Supplementary Information

Figure legends

Supplementary Figure 1: Single cell data demonstration from Segerstolpe et al.

a. Example of cross-subject and cross-cell variation in cell type specific gene expression. The boxplot contains 4 cell types: alpha, beta, gamma, and delta cells from Segerstolpe et al. single-cell RNA-seq data. The x-axis is the log transformed average relative abundance across cells from the same cell type, and the y-axis is the subject label. The relative abundance of gene *GC* is widely spread across the x-axis while the relative abundance of gene *TTR* is more concentrated across subjects. We consider gene *GC* as non-informative and *TTR* as informative. **b.** Comparison of log transformed relative abundance levels between real bulk tissue RNA-seq data and artificially constructed bulk RNA-seq data for the same subject. Single-cell and bulk tissue RNA-seq data are both from Segerstolpe et al. Each dot represents a gene and the gray line is x=y. **c.** Heatmap of true and estimated cell type proportions. In addition to the four methods described in the main text, we also evaluated the estimates given by MuSiC and NNLS when using only the marker genes used in BSEQ-sc.

Supplementary Figure 2: Heatmaps of true and estimated cell type proportions of artificial bulk data constructed using single-cell RNA-seq data by Xin et al.

a. Deconvolution results when the single-cell reference is from the 6 healthy subjects of Segerstolpe et al. with leave-one-out, i.e., for each subject under deconvolution, only single-cell data from the remaining 5 subjects were used as single-cell reference. **b.** Deconvolution results when the single-cell reference is from the 12 healthy subjects of Xin et al. with leave-one-out, i.e., for each subject under deconvolution, only single-cell data from the remaining 11 subjects were used as single-cell reference. **c.** The cell type proportions for the artificial bulk data are manually adjusted so that beta cells are the dominant cell type, as expected in real bulk tissue. Alpha cells dominate in the scRNA-seq data due to dissociation and capture bias. Thus, this analysis mirrors the real data analysis scenario where cell type proportions differ substantially between scRNA-seq reference and bulk tissue. In more detail, we combined cells from two subjects as one artificial bulk tissue RNA-seq dataset, for example, H1.2 combined cells from subject H1 and H2. Then we dropped 75% of the alpha cells at random. The single-cell reference is from the 6 healthy subjects of Segerstolpe et al. Here, all methods that rely on pre-selected marker genes from CIBERSORT are heavily biased by the cell type proportions in the single cell reference, and miss the true cell type proportions in the bulk tissue data. In comparison, MuSiC is able to adjust to the difference between scRNA-seq reference and bulk data.

Supplementary Figure 3: Heatmaps of true and estimated cell type proportions with missing cell types in single-cell reference. The artificial bulk data and the single-cell reference are both from Segerstolpe et al. We constrained our analysis to the 6 major cell types: alpha, beta, delta, gamma, acinar and ductal cells. The artificial bulk data is constructed by summing read counts from the 6 major cell types while the single-cell reference contains only 5 cell types (the column header shows the cell type that is missing in the single-cell reference). The x-axis labels cell types used in the single-cell reference and the y-axis shows the subject label. The top panel

shows the true composition, while panels below it show the results from each method. See **Supplementary Table 4** for detailed evaluation results.

Supplementary Figure 4: Benchmark evaluation using mouse kidney single-cell RNA-seq data from Park et al. The artificial bulk RNA-seq data is constructed by summing read counts across cells in all 16 cell types while the single-cell reference only consists of 13 cell types. The other 3 cell types were discarded in the single-cell reference because they are too rare.

a. Heatmap of estimated cell type proportions and evaluation results. **b.** Scatter plot of real cell type proportions versus estimated cell type proportions.

Supplementary Figure 5: Estimated cell type proportions of the 13 cell types in three real mouse bulk RNA-seq datasets.

a. Boxplot of estimated cell type proportions of 10 mice (4 APOL1 disease mice and 6 control mice) from Beckerman et al. **b.** Line plot of cell type proportion changes after FA induction (Craciun et al.) at 6 time points. There are 3 replicates at each time point and the average proportions are plotted. N: normal. **c.** Line plot of cell type proportions of control (Sham operated mice), 2 days and 8 days after UUO (Arvaniti et al.).

Supplementary Figure 6: Estimated cell type proportions and correlation of the estimated cell type proportions across samples for bulk RNA-seq data of rat renal tubule segments (Lee et al.). Park et al. mouse single-cell RNA-seq data are used as reference. **a.** NNLS. **b.** BSEQ-sc. **c.** CIBERSORT.

Supplementary Notes

Supplementary Note 1: Artificial bulk tissue RNA-seq data construction

We construct artificial bulk tissue RNA-seq data by summing up read counts across all cells from the same subject in the single cell RNA-seq data. By way of construction, the cell type proportions of the artificial bulk data are equal to the observed cell type proportions in the single-cell data, and this allows us to compare estimated cell type proportions from various methods with the true proportions. **Supplementary Figure 1b** shows that the artificial bulk tissue RNA-seq data have similar gene expression as the real bulk RNA-seq data generated from the same subjects.

Supplementary Note 2: Impact of varying cell type proportions of artificial bulk data in deconvolution.

Figure 2b in the main text shows the deconvolution results from MuSiC, NNLS, BSEQ-sc and CIBERSORT, and these results indicate that the alpha cell proportion is over-estimated by all methods except for MuSiC. To evaluate the impact of different cell type proportions in the bulk data on deconvolution estimates, we generated additional artificial bulk data to show that MuSiC can still reliably estimate cell type proportions even when the true cell type proportions in the bulk data are very different from the cell type proportions in the single-cell reference. In this newly constructed benchmark data, the single-cell reference stays the same while we construct the artificial bulk data from Xin et al. by combining cells from 2 subjects with 75% alpha cells dropped. In this way, beta cells become the dominant cell type in the artificial bulk data, as

expected for real bulk tissue. **Supplementary Figure 2c** shows that only MuSiC recovers the true cell type composition, revealing that beta cells are the major cell type in the artificial bulk data, whereas the other methods overestimate the alpha cell proportion, indicating that these methods are more likely to be influenced by the cell type proportions in the single-cell reference. This analysis also gives the likely explanation for why, in the Fadista et al. data, all methods that rely on CIBERSORT marker genes grossly overestimate alpha cell proportion.

Supplementary Note 3: Impact of missing cell types in single-cell reference on deconvolution.

One of the limitations of single cell RNA-seq is cell loss during cell dissociation. This not only biases cell type proportions, but also leads to failure of detecting certain cell types, especially those rare cell types. In practice, the single-cell reference dataset might be incomplete, and not every cell type present in the bulk data is included in the single-cell reference. Since the deconvolution methods rely on observed cell types in the single-cell reference, it is important to evaluate whether cell type proportions can be reliably estimated when some cell types are missing in the single-cell reference.

We evaluate MuSiC, NNLS, BSEQ-sc and CIBERSORT with missing cell types (**Supplementary Figure 3, Supplementary Table 4**). The artificial bulk data consist of 6 cell types while the single-cell reference only consists of 5 cell types. The evaluation shows that when major cell types are missed, none of the methods can give accurate estimates. However, the cell type proportions are estimated accurately by MuSiC when the missing cell type is not the dominant cell type in the bulk tissue.

Supplementary Tables

Supplementary Table 1: Pancreatic islet datasets

Name	Journal	Year	Session #	Tissue Type	Data type	Protocol	# samples	# cells	# genes	# cell types
Segerstolpe et al.	Cell Metabolism	2016	E-MTAB-5061	Pancreatic islet	Single-cell RNA-seq	Smart-seq2	10 (6 H + 4 T2D)	2209	25453	14 + 1 NA
Segerstolpe et al.	Cell Metabolism	2016	E-MTAB-5060	Pancreatic islet	Bulk RNA-seq	Smart-seq2	7 (3H + 4 T2D)	NA	25453	NA
Xin et al.	Cell Metabolism	2016	GSE81608	Islet: endocrine	Single-cell RNA-seq	Illumina HiSeq 2500	18 (12H + 6 T2D)	1492	39849	4
Fadista et al.	PNAS	2014	GSE50244	Pancreatic islet	Bulk RNA-seq	Illumina HiSeq 2000	89	NA	56638	NA

Supplementary Table 2: Mouse/Rat kidney datasets

Name	Journal	Year	Session #	Tissue Type	Data type	Protocol	# samples	# cells	# genes	# cell types
Park et al.	Science	2018	GSE107585	Kidney	Single-cell RNA-seq	10x	7 health, male	43745	16273	14 + 2 novel
Beckerman et al.	Nature Medicine	2017	GSE81492	Kidney	Bulk RNA-seq	Illumina HiSeq 2500	10 (6 control + 4 APOL1)	NA	19033	NA
Lee et al.	JASN	2015	GSE56743	Kidney tubule	Bulk RNA-seq	Illumina HiSeq 2000	118 replicates (14 segments)	NA	10903	NA
Craciun et al.	JASN	2015	GSE65267	Kidney	Bulk RNA-seq	Illumina HiSeq 2000	18 replicates (6 time points)	NA	25219	NA

Arvaniti et al.	Scientific Reports	2016	GSE79443	Kidney	Bulk RNA-seq	Illumina HiSeq 2000	10 replicates (Sham + 2 time points)	NA	38683	NA

Supplementary Table 3: Linear regression to examine the relationship between estimated cell type proportions and HbA1c level. The fitted linear model is est. proportions ~ HbA1c + Age + BMI + Gender. Significant results (p value < 0.05) are highlighted.

Cell type		MuSiC			BSEQ-sc		
		Estimate	Std.Error	P value	Estimate	Std.Error	P value
alpha	(Intercept)	0.380382	0.207754	0.07125	1.351464	0.240052	3.26E-07
	HbA1c	-0.00203	0.027737	0.941834	-0.07377	0.032049	0.024249
	Age	-0.00097	0.001935	0.617836	0.002753	0.002236	0.222198
	BMI	-0.00167	0.007945	0.834127	-0.01711	0.00918	0.066449
	Gender	0.033135	0.042881	0.442221	-0.00638	0.049548	0.897869
beta	(Intercept)	0.877022	0.190276	1.71E-05	0.065847	0.046433	0.16047
	HbA1c	-0.0614	0.025403	0.01819	-0.00295	0.006199	0.635957
	Age	0.002639	0.001772	0.140873	0.000576	0.000433	0.187339
	BMI	-0.01362	0.007276	0.065293	-0.00162	0.001776	0.365258
	Gender	-0.07987	0.039274	0.04566	-0.00541	0.009584	0.574159
gamma	(Intercept)	0.008556	0.010504	0.417988	0.102201	0.024366	7.69E-05
	HbA1c	0.001047	0.001402	0.457785	-0.00278	0.003253	0.396334
	Age	9.21E-05	9.78E-05	0.349431	-0.00013	0.000227	0.570225
	BMI	-0.00057	0.000402	0.160731	-0.00207	0.000932	0.029738
	Gender	-0.00165	0.002168	0.450416	-0.00092	0.005029	0.855252
delta	(Intercept)	0.057678	0.010592	6.81E-07	0.015539	0.018715	0.409122
	HbA1c	-0.00106	0.001414	0.455427	0.002017	0.002499	0.422131
	Age	-0.00016	9.87E-05	0.12039	9.99E-05	0.000174	0.568316
	BMI	-0.0011	0.000405	0.008142	-0.00103	0.000716	0.154263
	Gender	0.000424	0.002186	0.846817	-0.00254	0.003863	0.512616
acinar	(Intercept)	-0.10619	0.131102	0.420638	-0.14553	0.052092	0.006672
	HbA1c	0.034967	0.017503	0.049519	0.019075	0.006955	0.007684
	Age	-0.00247	0.001221	0.046841	0.00066	0.000485	0.178153
	BMI	0.00662	0.005013	0.190883	0.002008	0.001992	0.316847
	Gender	0.05332	0.02706	0.052632	-0.02338	0.010752	0.032985
ductal	(Intercept)	-0.21745	0.141008	0.127428	-0.38952	0.232841	0.098686
	HbA1c	0.028474	0.018826	0.134781	0.058397	0.031086	0.064353
	Age	0.000863	0.001313	0.513005	-0.00396	0.002169	0.072066
	BMI	0.010341	0.005392	0.059097	0.019814	0.008904	0.029191
	Gender	-0.00536	0.029105	0.854406	0.038631	0.048059	0.424144

Supplementary Table 4: Evaluation of deconvolution methods when there are missing cell types in the single-cell reference. The missing cell type is shown in bold and the proportions in the bulk tissue data are shown in parentheses.

alpha (0.447)	RMSD	mAD	R	**beta** (0.137)	RMSD	mAD	R
MuSiC	0.13	0.09	0.72	MuSiC	0.04	0.03	0.98

NNLS	0.27	0.18	0.42	NNLS	0.12	0.08	0.86
BSEQ-sc	0.17	0.12	0.58	BSEQ-sc	0.12	0.08	0.87
CIBERSORT	0.12	0.09	0.77	CIBERSORT	0.09	0.06	0.91
delta (0.092)	RMSD	mAD	R	**gamma** (0.062)	RMSD	mAD	R
MuSiC	0.04	0.03	0.98	MuSiC	0.05	0.038	0.97
NNLS	0.12	0.08	0.82	NNLS	0.12	0.081	0.84
BSEQ-sc	0.12	0.08	0.85	BSEQ-sc	0.12	0.083	0.86
CIBERSORT	0.10	0.07	0.90	CIBERSORT	0.10	0.070	0.90
acinar (0.084)	RMSD	mAD	R	**ductal** (0.177)	RMSD	mAD	R
MuSiC	0.05	0.04	0.97	MuSiC	0.050	0.037	0.97
NNLS	0.11	0.07	0.85	NNLS	0.067	0.046	0.96
BSEQ-sc	0.14	0.10	0.79	BSEQ-sc	0.084	0.064	0.93
CIBERSORT	0.07	0.05	0.93	CIBERSORT	0.076	0.058	0.94

Supplementary Table 5: Summary of cell types of Park et al. single-cell dataset. Park et al. sequenced 57,979 cells from healthy mouse kidneys and identified 16 cell types. As suggested in Park et al., we limited our consideration to the 13 confidently characterized cell types and eliminated CD-Trans and 2 novel cell types in our deconvolution analyses.

Cell Type	Abbr.	# Cell	% Cell	Cell Type	Abbr.	# Cell	% Cell
Endothelial	Endo	1,001	2.29	Fibroblast	Fib	549	1.26
Podocyte	Podo	78	0.18	Macrophage	Macro	228	0.52
Proximal tubule	PT	26,482	60.54	Neutrophil	Neutro	74	0.17
Loop of Henle	LOH	1,581	3.61	B lymphocyte	B lymph	235	0.54
Distal convoluted tubule	DCT	8,544	19.53	T lymphocyte	T lymph	1,308	2.99
Collecting duct principal cell	CD-PC	870	1.99	Natural killer cell	NK	313	0.72
Collecting duct intercalated cell	CD-IC	1729	3.95	Novel cell type 1	Novel 1	601	1.37
Collecting duct transitional cell	CD-Trans	110	0.25	Novel cell type 2	Novel 2	42	0.10