a probabalistic model of deconvolution via maximum likeli-

hood estimation. We demonstrate its highly capable performance across a diverse array of scenarios, including different organs, cell types, and practical challenges. We then highlight a newly opened avenue for continued development made feasible by our generative framework, including general hypotheses tests and confidence regions.

Results

Method Overview

Although a single run of bulk RNA-seq produces only a solitary gene expression vector, myriad cells contribute to this measurement. Hence the obtained snapshot is truly the composition of the gene expression levels of numerous individual, putatively independent cells. When coupled with an assumption that cells of the same type behave similarly, this large number of cells permits application of the central limit theorem (CLT). Conveniently, the marginal distribution for gene expression of an arbitrary cell in the bulk sample is a straightforward mixture distribution (see Equation (2) in Methods). The resulting CLT-derived likelihood consequently depends only on the first two moments of gene expression for each cell type and the respective cell type proportions in the bulk sample, the latter of which we make our goal to infer computationally. To estimate the requisite cell type-specific moments, RNA-Sieve uses gene expression measurements from scRNA-seq experiments. We further model the estimation error of the computed moments by once again invoking the CLT, and the combination of these two approximations yields a full, composite likelihood built using normal distributions. We subsequently infer cell type proportions via a custom-made maximum likelihood estimation procedure. Several features of our algorithm ensure accurate and robust results. Our alternating optimization scheme is split into two components to better avoid sub-optimal local minima, with a final projection step handling flat extrema to avoid slow convergence. We also incorporate a gene filtering procedure explicitly devised to improve cross-protocol stability, a crucial concern given that single-cell and bulk experiments will always be performed with different technological platforms. Our algorithm can also perform joint deconvolutions, leveraging multiple samples to produce more reliable estimates while parallelizing much of the optimization. Finally, we wish to emphasize that our likelihood-based model allows us to pursue extensions which are infeasible using prior approaches. A notable example includes confidence regions for estimates (see Extension to Confidence Regions), among others (see Discussion). Full mathematical and computational details are presented in Methods and a schematic is displayed in Figure 1.

Performance in Pseudobulk Experiments

To establish RNA-Sieve's effectiveness, we performed *in silico* experiments using scRNA-seq data from the *Tabula Muris Senis* Consortium (Pisco et al., 2020). In these experiments we built "pseudobulks" by aggregating reads from labeled cells in known proportions to use for deconvolution. We considered thirteen organs with two to eleven cell types per organ. Moreover, counts generated via both the Smart-Seq2 and 10x Chromium protocols are available for each organ, enabling convenient cross-protocol comparisons. These are particularly important given that bulk and single-cell RNA-seq samples are always processed using different techniques. To evaluate RNA-Sieve, we compared its performance to that of six recently published methods as well as the naive solution of non-negative least squares (NNLS). Performance was assessed for each organ by computing the L_1 distance between inferred and true proportions and dividing by the number of cell types present. We found that RNA-Sieve produced the smallest mean error in both possible reference/bulk configurations (Figure 2 and Table 1; full results in Table S1). To better understand performance,

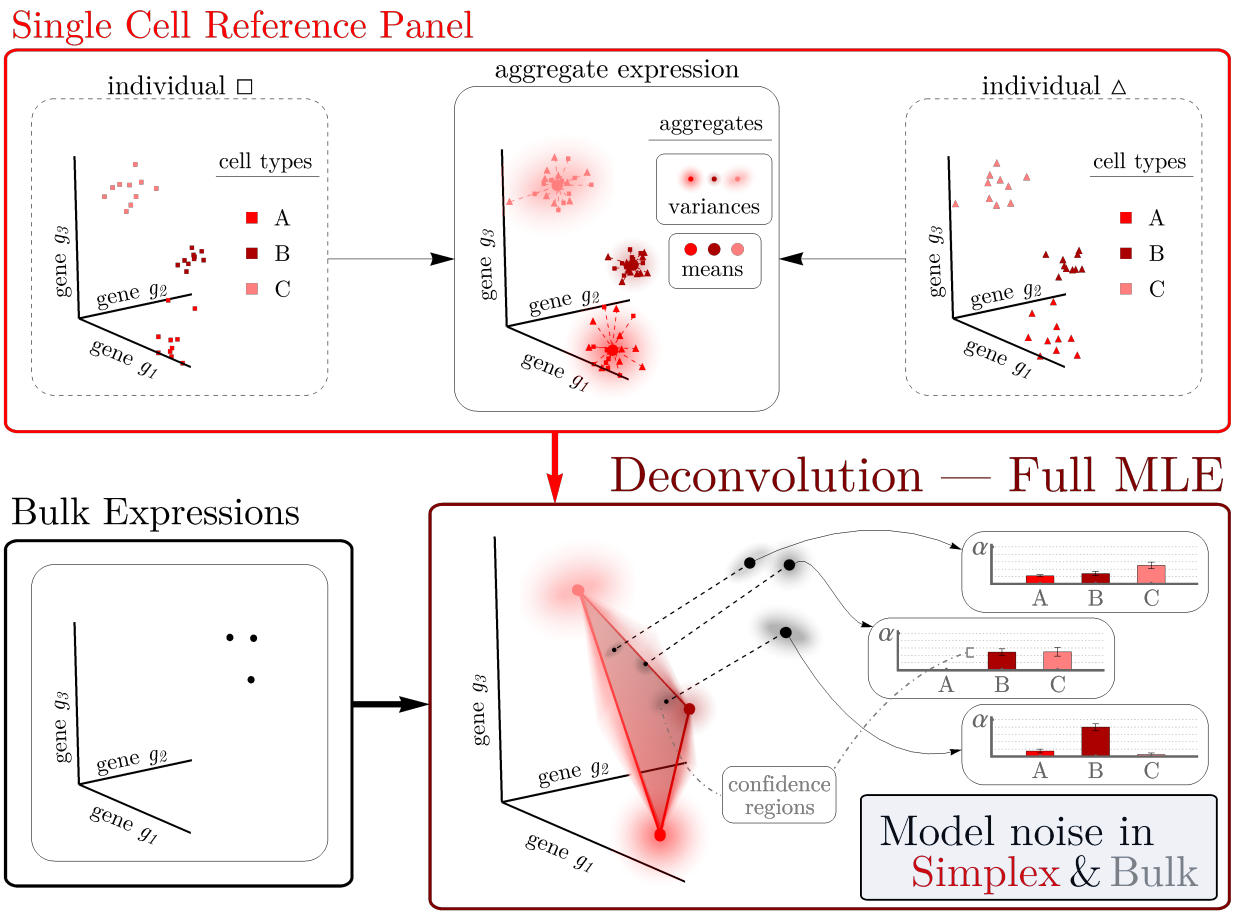


Figure 1: **The RNA-Sieve pipeline.** After applying a filtering procedure to scRNA-seq data, RNA-Sieve builds reference matrices for the mean and variance of expression for each gene across cell types. Using these estimates and bulk RNA-seq data, it performs joint deconvolution via maximum likelihood estimation by expressly modeling noise both in the reference and bulk data, yielding cell type proportion estimates for each sample.

(a) Smart-Seq2 reference and 10x Chromium pseudobulk.

	RNA-Sieve	Bisque	CIBERSORTx	DWLS	MuSiC	NNLS	Scaden	SCDC
Mean	**0.069**	0.127	0.086	0.112	0.101	0.305	0.086	0.087
Median	**0.061**	0.101	0.071	0.072	0.106	0.310	0.072	0.083
IQR	0.077	0.052	0.034	0.069	0.060	0.153	0.041	0.079

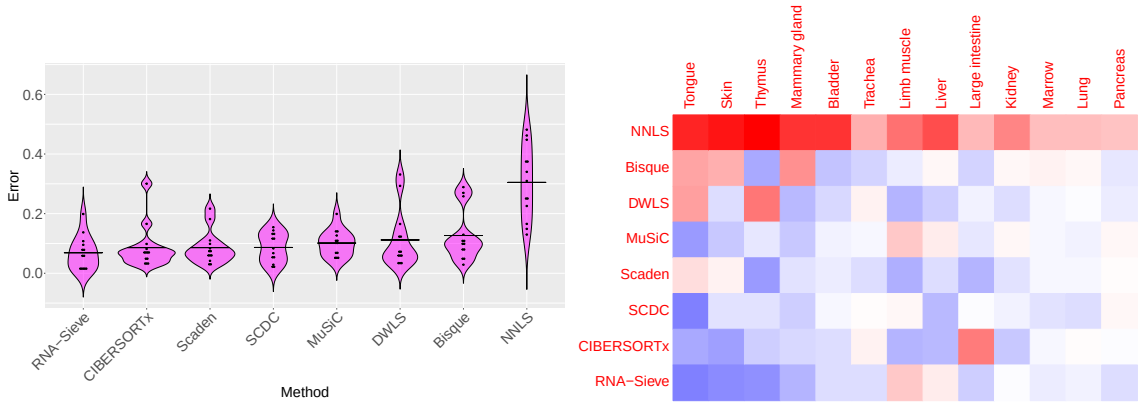
(b) 10x Chromium reference and Smart-Seq2 pseudobulk.

	RNA-Sieve	Bisque	CIBERSORTx	DWLS	MuSiC	NNLS	Scaden	SCDC
Mean	**0.067**	0.125	0.074	0.076	0.187	0.264	0.105	0.181
Median	0.082	0.110	0.062	**0.054**	0.157	0.164	0.105	0.157
IQR	0.099	0.089	0.078	0.069	0.046	0.170	0.046	0.121

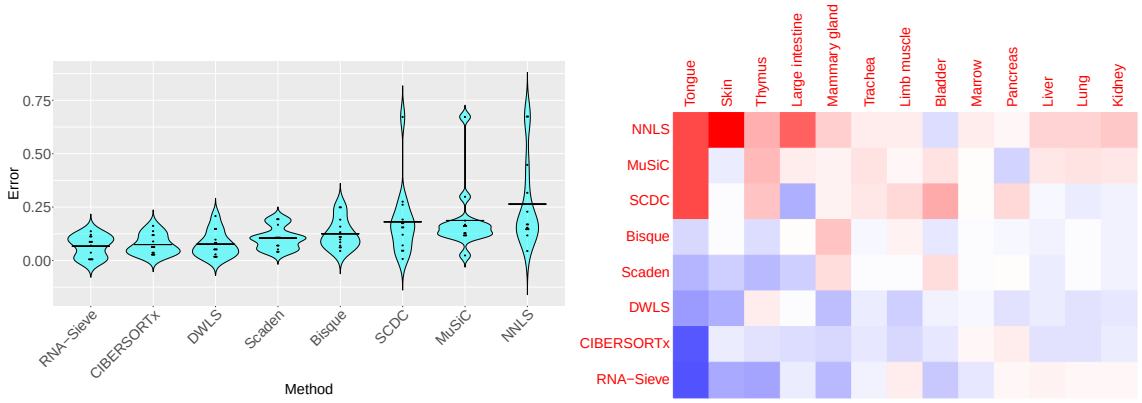
Table 1: **Summary of deconvolution errors for each considered method.** Errors were computed as the L1 distance between the inferred and true proportions averaged over the number of present cell types per organ. Single-cell RNA-seq data for the references and pseudobulks were taken from the *Tabula Muris Senis* experiment. The mean, median, and interquartile range are displayed for the results in thirteen different organs; see Methods for additional details.

we also visualized errors when aggregating by organ (that is, the column-wise distributions of the checkerboard plots in Figure 2, see Figure S1). This demonstrated that RNA-Sieve performed at least as well as all competitors, and often notably better, in nearly all organs. Our strong performance across organs regardless of the number of cell types or similarities among them suggests that RNA-Sieve is versatile over a range of scenarios. Finally, we directly compared each method's errors to those of RNA-Sieve on the same deconvolution tasks (given by the row-wise distributions of the checkerboards in Figure 2, see Figure S2). In each case, RNA-Sieve produced smaller errors than the other methods a majority of the time.

In practice, complications to the generic deconvolution problem may arise. For example, the scRNA-seq reference data may lack one or more cell types found in the bulk sample, or may even contain superfluous ones. Such problems are more likely to occur when performing cross-experimental or cross-subject deconvolutions, as we often need to do. It is thus important to examine how algorithms perform in these situations. We further recognize the necessity to demonstrate robustness to misspecification for a model-based approach like RNA-Sieve. To do so, we selected the kidney, limb muscle, liver, and marrow due to their representative ranges of cell type number and dissimiliarities, and considered all possible configurations containing one extra or missing cell type in the single-cell references. When the reference contains too many cell types, any deconvolution scheme should assign little to no mass to that cell type in the inferred proportions. We found that to be the case with RNA-Sieve (Figures 3 and S3) as long as the missing cell type is sufficiently distinct from those which are still present in the reference. If cell types are highly similar, deconvolution proportions become unidentifiable, and so are not expected to change meaningfully upon removal of one such cell type from the bulk. Meanwhile, when a cell type is missing from the reference, a very realistic possibility, the deconvolution problem becomes overdetermined. Any deconvolution algorithm should either detect a singularity, or if it cannot do so, move the weight of the removed cell type to those most similar to it. Our empirical results (Figures 4 and S4) indicate that RNA-Sieve tends to do precisely this–it assign the mass of missing cell types to other cell types contained in the reference which are transcriptomically similar. In some cases, this means mass transfer to one single cell type, while in others the weight is shared among multiple. This result suggests that



(a) **Smart-Seq2 reference and 10x Chromium pseudobulk**



(b) **10x Chromium reference and Smart-Seq2 pseudobulk**

Figure 2: **Distribution of errors for each method in pseudobulk experiments.** Pseudobulk experiments were performed in 13 different organs using data from the *Tabula Muris Senis* experiment. Errors were computed as the average L_1 error across cell types in each organ. In the violin plots, horizontal black bars correspond to the mean error and methods are ordered left to right from lowest to greatest mean error. In the grid plots, methods and organs were ordered using SVD-induced clustering. Roughly speaking, the methods from top to bottom are characterized by improving performance while the organs from left to right are characterized by decreasing variability in different methods' performances. Color indicates the difference between the average error across methods in that organ; deeper shades of red (blue) indicate poor (good) relative performance.

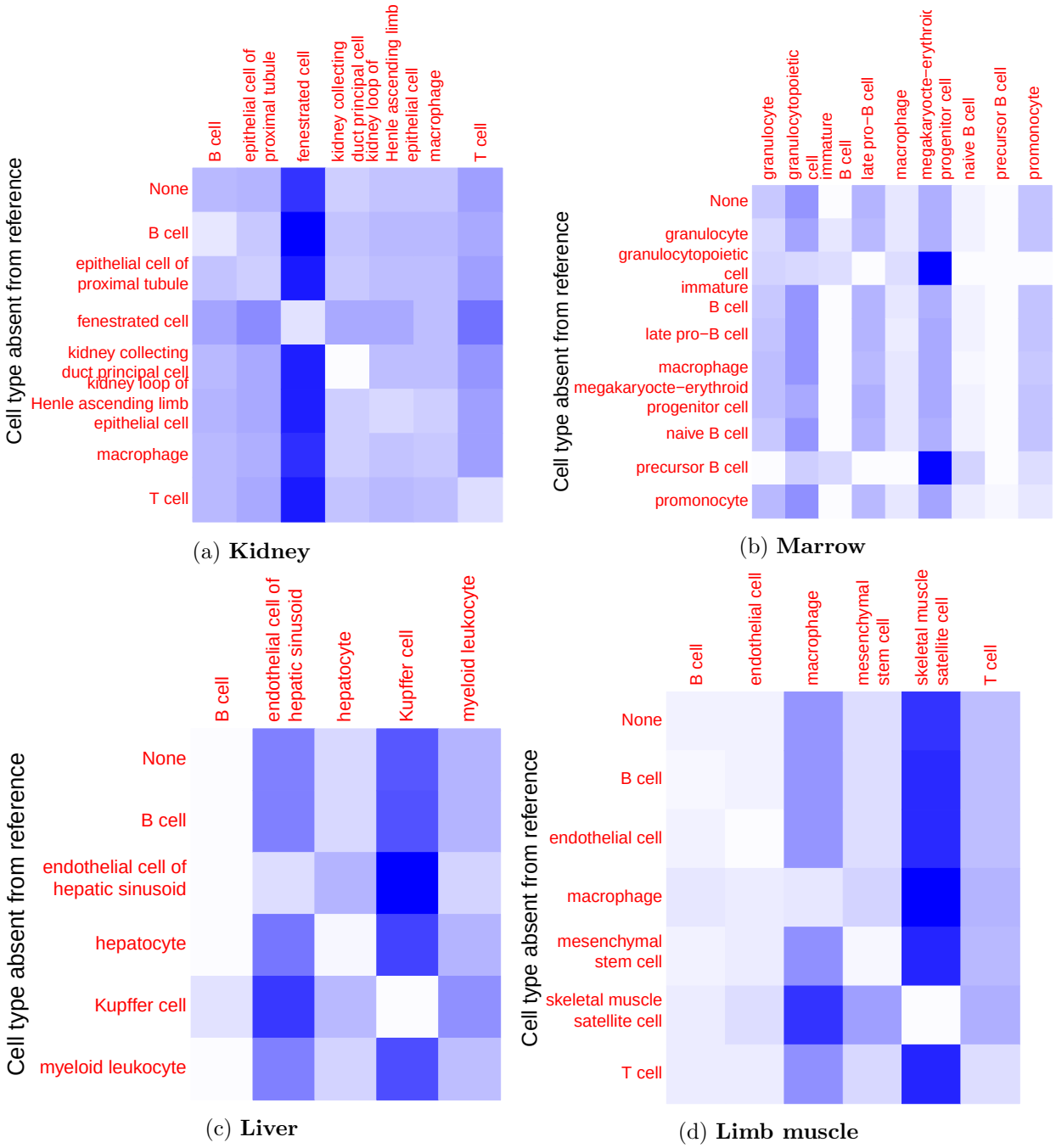


Figure 3: **Deconvolution with extra cell types in the reference matrix.** Deconvolution was performed in pseudobulk experiments in four different organs. For each organ, we followed a leave-one-out procedure in which one cell type is removed from the pseudobulk at a time. Deconvolution was then performed with this extra cell type in the reference in order to examine RNA-Sieve's specificity. The top row shows the inferred proportions with no extra reference cell types. Darker colors indicate a higher estimated proportion value. Here we used Smart-Seq2 data for the reference and 10x Chromium for the pseudobulk.

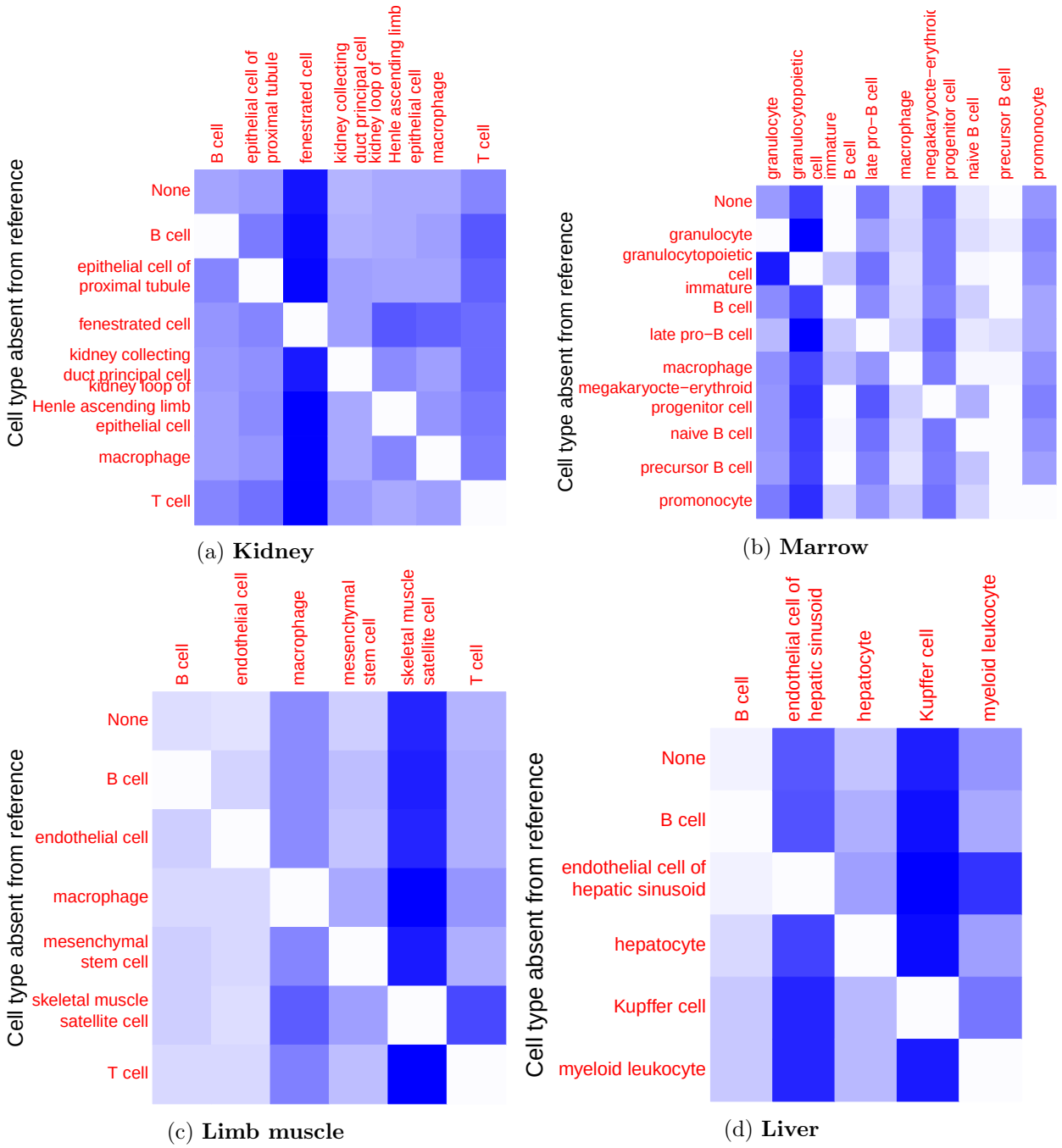


Figure 4: **Deconvolution with missing cell types in the reference matrix.** Deconvolution was performed in pseudobulk experiments in four different organs. For each organ, we followed a leave-one-out procedure in which one cell type is removed from the reference at a time. Deconvolution was then performed with an extra cell type in the pseudobulk in order to examine RNA-Sieve's ability to handle such a misspecification. The top row shows the inferred proportions with no missing reference cell types. Darker colors indicate a higher estimated proportion value. Here we used Smart-Seq2 data for the reference and 10x Chromium for the pseudobulk.

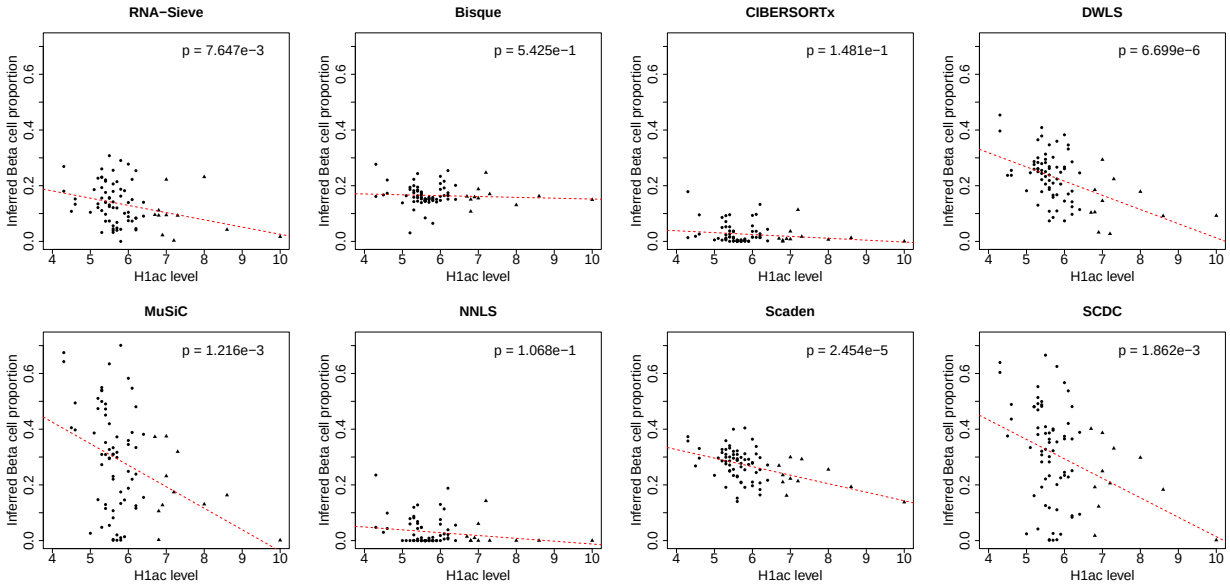


Figure 5: **Deconvolution results on validation data.** Single-cell expression data in pancreatic islets from Xin et al. (2016) was used as reference to deconvolve bulk RNA-seq data from Fadista et al. (2014). Each point represents the estimated beta pancreatic islet cell proportion one of 77 bulks with recorded HbA1c levels. The p-value is for a univariate regression on the estimated proportions. Circles correspond to healthy samples while triangles represent samples from diabetic patients.

in the case of misspecification, RNA-Sieve will still achieve sensible solutions as long as sufficiently representative cell types are captured in the reference set. We note that given the generative nature of our model, a hypothesis test to detect missing cell types is, as opposed to existing methods, within the capabilities of our framework (see Discussion).

Validation with Human Pancreatic Islet Data

To validate RNA-Sieve's performance outside of pseudobulk deconvolutions, we chose to analyze samples from the pancreatic islets region of the human pancreas. This small area primarily consists of different types of endocrine cells and is important for maintaining stable levels of blood glucose. Among these cell types is the pancreatic beta cell, which produces and secretes the hormones insulin and amylin. In patients with type 2 diabetes (T2D), beta cells become less prevalent and functional, so their proportion in bulk samples will correlate with disease status. The gene Hb1Ac (hemoglobin A1c) serves as a marker gene for T2D, and thus high Hb1Ac expression levels should be negatively associated with beta cell proportions. By examining the degree to which we recapitulate this known clinical relationship, we can assess our model's performance on real bulk RNA-seq data. To do so, we relied on previously published scRNA-seq and bulk RNA-seq samples taken from healthy and T2D patients. These same data have also been utilized as validation for other deconvolution methods and thus represent a standard and accepted basis for this task. Deconvolution via RNA-Sieve indeed yielded the expected negative correlation between HbA1c levels and estimated beta pancreatic islet cell proportions (Figure 5). Moreover, the p-value for the slope of the simple linear regression model which predicts Beta cell proportion from HbA1c levels is competitive with those produced by current state-of-the-art approaches.

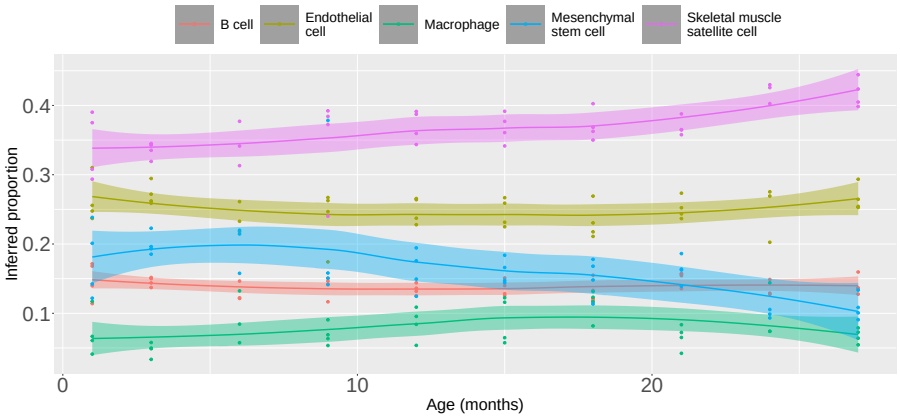
Analysis of Real Bulk Organ Samples

We next applied RNA-Sieve to real bulk RNA-seq data to look for interesting patterns in organ composition. We chose to continue working with the *Tabula Muris Senis* dataset as it contains many bulk RNA-seq samples in addition to the scRNA-seq data previously described. Due to its expansive experimental design across organs and ages, this resource is uniquely suited to interrogate changes in cell type compositions associated with the process of aging. By identifying changes in the balance of cell classes, we hope to provide insight into shifts in the mechanisms driving organ functions at different stages of life. In general, aging represents one of the more complicated biological processes, and one which occurs in every person or organism. Due to its ubiquity and significant effects on quality of life, improved understanding of the etiologies underlying age-associated functional deficits holds great potential therapeutic value. Degradation of the musculoskeletal and immune systems are among the most apparent phenotypic changes occurring during mammalian aging. Here we highlight intriguing results from three organs with roles in these bodily systems–the limb muscle in the former, and the spleen and bone marrow in the latter. In the absence of ground truth, we judge the reliability of our estimates by the relative consistency of inferred proportions within and across age groups.

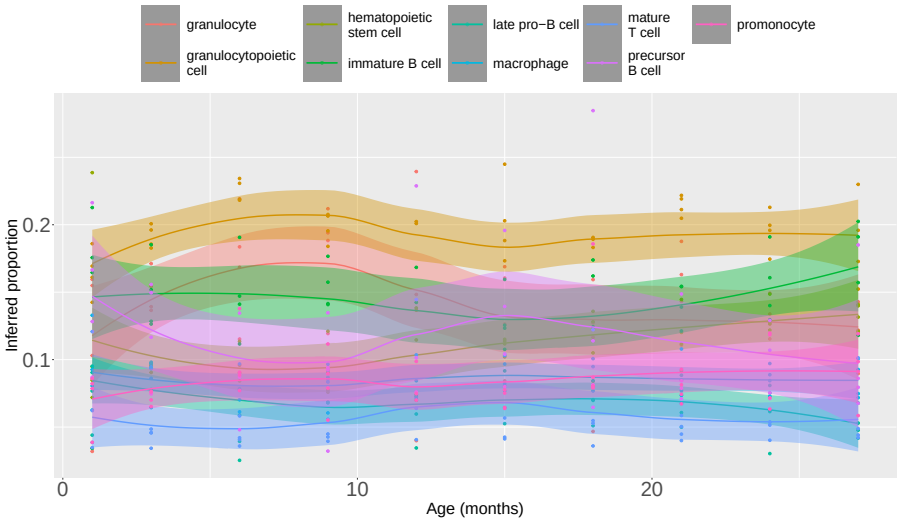
Limb muscle in the arms and legs provides support and locomotion. It primarily consists of skeletal muscle, stromal, immune, and endothelial cells. Differentiated muscle cells contract to produce the aforementioned support and motion, while satellite cells serve in myogenesis and muscular repair (Christov et al., 2007). Stromal cells comprise the connective organ which binds sarcomeres together and connects muscles to bones in addition to displaying certain regenerative capabilities (Kasprzycka et al., 2019). Immune response is often noted in muscle cells after mechanical stress-induced damage to muscle fibers, when the tissue becomes inflamed due to tearing (Novak et al., 2014; Burzyn et al., 2013). Endothelial cells are present due to the often high degree of vascularization needed to support muscle function (Christov et al., 2007). Upon application of RNA-Sieve to the available bulk muscle samples, we observed a noticeable increase in skeletal muscle satellite cells and a substantial decrease in the mesenchymal stem cell proportion in older mice (Figure 6a). These trends are present, albeit fairly gradual until around 21 months old with more sudden changes apparent thereafter. There was also an apparent increase in macrophage proportions up until 15 months of age followed by a slow decline for the remainder of life. Each of these three cell types has been demonstrated to function in muscle fiber repair through different mechanisms (Snijders et al., 2015). This behavior may thus indicate changes in the relative use of different regenerative pathways as individuals age.

The bone marrow is a vital component of the immune system which executes the bulk of hematopoiesis and contains numerous constituent cell types ranging from various stem cells to more mature cell classes (Gurkan and Akkus, 2008). This rich combination yielded several age-associated trends in cell type composition, and we choose to focus on two. First, an effectively linear growth in the number of hematopoietic stem cells was observed with increasing age. Though this may seem surprising given reduced adaptive immunity with age, this exact phenomenon has been previously observed in both mice and humans (Pang et al., 2011), and it is accompanied by a decrease in functionality of these cells. Conversely, granulocyte proportion appeared to decrease after roughly 9 months of age. Further examination reveals that the granulocyte fraction tends to mirror that of granulocytopoietic cells, but with an increasing deficit between the two as age increases. Such a pattern is suggestive of the reduced potency of granulocytopoiesis that we would expect with age. Hence, in the marrow we are able to identify known patterns of cell type composition variation despite the presence of many transcriptomically similar cell types.

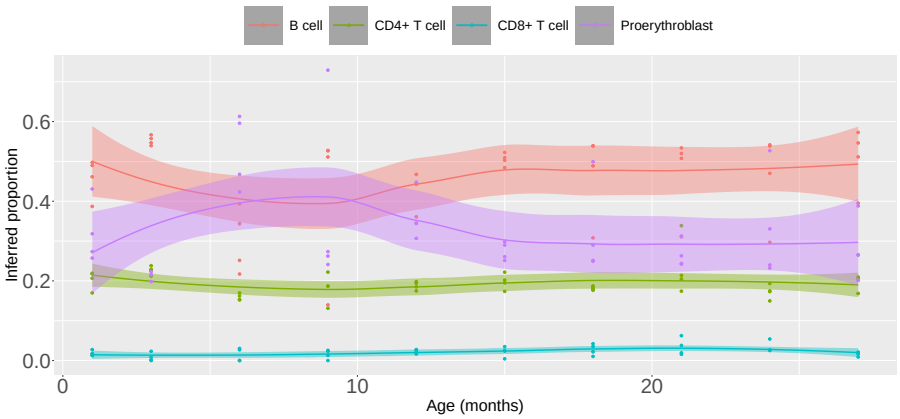
The spleen occupies a central role in the lymphatic system and is important, though not strictly



(a) Limb muscle



(b) Bone marrow



(c) Spleen

Figure 6: **Deconvolution results for real bulks from the *Tabula Muris Senis*.** Roughly forty real bulk samples across ten ages were deconvolved using RNA-Sieve for the limb muscle, bone marrow, and spleen. In all cases, Smart-Seq2 data were used as the reference. Each point represents the inferred proportion for a given cell type in a bulk sample. Lines display the smoothed trend of proportions as a function of age, with uncertainty shown by the shaded intervals.

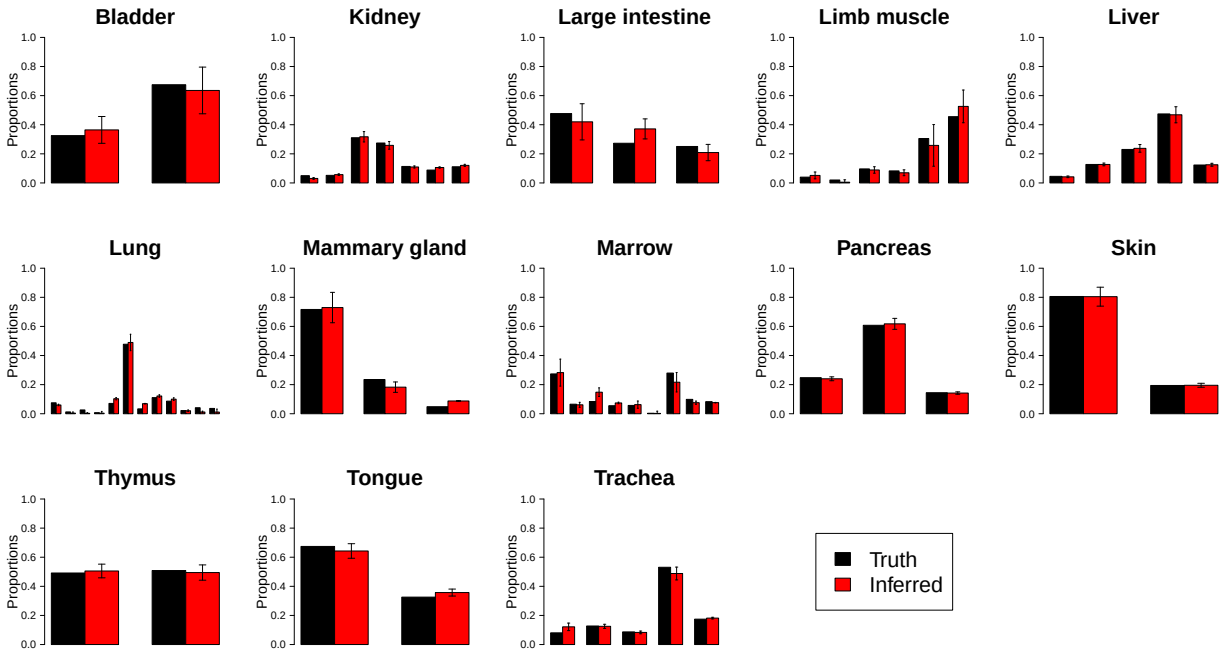
essential, for proper functioning of the immune system and red blood cell recycling. This organ is split into two pieces—the red pulp, which contains blood cells, and the white pulp, which is primarily lymphatic (Mebius and Kraal, 2005). Typically, the large majority of present immune cells are B and T cells, with smaller quantities of other cell types (Hensel et al., 2019). Various progenitor cells may be present to spur production of immune and red blood cells, though these processes are primarily performed in the bone marrow (Mumau et al., 2018). It is notably difficult to distinguish among these progenitor cells in their early stages, making it possible that several varieties are captured within the single label of proerythroblasts. Upon deconvolution, we found that our inferred proportions for B and T cells matched accepted ranges (Hensel et al., 2019). More interestingly, we noticed an unexpected and transient spike in the proportion of proerythroblasts peaking at roughly nine months of age (Figure 6c). Importantly, this increase is observed in all four of the 9-month-old individuals, and is thus not an artifact of outlying samples. Mice at this age are roughly analogous to humans of between 30-40 years of age, and as hematopoiesis is generally restricted to the marrow at this age except under stress conditions, it is unclear whether a programmed hematopoeitic process is occurring or if we are capturing the behavior of a cell type not enumerated in the reference set.

Extension to Confidence Regions

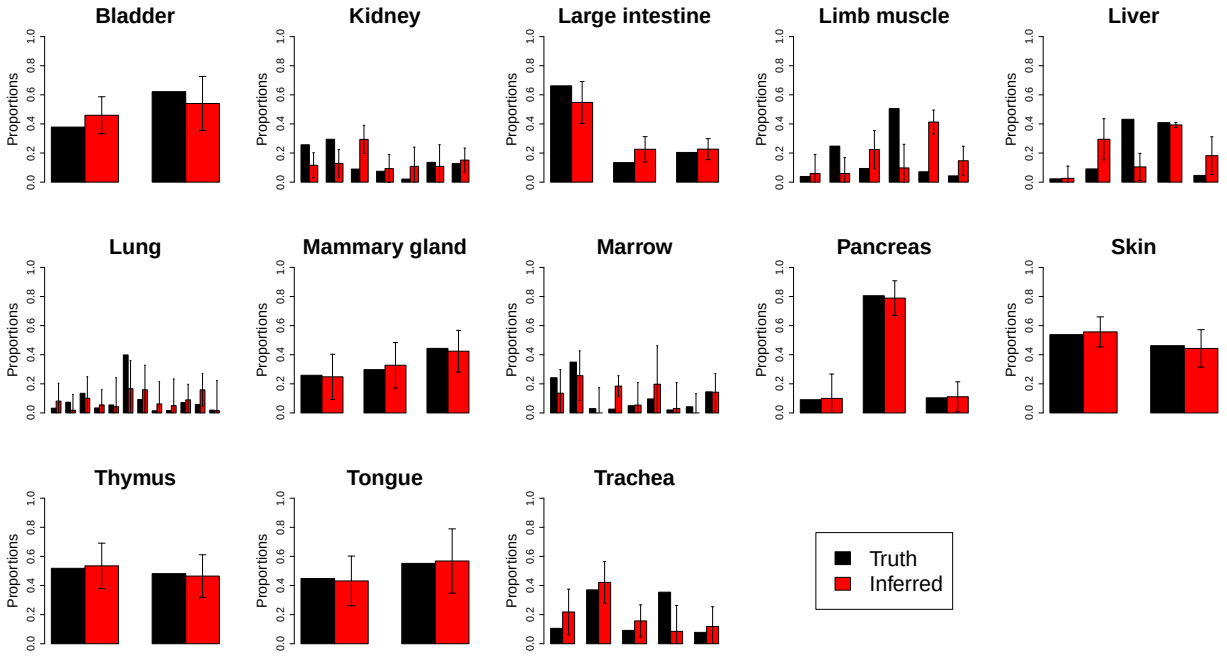
Beyond deconvolution, the generative framework of RNA-Sieve permits extensions which remain out of reach using prior approaches. One such possibility is the computation of confidence regions for inferred cell type proportions. Despite its clear importance, error quantification in deconvolution is challenging and has received relatively scant attention, leaving users to only guess at the reliability of their results. As deconvolution is sometimes performed upstream of tasks such as differential expression or eQTL detection, it is critical to know whether inferred proportions are precise. Because RNA-Sieve infers these proportions via maximum likelihood estimation, we can directly tap into the wide array of theory on asymptotic confidence bounds. Specifically, we construct confidence regions for inferred proportion values through numerical computation of the inverse Fisher information matrix (see Confidence Intervals in Methods). We demonstrate RNA-Sieve's ability to produce well-calibrated confidence regions by constructing them for the cross-protocol deconvolutions performed in Performance in Pseudobulk Experiments in addition to within-protocol experiments of similar kind.

For the within-protocol comparisons of Figures 7a and S5a we obtain narrow, yet well-calibrated, confidence intervals, which indicate the effectiveness of our procedure when all modeling assumptions are met. However, the typical deconvolution setting will present complications in the form of protocol differences in the scRNA-seq reference and bulk RNA-seq data. With the addition of an automated procedure to ensure robustness (see Confidence Intervals), we find that we still achieve good performance in spite of protocol mismatch (Figures 7b and S5b). To ensure that we obtain sensible results with real bulk RNA-seq data, we also generated confidence intervals for the samples analyzed in Analysis of Real Bulk Organ Samples. Though assessing their accuracy is impossible absent ground truth, we found that the interval widths appear reasonable and similar to those we obtained with scRNA-seq pseudobulks (Figure 8). The distribution of confidence interval half-widths for cell type proportions were also generally consistent across samples (Figure S6). We note that MuSiC presents a quantity which seemingly corresponds to the variance in proportion estimates, though it was not emphasized in their manuscript and we generally found the produced values to be overly small in our trials.

In principle, the widths of confidence intervals should depend on the number of cells and genes in the reference, similarity among cell types in the reference, and agreement between the reference

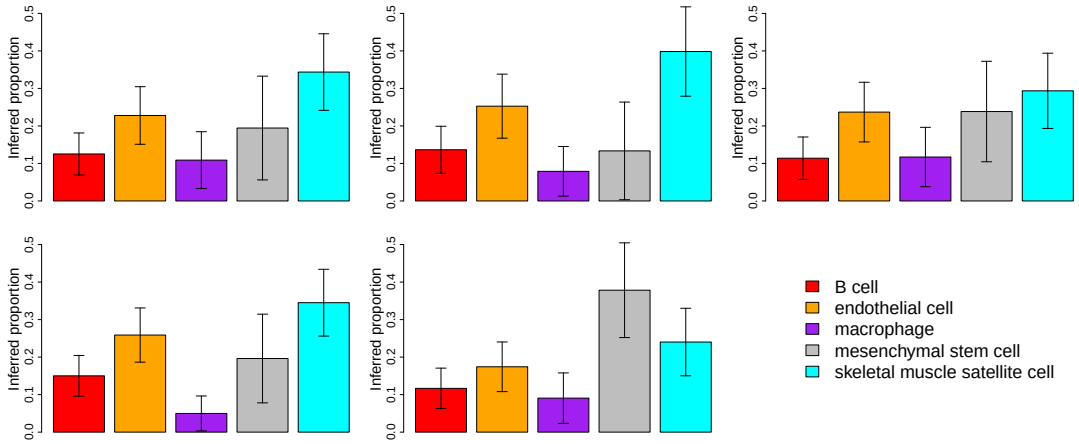


(a) **Within-protocol – Smart-Seq2 pseudobulk and reference**

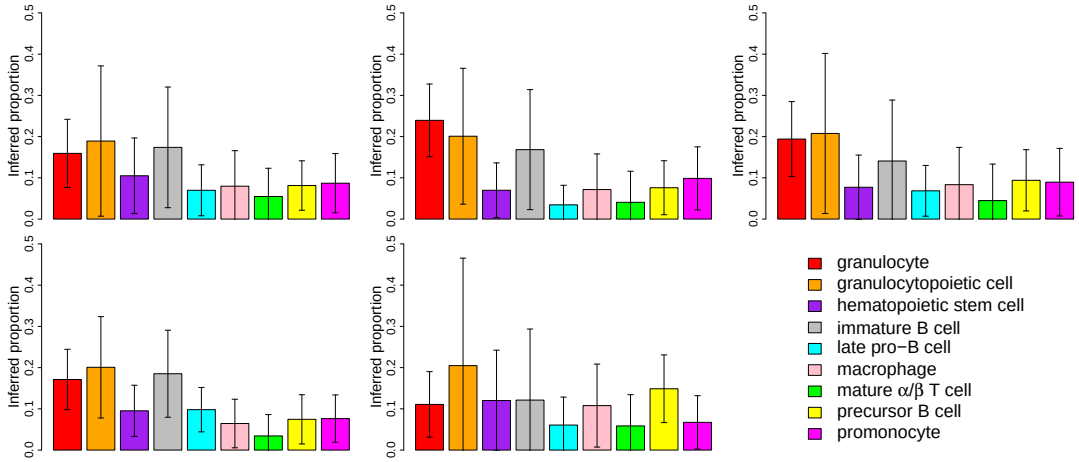


(b) **Cross-protocol – Smart-Seq2 reference and 10x Chromium pseudobulk**

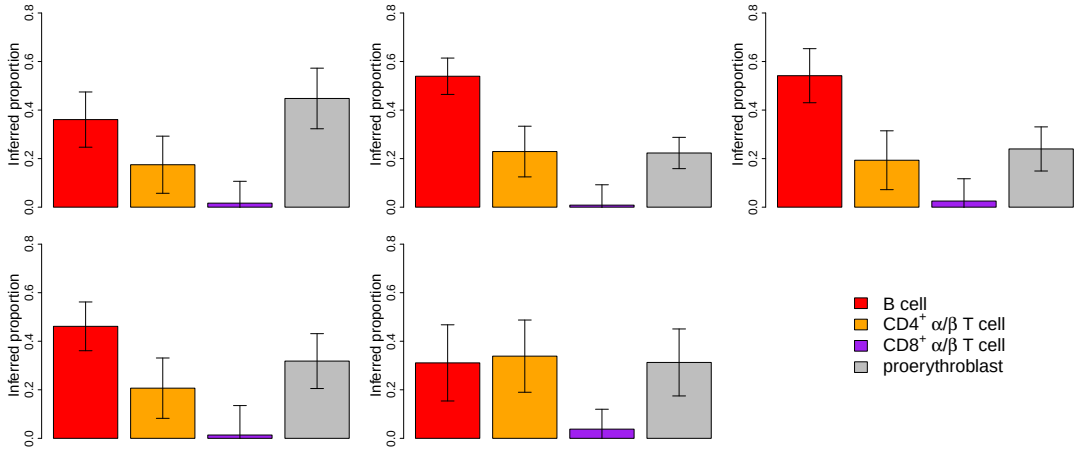
Figure 7: **RNA-Sieve results with confidence intervals in pseudobulk experiments.** Inferred cell type proportions in pseudobulk experiments using data from the *Tabula Muris Senis* experiment. The black error bars on inferred proportions show the marginal 95% confidence intervals as computed from the estimated Fisher information produced by RNA-Sieve. See Confidence Intervals for mathematical details. Table S2 contains the cell types in each organ, which could not be displayed because of space constraints.



(a) **Limb muscle**



(b) **Marrow**



(c) **Spleen**

Figure 8: **Confidence intervals with real bulk samples** For all of each organ's samples, we produced estimated cell type proportions with 95% confidence intervals using RNA-Sieve. Smart-Seq2 data were used as the reference. Here we present five randomly chosen samples for each organ (out of ∼40); histograms showing the typical radii are displayed in Figure S6.

and bulk measurements. Our empirical results suggest that these eminently logical factors do, in fact, drive the widths of our intervals. For example, the confidence intervals in cross-protocol deconvolutions are a bit wider than their within-protocol counterparts, due to our adaptive procedure's conservative nature when it detects differences between the reference and bulk. This arises in part because we deem fewer genes reliable when compared to within-protocol experiments. Moreover, evidence of the contribution of reference sample size is present in a few organs, most notably the lung with its many low-frequency cell types.

Discussion

Here we have introduced our method for supervised bulk gene expression deconvolution, RNA-Sieve, and illustrated its robust performance in a variety of settings. Unlike methods which rely on variants of least squares or the application of black-box machine learning algorithms, we place the deconvolution problem into a generative probabilistic framework by relying on distributional results from asymptotic theory. Through simulations and applications to real data, we demonstrated the broad applicability of our method and its utility to investigate biological questions of interest.

It is valuable to understand how RNA-Sieve differs from other approaches and to consider the consequences of these divergent design choices. Least squares-based solutions such as MuSiC (Wang et al., 2019), SCDC (Dong et al., 2020), and DWLS (Tsoucas et al., 2019) devise their own implementations of weighted non-negative least squares (W-NNLS). These methods aim to handle heteroskedasticity across genes by re-weighting them according to their variability and specificity, allowing genes which are putatively more informative to carry increased importance in the regression task. Alternatively, Bisque (Jew et al., 2020) uses NNLS after applying a transformation to bring the reference and bulk data into better distributional agreement. From a modeling perspective, least squares-based solutions generally address uncertainty in the bulk only, leaving stochasticity in the single-cell reference unaccounted for. With RNA-Sieve, rather than devising a specialized gene-weighting scheme, we naturally emphasize some genes more than others via variances resulting from an explicit generative model incorporating noise in both single cells and the bulk. We also do not explicitly attempt to bring reference and bulk data into better agreement a là Bisque, instead relying on a filtering protocol to remove genes which display signs of significant deviation from our assumptions. Integrating an explicit transformation is an interesting possibility for RNA-Sieve, and should only boost its performance by further aligning data to our model assumptions. Other methods employ machine learning techniques, such as CIBERSORTx (Newman et al., 2019), which uses ν-support vector regression, and Scaden (Menden et al., 2020), which utilizes deep neural networks. These approaches can be opaque to the user due to their reliance on black box algorithms. Comparatively, our formulation of RNA-Sieve as the MLE of an explicit generative model is transparent in both parameter interpretation and performance guarantees. The parameters updated during optimization have explicit biological meanings and tracing their values allows for a deeper interrogation of the predictions RNA-Sieve generates. This is a useful feature when providing context to inferred cell type proportions as well as exploring the theoretical limits of deconvolution as a function of cell type properties. Like MuSiC, SCDC, Bisque, and Scaden, we do not select marker genes in RNA-Sieve. This helps us maintain computational efficiency, while simultaneously providing robustness with respect to outlier fluctuations in gene expression. We also parallelize our optimization steps and jointly update parameters when deconvolving multiple bulk samples. This yields significant speedups relative to serial runs and allows us to share statistical strength across all bulks.

RNA-Sieve is embedded in a flexible generative framework, which can be adapted to a variety of

situations to make deconvolution performance more effective. One of these is the modeling of further sources of variation. For instance, if gene expression distributions are expected to differ drastically across individuals from which samples are taken, this knowledge can be explicitly incorporated into our likelihood. Without such modification, RNA-Sieve implicitly follows the paradigm of MuSiC, SCDC, and Bisque in penalizing genes of large inter-individual variance via the marginal variances resulting from estimation of the reference panel. A similar notion applies to mitigating potential batch effects or effectively combining disparate references, which both enjoy natural interpretations when considering likelihoods.

A principal motivation of this work was to expand the scope of accessible questions in the deconvolution setting. Our likelihood-based approach facilitates extensions which are intractable with current algorithms. As a first step, we have chosen to demonstrate our ability to explicitly construct confidence regions for inferred proportions, producing a mathematically rigorous quantification of the uncertainty in our estimates. The necessity of these bounds is plainly substantiated by the use of deconvolution upstream of tasks ranging from cell type-specific differential expression to eQTL detection using heterogeneous RNA-seq organ samples. The credibility of any such analyses is predicated on the accuracy of deconvolution, because any errors in this initial step will propagate through to the final result. Consequently, we anticipate that our confidence regions will encourage improved assessment of the reliability of results obtained through these types of analyses. Our confidence intervals are also of obvious inherent value when using deconvolution results to infer differences in cell type composition between samples, whether due to disease status or other factors. Beyond error quantification via confidence intervals, potent possibilities lie in hypothesis testing. Currently, CIBERSORTx does propose one type of test, though our understanding is that it tests whether *any* of the bulk cell types were found in the reference. This is rather restrictive, so we hope to develop procedures with broader utility. One example with clinical impact is a test to determine whether the reference panel is missing cell types present in the bulk sample. Even though we have demonstrated that RNA-Sieve is robust with respect to such misspecification (see Performance in Pseudobulk Experiments), it is nonetheless beneficial to know whether the deconvolution performed was sufficiently valid using a principled approach. Such a test can be directly developed in our framework by examining the residuals produced by our maximum likelihood estimate, and work in this direction is underway.

Despite the flurry of recently developed methods, the question of statistical deconvolution of gene expression data remains far from solved. RNA-Sieve illustrates the efficacy as well as promise of generative modeling in this setting, and we hope it spurs continued development within other methodological paradigms. In particular, notions of error quantification and hypothesis testing merit further attention.

Methods

Notation

To ease parsing of technical equations, we briefly introduce our notation here. We generally refer to vector quantities with boldfaced lowercase letters, while plain lower- and uppercase symbols are reserved for scalars (or scalar functions) and matrices, respectively. The kth column vector of a matrix $A = (a_{ij})_{ij}$ is written as $\boldsymbol{a}_k$, and inner products between vectors $\boldsymbol{v}, \boldsymbol{w}$ are typically denoted $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$. To distinguish observed, random quantities from their underlying deterministic, ground truth objects we add tildes to the former and asterisks to the latter; e.g., $\tilde{\boldsymbol{b}}$ are observed bulk gene expressions, while $\boldsymbol{b}^*$ are the true bulk gene expression means. Finally, estimates of latent parameters carry hats; e.g., $\hat{\boldsymbol{\alpha}}$ is the vector of mixture weights inferred by our deconvolution

procedure.

Mathematical Model

We assume that for each gene $g \in \{1, \dots, G\} = [G]$, and cell type $k \in \{1, \dots, K\} = [K]$, there exists a distribution $\nu_{g,k}$ describing the expression of gene g in cell type k. As multiple cell types comprise any given organ, the expression of gene g in a cell drawn at random from a organ is governed by the mixture distribution

$$\rho_g = \sum_{k=1}^{K} \alpha_k^* \nu_{g,k}, \tag{2}$$

where $\boldsymbol{\alpha}^* = (\alpha_k^*)_{k\in[K]} \in \Delta^{K-1}$ contains the proportions of each cell type in the organ of interest. Despite the *a priori* infinite-dimensional setting, if $G > k$ and ρ_g, $\{\nu_{g,k}\}_{k\in[K]}$ are fully known and sufficiently distinct, the convex combination of (2) immediately implies that $\boldsymbol{\alpha}^*$ can be recovered as the unique solution of the finite-dimensional problem

$$\underbrace{\begin{bmatrix} f(\nu_{1,1}) & \dots & f(\nu_{1,K}) \\ \vdots & & \vdots \\ f(\nu_{G,1}) & \dots & f(\nu_{G,K}) \end{bmatrix}}_{M} \cdot \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}}_{\boldsymbol{\alpha}} = \underbrace{\begin{bmatrix} f(\rho_1) \\ \vdots \\ f(\rho_G) \end{bmatrix}}_{\boldsymbol{b}}, \tag{3}$$

where $f \in \mathcal{M}(\mathbb{R})^*$ is any suitable linear functional on the space of measures on $\mathbb{R}$. Natural f to consider include the family of point evaluations $\{\delta_x : \nu \mapsto F_\nu(x)\}_{x\in\mathbb{R}}$ (where F_ν denotes the cumulative distribution function (CDF) of ν) or its ith moments $\{\mu_i : \nu \mapsto \int x^i \nu(dx)\}_{i\in\mathbb{N}}$, both of which enjoy a wealth of statistical theory and proposed estimators.

In experimental settings, exact gene expression distributions are not accessible and instead must be estimated, so utilizing easily and robustly inferable f becomes crucial. In addition to not having direct access to $\{\rho_g\}_{g\in[G]}$, any analysis is further complicated by the fact that bulk sequencing only yields gene expression levels over whole samples and not for particular cells or cell types. That is, the output is effectively a random variable $X_g = \sum_{i=1}^n X_{g,i}$ where $X_{g,i} \overset{iid}{\sim} \rho_g$ gives the measured expression of gene g aggregated over the $n \in \mathbb{N}$ individual cells comprising the sample. Since $X_g \sim (\rho_g)^{\circledast n}$ is an n-fold convolution, it is expedient to choose an f in (3) that is not only linear on the space of measures, but also on the corresponding space of measurable functions. In particular, setting $f = \mu_1 = \mathbb{E}$ turns (3) into

$$M\boldsymbol{\alpha} = \boldsymbol{b}/n, \qquad \text{where} \qquad M_{g,k} = \mathbb{E}_{Y\sim\nu_{g,k}} Y \quad \text{and} \quad b_g = \mathbb{E}X_g. \tag{4}$$

Incorporating the fact that we only observe noisy bulk samples X_g instead of $\boldsymbol{b}$ directly results in

$$\tilde{\boldsymbol{b}}/n = (\boldsymbol{b} + \varepsilon_{\boldsymbol{b}})/n = M\boldsymbol{\alpha} + \varepsilon_{\boldsymbol{b}}/n, \tag{5}$$

where $(\varepsilon_{\boldsymbol{b}})_g \sim X_g - b_g = \sum_{i=1}^n (X_{g,i} - b_g/n) \sim \mathcal{N}(0, n \cdot \sigma_g^2(M, \boldsymbol{\alpha}, S))$ for large n by the central limit theorem (CLT), with $\sigma_g^2(M, \boldsymbol{\alpha}, S) = \mathrm{Var}(\rho_g)$ as a function of $S = (s_{g,k})_{g,k} = \mathrm{Var}(\nu_{g,k})$. When ignoring the dependence of σ_g^2 on $\boldsymbol{\alpha}$, (5) lends itself to a simple (weighted) non-negative least squares scheme solving

$$M\boldsymbol{\alpha} = \tilde{\boldsymbol{b}}. \tag{6}$$

This yields a solution $\hat{\boldsymbol{\alpha}}_{LS}$ of roughly $\|\hat{\boldsymbol{\alpha}}_{LS}\|_1 \approx n$ that simply requires re-scaling to fit onto the simplex. This together with data-driven modifications is the approach pursued in Dong et al.

(2020); Tsoucas et al. (2019); Wang et al. (2019), where it is argued that (6) outperforms previous methods. It is thus natural to ask whether incorporating the dependence of σ_g^2 on $\boldsymbol{\alpha}$ can improve accuracy even further.

To answer this question, we first make explicit this $\boldsymbol{\alpha}$-dependence by computing

$$\begin{aligned}\sigma_g^2 = \sigma_g^2(M, \boldsymbol{\alpha}, S) = \mathrm{Var}(\rho_g) &= \mu_2(\rho_g) - \mu_1^2(\rho_g) \\ &= \left(\sum_{k=1}^{K} \alpha_k^* \mu_2(\nu_{g,k})\right) - b_g^2 \\ &= \left(\sum_{k=1}^{K} \alpha_k^* \Big[s_{g,k} + m_{g,k}^2\Big]\right) - b_g^2. \end{aligned} \tag{7}$$

The likelihood of observing data $\hat{\boldsymbol{b}}$ then follows straightforwardly from the central limit theorem:

$$\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n}\left(\tilde{\boldsymbol{b}} \in d\boldsymbol{p}\right) = \prod_{g=1}^{G} \frac{1}{\sqrt{2\pi n \sigma^g(M, \boldsymbol{\alpha}, S)}} \exp\left\{\frac{-(p_g - (M\boldsymbol{\alpha})_g)^2}{2n\sigma_g^2(M, \boldsymbol{\alpha}, S)}\right\} \tag{8}$$

The above assumes exact knowledge of the individual distributions $\nu_{g,k}$ (or rather their expectations $m_{g,k}$), which is implausible in experimental settings. Instead, M needs to be estimated from data through some estimator $\tilde{M}$, which we conveniently take to be the sample mean of expression across cells within each cell type, $\tilde{m}_{g,k} = \frac{1}{c_k}\sum_{i=1}^{c_k} C_{g,k}^i$, where $C_{g,k}^i \overset{iid}{\sim} \nu_{g,k}$, and c_k denotes the number of single cells of cell type k. With this additional correction, (5) becomes

$$\tilde{\boldsymbol{b}}/n = \Omega\boldsymbol{\alpha} + \varepsilon_{\boldsymbol{b}}/n \quad \text{and} \quad \tilde{M} = M + \varepsilon_M \tag{9}$$

where ε_M is a matrix of entries $(\varepsilon_M)_{g,k}$ independently following $\mathcal{N}(0, s_{g,k}/c_k)$ distributions. The new likelihood thus becomes

$$\begin{aligned}\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n,\boldsymbol{c}}\left(\tilde{\boldsymbol{b}} \in d\boldsymbol{p}, \tilde{M} \in dO\right) = &\prod_{g=1}^{G} \frac{1}{\sqrt{2\pi n \sigma_g^2(M, \boldsymbol{\alpha}, S)}} \exp\left\{\frac{-(p_g - (M\boldsymbol{\alpha})_g)^2}{2n\sigma_g^2(M, \boldsymbol{\alpha}, S)}\right\} \\ &\times \prod_{g\in[G], k\in[K]} \frac{1}{\sqrt{2\pi s_{g,k}/c_k}} \exp\left\{\frac{-(o_{g,k} - m_{g,k})^2}{2s_{g,k}/c_k}\right\}. \end{aligned} \tag{10}$$

This is the likelihood with which we work in our model, whose suitability depends on a handful of implicit assumptions which are worth examining. The first is that the large number of cells assayed in an experiment permits us to use asymptotic theory and apply the classical CLT. As a result, we can write down a likelihood for our observations using normal distributions as long as $\mu_2(\nu_{g,k}) < \infty$, which is true since gene expression profiles are necessarily bounded. Secondly, we suppose that the errors arising from estimating $\boldsymbol{b}$ and M are independent. This is appropriate as the bulk and single-cell experiments are performed separately. We additionally presume that expression levels in different genes are independent, as are those in different cells. It is unclear whether the latter is completely true in practice, though there is little evidence to the contrary. On the other hand, expression levels across genes within samples (either bulk or individual cells) are liable to be somewhat dependent due to expression co-regulation and the nature of the sampling process performed in RNA-seq. Given the large number of genes assayed, the latter co-dependence is apt to be fairly small. Meanwhile, co-expression estimation in single cells remains an open problem independent of deconvolution tasks, and so is not accounted for in RNA-Sieve. Once correlation structure is known however, it is straightforwardly incorporated into the likelihood we propose.

Remark (Joint deconvolution of multiple bulk samples)**.** *If it is known that multiple bulk gene expression vectors share the same constituent cell type expression profiles, we can gain statistical strength and decrease the computational burden by inferring their mixture proportions jointly rather than individually. Assuming statistical independence of the bulk sample observations, we must simply augment the likelihood in* (10) *by including the* $N-1$ *additional mixtures in* $A = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) \in \mathbb{R}^{K\times N}, \tilde{B} = (\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_N) \in \mathbb{R}^{G\times N}$*:*

$$\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n,\boldsymbol{c}}\left(\tilde{B} \in dP, \tilde{M} \in dO\right) = \prod_{b=1}^{N}\prod_{g=1}^{G} \frac{1}{\sqrt{2\pi n \sigma_g^2(M, \boldsymbol{\alpha}_b, S)}} \exp\left\{\frac{-((\boldsymbol{b}_b)_g - (M\boldsymbol{\alpha}_b)_g)^2}{2n\sigma_g^2(M, \boldsymbol{\alpha}_b, S)}\right\}$$
$$\times \prod_{g\in[G], k\in[K]} \frac{1}{\sqrt{2\pi s_{g,k}/c_k}} \exp\left\{\frac{-(o_{g,k} - m_{g,k})^2}{2s_{g,k}/c_k}\right\}. \tag{11}$$

Data Pre-processing Procedure

Due to the well-known influence of technical variability in scRNA-seq data, we suggest that users of RNA-Sieve perform their own quality control filtering of cells and genes prior to running our software in addition to their preferred normalization. Given the potential complexity of these patterns in general, we feel that manual cleaning is more reliable than automated procedures. Nonetheless, we implement a simple, largely optional, cell filtering and normalization scheme to ensure the accuracy of results when the user has chosen not to perform their own quality control. Our procedure attempts to do the following:

1. Remove low-quality cells with anomalously low or high library sizes (≥ 3 median absolute deviations away from the median value of total number of reads per cell in each cell type)
2. Normalize read counts in cells (re-scale reads so that all cells have the median number of reads from across all cells);
3. Identify and remove cells which may be mislabeled or are simply extremely different from other cells with the same cell type label (≥ 3 median absolute deviations away from the median value of inter-cellular pairwise distances in each cell type)
4. Identify and retain genes which are expressed sufficiently often ($\geq 20\%$ in at least one cell type).

We note that the first three steps are optional whereas step 4 is necessary to remove lowly expressed genes, whose presence may result in poor optimization outcomes due to creating biologically implausible expressions (a non-zero bulk expression can never be realized as a convex combination of zero or almost zero, low variance, single cell expressions).

We implemented two additional layers of gene filtering which we found improved robustness to cross-protocol differences in reference and bulk gene expression measurements. The motivation behind these steps is as follows:

1. By virtue of being a convex combination of expression levels from different cell types (under our generative model (10)), a gene's *true* expression b_g must necessarily lie between its smallest and largest corresponding expressions $m_{g,k}$ across cell types $k \in [K]$. That is,

$$b_g \in \left[\min_{k\in[K]} m_{g,k}, \max_{k\in[K]} m_{g,k}\right], \tag{12}$$

which naturally motivates a filtering scheme based on violations of these constraints. Of course, these inequalities do not necessarily hold in the presence of observational noise, which may push a gene's bulk expression outside of its theoretical extremes. However, a stochastic version of (12) persists in that

$$\mathbb{P}_{M,S}^{\boldsymbol{\alpha},n,\boldsymbol{c}}\left[\delta\left(\tilde{b}_g, \left[\min_{k\in[K]} \tilde{m}_{g,k}, \max_{k\in[K]} \tilde{m}_{g,k}\right]\right) \geq t\right] \tag{13}$$

decays in t with sub-Gaussian tails (with constants depending on $\{\sigma_{g,k}\}_{k\in[K]}$), where $\delta(p, A) = \inf_{a\in A} |p-a|$ is the shortest distance of the point p to a set A. It is thus plausible to filter out all genes for which (13) is sufficiently small (in principle, computing the precise tail bounds (13) requires access to the true parameter $\boldsymbol{\alpha}$, which prior to deconvolution is not available; however, reasonable upper bounds of (13) can be calculated independently of $\boldsymbol{\alpha}$).

2. Gene expression profiles may experience (occasionally drastic) shifts when measured with distinct protocols. For example, mean and variance estimates of some gene expression levels may correlate little, or even not at all, across data generated using Smart-Seq2 and UMI-based technologies. To identify and remove these genes, we resort to a handful of empirically effective filtering steps. Specifically, we remove a gene if it presents as an outlier (as measured by median absolute deviations from the median) in any of the following summary statistics:

$$T_M(g) = \max_{k\in[K]} \hat{m}_{g,k}, \qquad T_S(g) = \max_{k\in[K]} \hat{s}_{g,k},$$

$$R_{S/M}(g) = \max_{k\in[K]} \frac{\hat{s}_{g,k}^{\frac{1}{2}}}{\hat{m}_{g,k}} = \max_{k\in[K]} c_V(C_{g,k}), \qquad R_{\boldsymbol{b}/M}(g) = \min_{k\in[K]} \frac{\hat{b}_g}{\hat{m}_{g,k}}, \tag{14}$$

where $c_V(C_{g,k})$ denotes the coefficient of variation associated with expression profiles of gene g in cell type k. While the choice of these summary statistics was primarily guided by empirical considerations, they do reveal intuitively plausible and previously observed patterns: T_M, T_S and $R_{S/M}$ reflect the fact that severe over- or under-expression, or high degrees of variability in expression are not well-preserved across protocols, whereas $R_{\boldsymbol{b}/M}(g) = \hat{b}_g / \max_{k\in[K]} \hat{m}_{g,k}$ directly assesses any abnormal conversion factors between bulk and reference protocols.

In our experience, applying these filters based on (13) and (14) on top of the basic cell filter leaves an average of around $8,000$ genes on which to perform the deconvolution task.

Optimization and Estimation

We estimate $\boldsymbol{\alpha}$, the cell type proportions for a given bulk sample, using the MLE which arises from maximizing (10). Given the number of free parameters $(GK+K+1)$ and structure of the likelihood, this is non-trivial, with standard optimization schemes commonly failing or returning sub-optimal solutions. We thus propose an alternating maximization scheme which iteratively estimates and updates $\boldsymbol{\alpha}$, M, and n (and consequently σ_g^2) via a combination of quadratic programming and gradient descent. Despite the increased computational burden relative to W-NNLS or similar techniques, we find that convergence remains fairly fast, requiring no more than two to five minutes on typical data sets of $10,000+$ genes and 6 cell types using a modern laptop computer. We sketch an overview of our optimization procedure below in Algorithm 1 (where $\mathbb{P}_{M,S|\boldsymbol{\sigma}^2(M',\boldsymbol{\alpha}',S)}^{\boldsymbol{\alpha},n,\boldsymbol{c}}$ refers to (10) with $\boldsymbol{\sigma}^2(M, \boldsymbol{\alpha}, S)$ kept fixed at $\boldsymbol{\sigma}^2(M', \boldsymbol{\alpha}', S)$).

Implementations of Algorithm 1 are currently available in Python and Mathematica. While these implementations differ slightly, they both agree on the following fundamental design choices:

Algorithm 1: Find MLE of $\boldsymbol{\alpha}$

Data: Single cell expression vectors $\{\tilde{v}_{k,i}\}_{k\in[K],i\in[c_k]} \subset \mathbb{R}^G$, bulk gene expression vector $\tilde{\boldsymbol{b}} \in \mathbb{R}^G$

Result: Mixture proportions $\{\hat{\alpha}_k\}_{k\in[K]}$ of cell types in the bulk, number of cells $\hat{n} \in \mathbb{R}_+$ in bulk, mean expression $\hat{M} \in \mathbb{R}^{G\times K}$ of cell types

1 **begin**

2 $\quad \tilde{m}_{g,k} \longleftarrow \frac{1}{c_k}\sum_{i=1}^{c_k}(\tilde{v}_{g,k})_i,\ s_{g,k} \longleftarrow \frac{1}{c_k}\sum_{i=1}^{c_k}\left((v_{g,k})_i - \tilde{m}_{g,k}\right)^2$

3 $\quad \boldsymbol{\alpha}_0 \longleftarrow \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}_+^K} \|\tilde{M}\boldsymbol{\alpha} - \tilde{\boldsymbol{b}}\|_2^2,\ n_0 \longleftarrow \|\boldsymbol{\alpha}_0\|_1,\ \boldsymbol{\alpha}_0 \longleftarrow \boldsymbol{\alpha}_0/\|\boldsymbol{\alpha}_0\|_1\ \ M_0 \longleftarrow \tilde{M}$

4 $\quad$ **while** $\mathbb{P}^{\boldsymbol{\alpha}_{j+1},n_{j+1},\boldsymbol{c}}_{M_{j+1},S}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right) - \mathbb{P}^{\boldsymbol{\alpha}_j,n_j,\boldsymbol{c}}_{M_j,S}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right) > \delta$ **do**

5 $\quad\quad M_{j+1} \longleftarrow \arg\max_M \mathbb{P}^{\boldsymbol{\alpha}_j,n_j,\boldsymbol{c}}_{M,S|\boldsymbol{\sigma}^2(M_j,\boldsymbol{\alpha}_j,S)}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right)$

6 $\quad\quad \boldsymbol{\alpha}_{j+1} \longleftarrow \arg\max_{\boldsymbol{\alpha}\in\Delta^{K-1}} \mathbb{P}^{\boldsymbol{\alpha},n_j,\boldsymbol{c}}_{M_{j+1},S|\boldsymbol{\sigma}^2(M_{j+1},\boldsymbol{\alpha}_j,S)}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right)$

7 $\quad\quad n_{j+1} \longleftarrow \arg\max_{n\in\mathbb{R}_+} \mathbb{P}^{\boldsymbol{\alpha}_{j+1},n,\boldsymbol{c}}_{M_{j+1},S}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right)$

8 $\quad$ **end**

9 $\quad (\boldsymbol{\alpha}_\ell, M_\ell, n_\ell) \longleftarrow$ Last $(\boldsymbol{\alpha}_j, M_j, n_j)$ iterate returned in line 7

10 $\quad$ **while** $\mathbb{P}^{\boldsymbol{\alpha}_{\ell+1},n_{\ell+1},\boldsymbol{c}}_{M_{\ell+1},S}\left(d\tilde{\Omega}, d\tilde{\boldsymbol{b}}\right) - \mathbb{P}^{\boldsymbol{\alpha}_\ell,n_\ell,\boldsymbol{c}}_{M_\ell,S}\left(d\tilde{\Omega}, d\tilde{\boldsymbol{b}}\right) > \delta$ **do**

11 $\quad\quad M_{\ell+1} \longleftarrow \arg\max_M \mathbb{P}^{\boldsymbol{\alpha}_\ell,n_\ell,\boldsymbol{c}}_{M,S}\left(d\tilde{\Omega}, d\tilde{\boldsymbol{b}}\right)$

12 $\quad\quad \boldsymbol{\alpha}_{\ell+1} \longleftarrow \arg\max_{\boldsymbol{\alpha}\in\Delta^{K-1}} \mathbb{P}^{\boldsymbol{\alpha},n_\ell,\boldsymbol{c}}_{M_{\ell+1},S}\left(d\tilde{\Omega}, d\tilde{\boldsymbol{b}}\right)$

13 $\quad\quad n_{\ell+1} \longleftarrow \arg\max_{n\in\mathbb{R}_+} \mathbb{P}^{\boldsymbol{\alpha}_{\ell+1},n,\boldsymbol{c}}_{M_{\ell+1},S}\left(d\tilde{M}, d\tilde{\boldsymbol{b}}\right)$

14 $\quad$ **end**

15 $\quad (\hat{\boldsymbol{\alpha}}, \hat{M}, \hat{n}) \longleftarrow$ Last iterate returned in line 13

16 $\quad \hat{\boldsymbol{\alpha}} \longleftarrow \arg\min_{\boldsymbol{\alpha}\in\Delta^{K-1}} \|\hat{M}\boldsymbol{\alpha} - \tilde{\boldsymbol{b}}\|^2_{\boldsymbol{\sigma}^2(\hat{M},\hat{\boldsymbol{\alpha}},S)}$

17 $\quad$ Return $(\hat{\boldsymbol{\alpha}}, \hat{M}, \hat{n})$.

18 **end**

1. Instead of maximizing $\mathbb{P}^{\boldsymbol{\alpha},n,\boldsymbol{c}}_{M,S}$, we minimize $-\log \mathbb{P}^{\boldsymbol{\alpha},n,\boldsymbol{c}}_{M,S}$, rendering lines 5 and 6 as quadratic programs which can be solved efficiently.

2. Lines 7 and 13 can be solved explicitly by differentiating (10) and finding the zeros of the resulting algebraic fractions in n. Thus, these steps do not require any explicit optimization scheme.

3. The optimizations in lines 11 through 13 proceed via gradient descent (or a variation thereof), and so could possibly require long run-times. However, the coarser maximization (minimization, cf. item 1) in lines 5–7 typically improves the objective function to such an extent that only two or three more iterations are required. Moreover, both sets of optimizations are amenable to parallelization.

4. Algorithm 1 straightforwardly generalizes to the setting of jointly inferring mixture proportions in an arbitrary number N of bulk samples (cf. the remark in Mathematical Model). Both of our implementations support this generalized deconvolution.

Lastly, we note that although the alternating optimization in lines 5–7 is not guaranteed to converge, the second round of maximization in lines 11–13 is a proper coordinate descent and is therefore guaranteed to reach a local minimum.

Confidence Intervals

As indicated in Extension to Confidence Regions, the explicit generative modeling of (10) allows us to not only compute precise point estimators of $\boldsymbol{\alpha}$ and n, but also to quantify this precision through confidence regions. More concretely, since our model is well-behaved in the sense of satisfying all assumptions in, say, Theorem 9.14 of Keener (2011), we expect our estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{n}$ to be distributed normally around the true configuration $\boldsymbol{\alpha}^*, n^*$ with covariance matrix given by the inverse of the Fisher information $I^{\boldsymbol{c}}_{M,S}(\boldsymbol{\alpha}^*, n^*) \approx I^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$. Given such asymptotic normality, it is straightforward to construct both marginal confidence intervals (from the diagonal entries of $\left[I^{\boldsymbol{c}}_{\hat{M},S}\right]^{-1}$) as well as K-dimensional confidence regions around $\hat{\boldsymbol{\alpha}}$. Generically, there are infinitely many possibilities for choosing such confidence regions from $I^{n}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$, which we acknowledge by providing the entire (inverse) Fisher information to the user to allow computation of the confidence volume of their choosing. One option is to calculate the canonical (that is, Lebesgue volume-minimizing) q-confidence region $C_q = \{\boldsymbol{\alpha} \in \mathbb{R}^K : \|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|^2_{I^{\boldsymbol{c}}_{\hat{M},S}} \leq F^{-1}_{\chi^2_K}(q)\}$, where $\|\boldsymbol{v}\|_\Sigma = \langle \boldsymbol{v}, \Sigma^{-1}\boldsymbol{v}\rangle$ is the Mahalanobis norm of $\boldsymbol{v}$ associated with covariance matrix Σ and $F_{\chi^2_K}$ denotes the CDF of a χ^2 variable with K degrees of freedom; this option is included as the default.

Confidence intervals derived in this manner are, as a consequence of the aforementioned Theorem 9.14 in Keener (2011), necessarily well-calibrated *if* data adhere to our generative model (10). As we have observed in Data Pre-processing Procedure, this may not hold when the use of different protocols in the reference and bulk experiments induce significant distributional shifts. Nonetheless, we can still provide conservative, yet well-calibrated, confidence regions by thinning our data in the following sense: Under our generative model (10), the quantities $\left[(\hat{M}\hat{\boldsymbol{\alpha}})_g - \hat{b}_g\right]/\sigma_g(\hat{M}, \hat{\boldsymbol{\alpha}}, S)$ should be distributed as *i.i.d.* standard normal random variables, so detecting deviations from normality should signal the extent of model misspecification. Larger degrees of model misspecification necessitate more conservative confidence regions, motivating the use of a measure of departure from normality (essentially, the empirical distribution's tails) to determine the number of genes included (via uniform sampling) in our computation of $I^{\boldsymbol{c}}_{\hat{M},S}(\hat{\boldsymbol{\alpha}}, \hat{n})$.

Simulations both across and within protocols confirm the utility of confidence regions assessed in this manner (cf. Figure 7).

A Note on n

One of the parameters inferred by our model is n, the number of cells in the bulk sample. This parameter is accurate and physically meaningful in within-protocol deconvolutions, or cross-protocol experiments were relative amplification factors are explicitly known. However, it loses interpretability when the relative scales across protocols are unclear, and so we sought to verify that both our inferred proportions and computed confidence intervals are robust even in such situation. Numerical experiments in which we re-scaled the bulk samples to artificially manipulate the inference of n showed no degradation in performance over a wide range of values, providing additional support beyond the observed high-quality results in cross-protocol deconvolutions (Figure S7). Moreover, theoretical computations suggest a fairly weak dependence of confidence intervals on n, which are instead driven primarily by the total number of genes available for deconvolution.

Computational Methods

We relied upon pseudobulk experiments to benchmark RNA-Sieve against other methods. This common approach aggregates scRNA-seq measurements across cells in order to construct gene expression mixtures which we can treat as bulk samples. As the proportions of contributing cell types are known, we can then evaluate methods based on their accuracy with respect to the ground truth. We primarily used data from the *Tabula Muris Senis* Consortium, which covers many organs and cell types in two different single-cell experimental protocols. For within-protocol comparisons, individual mice were assigned to either the reference or bulk groups such that roughly equal numbers of cells were present in each. For cross-protocol comparisons, all eligible cells from each protocol were used. The same data splits were used for all algorithms, and all were run as described in their respective tutorials using default settings.

Data Availability

All data used in this paper are publicly available. The Gene Expression Omnibus accession code for single-cell expression data in pancreatic islets from Xin et al. (2016) is GSE81608, while the accession code for bulk RNA-seq data from Fadista et al. (2014) is GSE50244. The GEO accession code for the *Tabula Muris Senis* data is GSE132042.

Software Availability

RNA-Sieve is implemented in both Python and Mathematica. The source code will be made publicly available on the Song Lab GitHub repository: `https://github.com/songlab-cal/rna-sieve`.

Acknowledgments

This research is supported in part by an NIH grant no R35-GM134922. Y.S.S. is a Chan Zuckerberg Biohub Investigator.

References

Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K., 2018. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, **34**(11):1969–1979.

Bense, R. D., Sotiriou, C., Piccart-Gebhart, M. J., Haanen, J. B., van Vugt, M. A., de Vries, E. G., Schröder, C. P., and Fehrmann, R. S., 2017. Relevance of tumor-infiltrating immune cell composition and functionality for disease outcome in breast cancer. *JNCI: Journal of the National Cancer Institute*, **109**(1).

Bremnes, R. M., Busund, L.-T., Kilvær, T. L., Andersen, S., Richardsen, E., Paulsen, E. E., Hald, S., Khanehkenari, M. R., Cooper, W. A., Kao, S. C., *et al.*, 2016. The role of tumor-infiltrating lymphocytes in development, progression, and prognosis of non–small cell lung cancer. *Journal of Thoracic Oncology*, **11**(6):789–800.

Burzyn, D., Kuswanto, W., Kolodin, D., Shadrach, J. L., Cerletti, M., Jang, Y., Sefik, E., Tan, T. G., Wagers, A. J., Benoist, C., *et al.*, 2013. A special population of regulatory t cells potentiates muscle repair. *Cell*, **155**(6):1282–1295.

Cabrera, O., Berman, D. M., Kenyon, N. S., Ricordi, C., Berggren, P.-O., and Caicedo, A., 2006. The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proceedings of the National Academy of Sciences*, **103**(7):2334–2339.

Christov, C., Chrétien, F., Abou-Khalil, R., Bassez, G., Vallet, G., Authier, F.-J., Bassaglia, Y., Shinin, V., Tajbakhsh, S., Chazaud, B., *et al.*, 2007. Muscle satellite cells and endothelial cells: close neighbors and privileged partners. *Molecular biology of the cell*, **18**(4):1397–1409.

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., and Jiang, Y., 2020. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*, .

Fadista, J., Vikman, P., Laakso, E. O., Mollet, I. G., Esguerra, J. L., Taneera, J., Storm, P., Osmark, P., Ladenvall, C., Prasad, R. B., *et al.*, 2014. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences*, **111**(38):13924–13929.

Funada, Y., Noguchi, T., Kikuchi, R., Takeno, S., Uchida, Y., and Gabbert, H. E., 2003. Prognostic significance of cd8+ t cell and macrophage peritumoral infiltration in colorectal cancer. *Oncology reports*, **10**(2):309–313.

Goldman, S. L., MacKay, M., Afshinnekoo, E., Melnick, A. M., Wu, S., and Mason, C. E., 2019. The impact of heterogeneity on single-cell sequencing. *Frontiers in Genetics*, **10**:8.

Gurkan, U. A. and Akkus, O., 2008. The mechanical environment of bone marrow: a review. *Annals of biomedical engineering*, **36**(12):1978–1991.

Hagenauer, M. H., Schulmann, A., Li, J. Z., Vawter, M. P., Walsh, D. M., Thompson, R. C., Turner, C. A., Bunney, W. E., Myers, R. M., Barchas, J. D., *et al.*, 2018. Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. *PLOS One*, **13**(7):e0200003.

Hensel, J. A., Khattar, V., and Ponnazhagan, S., 2019. Characterization of immune cell subtypes in three commonly used mouse strains reveals gender and strain-specific variations. *Laboratory Investigation*, **99**(1):93–106.

Hu, P., Fabyanic, E., Kwon, D. Y., Tang, S., Zhou, Z., and Wu, H., 2017. Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-seq. *Molecular Cell*, **68**(5):1006–1015.

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., Sul, J. H., Pietiläinen, K. H., Pajukanta, P., and Halperin, E., *et al.*, 2020. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications*, **11**(1):1–11.

Kalisky, T., Rajendran, P. S., Sahoo, D., Sim, S., Okamoto, J., Miranda, S. P., Johnston, D. M., Clarke, M. F., Quake, S. R., and Dalerba, P., *et al.*, 2013. Analysis of human colon tissue cell composition using single-cell gene-expression PCR. *Journal of Biomolecular Techniques*, **24**(Suppl):S11.

Kasprzycka, P., Archacka, K., Kowalski, K., Mierzejewski, B., Zimowska, M., Grabowska, I., Piotrowski, M., Rafałko, M., Ryżko, A., Irhashava, A., *et al.*, 2019. The factors present in regenerating muscles impact bone marrow-derived mesenchymal stromal/stem cell fusion with myoblasts. *Stem cell research & therapy*, **10**(1):1–17.

Keener, R. W., 2011. *Theoretical Statistics: Topics for a Core Course*. Springer.

Lowe, R. and Rakyan, V. K., 2014. Correcting for cell-type composition bias in epigenome-wide association studies. *Genome Medicine*, **6**(3):23.

Mebius, R. E. and Kraal, G., 2005. Structure and function of the spleen. *Nature reviews immunology*, **5**(8):606–616.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., *et al.*, 2015. The human transcriptome across tissues and individuals. *Science*, **348**(6235):660–665.

Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D. S., Kloiber, K., Heutink, P., and Bonn, S., 2020. Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances*, **6**(30):eaba2619.

Mumau, M. D., Vanderbeck, A. N., Lynch, E. D., Golec, S. B., Emerson, S. G., and Punt, J. A., 2018. Identification of a multipotent progenitor population in the spleen that is regulated by nr4a1. *The Journal of Immunology*, **200**(3):1078–1087.

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., *et al.*, 2019. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, **37**(7):773–782.

Novak, M. L., Weinheimer-Haus, E. M., and Koh, T. J., 2014. Macrophage activation and skeletal muscle healing following traumatic injury. *The Journal of pathology*, **232**(3):344–355.

Pang, W. W., Price, E. A., Sahoo, D., Beerman, I., Maloney, W. J., Rossi, D. J., Schrier, S. L., and Weissman, I. L., 2011. Human bone marrow hematopoietic stem cells are increased in frequency

and myeloid-biased with age. *Proceedings of the National Academy of Sciences*, **108**(50):20012–20017.

Pisco, A. O., Schaum, N., McGeever, A., Karkanias, J., Neff, N. F., Darmanis, S., Wyss-Coray, T., Quake, S. R., et al., 2020. A single cell transcriptomic atlas characterizes aging tissues in the mouse. *BioRxiv*, :661728.

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J., 2014. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, **42**(14):8845–8860.

Shiwa, Y., Hachiya, T., Furukawa, R., Ohmomo, H., Ono, K., Kudo, H., Hata, J., Hozawa, A., Iwasaki, M., Matsuda, K., *et al.*, 2016. Adjustment of cell-type composition minimizes systematic bias in blood DNA methylation profiles derived by DNA collection protocols. *PLOS One*, **11**(1):e0147519.

Snijders, T., Nederveen, J. P., McKay, B. R., Joanisse, S., Verdijk, L. B., van Loon, L. J., and Parise, G., 2015. Satellite cells in human skeletal muscle plasticity. *Frontiers in Physiology*, **6**:283.

Stout, M. B., Justice, J. N., Nicklas, B. J., and Kirkland, J. L., 2017. Physiological aging: links among adipose tissue dysfunction, diabetes, and frailty. *Physiology*, **32**(1):9–19.

Sudmant, P. H., Alexis, M. S., and Burge, C. B., 2015. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biology*, **16**(1):1–11.

Trapnell, C., 2015. Defining cell types and states with single-cell genomics. *Genome Research*, **25**(10):1491–1498.

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G.-C., 2019. Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, **10**(1):1–9.

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M., 2019. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, **10**(1):380.

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A. J., Yancopoulos, G. D., Lin, C., and Gromada, J., *et al.*, 2016. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell metabolism*, **24**(4):608–615.

Yu, Q. and He, Z., 2017. Comprehensive investigation of temporal and autism-associated cell type composition-dependent and independent gene expression changes in human brains. *Scientific Reports*, **7**(1):4121.

Zhou, R., Zhang, J., Zeng, D., Sun, H., Rong, X., Shi, M., Bin, J., Liao, Y., and Liao, W., 2019. Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I-III colon cancer. *Cancer Immunology, Immunotherapy*, **68**(3):433–442.

Supplementary Material

RNA-Sieve: A likelihood-based deconvolution of bulk gene expression data using single-cell references

Dan D. Erdmann-Pham*, Jonathan Fischer*, Justin Hong, Yun S. Song

October 1, 2020

This file contains:

- Supplementary Tables S1–S2
- Supplementary Figures S1–S7

(a) Smart-Seq2 reference and 10x Chromium pseudobulk

	RNA-Sieve	Bisque	CIBERSORTx	DWLS	MuSiC	NNLS	Scaden	SCDC
Bladder	0.081	**0.047**	0.082	0.072	0.106	0.378	0.099	0.113
Kidney	0.095	0.109	**0.028**	0.055	0.110	0.249	0.062	0.083
Large intestine	0.076	0.082	0.300	0.123	0.108	0.226	**0.042**	0.136
Limb muscle	0.199	0.108	0.037	0.039	0.199	0.310	**0.030**	0.144
Liver	0.137	0.129	0.030	0.054	0.139	0.340	0.076	**0.027**
Lung	0.056	0.078	0.071	0.064	0.056	0.149	0.057	**0.029**
Mammary gland	**0.020**	0.258	0.072	0.029	0.047	0.371	0.083	0.058
Marrow	0.061	0.101	0.071	0.073	0.070	0.166	0.072	**0.049**
Pancreas	**0.011**	0.029	0.050	0.030	0.067	0.130	0.059	0.067
Skin	**0.019**	0.270	0.048	0.123	0.098	0.462	0.182	0.128
Thymus	**0.017**	0.050	0.098	0.331	0.127	0.482	0.030	0.120
Tongue	**0.016**	0.289	0.068	0.293	0.047	0.448	0.217	0.017
Trachea	0.108	**0.097**	0.166	0.165	0.142	0.252	0.110	0.154

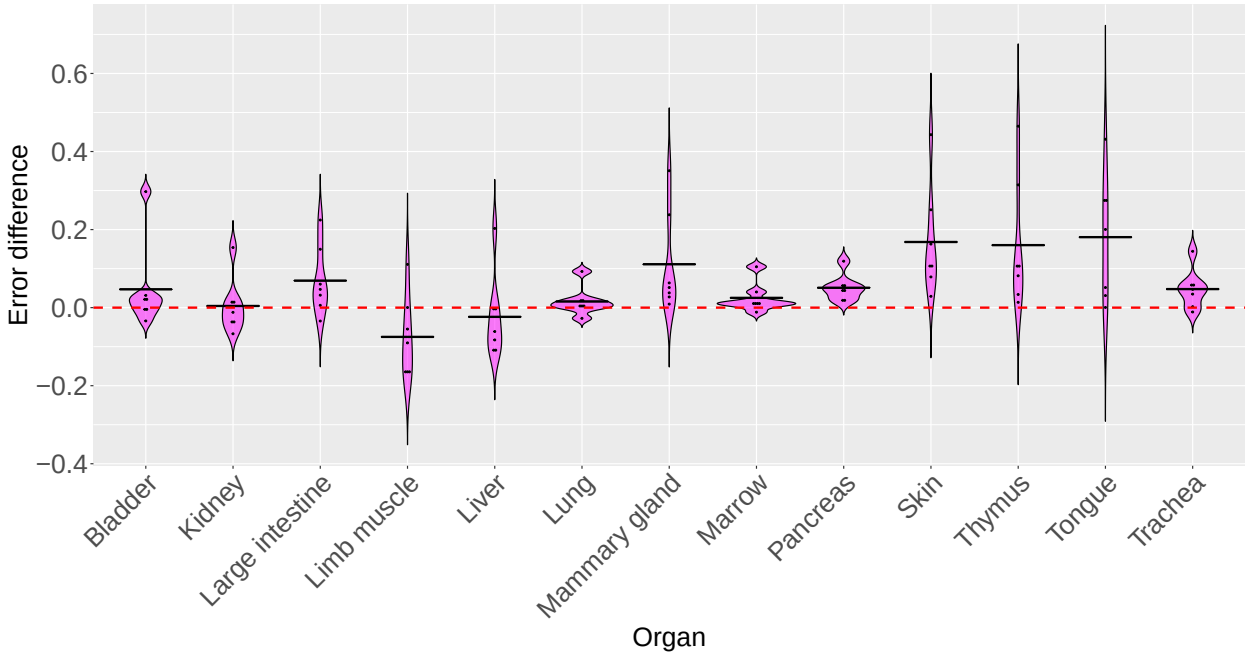
(b) 10x Chromium reference and Smart-Seq2 pseudobulk

	RNA-Sieve	Bisque	CIBERSORTx	DWLS	MuSiC	NNLS	Scaden	SCDC
Bladder	**0.002**	0.066	0.059	0.085	0.156	0.044	0.167	0.261
Kidney	0.082	0.045	0.036	**0.028**	0.113	0.173	0.044	0.046
Large intestine	0.117	0.158	0.089	0.152	0.186	0.448	0.066	**0.007**
Limb muscle	0.137	0.132	0.037	**0.013**	0.122	0.142	0.102	0.177
Liver	0.107	0.056	**0.032**	0.050	0.126	0.164	0.052	0.070
Lung	0.092	0.069	0.029	**0.021**	0.130	0.153	0.074	0.045
Mammary gland	**0.009**	0.244	0.062	0.013	0.160	0.228	0.196	0.157
Marrow	**0.070**	0.110	0.124	0.097	0.113	0.147	0.111	0.121
Pancreas	0.121	0.085	0.137	0.054	**0.023**	0.117	0.111	0.173
Skin	**0.037**	0.191	0.162	0.050	0.168	0.676	0.109	0.192
Thymus	**0.002**	0.110	0.114	0.208	0.298	0.317	0.036	0.275
Tongue	**0.006**	0.254	0.022	0.143	0.672	0.672	0.191	0.672
Trachea	0.092	0.098	**0.067**	0.080	0.166	0.151	0.105	0.153

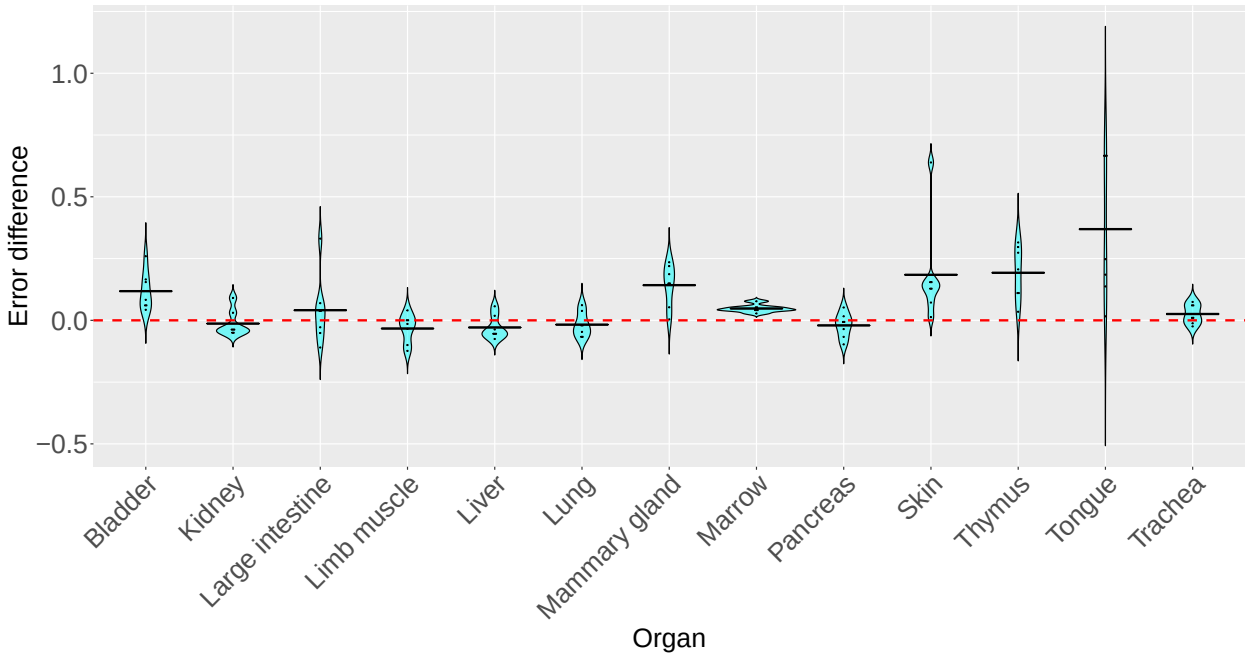
Table S1: **Deconvolution errors for different algorithms in pseudobulk experiments.** Deconvolutions were performed using the specified methods in thirteen organs using both Smart-Seq2 and 10x Chromium data from the *Tabula Muris Senis* experiment. Presented errors show the L_1 distance between the ground truth and inferred values divided by the number of present cell types. These values correspond to Table 1 and Figure 2 of the main text.

Organs	**# cell types**	**Cell types**
Bladder	2	bladder cell, bladder urothelial cell
Kidney	7	B cell, epithelial cell of proximal tubule, fenestrated cell, kidney collecting duct principal cell, kidney loop of Henle ascending limb epithelial cell, macrophage, T cell
Large intestine	3	enterocyte of epithelium of large intestine, epithelial cell of large intestine, intestinal crypt stem cell
Limb muscle	6	B cell, endothelial cell, macrophage, mesenchymal stem cell, skeletal muscle satellite cell, T cell
Liver	5	B cell, endothelial cell of hepatic sinusoid, hepatocyte, Kupffer cell, myeloid leukocyte
Lung	12	adventitial cell, B cell, bronchial smooth muscle cell, CD4+ $\alpha\beta$ T cell, CD8+ $\alpha\beta$ T cell, classical monocyte, fibroblast of lung, myeloid dendritic cell, neutrophil, natural killer cell, non-classical monocyte, vein endothelial cell
Mammary gland	3	basal cell, luminal epithelial cell of mammary gland, stromal cell
Marrow	9	granulocyte, granulocytopoietic cell, immature B cell, late pro-B cell, macrophage, megakaryocyte-erythroid progenitor cell, naive B cell, precursor B cell, promonocyte
Pancreas	3	pancreatic A cell, pancreatic B cell, pancreatic D cell
Skin	2	basal cell of epidermis, epidermal cell
Thymus	2	DN4 thymocyte, thymocyte
Tongue	2	basal cell of epidermis, keratinocyte
Trachea	5	basal epithelial cell of tracheobronchial tree, chondrocyte, endothelial cell, fibroblast, macrophage

Table S2: **Cell types for each organ in pseudobulk experiments.** These were the cell types used in pseudobulk experiments with the *Tabula Muris Senis* data. The order in which they are listed here matches their order in any figures based off of these experiments.

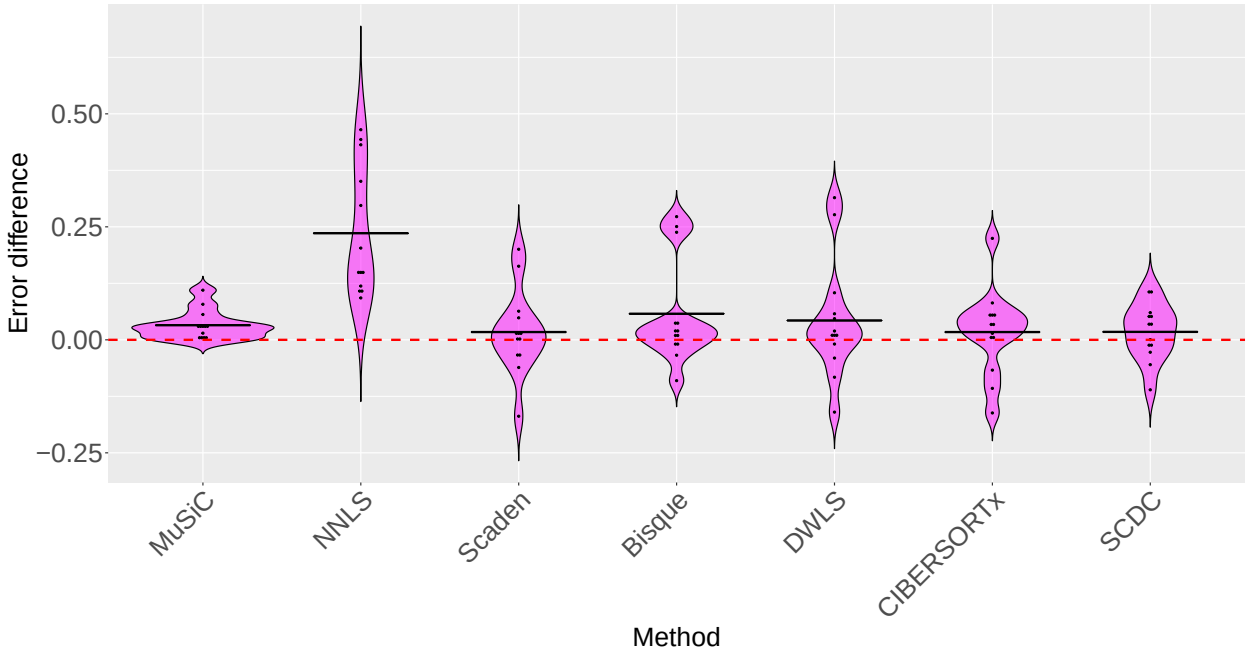


(a) **Smart-Seq2 reference and 10x Chromium pseudobulk**

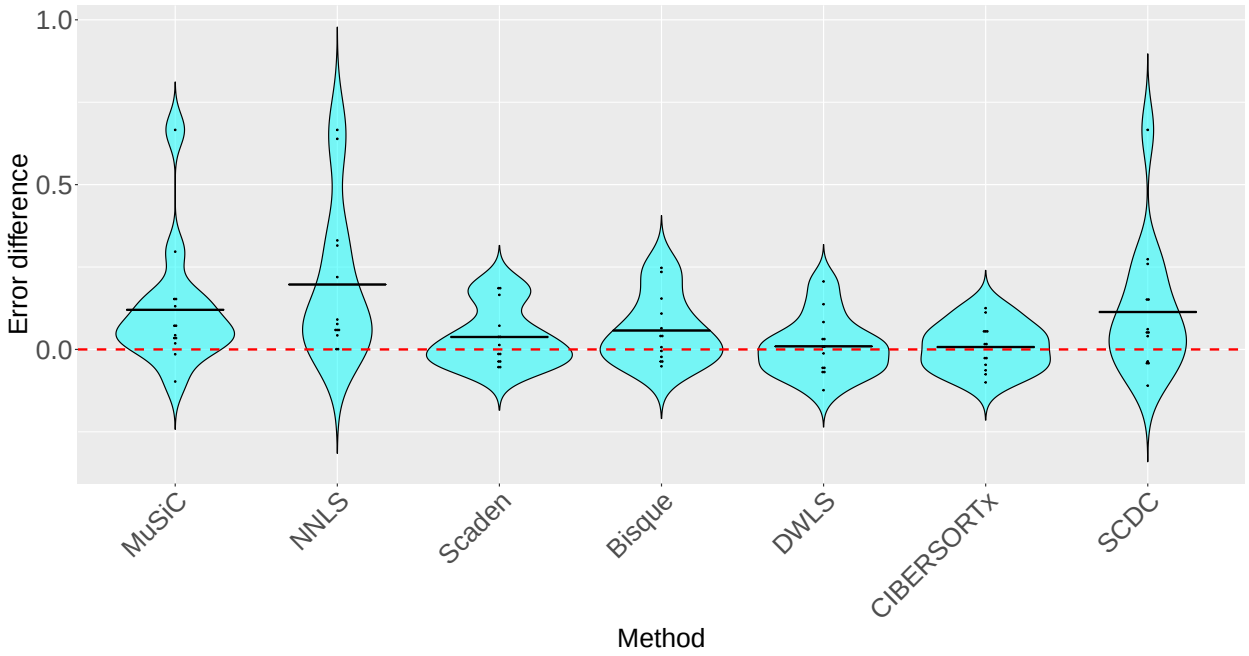


(b) **10x Chromium reference and Smart-Seq2 pseudobulk**

Figure S1: **Comparison of other methods to RNA-Sieve across 13 murine organs.** Pseudobulk experiments were performed in 13 different organs using data from the *Tabula Muris Senis* experiment. Errors were computed as the average L_1 error across cell types in each organ. For each organ, the difference in errors was computed between other methods and RNA-Sieve. Horizontal black bars correspond to the mean difference in error, and positive values indicate better comparative performance for RNA-Sieve.



(a) **Smart-Seq2 reference and 10x Chromium pseudobulk**



(b) **10x Chromium reference and Smart-Seq2 pseudobulk**

Figure S2: **Direct comparison of other methods to RNA-Sieve** Pseudobulk experiments were performed in 13 different organs using data from the *Tabula Muris Senis* experiment. Errors were computed as the average L_1 error across cell types in each organ. For each method, the difference in errors was computed between it and RNA-Sieve across each of the 13 organs. Horizontal black bars correspond to the mean difference in error, and positive values indicate better comparative performance for RNA-Sieve.

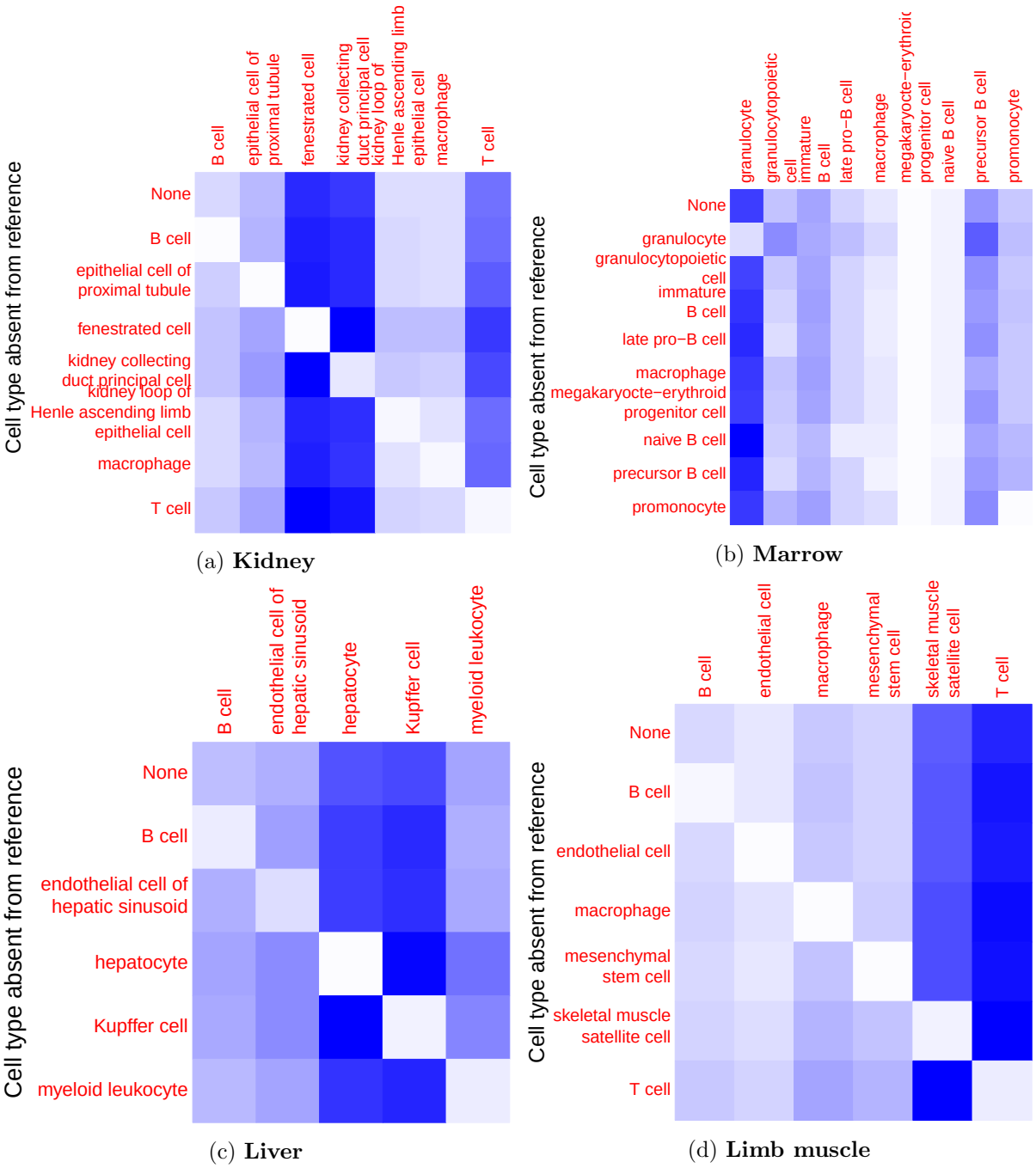


Figure S3: **Deconvolution with extra cell types in the reference matrix.** Deconvolution was performed in pseudobulk experiments in four different organs. For each organ, we followed a leave-one-out procedure in which one cell type is removed from the pseudobulk at a time. Deconvolution was then performed with this extra cell type in the reference in order to examine RNA-Sieve's specificity. The top row shows the inferred proportions with no extra reference cell types. Darker colors indicate a higher estimated proportion value. Here we used 10x Chromium data for the reference and Smart-Seq2 for the pseudobulk.

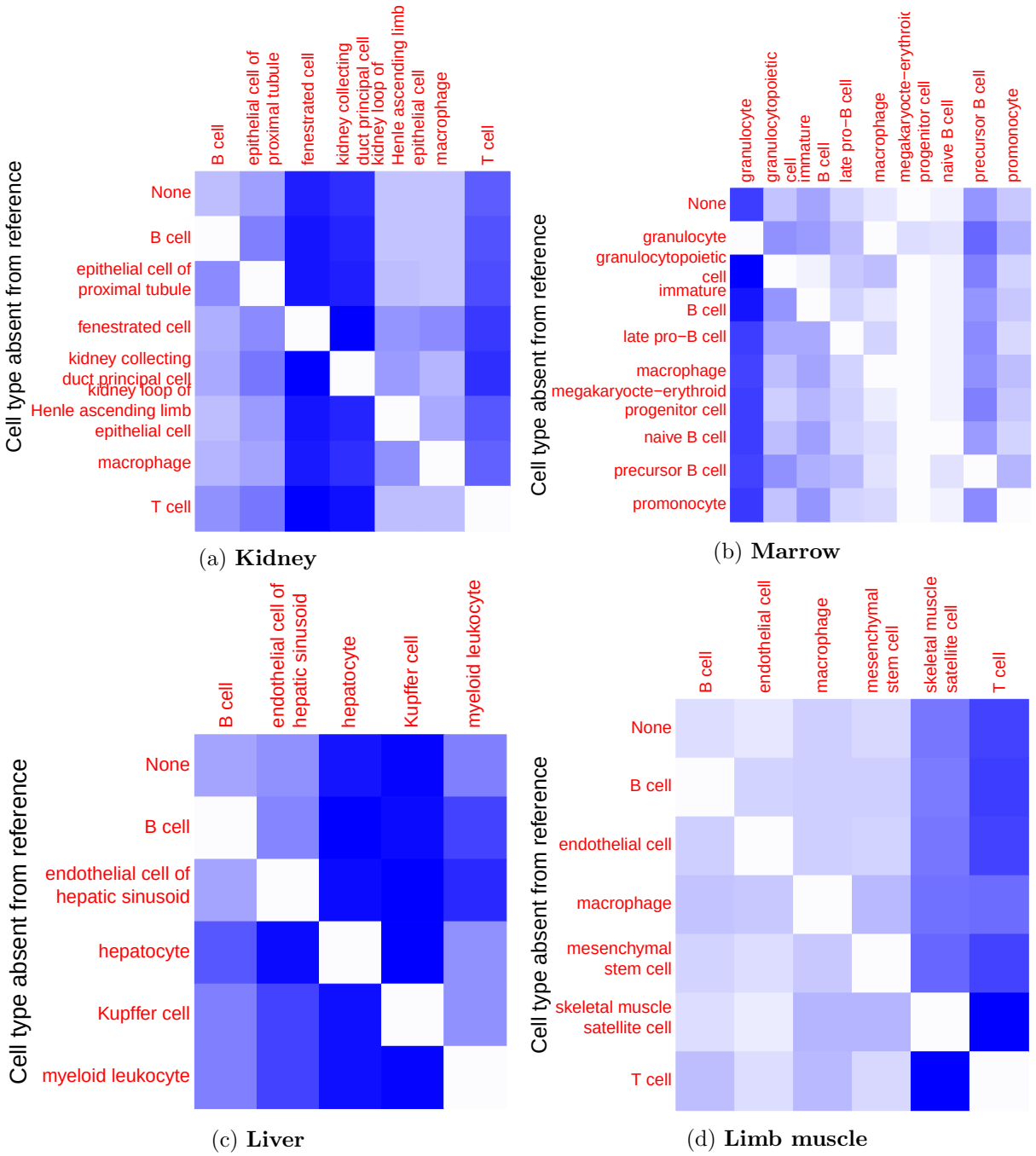
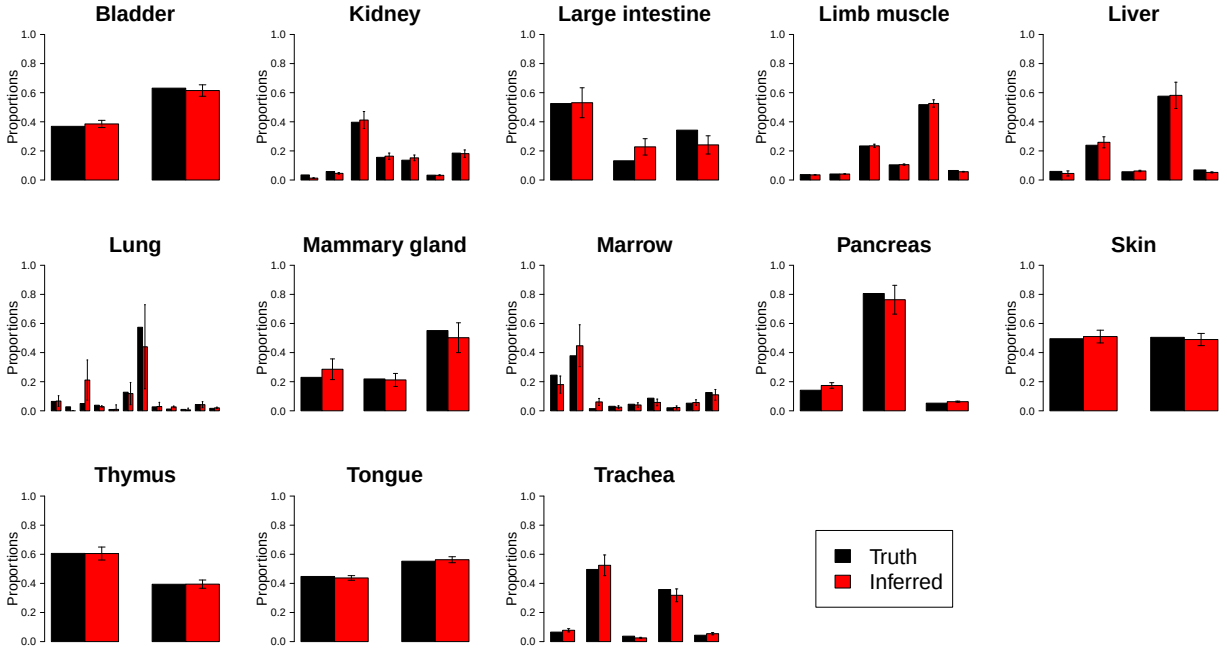
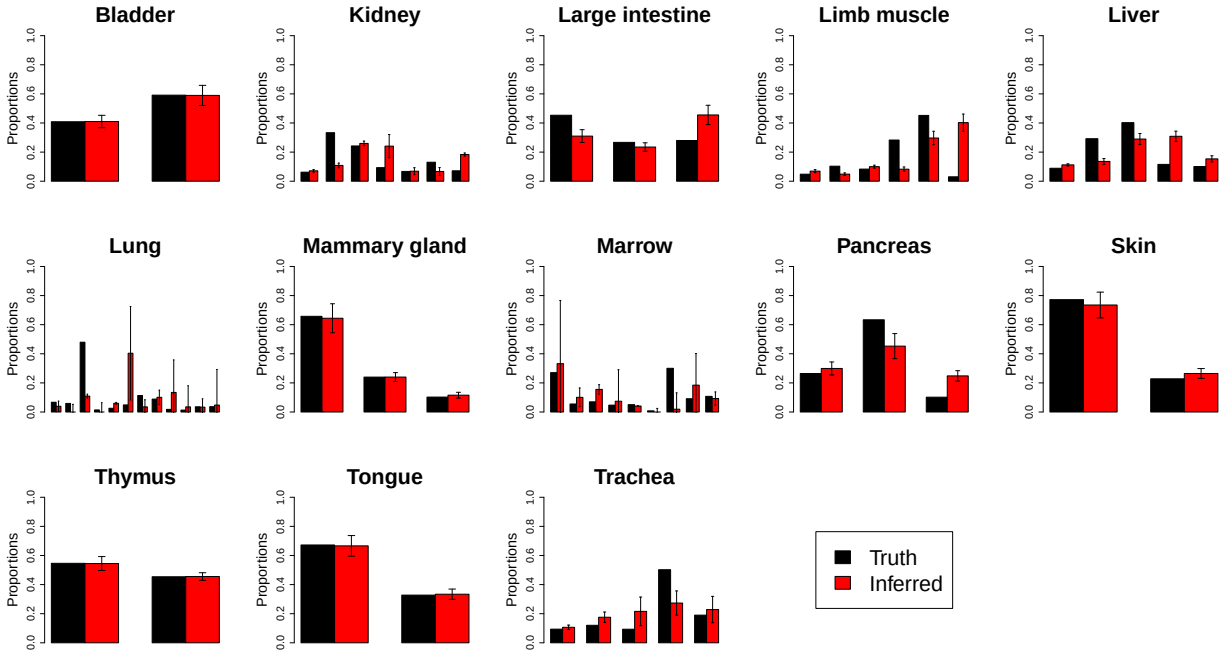


Figure S4: **Deconvolution with missing cell types in the reference matrix.** Deconvolution was performed in pseudobulk experiments in four different organs. For each organ, we followed a leave-one-out procedure in which one cell type is removed from the reference at a time. Deconvolution was then performed with an extra cell type in the pseudobulk in order to examine RNA-Sieve's ability to handle such a misspecification. The top row shows the inferred proportions with no missing reference cell types. Darker colors indicate a higher estimated proportion value. Here we used 10x Chromium data for the reference and Smart-Seq2 for the pseudobulk.

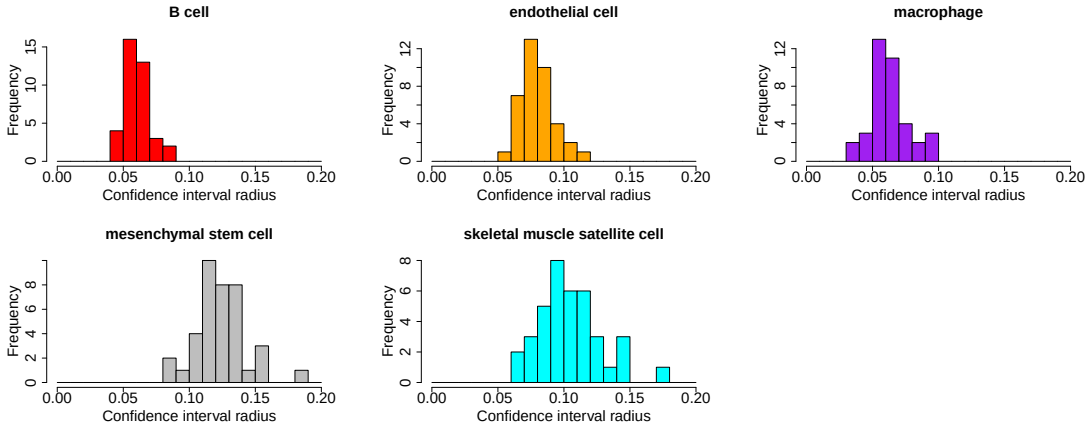


(a) **Within-protocol – 10x Chromium pseudobulk and reference**

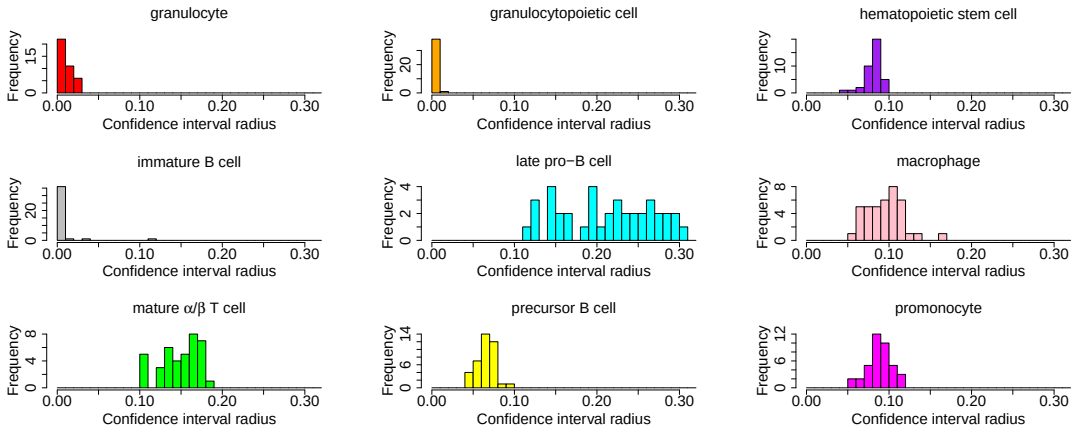


(b) **Cross-protocol – 10x Chromium reference and Smart-Seq2 pseudobulk**

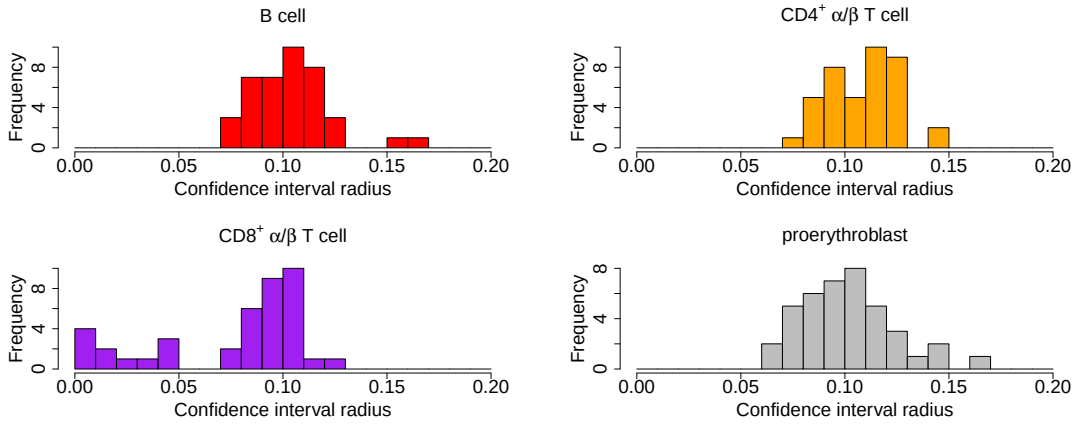
Figure S5: **RNA-Sieve results with confidence intervals in pseudobulk experiments.** Inferred cell type proportions in pseudobulk experiments using data from the *Tabula Muris Senis* experiment. The black error bars on inferred proportions show the marginal 95% confidence intervals as computed from the estimated Fisher information produced by RNA-Sieve. Table S2 contains the cell types in each organ, which could not be displayed because of space constraints.



(a) **Limb muscle**

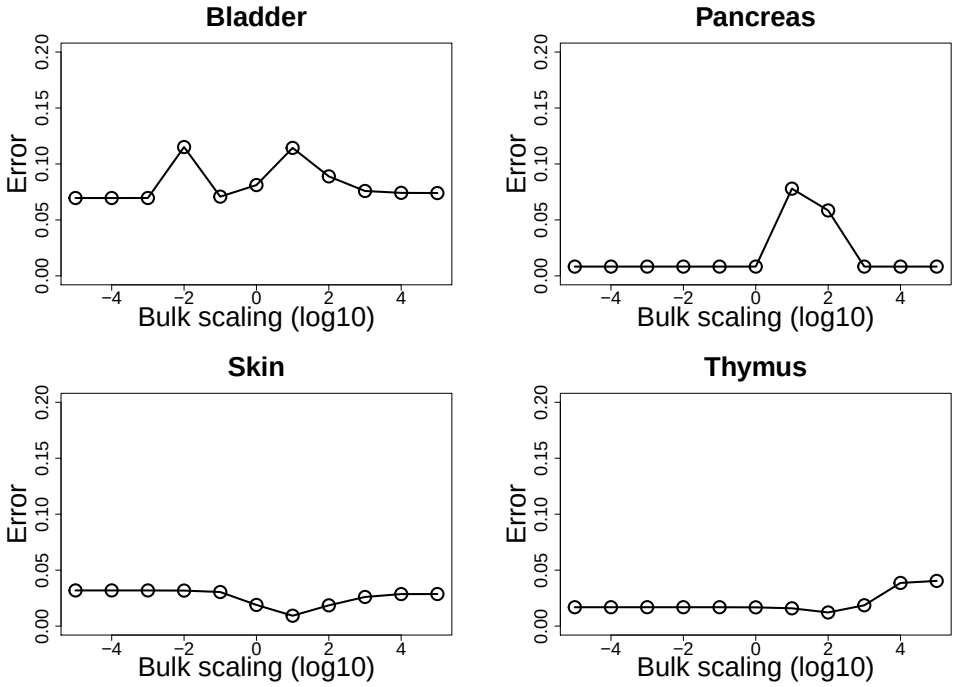


(b) **Marrow**

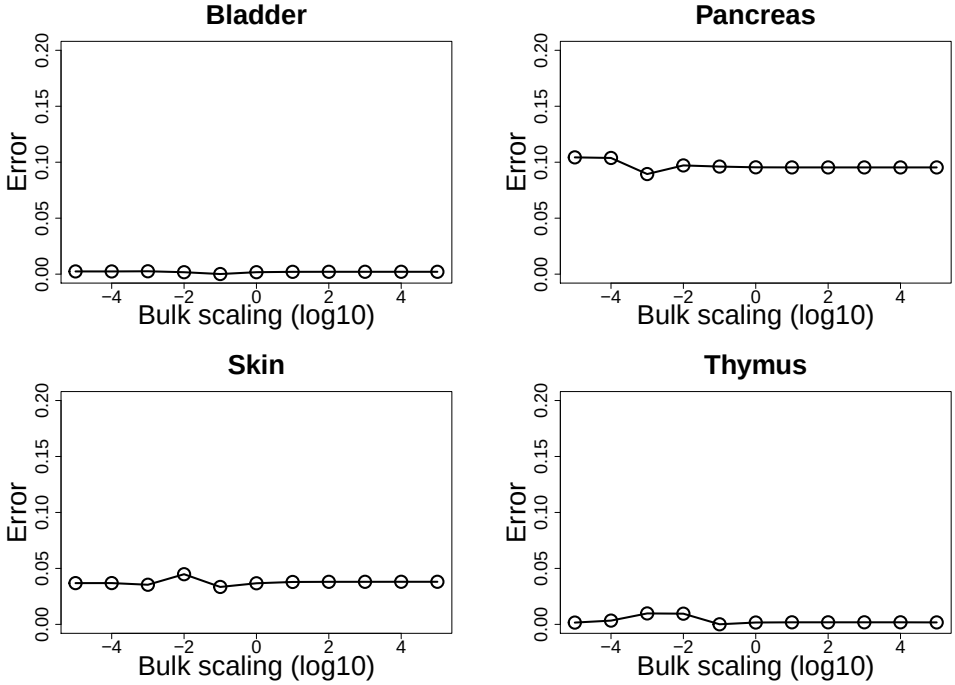


(c) **Spleen**

Figure S6: **Histograms of CI radii with real bulk samples.** The radius of the 95% confidence interval for inferred cell type proportions was computed using RNA-Sieve for each real bulk sample in the listed organs (∼40 per organ).



(a) **Smart-Seq 2 reference and 10x Chromium pseudobulk**



(b) **10x Chromium reference and Smart-Seq2 pseudobulk**

Figure S7: **Effects of different bulk scalings.** To determine RNA-Sieve's sensitivity to the the application of different scalings to bulk samples, we multiplied pseudobulk counts by different values across a wide range and performed deconvolution. We then computed RNA-Sieve's error for each value. Here we show representative results from four organs.