# Cell type-specific weighting-factors for accurate virtual single-cell RNA-sequencing of diverse organs

Kengo Tejima[1,2,3,*], Satoshi Kozawa[1,2,3,*], Thomas N. Sato[1,2,3,4¶]

ERATO Sato Live Bio-Forecasting Project, Kyoto, Japan[1]; The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International, Kyoto, Japan[2]; Karydo TherapeutiX, Inc., Kyoto, Japan[3]; V-iClinix Laboratory, Nara Medical University, Nara, Japan[4]

¶Corresponding Authors:

Thomas N. Sato

The Thomas N. Sato BioMEC-X Laboratories

Advanced Telecommunications Research Institute International (ATR)

Karydo TherapeutiX, Inc.

2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0288

Email: island1005@gmail.com

TEL: 81-774-95-2311

FAX: 81-774-95-2329

*These two authors contributed equally to this work

**Abstract**

Computational deconvolution of transcriptome data of organs/tissues uncovers their structural and functional complexities at cellular resolution without performing single-cell RNA-sequencing experiments. However, the deconvolution of highly heterogenous diverse organs/tissues remains a challenge. Herein, we report "cell type-specific weighting-factors" that are essential for accurate deconvolution, but critically lacking in the existing methods. We computed such weighting-factors for 97 cell-types across 10 mouse organs and demonstrate their effective usage in the Bayesian framework to generate their virtual single-cell RNA-sequencing data, hence accurately estimating both cell-type ratios and the complete transcriptome of each cell-type in these organs. The method also efficiently detects the temporal changes of such cell type-profiles during organ pathogenesis in disease models. Furthermore, we present its potential utility for human organ/bulk-tissue deconvolution. Taken together, the weighting-factors reported herein and their computation for new cell-types and/or new species such as human are essential tools/resources for studying high-resolution biology and disease.

41 **Introduction**

42 Organs consist of ca. $1x10^8$ – $3x10^{12}$ cells in human and their functional interactions are critical for
43 organ structure and function. Changes of the cellular composition and/or the cellular phenotype of the
44 organs are tightly linked to their dysfunction, remodeling and regeneration. Single-cell RNA-
45 sequencing (scRNA-seq) analyzes comprehensive gene expression profiles of individual composite
46 cells of organs, providing an opportunity to unveil such changes at single cell resolution (Deng et al.,
47 2014; Han et al., 2018; Regev et al., 2017; Sandberg, 2014; Tabula Muris et al., 2018). Hence, scRNA-
48 seq is a powerful method to generate unprecedentedly detailed molecular cell atlas of normal and
49 pathological organs.

50 However, scRNA-seq has limitations. First, prior to subjecting to scRNA-seq, the organs must
51 be dissociated to individual cells by enzymatic and physical means and the living cells must be
52 individually collected. Hence, the major premise is the availability of fresh organs. This is often a
53 problem in the case of studying clinical human organs. In many cases, they are stored frozen for months
54 and years and that they cannot be subjected to scRNA-seq analyses.

55 Furthermore, rare pathological samples are difficult to be freshly obtained and that the analyses
56 must be conducted with frozen samples of limited availability. In addition, clinical human samples,
57 such as biopsies, are often small parts of the whole organ. Even in the case of autopsies, dissociating a
58 large whole-organ such as the heart and the brain into individual cells for scRNA-seq analyses is a
59 formidable task, if not an impossible one.

60 Another limitation is the difficulty of analyzing multiple organs from the identical individual
61 subjects. Due to the cell-dissociation/-purification steps, it is a difficult, and often an impractical, task
62 to conduct scRNA-seq of multiple organs from the identical experimental model organism. This is often
63 necessary in many experimental situations such as the study of the drug effects and/or pathogenesis
64 on/in multiple organs in the same subject (Kozawa et al., 2018).

65 Procedural artifacts in gene expression are also problems with scRNA-seq. It has been reported
66 that cell dissociation steps induce aberrant gene expression in the cells (van den Brink et al., 2017). The
67 cost and labor are also major hurdles. The scRNA-seq analyses of many organs and samples derived

68 from many disease subjects and disease-models of humans and experimental animal models,
69 respectively, are cost-/labor-intensive.

70 A solution to these limitations and problems is computational deconvolution of whole-organ
71 and bulk-tissue RNA datasets. The whole-organ/bulk-tissue RNA can be more easily purified from
72 fresh or frozen samples. Simultaneous purification of RNA from a large number of organs/tissues are
73 also possible. The cost and labor with whole-organ/bulk-tissue RNA analyses are much less than those
74 of scRNA-seq.

75 Several computational methods to deconvolve whole-organ/bulk-tissue RNA-seq data are
76 reported (Abbas et al., 2009; Avila Cobos et al., 2018; Gaujoux and Seoighe, 2012; Gong et al., 2011;
77 Gong and Szustakowski, 2013; Kang et al., 2019; Li et al., 2016; Newman et al., 2015; Repsilber et al.,
78 2010; Shen-Orr and Gaujoux, 2013; Tsoucas et al., 2019; Wang et al., 2015; Wang et al., 2019; Zhong
79 et al., 2013). They compute cell type compositions of the source tissues using the corresponding bulk-
80 tissues RNA-seq data. Three most recently reported methods claim that they outperform the other
81 methods published in previous years. They are MUlti-Subject Single Cell deconvolution (MuSiC),
82 Dampened Weighted Least Squares (DWLS) and Complete Deconvolution for Sequencing data
83 (CDSeq) methods (Kang et al., 2019; Tsoucas et al., 2019; Wang et al., 2019). However, their
84 validations are limited to synthetic datasets, RNA-seq data derived from cell cultures and/or the
85 mixtures of a few tissues, and/or 1 – 4 real organs. Thus, their applicability to more diverse real-organs
86 remain unclear.

87 In this study, we examined these existing methods for deconvolving 14 mouse whole-organs.
88 We evaluated mouse organs, instead of human organs, as many organs and paired single-cell RNA-seq
89 datasets are readily available (Han et al., 2018; Kozawa et al., 2018; Tabula Muris et al., 2018).
90 Availability of such exactly paired datasets for human organs is limited. Our analyses of the existing
91 deconvolution methods with these paired mouse datasets found that they fail to recapitulate their known
92 composite cell-type ratios. Hence, we examined a cause/causes for their failure and found "cell type-
93 specific weighting-factors" that could overcome their defects. Hence, we computed 97 such wighting-
94 factors of 97 cell-types across 10 mouse organs. Using them in the Bayesian framework, we generate
95 virtual scRNA-seq data of these organs, accurately estimating their cell-type compositions and the

96 complete transcriptome profiles of their composite cell-types. Furthermore, we demonstrate that the
97 approach effectively detects the temporal changes of cell-type compositions and gene expression
98 patterns of the composite cell-types during the cardiac pathogenesis in the myocardial infarction model.
99 In addition to these studies with mouse datasets, we also show a potential usage of this approach for
100 human bulk-tissue deconvolution.
101
102 **Results**
103 **The existing deconvolution methods fail to accurately estimate the cell-type-ratios of mouse**
104 **organs**
105 Several deconvolution methods are reported (Abbas et al., 2009; Avila Cobos et al., 2018; Gaujoux and
106 Seoighe, 2012; Gong et al., 2011; Gong and Szustakowski, 2013; Kang et al., 2019; Li et al., 2016;
107 Newman et al., 2015; Repsilber et al., 2010; Shen-Orr and Gaujoux, 2013; Tsoucas et al., 2019; Wang
108 et al., 2015; Wang et al., 2019; Zhong et al., 2013). However, their applicability to whole-organs/bulk-
109 tissues deconvolution limited to only a few organs/tissues. Hence, we evaluated these previously
110 reported methods with 14 real mouse organs (Figure 1). We chose MuSiC and DWLS methods, as they
111 appear to outperform all the other methods in deconvoluting 1 – 4 organs/bulk-tissues (Tsoucas et al.,
112 2019; Wang et al., 2019). The result shows that the computed cell-type ratios significantly deviate from
113 the ones experimentally estimated by real scRNA-seq studies to various degrees in most of these organs
114 (Figure 1).
115
116 **Cell type-specific weighting-factors to offset the gap between the synthetic and real whole-organ**
117 **RNA-seq data**
118 To identify a cause/causes for the failed performance of the existing methods (Figure 1), we studied
119 their algorithm and computational framework. We found that they are built on the common assumption
120 that the whole-organ/bulk-tissue RNA-seq is the sum of gene expression of each composite cell-type
121 weighted by cell-type ratios. Based on this assumption, the synthetic whole-organ/bulk-tissue data
122 constructed from single-cell transcript counts combined according to the known fraction of each cell-
123 type in the organ/tissue (i.e., experimentally determined by real scRNA-seq) must be approximately

124 equal to the real whole-organ/bulk-tissue RNA-seq data. We evaluated this assumption by comparing
125 the gene expression profiles of the synthetic whole-organ RNA-seq data and the real whole-organ RNA-
126 seq data of the same organs (Figure 2A). The synthetic whole-organ RNA-seq data is calculated as the
127 sum of the normalized transcript counts weighted by the known ratios of the composite cell-types of
128 the organs (see Materials and methods). Shown are the results by using the indicated number of the top
129 ranked signature genes for distinguishing the cell-types for each organ as computed by random forest
130 (RF) (Figure 2A, see Materials and methods, the list of all signature genes can be found in Figure 2 -
131 Figure supplement 1 – Source data 1). The similarities/dissimilarities for the 14 organs are shown by
132 Pearson correlation coefficients (Figure 2A). The results show that the Pearson correlation coefficients
133 are less than 0.75 for 10 organs (aorta, brain, heart, large intestine, liver, lung, pancreas, skin, skeletal
134 muscle, thymus) (Figure 2A). Such large gaps are unlikely due to sex or strain differences, as both the
135 whole-organ and single-cell RNA-seq are derived from the same sex (male) and strain, C57BL6, (Han
136 et al., 2018; Kozawa et al., 2019; Kozawa et al., 2018; Tabula Muris et al., 2018), except for the skeletal
137 muscle where its scRNA-seq data are derived from female mice. This result suggests the common
138 assumption of the other methods that the whole-organ/bulk-tissue RNA-seq is the sum of the
139 normalized transcript counts of each composite cell-type weighted by cell-type ratios must be re-
140 evaluated.

141 On the basis of the apparent gap between the synthetic and real organ RNA-seq data (Figure
142 2A), we postulated the presence of a putative variation among different cell-types that are not reflected
143 on real single cell RNA-seq data. One such variation could be total cellular RNA contents. For cultured
144 cells, it is experimentally shown that the total RNA contents vary between 50,000 – 300,000 transcripts
145 per cell (Marinov et al., 2014). While such cellular RNA content-variations in heterogenous solid
146 organs/tissues remain elusive, there are at least some evidence. It is reported that cell size regulates
147 cellular RNA abundance (Schmidt and Schibler, 1995). In this regard, in the heart, the volume of the
148 cardiac muscle cell is 20 – 25 times of that of non-muscle cells such as endothelial cells and fibroblasts
149 (Anversa et al., 1980; Dammrich and Pfeifer, 1983; Mattfeldt et al., 1986; Tang et al., 2009; Zhou and
150 Pu, 2016). Hence, it is possible that the total RNA content per cells may be significantly different
151 between muscle and non-muscle cell types. Interestingly, while the muscle cells consist of only 25%

152 and 10% of the heart and the skeletal muscle, respectively (reference bar plots in Figure 1) (Banerjee
153 et al., 2007; Bergmann et al., 2015; Nag, 1980; Pinto et al., 2016; Raulf et al., 2015; Walsh et al., 2010;
154 Zhou and Pu, 2016), the existing deconvolution methods (MuSiC, DWLS) computed that the muscle
155 cell-type is the most abundant cell type consisting of 80-90% in the heart and the skeletal muscle
156 (MuSiC and DWLS bar plots in Figure 1).

157 To account for this and other possible variations among different cell-types, we computed and
158 determined the cell type-specific weighting-factors for each organ (Figure 2B, see also Materials and
159 methods). The concept is as follows: The total transcript counts per cell in scRNA-seq data are
160 normalized so that the total counts per cell become the same. This results in the loss of the putative
161 differences of total cellular RNA contents among various cell-types. Hence, the cell type-specific
162 weighting-factors ($w_1$, $w_2$, $w_3$, etc.) reflecting the relative differences of the total cellular RNA contents
163 among various cell-types compensate such loss, allows for more accurate deconvolution (Figure 2B,
164 see also Materials and methods). Based on this concept, we computed at least 100 weighting-factors for
165 each "virtual-cell/cell-subject". To introduce a Bayesian framework for the subsequent analyses, we
166 made them follow Gaussian distribution and calculated their mean and variance (see Materials and
167 methods for the detailed calculation method). The result is shown as boxplots of the means of individual
168 virtual-cells/cell-subjects for each cell-type in each organ (Figure 2 – Figure Supplement 1). The total
169 number of virtual-cells/cell-subjects is 100 or more and adjusted according to the cell-type ratios of the
170 reference data (i.e., the known real cell-type ratios) for each organ (e.g., 100 cell type As and 200 cell
171 type Bs when the ratio of A and B cell types is 1:2 in an organ). Using the mean weighting-factor value
172 for each cell-type (i.e., the mean of the means of all virtual-cells/cell-subjects for each cell-type), we
173 weighted the transcript counts for each cell-type (see Materials and methods). We then summed these
174 weighted transcript counts for all composite cell-types of each organ according to their ratios to
175 construct virtual organ RNA-seq data (Figure 2B, see Materials and methods). These updated virtual
176 organ RNA-seq are compared to the corresponding real organ RNA-seq data (Figure 2C). The results
177 show the significant improvement (Pearson correlation coefficients = 0.8 – 1.0) in the similarities of
178 gene expression profiles for all organs, except for the skin and the thymus (Compare Figure 2C to Figure
179 2A).

180

181 **Computation of the cell-type ratios and the complete transcriptomes of all composite cell-types**

182 Using these weighting-factors, we designed a Bayesian framework to simultaneously compute both
183 cell-type ratios and their gene expression patterns (Figure 3A, see also Materials and methods). In this
184 framework, using the mean of the cell type-specific weighting-factors ($w_1$, $w_2$, $w_3$ in Figure 3A) for
185 each cell type (indicated as 1, 2, 3 in Figure 3A), we compute virtual transcript counts ("x" in Figure
186 3A) for each gene (indicated as wiggling lines of different colors in Figure 3A) in each cell-type and
187 the ratio (indicated as "r" in Figure 3A) of each cell-type. The computation is iteratively performed in
188 the Bayesian framework with two hyperparameters, $\alpha$ (0.001 – 1,000) and $\beta$ (0.001 – 1,000), to account
189 for the combinatorial influence of the cell-type ratios and the gene expression patterns of each cell-type
190 on the overall organ-level transcriptome, respectively (Figure 3A, see also Materials and methods). For
191 the iteration (t = 1, …, T), the initial condition is set as $x_1^{(0)}$ and $r_1^{(0)}$ for the transcript counts and the
192 ratio of cell-type 1 (see the diagram labels as "Prediction" in Figure 3A). Those for cell-types 2 and 3
193 are indicated accordingly (see the diagram labels as "Prediction" in Figure 3A). Such initial conditions
194 are the normalized transcript count for each gene and the reference cell-type ratio, respectively. The
195 iteration is repeated up to 1,001 times so that the synthetic whole organ transcript counts calculated by
196 the virtual counts (weighted by the cell type-specific weighting-factors) weighted by the virtual cell-
197 type ratio (labeled as "Virtual Whole Organ RNAseq" in Figure 3A) is as close as those of the
198 corresponding real organ (Figure 3A, see Materials and methods).

199 By this approach, we estimate the cell-type ratios of 10 real organs (aorta, fat, heart, kidney,
200 large intestine, liver, lung, marrow, skeletal muscle, spleen) (Figure 3B). From the 14 organs evaluated
201 in Figures 1 and 2, we omitted the brain, the pancreas, the skin and the thymus for the following reasons:
202 1) Their real cell-type ratios are not available – we used the synthetic ratios in Figures 1 and 2 (see the
203 Materials and methods), 2) The real cell-type ratios of islet tissues are available, but not those of the
204 real "whole pancreas"– the reference in Figures 1 and 2 is that of the islet (Tabula Muris et al., 2018),
205 3) For the skin or the thymus, the weighting-factors could not improve the Pearson correlation
206 coefficients above 0.8 (Figure 2C).

207 The computed cell-type ratios of the 10 organs are similar to the reference ratios that are
208 experimentally determined by real scRNA-seq studies (Figure 3C). Notably, our method corrected the
209 abnormally large ratios of cardiac muscle and muscles in the heart and the skeletal muscle, respectively,
210 that are estimated by MuSiC and DWLS (Figure 3B). Their mean squared errors (MSEs) in reference
211 to the experimentally determined ratios indicate that our method outperforms the others for 5 real organs
212 (fat, heart, large intestine, liver, skeletal muscle) (Figure 3C).
213 We also computed the virtual transcriptome of 23,131 (all organs, except the skeletal muscle)
214 and 14,323 (the skeletal muscle) genes in each of the "virtual" cell-type and compared them to those of
215 the real cell-type for the 10 organs (Figure 4, see Materials and methods). Our approach computes a
216 posterior distribution and its Expected A Posteriori (EAP) for counts of each transcript for each cell-
217 type. Hence, we calculated Pearson correlation coefficients, Spearman correlation coefficients and MSE
218 for each cell-type, using the EAPs of all genes in virtual and real cell-types for each organ (Figure 4A
219 and Figure 4 – Source data 1). Their Pearson correlation coefficients indicate that the virtual
220 transcriptomes are comparable to the real ones for all cell-types and organs (Figure 4). They also show
221 similarities and relatedness of the same/related cell type annotations across different organs (Figure 4A).
222 For example, the similarity of endothelial cells in multiple organ such as aorta, fat, heart, liver and lung
223 is recapitulated in the virtual transcriptomes among these organs (Figure 4A). The relatedness of each
224 of the immune system cells across multiple organs are also represented (Figure 4A). The t-distributed
225 Stochastic Neighbor Embedding (t-SNE) analyses of all 10 virtual scRNAseq data show that all
226 composite cell-types in each organ can be classified according to their gene expression profiles (Figure
227 4B).
228
229 **Computing the changes of cell-type ratios and gene expression in disease**
230 We next evaluated whether our method can detect the temporal changes of cell-type ratios and gene
231 expression in each cell-type that accompany disease processes. Cardiovascular diseases are the number
232 one death cause worldwide (https://www.who.int/news-room/fact-sheets/detail/cardiovascular-
233 diseases-(cvds)). In addition, extensive experimental studies characterizing the temporal changes in the
234 cell-type compositions during heart disease exist (Schirone et al., 2017; Shah and Brownlee, 2016).

235 Moreover, we have shown that the heart is an organ that our method, but not other previously published
236 methods, can effectively compute both the cell type ratios and their gene expression patterns of the
237 composite cell-types (Figures 3&4). Hence, we examined a mouse model for myocardial infarction
238 (MI) (Figure 5).

239 Using this model, we investigated whether our computational method can detect both the
240 known temporal changes of the cardiac cell type ratios and of the cardiac cell-dependent gene
241 expression (Figure 5). Three stages of myocardial infarction were studied (Kozawa et al., 2018;
242 Murakoshi et al., 2013): 1) one-day post coronary artery-ligation (E-MI, early myocardial infarction
243 stage), 2) 7-days post coronary artery-ligation (M-MI, early fibrosis stage), 3) 8-weeks post coronary
244 artery-ligation (L-MI, cardiac remodeling stage).

245 First, we evaluated whether the method detects the known changes of the cell-type ratio that
246 accompanies the pathological changes (Figure 5A). For this analysis, we used the cell type-specific
247 weighting-factors calculated by using the RNA-seq data of the sham controls (E-sham, M-sham, L-
248 sham) and the reference cell-type ratios of the normal mouse heart for each cell-type and each stage (E,
249 M, L). The result shows that the method can detect the expected two-major MI associated cell-type
250 composition changes (Leask, 2010; van den Borne et al., 2010): Decreased cardiac muscle cells and
251 increased fibroblasts (Figure 5A). The method also detected the increased myofibroblasts during M-MI
252 stage (early fibrosis stage) (Leask, 2010; van den Borne et al., 2010) (Figure 5A). This is also in
253 agreement with the published experimental observations (Leask, 2010; van den Borne et al., 2010).

254 Next, we studied whether the method finds the known gene expression changes in these cell-
255 types (Figure 5B). The result shows that the method identifies gene expression changes that are
256 previously reported in experimental studies (Cleutjens et al., 1995; Fraccarollo et al., 2002; McCurdy
257 et al., 2011; Nakao et al., 1997; Sergeeva and Christoffels, 2013) (Figure 5B). In the cardiac muscle
258 cells, the statistically significant (adjusted p values $< 0.001$) upregulation ($\log_2$fold changes $> 0.7$) of
259 Nppb, Nppa, Sparc, Col4a1 and downregulation ($\log_2$fold changes $< 0.7$) of Myh6 were detected (Figure
260 5B). In the fibroblasts, the statistically significant (adjusted p values $< 0.001$) upregulation ($\log_2$fold
261 changes $> 0.7$) of Col4a1, Col1a1 and Sparc were detected (Figure 5B). In addition to these previously

262 known landmark genes for MI pathogenesis, the computation found many other differentially regulated
263 genes in each cell type (Figure 5B).
264
265 **Application to human bulk-tissue deconvolution**
266 Unlike mouse organs, many of the human datasets are derived from different parts of the same
267 organs/tissues among different studies and even in the same studies. Furthermore, for technical and
268 ethical reasons, it would not be easy to obtain the exactly paired whole-organ/bulk-tissue RNA-seq and
269 scRNA-seq data from the same individual subjects. Hence, the evaluation of our method for human
270 bulk-tissue is conducted with the following limitations: 1) The human bulk-tissue RNA-seq and the
271 scRNA-seq data of the heart and the kidney are not paired. The human bulk-tissue data are from "The
272 Human Protein Atlas" (https://www.proteinatlas.org/) and the scRNA-seq data of the human heart and
273 kidney are derived from pieces of tissues excised from the human left ventricle (LV) and the left
274 auricle/left atrial appendages (LA) (Wang et al., 2020) and a small part of the human kidney (Liao et
275 al., 2020), respectively, 2) As the result of 1), we used the mouse weighting-factors (the paired bulk-
276 tissue and single-cell RNA-seq data are necessary to compute the human weighting-factors), 3) As the
277 result of 1), the reference cell-type ratios we used are those of the corresponding mouse organs modified
278 by the other published human data (see Materials and methods), 4) As the result of 1), the cell-type
279 classification of the virtual scRNA-seq data is related to, but not identical to that of these real scRNA-
280 seq data.
281 Despite these limitations, we computed the cell-type ratios and transcriptome profiles (Figure
282 6) of these bulk-tissues by using 16,144 human orthologs of mouse genes present in the human bulk-
283 tissue RNA-seq data. For the computation of the heart cell-type ratio, the ratio computed by our method
284 (V-scRNAseq in Figure 6A) is comparable to the reference (Reference in Figure 6A). In contrast, the
285 existing methods, MuSiC and DWLS, failed to recapitulate the reference value, resulting in the
286 abnormally large ratio of the cardiac muscle (MuSiC and DWLS in Figure 6A), as observed in the
287 mouse heart deconvolution (Figure 3B, C). For the kidney cell-type computation, the calculations by
288 all three methods resulted in the similar ratios that are comparable to the reference values (Figure 6A),

289 as observed in the mouse kidney deconvolution (Figure 3B, C). These results suggest an applicability
290 of our method to cell-type deconvolution of human tissues.

291 Next, the gene expression pattern of each virtual human cell-type was analyzed (Figure 6B).
292 This analysis was conducted by using the gene sets that are detected in both the virtual and the real cell-
293 types in each comparison for each organ (Figure 6B, see Materials and methods). The Spearman
294 analysis shows related transcriptome patterns between the virtual and the real cell-types in all the
295 comparisons (Figure 6B), with weaker correlation between the virtual and real collecting duct cells of
296 the kidney (R-hsCDI vs. V-CD and R-hsCDP vs. V-CD in Figure 6B). The Pearson analysis shows
297 weaker correlation, but the relative pattern is similar to that of the Spearman (Figure 6B). The MSEs
298 varied between 0.022 – 0.085 (Figure 6B). Despite the use of these unpaired datasets, the results show
299 utilities of our approach for estimating the gene expression pattern of each cell-type comprising the
300 bulk human tissues.

301

302 **Discussion**

303 In this study, we developed a computational method that allows for the cell-type and the gene-
304 expression deconvolution of 10 mouse whole-organs RNA-seq data. This is realized by computing and
305 utilizing 97 cell type-specific weighting-factors across these organs in the Bayesian framework to
306 correct the gap between the synthetic and real organ RNA-seq data (Figures 2 – 4). This approach
307 allowed us to more accurately estimate the ratios of the composite cell-types across 10 mouse whole-
308 organs and also compute their complete transcriptome (Figures 3B&C, 4).

309 One possible biological source of these weighting-factors is the relative differences of total
310 cellular RNA contents among various cell-types as described in Result section. Interestingly, the
311 importance of putative differential cellular RNA contents is recently implicated in deconvolving mixed
312 cell types in blood (Zaitsev et al., 2019). In their study, they suggest a "spike-in" approach using
313 "cultured cells of the same cell-types" in the experimental design to normalize such putative differential
314 cellular RNA contents (Zaitsev et al., 2019). However, this approach is not suitable for organ/bulk-
315 tissue deconvolution for the following reasons: 1) The cellular RNA contents of culture cells are not
316 necessarily the same as those of the real cells in the solid organs, 2) It is nearly impossible to gather all

317 cell-types that are present in the highly complex solid organs of interest as culture cells, 3) It is also
318 impractical to perform such "spike-in" experiments for many distinct cell-types of diverse cellular RNA
319 contents.

320 Hence, we propose to use the 97 cell type-specific weighting-factors presented in this paper
321 and also the method reported herein to compute additional ones to account for such putative differential
322 cellular RNA-contents among diverse cell-types in the organ/bulk-tissues for their deconvolution.
323 Furthermore, the cell type-specific weighting-factors may also be correcting other yet unknown factors
324 contributing to the gap between the real and virtual data. This cannot be dealt with by the spike-in
325 approach.

326 Another advantage of using our approach is its ability to simultaneously compute both the ratios
327 and the complete transcriptome of each cell-type across diverse real individual organs and also to detect
328 their disease-associated changes (Figures 3 – 5). Many of the previously reported deconvolution
329 methods are primarily specialized for estimating cell-type compositions, but not for computing cell
330 type-specific gene expression patterns. MuSiC and DWLS can estimate only the ratio changes of the
331 cell-types in organs induced by external perturbations, but not the gene expression changes (Tsoucas et
332 al., 2019; Wang et al., 2019). Exception to this is CDseq (Kang et al., 2019). However, the calculation
333 of only the overall gene expression patterns of each cell-type using only synthetic, but not real, datasets
334 are shown for CDseq (Kang et al., 2019). Our approach detected not only several previously known
335 gene expression changes (Figure 5B), but also many others in each cell-type at each stage in the MI-
336 heart (Figure 5B). Hence, their future experimental validation could facilitate the identification of
337 therapeutic targets and/or biomarkers for MI pathogenesis.

338 While we were unable to definitively demonstrate how well our method work for the human
339 bulk-tissue RNA-seq deconvolution due to the lack of the exactly paired scRNA-seq data of the
340 corresponding bulk-tissues, we provided some evidence to support this possibility (Figure 6). The future
341 availability of such paired datasets for human organs/bulk-tissues would provide an opportunity to
342 validate this possibility. Various large databases of human organ RNA-seq (normal and pathological)
343 are available (e.g., https://gtexportal.org/home/, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-
344 bin/study.cgi?study_id=phs001203.v1.p1). Their deconvolution is an efficient use of the existing

345 resource and also provide quick and useful working-hypotheses for targeted experimental validations.
346 This is complementary to expensive and labor-intensive real scRNA-seq analyses of all human 30
347 trillion cells of healthy and disease subjects.

348 Our method is not just a "poor-person's tool kit" in place of a real organ scRNA-seq. It is a
349 complementary approach providing opportunities for efficient and effective use of the existing datasets
350 and bioresources that are difficult, time-consuming and/or labor/cost-intensive to be replicated and/or
351 expanded. It is also an in silico tool to generate biologically relevant and experimentally testable
352 hypothesis for biological and/or medical questions.

353 While we demonstrate the usefulness of our approach, some future improvements would further
354 accelerate its applications. Even with the cell type-specific weighting-factors, we were unable to offset
355 the discrepancies between the real and the virtual whole-organ transcriptome in the skin and the thymus
356 (Figure 2C). While the sex and the strain are matched between the whole-organ and scRNA-seq in these
357 organs (Kozawa et al., 2019; Kozawa et al., 2018; Tabula Muris et al., 2018), some procedural artifacts
358 such as cell-dissociation methods of these organs or the subtle differences in the housing environment
359 of the mice and/or the site of tissue harvesting (especially the skin) could have influenced the lack of
360 significant improvements by the weighting-factors. These possibilities could be solved by obtaining the
361 exactly pair-matched organs/tissues for the whole-organ and scRNA-seq data. Other potential
362 improvements include the development of a completely reference-free deconvolution algorithm (i.e.,
363 the complete deconvolution) and the bias/hypothesis-free detection of rare unexpected cell-types.
364 Importantly, even with these future improvements, the cell type-specific weighting-factors and also
365 their calculation method described herein remain essential.

366

367

368 **Materials and methods**

369 **Data preprocessing & count-normalization**

370 All data processing and analyses were performed using R version 3.6.1. All cell-type labels are as
371 described in each of the original scRNA-seq studies(Han et al., 2018; Tabula Muris et al., 2018). Gene
372 symbols in the scRNA-seq data were converted to the corresponding ones in the whole organ RNA-seq
373 data by using entrez gene IDs derived from "org.Mm.egALIAS2EG" in R package of "org.Mm.eg.db".
374 ERCC-labelled genes were removed as they are spike-in genes. In addition, the counts of three genes,
375 *Rn45s*, *Akap5* and *Lrrc17*, were also removed as they are non-mRNA artifacts significantly influencing
376 the total counts. We then adjusted the total counts of each cell in the scRNA-seq datasets to 100 by
377 normalizing those of each gene. We chose arbitrarily 100 herein but lager numbers result in difficulty
378 treating the resulting inverse matrix. On the other hand, with the smaller numbers, the transcript counts
379 of individual genes become too small. The same normalization process was also applied to each of the
380 whole-organ RNA-seq datasets.

381

382 **Calculation of reference cell-type ratios**

383 For 11 organs (aorta, fat, kidney, large intestine, liver, lung, marrow, pancreas, skin, spleen, & thymus),
384 the cell-type ratios described in *Tabula Muris*(Tabula Muris et al., 2018) were used as the references.
385 For the skeletal muscle, the cell type ratios described in Mouse Cell Atlas(Han et al., 2018) were used
386 as the reference. For the heart, the cell-type ratios described in *Tabula Muris*(Tabula Muris et al., 2018)
387 were modified as the cardiac muscles and the non-muscle cells were separately analyzed and the
388 reported ratio of cardiac muscle cells (3.1%) appears significantly smaller than that (30 – 40%) is
389 generally agreed in the field on the basis of various previous studies (Banerjee et al., 2007; Bergmann
390 et al., 2015; Nag, 1980; Pinto et al., 2016; Raulf et al., 2015; Walsh et al., 2010; Zhou and Pu, 2016).
391 For this modification, we set the ratio of the cardiac-muscle cells at 30% and the remaining 70% was
392 divided by the non-muscle cell types by maintaining their ratios among them the same. For the brain,
393 we estimated the reference cell-type ratio based on the reports in NIH
394 (http://www.nervenet.org/papers/BrainRev99.html#Numbers) using the cell subject numbers of the
395 cell-types in scRNA-seq data described in *Tabula Muris* (Tabula Muris et al., 2018). First, we divided

396 the brain cell-type classes into four classes: “neurons”, “glial cells”, “endothelial cells”, “others”, and
397 then set the ratio of these classes at 75 : 23 : 7 : 4, respectively, accordingly to their estimated ratios in
398 murine brain (http://www.nervenet.org/papers/BrainRev99.html#Numbers). Second, we further
399 divided “neurons”, “glial cells”, and “others” into more cell-type classes according to *Tabula Muris*
400 (Tabula Muris et al., 2018). The “neurons” were further divided into “neuron-excitatory neurons and
401 some neuronal stem cells” and “neuron-inhibitory neurons”. The “others” class was divided into “brain
402 pericyte-NA” and “oligodendrocyte precursor cell-NA”. For the “glial cells” class, we introduced three
403 assumptions: 1) the “glial cells” can be classified into four cell-types, “microglial cell-NA”, “astrocyte-
404 NA”, “Bergmann glial cells-NA”, and “oligodendrocyte-NA”, according to *Tabula Muris*(Tabula
405 Muris et al., 2018); 2) the ratios of these four glial cell types follow those of *Tabula Muris* (Tabula
406 Muris et al., 2018) among them, and 3) the ratio of “microglial cells-NA” in the whole brain is 0.1, as
407 it is reported that the “microglial cells” account for 10-15 % of the whole brain cells. On the basis of
408 these assumptions, the ratios of each brain cell type was estimated as “macrophage-NA” (ca. 0.2%),
409 “microglial cell -NA” (10.00%), “astrocyte-NA” (ca. 2.2%), “Bergmann glial cell-NA” (ca. 2.1%),
410 “brain pericyte-NA” (ca. 1.5%), “endothelial cell-NA” (ca. 6.4%), “neuron-excitatory neurons and
411 some neuronal stem cells” (ca. 47.5%), “neuron-inhibitory neurons” (ca. 21.3%), oligodendrocyte-NA”
412 (ca. 8.7%) and “oligodendrocyte precursor cell-NA” (ca. 1.9%) and used as the reference. For human
413 organs, we used the mouse cell-type ratios for the heart and the kidney, as the references.
414

415 **Selection of signature gene sets for cell-type distinction**

416 Random Forest (RF) was performed to compute signature genes for distinguishing cell-types from one
417 another in each organ using the scRNA-seq count data derived from the cell-type reference datasets
418 described in the above section. In this study, ‘randomForest’ package in R was used for tuning and
419 producing a classifier by RF. The scRNA-seq data were first divided into two: the training data for
420 producing a classifier by RF and the test data for F1-score calculation to test accuracy in the classifier.
421 The cell-type ratios as described in the previous section were used for the RF analyses. Following the
422 classifier production, the important features in the classifier were extracted as the signature genes for

423 each cell-type, where we used "Mean Decrease in Gini" values as the importance indicator for each
424 gene.
425
426 **Computation of the cell type-specific weighting-factors**
427 The apparent gap between the "synthetic" and "real" whole-organ RNA-seq count data across many
428 organs (Figure 2A) indicated the necessity to account for putative cell type-specific differences. One
429 possible such variation could be total cellular RNA contents (Marinov et al., 2014; Schmidt and Schibler,
430 1995)(see also the detailed description in Results section). Hence, we computed weighting-factors to
431 correct such differences. The concept of computation strategy for the weighting-factors is summarized
432 in Figure 2B and its legend. Briefly, the scRNA-seq counts are normalized (i.e., the total counts per cell
433 is the same for all individual cells), hence the putative differential total cellular RNA contents among
434 various cell-types are lost from the count data. With the cell type-specific weighting-factors ($w_1$, $w_2$,
435 $w_3$, etc.), the RNA-sequencing data recover such loss, hence leading to more accurate deconvolution.
436 In the computation, we calculated at least 100 weighting-factors for each virtual-cell/cell-subject of
437 each cell-type, instead of one weighting-factor per cell-type and made them follow Gaussian
438 distribution. Based on this concept, we developed a model as follows:

439 $$m\mathbf{y} = \sum_{j} w_j \mathbf{x}_j \ \ (j \in C^m) \qquad (1)$$

440 where $\mathbf{y} \in \mathbb{R}^n$, $w_j > 0$ and $\mathbf{x}_j \in \mathbb{R}^n$ denote the normalized whole-organ RNA-seq counts vector, the
441 weighting-factor for each cell subject $j$, and the normalized scRNA-seq counts matrix, respectively.
442 The $n$ is the total number of the signature genes for the organs. In this study, 100, 300, 2,000 signature
443 genes were selected according to the ranking on the basis of "Mean Decrease in Gini" values calculated
444 by the RF analyses. When the total number of signature gene derived by the RF analyses is less than
445 2,000, all signature genes were used (1,577 for the aorta and 1,461 for the kidney). The combination of
446 cell subjects (i.e. the virtual cells), $C^m$ , was selected at random while maintaining the cell-type ratio of
447 the reference data with the total number of cell subjects, $m$, which is set to become equal to or less than
448 $n$ (the latter in the case when the number of the selected signature genes are too large to collect the

equal number of the cell subjects). With this model, $w_j$ was calculated by solving a quadratic problem under the constraint that the resulting value is over 0.01 as follows:

$$\widetilde{w}_j = \mathrm{argmin}_{w_j} \sum_i^S \left| m\mathbf{y}_i - \sum_j w_j \mathbf{x}_j \right|^2 \quad s.t.\ w_j \geq 0.01. \tag{2}$$

Here, $S$ is the number of whole-organ RNA-seq count datasets. In this study, we used biological duplicate count datasets of the corresponding whole-organ RNA-seq, thus $S$ is 2. This quadratic problem was solved by using "quadprog" package in R. The process of both random-selection of the cell subject combination and $\widetilde{w}_j$ calculation was recursively performed until more than 100 $w_j$ were generated for all selected cell subject. Then, by assuming that the distribution of the weighting-factors follows Gaussian distribution, we calculated the mean and variance of the weighted counts for each cell subject as follows:

$$\tilde{\mathbf{x}}_j \sim \mathcal{N}\left(\mu_{w_j}\mathbf{x}_j, \sigma_{w_j}^2 \mathrm{diag}(\mathbf{x}_j \odot \mathbf{x}_j)\right) \tag{3}$$

where, $\tilde{\mathbf{x}}_j$, $\mu_{w_j}$, and $\sigma_{w_j}^2$ denote the weighted count vector for the cell subject $j$, the mean of the weighting-factors for the cell subject $j$, and the variance of the weighting factors for the cell subject $j$, respectively. Operator $\odot$ represents the element-wise product between two vectors. On the basis of the computed mean and variance of the weighting-factors for each cell subjects, the weighted count vector of cell-type $k$ was calculated by assuming that they follow Gaussian distribution as follows:

$$\bar{\mathbf{x}}_k \sim \mathcal{N}\left(\frac{1}{N_k}\sum_j \mu_{w_j}\mathbf{x}_j, \frac{1}{N_k}\sum_j \sigma_{w_j}^2 \mathrm{diag}(\mathbf{x}_j \odot \mathbf{x}_j)\right) \ (j \in C^k). \tag{4}$$

where, $k$, $C^k$, and $N_k$ denote a cell-type, a group of subjects labeled with cell-type $k$, and the number of subjects in $C^k$, respectively.

**Computing the cell-type ratios & gene expression patterns**

The concept of our framework is described as (Figure 3A) and the model is as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{r} \tag{5}$$

where, $\mathbf{y}$, $\mathbf{X} = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_k, \ldots, \bar{\mathbf{x}}_K)$, and $\mathbf{r}$ are the whole-organ RNA-seq data vector, the matrix of the virtual transcript counts where the columns are the weighted-counts for each cell-type calculated as

above, and the coefficient vector corresponding to the cell-type ratio, respectively. To compute $\mathbf{X}$ and $\mathbf{r}$, we employed Bayes' theorem. To apply Bayes' theorem, noise $\mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$ was added to Equation (5) and a probabilistic model was developed as follows:

$$\mathrm{P}(\mathbf{y}|\mathbf{X}, \mathbf{r}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{r}, \beta^{-1}\mathbf{I}) \tag{6}$$

where, $\beta$ denotes a hyperparameter. According to the Bayes' theorem, the posterior distribution of $\mathbf{X}$ and $\mathbf{r}$ was obtained as

$$\mathrm{P}(\mathbf{X}, \mathbf{r}|\mathbf{y}) \propto \mathrm{P}(\mathbf{y}|\mathbf{X}, \mathbf{r})\mathrm{P}(\mathbf{X})\mathrm{P}(\mathbf{r}) \tag{7}$$

where, $\mathrm{P}(\mathbf{X})$ and $\mathrm{P}(\mathbf{r})$ denote prior distributions of $\mathbf{X}$ and $\mathbf{r}$, respectively. $\mathrm{P}(\mathbf{X})$ and $\mathrm{P}(\mathbf{r})$ are given as follows:

$$\mathrm{P}(\mathbf{X}) = \prod_{k=1}^{K} \mathrm{P}(\bar{\mathbf{x}}_k) = \prod_{k=1}^{K} \mathcal{N}\left(\bar{\mathbf{x}}_k \middle| \boldsymbol{\mu}'_k, \boldsymbol{\Sigma}_k^{-1}\right) \tag{8}$$

$$\mathrm{P}(\mathbf{r}) = \mathcal{N}(\mathbf{r}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{9}$$

where, $\alpha$ denotes a hyperparameter. In prior distribution $\mathrm{P}(\mathbf{X})$, $\mathcal{N}(\bar{\mathbf{x}}_k|\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}_k^{-1})$ is the same as Equation (4). In other words, $\boldsymbol{\mu}'_k = \frac{1}{N_k}\sum_j \mu_{w_j}\mathbf{x}_j$ , and $\boldsymbol{\Sigma}_k^{-1} = \frac{1}{N_k}\sum_j \sigma_{w_j}^2 \operatorname{diag}(\mathbf{x}_j \odot \mathbf{x}_j)$ . Instead of directly maximizing the posterior distribution, we employed iterative strategy that estimates $\mathbf{r}$ and $\bar{\mathbf{x}}_k$ that maximizes $\mathrm{P}(\mathbf{r}|\mathbf{y}, \mathbf{X})$ and $\mathrm{P}(\bar{\mathbf{x}}_k|\mathbf{y}, \mathbf{r}, \{\bar{\mathbf{x}}_l\}_{l\neq k})$ for each $\bar{\mathbf{x}}_k$.

For $\bar{\mathbf{x}}_k$ estimate, we employed the following probability equation:

$$\mathrm{P}(\bar{\mathbf{x}}_k|\mathbf{y}, \mathbf{r}, \{\bar{\mathbf{x}}_l\}_{l\neq k}) \propto \mathrm{P}(\mathbf{y}|\mathbf{X}, \mathbf{r})\mathrm{P}(\bar{\mathbf{x}}_k)\,. \tag{10}$$

Then, $\mathrm{P}(\bar{\mathbf{x}}_k|\mathbf{y}, \mathbf{r}, \{\bar{\mathbf{x}}_l\}_{l\neq k})$ follows Gaussian distribution and its mean and variance can be calculated by the following equations:

$$\mathbb{E}[\bar{\mathbf{x}}_k] = \left(\beta r_k^2\mathbf{I} + \boldsymbol{\Sigma}_k^{-1}\right)^{-1}\left(\beta r_k\mathbf{y} + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k^{'} - \beta\sum_{l\neq k} r_k r_l \bar{\mathbf{x}}_l\right) \tag{11}$$

$$\mathrm{Var}[\bar{\mathbf{x}}_k] = \beta r_k^2\mathbf{I} + \boldsymbol{\Sigma}_k^{-1}. \tag{12}$$

In the resulting $\mathbb{E}[\bar{\mathbf{x}}_k]$, all negative values were set to zero.

For $\boldsymbol{r}$ estimate, we employed the following probability equation:

$$\mathrm{P}(\mathbf{r}|\mathbf{y}, \mathbf{X}) \propto \mathrm{P}(\mathbf{y}|\mathbf{X}, \mathbf{r})\mathrm{P}(\mathbf{r}). \tag{13}$$

Then, $\mathrm{P}(\mathbf{r}|\mathbf{y}, \mathbf{X})$ follows Gaussian distribution and its mean can be calculated by the following equation:

499 $$\mathbb{E}[\mathbf{r}] = (\alpha\mathbf{I} + \beta\mathbf{X}^t\mathbf{X})^{-1}\beta\mathbf{X}^t\mathbf{y}. \quad (14)$$

500 For both resulting $\mathbb{E}[\mathbf{r}]$ and as $\mathbb{E}[\bar{\mathbf{x}}_k]$, all negative values were set to zero.

501 To estimate the gene expression pattern ($\{\bar{\mathbf{x}}_k^*\}_{k=1}^K$) and cell-type proportion ($\mathbf{r}^*$), we repeatedly
502 performed the computation of $\mathbf{X}$ and $\mathbf{r}$ alternately until both converge or, otherwise, the 1,001 iterations
503 were conducted. The convergence here is defined as the MSE of the iterations $t$ and $t-1$ is less than
504 $1e^{-10}$ for both $\mathbf{X}$ and $\mathbf{r}$. As initial $\mathbf{r}$ and $\mathbf{X}$ used the cell-type ratio and the mean of weighted counts of
505 the reference datasets. The hyperparameters $\alpha$ and $\beta$ were set as $10^{-3}, 10^{-2}, \ldots, 10^3$ for convenience.
506 The results with the combination of the number of signature gene-sets (100, 300, 2,000/1,577/1,461)
507 hyperparameters ($\alpha$ and $\beta$) that produced high similarity (based on the high Pearson and Spearman
508 correlation coefficients and low mean squared errors) to the real whole-organ RNAs-seq are selected as
509 the best predicted results and shown in each Figure.

510

511 **Virtual t-SNE analysis**

512 The virtual transcriptome data of each mouse organ was analyzed by t-SNE analyses. By setting the
513 total sampling size as 3,000, the number of the sampled cells for each cell type and the virtual single-
514 cell transcript counts for a cell-type k follow $\mathrm{P}^*(\mathbf{r})$ and $\mathrm{P}^*(\bar{\mathbf{x}}_k)$, respectively. This sampling process
515 was repeated until the total sampling size of 3,000 is achieved. Then, we applied t-SNE to these sampled
516 virtual scRNA-seq count data using R package 'Rtsne' with parameter perplexity = 50.

517

518 **Disease model computation**

519 For the disease model computation, we first calculated weighting-factors using the whole-organ RNA-
520 seq data of sham hearts (Kozawa et al., 2018) using the same reference cell-type ratio of the normal
521 mouse for each stage (E, M, L). We then computed the cell-type ratios and gene expression profiles for
522 each stage as described above using the whole heart RNA-seq of the sham/MI models (Kozawa et al.,
523 2018).

524

525 **Human bulk-tissue deconvolution**

526 The human heart and kidney RNA-seq data are from "The Human Protein Atlas"
527 (https://www.proteinatlas.org/). For their deconvolution, we used the cell type-dependent weighting-
528 factors for the corresponding mouse organ. The total counts of all genes of the human bulk-tissue RNA-
529 seq data were first normalized to 100. We then matched the mouse gene symbols to the ones of the
530 human by extracting genes with common names between mouse and human. This resulted in a total of
531 16,144 genes. Using these mouse-human common gene sets, we performed the computation procedures
532 as described for the mouse datasets, hence the cell type annotations of the virtual human data are
533 according to those of the mouse data. Since there no exactly paired scRNA-seq data are available for
534 the human bulk-tissue data, the virtual cell-type ratios are compared to the references used for the mouse
535 study in Figures 1 and 3. For the gene expression pattern comparisons, as no paired scRNA-seq data
536 are available, we used the scRNA-seq data derived from pieces of tissues excised from the human left
537 ventricle (LV) and the left auricle/left atrial appendages (LA) (Wang et al., 2020) (GSE109816) and a
538 small part of the human kidney (Liao et al., 2020) (GSE131685) for the heart and the kidney,
539 respectively. The quality control filtering for the heart count data is conducted according to the
540 publication, except that the alignment rate filtering was not performed since no information on this was
541 available in the published study. In addition, we used the human genes that are present in both this
542 human heart scRNA-seq and the human heart RNA-seq (derived from Human Protein Atlas). The
543 kidney data are exactly as published. The total counts for each cell-type are normalized to 100 and the
544 average counts for each gene were calculated from all samples (the heart = 12, the kidney = 3). For the
545 gene expression comparison, we selected the genes that are commonly detected in both the real and
546 virtual data for each of the matching cell-types, and then the total real and virtual counts are each
547 normalized to 100 for each matching cell-type. For the comparisons, the Pearson correlation, Spearman
548 correlation and mean squared error analyses are conducted. The results with the combination of the
549 number of signature gene-sets (Heart: 94, 275, 1,832; Kidney: 86, 262, 1,325) and the hyperparameters
550 ($\alpha$ and $\beta$) that produced high similarity (based on the high Pearson and Spearman correlation
551 coefficients and low MSEs) of the virtual to the real whole-organ RNAs-seq are selected as the best
552 predicted results and shown in Figure 6.
553

554 **Virtual differential gene expressing analysis**

555 The sum of the mean counts of all genes is normalized to a million. The count of each gene was rounded
556 off to integer and analyzed using R package, 'DESeq2 (version 1.24.0)'. The predicted counts per cell-
557 type was analyzed by comparing MI with sham at every stage and using 2 predicted count data-set in
558 every model (E,M,L-Sham/MI).

559

560 **MuSiC and DWLS**

561 MuSiC was conducted as described (Wang et al., 2019). For DWLS(Tsoucas et al., 2019), we replaced
562 solve.QP (R package: quadprog) by solve_osqp (R package: osqp) for quadratic problem solver.

563

564 **Datasets**

565 In this study, used are all publicly available datasets. Murine whole-organ RNA-seq data are from 11
566 weeks-old male C57BL6/N Jcl and derived from "i-organs.atr.jp" (Kozawa et al., 2020; Kozawa et al.,
567 2018). Human whole-organ RNA-seq data are from "The Human Protein Atlas"
568 (https://www.proteinatlas.org/; heart (ERR315328) and kidney (ERR315494)). The scRNA-seq data
569 are from *Tabula Muris* (aorta, brain, fat, heart, kidney, large intestine, liver, lung, marrow, pancreas,
570 skin, spleen, and thymus) and they are derived from 10-15 weeks-old male C57BL6/JN Mice (Tabula
571 Muris et al., 2018). The skeletal muscle scRNA-seq data are from 8-10 weeks-old female C57BL6/J
572 reported in "Mouse Cell Atlas"(Han et al., 2018). The adult human heart and kidney scRNA-seq data
573 are from GSE109816 and GSE131685, respectively.

574

575 **Data availability**

576 The public datasets are as indicated and the raw data are available as source data in this paper.

577

578 **Code availability**

579 The code for our algorithm is freely available (http://i-organs.atr.jp/Code_submit.zip).
580

581 **Acknowledgements**

582 We thank R. Takahashi, E. Kojima, K. Sugisaka for administrative assistance. We are also grateful to
583 the members of Sato lab at ATR and Karydo TherapeutiX, Inc. for advice and discussion throughout
584 the course of this work. This work was supported by funds from JST ERATO Grant Number
585 JPMJER1303 (T.N.S.) and Nakatani Foundation (T.N.S.).

586

587 **Author contributions**

588 T.N.S. conceived the idea of the project, designed the overall experiments and supervised the overall
589 research project. K.T. and S.K. performed all in silico analyses. T.N.S., K.T., S.K. wrote the manuscript.

590

591 **Completing interests**

592 T.N.S. is the inventor of the patents filed on the cell type-specific weighting-factors and holds shares
593 of Karydo TherapeutiX, Inc., which owns the right to this patent. K.T. and S.K. are partly employed by
594 Karydo TherapeutiX, Inc. Correspondence and requests for materials should be addressed to T.N.S.
595 (island1005@gmail.com).

596

597 **Additional files**

598 **Supplement file 1.** Information on the whole-organ/bulk-tissue RNA-seq and scRNA-seq datasets.

599

600

601 **References**

602 Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H.F. (2009). Deconvolution of
603 blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PloS one
604 *4*, e6098.

605 Anversa, P., Olivetti, G., Melissari, M., and Loud, A.V. (1980). Stereological measurement of cellular
606 and subcellular hypertrophy and hyperplasia in the papillary muscle of adult rat. Journal of molecular
607 and cellular cardiology *12*, 781-795.

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

610 Banerjee, I., Fuseler, J.W., Price, R.L., Borg, T.K., and Baudino, T.A. (2007). Determination of cell
611 types and numbers during cardiac development in the neonatal and adult rat and mouse. American
612 journal of physiology Heart and circulatory physiology *293*, H1883-1891.

613 Bergmann, O., Zdunek, S., Felker, A., Salehpour, M., Alkass, K., Bernard, S., Sjostrom, S.L.,
614 Szewczykowska, M., Jackowska, T., Dos Remedios, C.*, et al.* (2015). Dynamics of Cell Generation and
615 Turnover in the Human Heart. Cell *161*, 1566-1575.

616 Cleutjens, J.P., Verluyten, M.J., Smiths, J.F., and Daemen, M.J. (1995). Collagen remodeling after
617 myocardial infarction in the rat heart. The American journal of pathology *147*, 325-338.

618 Dammrich, J., and Pfeifer, U. (1983). Cardiac hypertrophy in rats after supravalvular aortic constriction.
619 I. Size and number of cardiomyocytes, endothelial and interstitial cells. Virchows Arch B Cell Pathol
620 Incl Mol Pathol *43*, 265-286.

621 Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic,
622 random monoallelic gene expression in mammalian cells. Science *343*, 193-196.

623 Fraccarollo, D., Galuppo, P., Bauersachs, J., and Ertl, G. (2002). Collagen accumulation after
624 myocardial infarction: effects of ETA receptor blockade and implications for early remodeling.
625 Cardiovascular research *54*, 559-567.

626 Gaujoux, R., and Seoighe, C. (2012). Semi-supervised Nonnegative Matrix Factorization for gene
627 expression deconvolution: a case study. Infect Genet Evol *12*, 913-921.

628 Gong, T., Hartmann, N., Kohane, I.S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., and
629 Szustakowski, J.D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic
630 programming with application to complex clinical blood samples. PloS one *6*, e27156.
631 Gong, T., and Szustakowski, J.D. (2013). DeconRNASeq: a statistical framework for deconvolution of
632 heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics *29*, 1083-1085.
633 Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F.*, et al.*
634 (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091-1107 e1017.
635 Kang, K., Meng, Q., Shats, I., Umbach, D.M., Li, M., Li, Y., Li, X., and Li, L. (2019). CDSeq: A novel
636 complete deconvolution method for dissecting heterogeneous samples using gene expression data.
637 PLoS computational biology *15*, e1007510.
638 Kozawa, S., Sagawa, F., Endo, S., De Almeida, G.M., Mitsuishi, Y., and Sato, T.N. (2020). Predicting
639 Human Clinical Outcomes Using Mouse Multi-Organ Transcriptome. iScience *23*, 100791.
640 Kozawa, S., Sagawa, F., Endo, S., Maria de Almeida, G., Mitsuishi, Y., and Sato, T.N. (2019).
641 Predicting Human Clinical Outcomes using Mouse Multi-Organ Transcriptome. iScience *In press*.
642 Kozawa, S., Ueda, R., Urayama, K., Sagawa, F., Endo, S., Shiizaki, K., Kurosu, H., Maria de Almeida,
643 G., Hasan, S.M., Nakazato, K.*, et al.* (2018). The Body-wide Transcriptome Landscape of Disease
644 Models. iScience *2*, 238-268.
645 Leask, A. (2010). Potential therapeutic targets for cardiac fibrosis: TGFbeta, angiotensin, endothelin,
646 CCN2, and PDGF, partners in fibroblast activation. Circulation research *106*, 1675-1680.
647 Li, B., Severson, E., Pignon, J.C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J.C., Rodig, S.*,*
648 *et al.* (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy.
649 Genome biology *17*, 174.
650 Liao, J., Yu, Z., Chen, Y., Bao, M., Zou, C., Zhang, H., Liu, D., Li, T., Zhang, Q., Li, J.*, et al.* (2020).
651 Single-cell RNA sequencing of human kidney. Sci Data *7*, 4.
652 Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J.
653 (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing.
654 Genome Res *24*, 496-510.

655 Mattfeldt, T., Kramer, K.L., Zeitz, R., and Mall, G. (1986). Stereology of myocardial hypertrophy
656 induced by physical exercise. Virchows Arch A Pathol Anat Histopathol *409*, 473-484.
657 McCurdy, S.M., Dai, Q., Zhang, J., Zamilpa, R., Ramirez, T.A., Dayah, T., Nguyen, N., Jin, Y.F.,
658 Bradshaw, A.D., and Lindsey, M.L. (2011). SPARC mediates early extracellular matrix remodeling
659 following myocardial infarction. American journal of physiology Heart and circulatory physiology *301*,
660 H497-505.
661 Murakoshi, M., Saiki, K., Urayama, K., and Sato, T.N. (2013). An anthelmintic drug, pyrvinium
662 pamoate, thwarts fibrosis and ameliorates myocardial contractile dysfunction in a mouse model of
663 myocardial infarction. PloS one *8*, e79374.
664 Nag, A.C. (1980). Study of non-muscle cells of the adult mammalian heart: a fine structural analysis
665 and distribution. Cytobios *28*, 41-61.
666 Nakao, K., Minobe, W., Roden, R., Bristow, M.R., and Leinwand, L.A. (1997). Myosin heavy chain
667 gene expression in human heart failure. The Journal of clinical investigation *100*, 2362-2370.
668 Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and
669 Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature
670 methods *12*, 453-457.
671 Pinto, A.R., Ilinykh, A., Ivey, M.J., Kuwabara, J.T., D'Antoni, M.L., Debuque, R., Chandran, A., Wang,
672 L., Arora, K., Rosenthal, N.A.*, et al.* (2016). Revisiting Cardiac Cellular Composition. Circulation
673 research *118*, 400-409.
674 Raulf, A., Horder, H., Tarnawski, L., Geisen, C., Ottersbach, A., Roll, W., Jovinge, S., Fleischmann,
675 B.K., and Hesse, M. (2015). Transgenic systems for unequivocal identification of cardiac myocyte
676 nuclei and analysis of cardiomyocyte cell cycle status. Basic Res Cardiol *110*, 33.
677 Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell,
678 P.J., Carninci, P., Clatworthy, M.*, et al.* (2017). Science Forum: The Human Cell Atlas. Elife *6*.
679 Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G.F., Selbig, J., Parida, S.K., Kaufmann, S.H., and
680 Jacobsen, M. (2010). Biomarker discovery in heterogeneous tissue samples -taking the in-silico
681 deconfounding approach. BMC bioinformatics *11*, 27.

682 Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. Nature
683 methods *11*, 22-24.
684 Schirone, L., Forte, M., Palmerio, S., Yee, D., Nocella, C., Angelini, F., Pagano, F., Schiavon, S.,
685 Bordin, A., Carrizzo, A.*, et al.* (2017). A Review of the Molecular Mechanisms Underlying the
686 Development and Progression of Cardiac Remodeling. Oxid Med Cell Longev *2017*, 3920195.
687 Schmidt, E.E., and Schibler, U. (1995). Cell size regulation, a mechanism that controls cellular RNA
688 accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-Y and
689 the liver-enriched transcription factor DBP. J Cell Biol *128*, 467-483.
690 Sergeeva, I.A., and Christoffels, V.M. (2013). Regulation of expression of atrial and brain natriuretic
691 peptide, biomarkers for heart development and disease. Biochimica et biophysica acta *1832*, 2403-2413.
692 Shah, M.S., and Brownlee, M. (2016). Molecular and Cellular Mechanisms of Cardiovascular Disorders
693 in Diabetes. Circulation research *118*, 1808-1829.
694 Shen-Orr, S.S., and Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific
695 information from heterogeneous samples. Curr Opin Immunol *25*, 571-578.
696 Tabula Muris, C., Overall, c., Logistical, c., Organ, c., processing, Library, p., sequencing,
697 Computational data, a., Cell type, a., Writing, g.*, et al.* (2018). Single-cell transcriptomics of 20 mouse
698 organs creates a Tabula Muris. Nature *562*, 367-372.
699 Tang, Y., Nyengaard, J.R., Andersen, J.B., Baandrup, U., and Gundersen, H.J. (2009). The application
700 of stereological methods for estimating structural parameters in the human heart. Anat Rec (Hoboken)
701 *292*, 1630-1647.
702 Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G.C. (2019). Accurate estimation of cell-
703 type composition from gene expression data. Nat Commun *10*, 2975.
704 van den Borne, S.W., Diez, J., Blankesteijn, W.M., Verjans, J., Hofstra, L., and Narula, J. (2010).
705 Myocardial remodeling after infarction: the role of myofibroblasts. Nature reviews Cardiology *7*, 30-
706 37.
707 van den Brink, S.C., Sage, F., Vertesy, A., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C.,
708 and van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression
709 in tissue subpopulations. Nature methods *14*, 935-936.

710 Walsh, S., Ponten, A., Fleischmann, B.K., and Jovinge, S. (2010). Cardiomyocyte cell cycle control
711 and growth estimation in vivo--an analysis based on cardiomyocyte nuclei. Cardiovascular research *86*,
712 365-373.
713 Wang, L., Yu, P., Zhou, B., Song, J., Li, Z., Zhang, M., Guo, G., Wang, Y., Chen, X., Han, L.*, et al.*
714 (2020). Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the
715 cellular landscape underlying cardiac function. Nature cell biology *22*, 108-119.
716 Wang, N., Gong, T., Clarke, R., Chen, L., Shih Ie, M., Zhang, Z., Levine, D.A., Xuan, J., and Wang,
717 Y. (2015). UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene
718 expressions in tumor samples. Bioinformatics *31*, 137-139.
719 Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution
720 with multi-subject single-cell expression reference. Nat Commun *10*, 380.
721 Zaitsev, K., Bambouskova, M., Swain, A., and Artyomov, M.N. (2019). Complete deconvolution of
722 cellular mixtures based on linearity of transcriptional signatures. Nat Commun *10*, 2209.
723 Zhong, Y., Wan, Y.W., Pang, K., Chow, L.M., and Liu, Z. (2013). Digital sorting of complex tissues
724 for cell type-specific gene expression profiles. BMC bioinformatics *14*, 89.
725 Zhou, P., and Pu, W.T. (2016). Recounting Cardiac Cellular Composition. Circulation research *118*,
726 368-370.
727
728

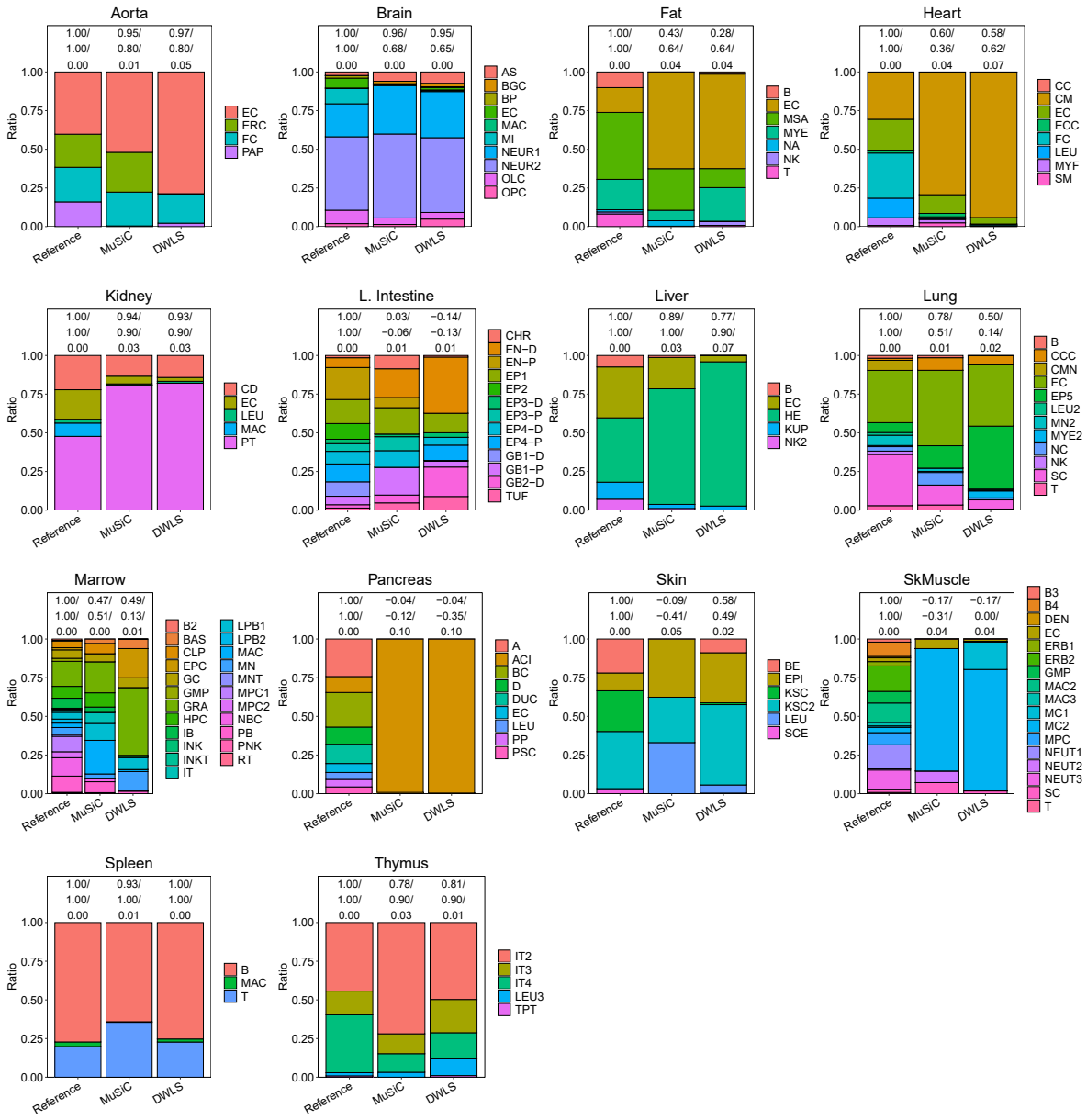


**Figure 1. Estimation of the composite cell type-ratios of 14 mouse organs by MuSiC and DWLS.** The whole-organ RNA-seq of 14 real mouse organs were deconvolved by MuSiC and DWLS and the computed ratios are shown. The reference ratios are as described in Materials and methods. Pearson and Spearman correlation coefficients and mean squared errors of the predicted cell type compositions in comparison to the reference compositions are indicated at the top of each bar in this order from the top to the bottom. The cell type annotations (color coded on the right) are indicated by abbreviations and their full annotations are available in Figure 1 – source data 1.

**Figure 1 – Source data 1.** Cell type abbreviations, cell-type ratios, comparison between methods.

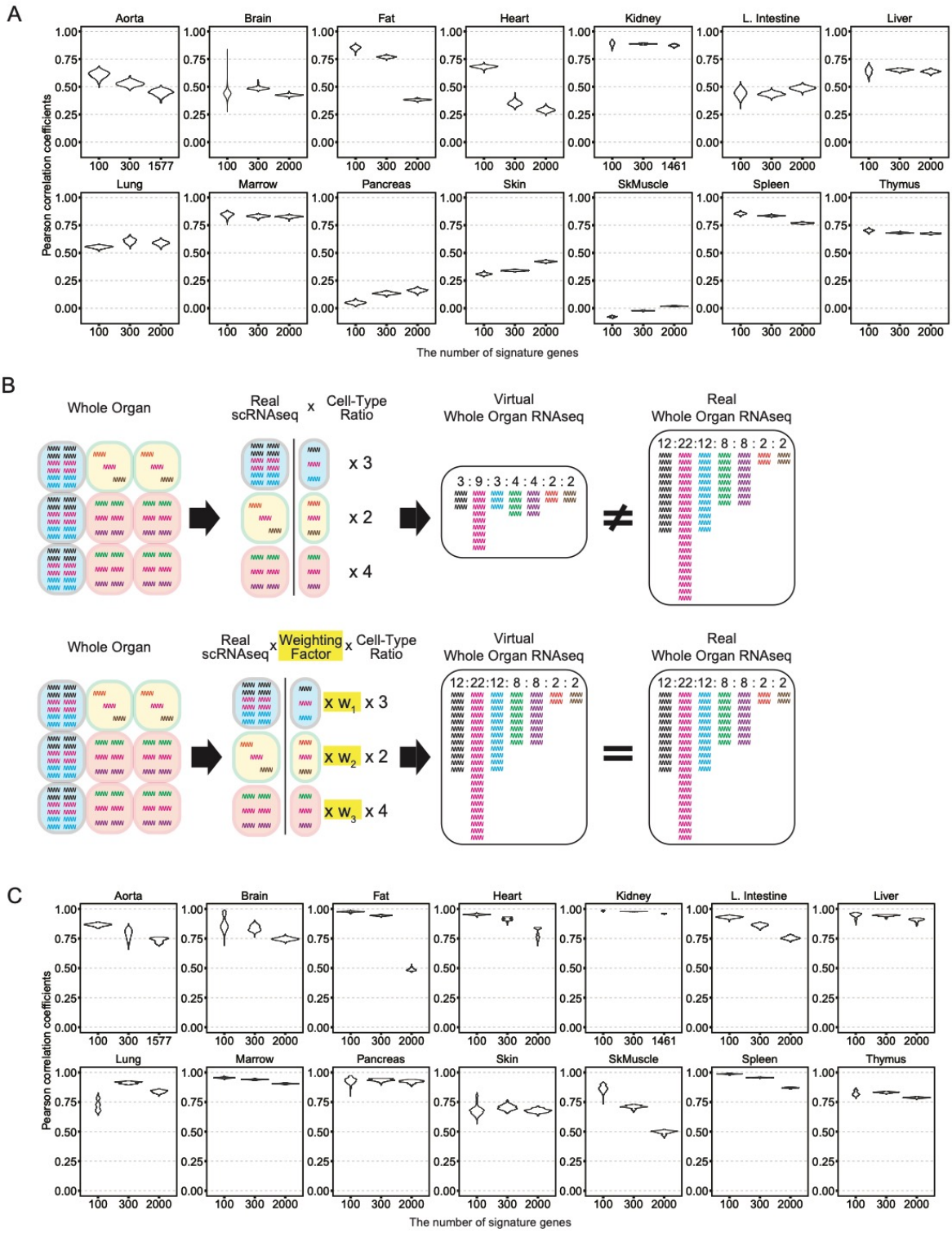


**Figure 2. Synthetic whole-organ RNA-seq by the cell type-specific weighting-factors.** **(A)** Sum of scRNA-seq transcriptomes of the composite cell-types weighted by their ratios. The x-axis: The number of the signature gene sets used for the calculation. The y-axis: Pearson correlation coefficients. **(B)** The Concept of "the cell type-specific weighting-factor". Shown is the case of an organ consisting of three cell types with 7 different genes (wiggling lines of different colors). The concept of "the cell type-specific weighting-factors (the lower panel) is compared to that of the existing other

747 methods (the upper raw) such as MuSiC, DWLS and others. In the existing other methods, it is
748 commonly assumed that the whole-organ (labeled as "Whole Organ") is the sum of the normalized
749 transcript counts of each cell-type (i.e., the total counts for each cell is the same) (labeled as "Real
750 scRNAseq") weighted by the cell-type ratios (labeled as "Cell Type Ratio"). Hence, the synthetic
751 whole-organ RNA-seq data (labeled as "Virtual Whole Organ RNAseq") differ significantly from the
752 real whole-organ RNA-seq data (labeled as "Real Whole Organ RNAseq"). This difference is fixed by
753 the cell type-specific weighting-factors ($w_1$, $w_2$, $w_3$, etc.) (labeled as "Weighting Factors") in our
754 method (the lower panel). Although we compute at least 100 weighting-factors for each "virtual-
755 cell/cell-subject" and calculated their mean and variance (Figure 2 – Figure supplement 1), the diagram
756 shows the mean weighting-factor value ($w_1$, $w_2$, $w_3$, etc.) for each cell-type. **(C)** The improved synthetic
757 whole-organ RNA-seq by the cell type-specific weighting-factors. The x-axis: The number of the
758 signature gene sets used for the calculation. The y-axis: Pearson correlation coefficients.
759
760 **Figure 2 – Source data 1.** Pearson correlation coefficients between synthetic and real whole-organ
761 RNA-seq.
762 **Figure 2 – Figure supplement 1.** The computed cell type-specific weighting-factors.
763 **Figure 2 – Figure supplement 1 – source data 1.** The number of all signature genes, the list of all
764 signature genes and their Mean Decrease in Gini, the weighting-factors for each cell-type and their
765 mean and variance.
766

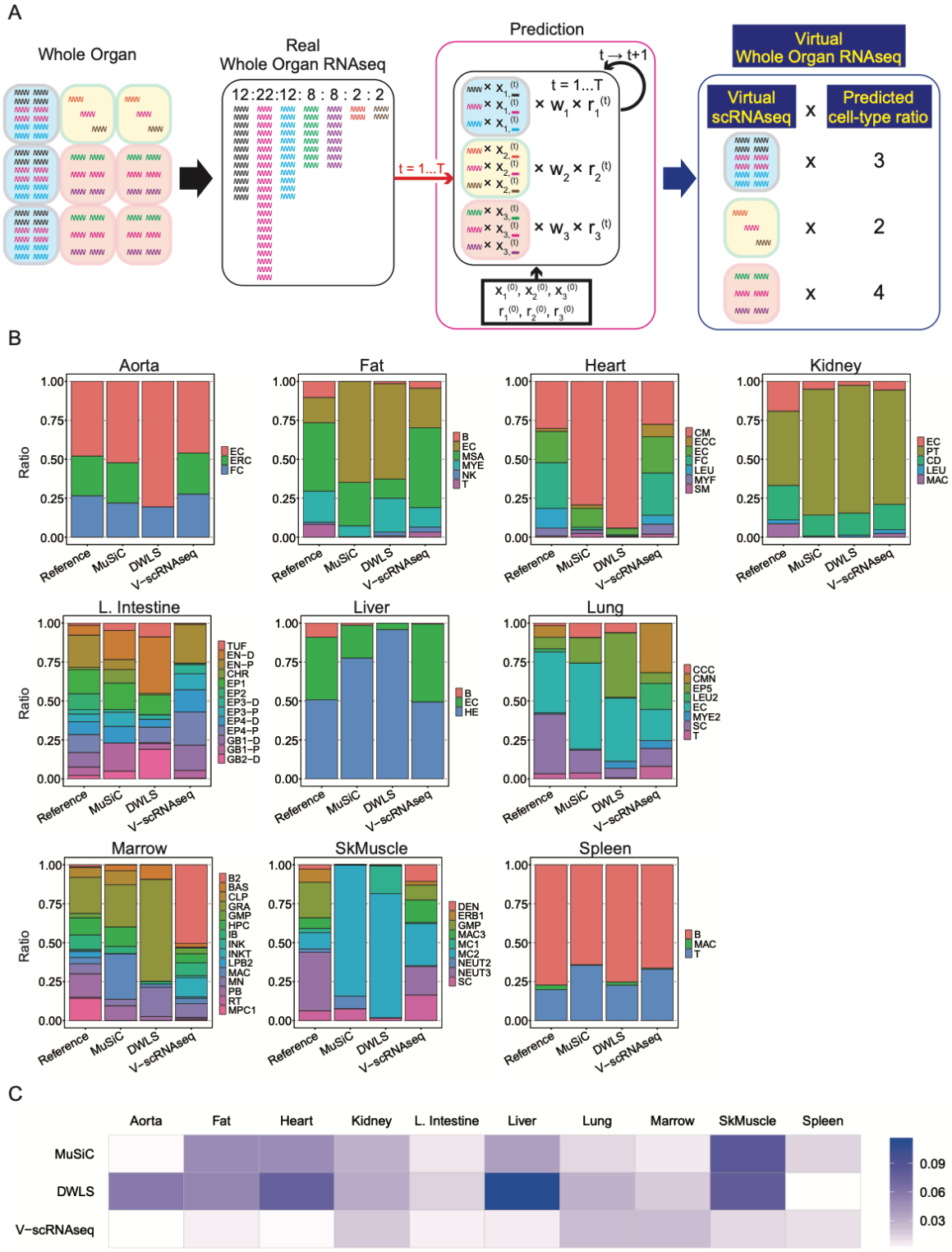


**Figure 3. The deconvolution framework and the computation of the cell type-ratios of 10 mouse organs.**

**(A)** The schematic diagram describing the concept of our deconvolution method. Shown is the case of an organ consisting of 3 cell-types (1, 2, 3) with 7 different genes (wiggling lines of different colors). The mean of the cell type-specific weighting-factors: $w_1$, $w_2$, $w_3$ for each cell-type 1, 2 and 3,

773 respectively. The virtual transcript counts ($x_1^{(t)}$, $x_2^{(t)}$, $x_3^{(t)}$ where the different genes are indicated by the
774 wiggling lines of different colors for each cell-type) and the ratio ($r_1^{(t)}$, $r_2^{(t)}$, $r_3^{(t)}$) of virtual cell-type 1, 2
775 and 3 at each iteration, t, are as indicated. Their initial conditions are $x_1^{(0)}$, $x_2^{(0)}$, $x_3^{(0)}$, $r_1^{(0)}$, $r_2^{(0)}$, $r_3^{(0)}$. The
776 iteration (t= 1, …, T) is performed until both **X** and **r** converge or, otherwise, the 1,001 iterations were
777 conducted (see Materials and methods). The convergence here is defined as the MSE of the iterations $t$
778 and $t-1$ is less than $1e^{-10}$ for both **X** and **r** (see Materials and methods). Upon completing the
779 iterations, the sum of the virtual transcript counts weighted by the cell type-specific weighting-factors
780 (indicated as "Virtual scRNAseq") weighted by the virtual cell-type ratios (indicated as "Predicted cell-
781 type ratio) generate the virtual whole-organ transcriptome (indicated as "Virtual Whole Organ
782 RNAseq", as shown at the far right of this panel. **(B)** Bar graph showing the cell type-ratios computed
783 by our method (V-scRNAseq) for each organ. The reference cell-type ratios and the ratios estimated by
784 MuSiC and DWLS are show as comparisons. The bar graphs are composed of the cell-types computed
785 to be present for each organ by our method. Those that are not detected by our method are omitted. **(C)**
786 The quantitative similarity of the computed ratios to the reference ratios are shown by heatmap of MSEs
787 for MuSiC, DWLS and our method (V-scRNAseq). MSEs: Mean Squared Errors.
788
789 **Figure 3 – Source data 1.** Cell-type ratios of the reference and those computed by each method,
790 MSE/Pearson/Spearman of the cell-type ratio comparison between the reference and each method,
791 MSE/Pearson/Spearman of the cell-type ratio and the transcriptome comparisons between the reference
792 and our method for each organ.
793

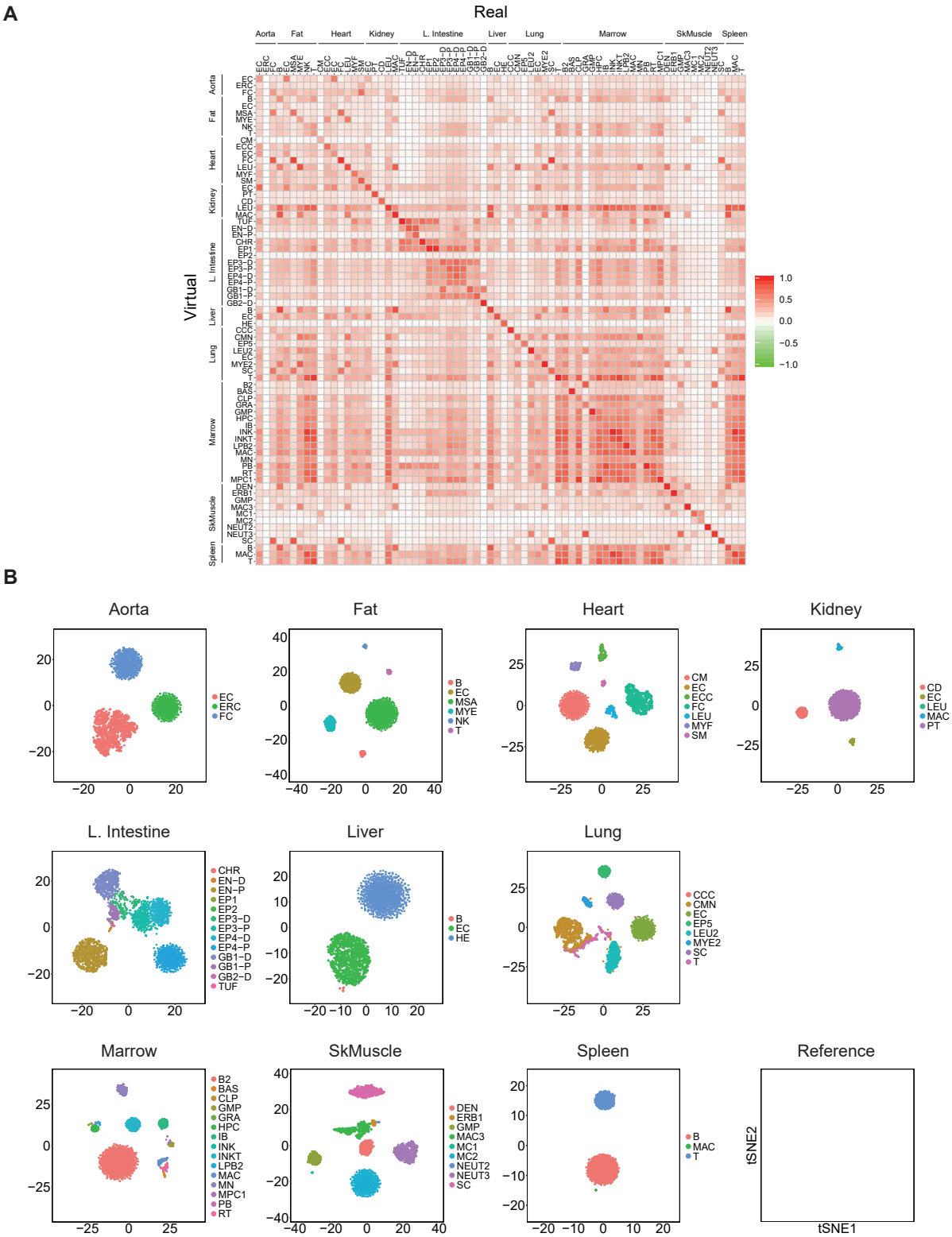


**Figure 4. Computation of the complete transcriptome of the composite cell-types of 10 mouse organs.**

**(A)** The heatmap of Pearson correlation coefficients for the transcriptome of all virtual and real cell-types across all 10 organs. The Pearson correlation coefficients are calculated for 23,131 (all organs, except the skeletal muscle) and 14,323 (the skeletal muscle) genes. The results with the same

800 combination of the hyperparameters ($\alpha$ and $\beta$) used for Figure 3B are shown. **(B)** The t-SNE for all
801 virtual organs. The labels for x- and y-axes are indicated as Reference at the bottom right.
802
803 **Figure 4 – Source data 1.** Mean and variance of virtual count data for all genes for each cell-type of
804 all organs.
805

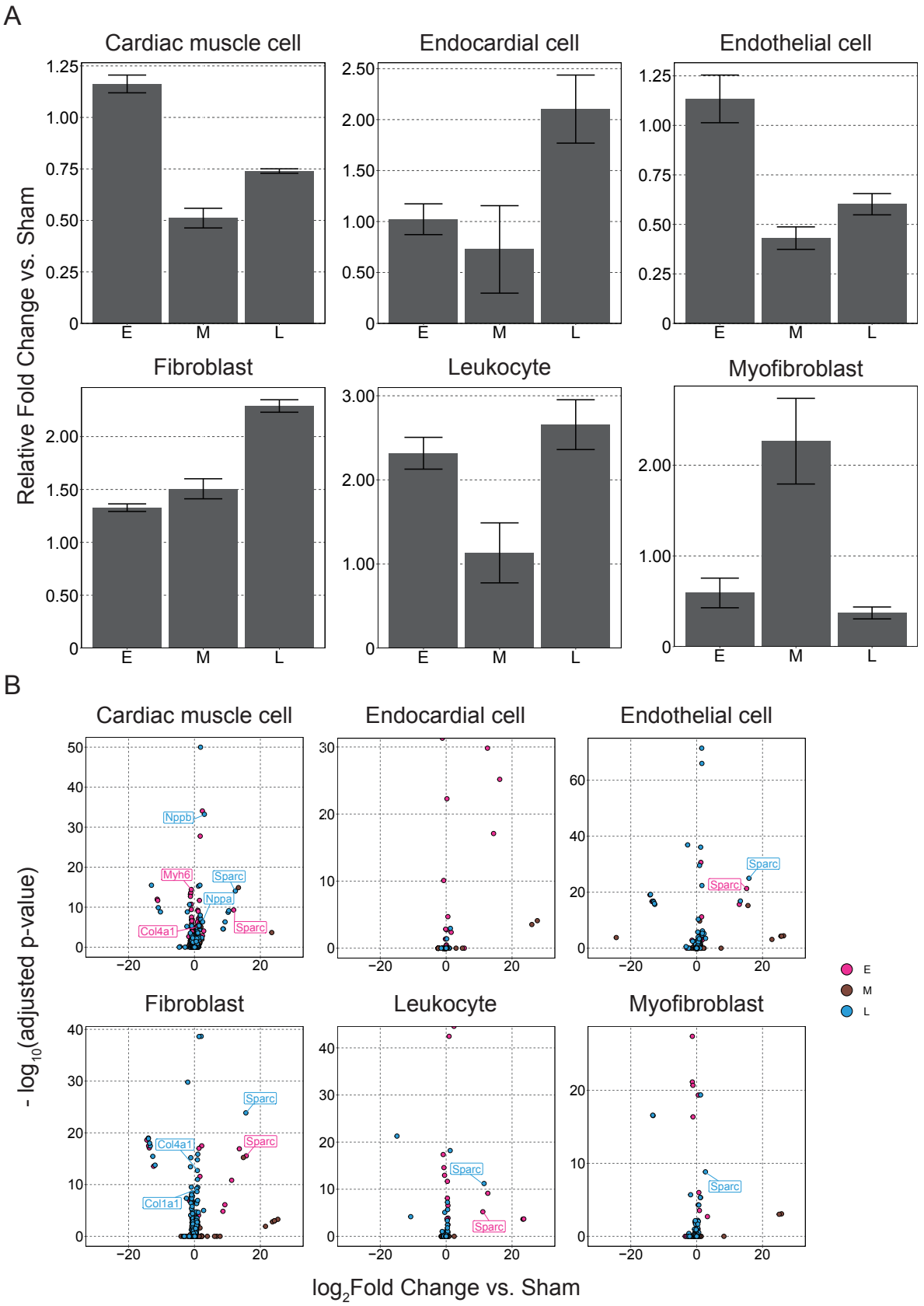


**Figure 5. Computation of both ratios and transcriptome changes of the composite cell types in the heart of a mouse model for myocardial infarction.**

810 **(A)** The estimated cell-type ratios in the MI heart relative to those of sham control heart at each stage
811 (E, M, L) are shown as bar graph. They are shown as mean and standard error bars. Deconvolution
812 results using sham (n=2) and MI (n=2 at each stage) are compared in all four combinations and plotted.
813 **(B)** The virtual differential gene expression for each cell type is shown as volcano plots. Sham (n=2)
814 and MI (n=2 at each stage). Each stage is color coded (shown on the right). The experimentally known
815 differentially expressed genes that are also identified by this analysis are indicated with gene
816 names/labels in each plot. The gene names/labels are also color coded according to the stage color codes.
817
818 **Figure 5 – Source data 1.** Weighting-factors, Pearson/Spearman/MSE of the cell-type ratio and the
819 transcriptome comparisons between the reference and our method with the selected hyperparameters,
820 the ratio of each cell-type, mean/SE of the cell-type ratio comparison (MI vs. Sham) for each cell-type.
821 **Figure 5 – Source data 2.** DESeq, the mean and variance of the counts for each cell-type.
822

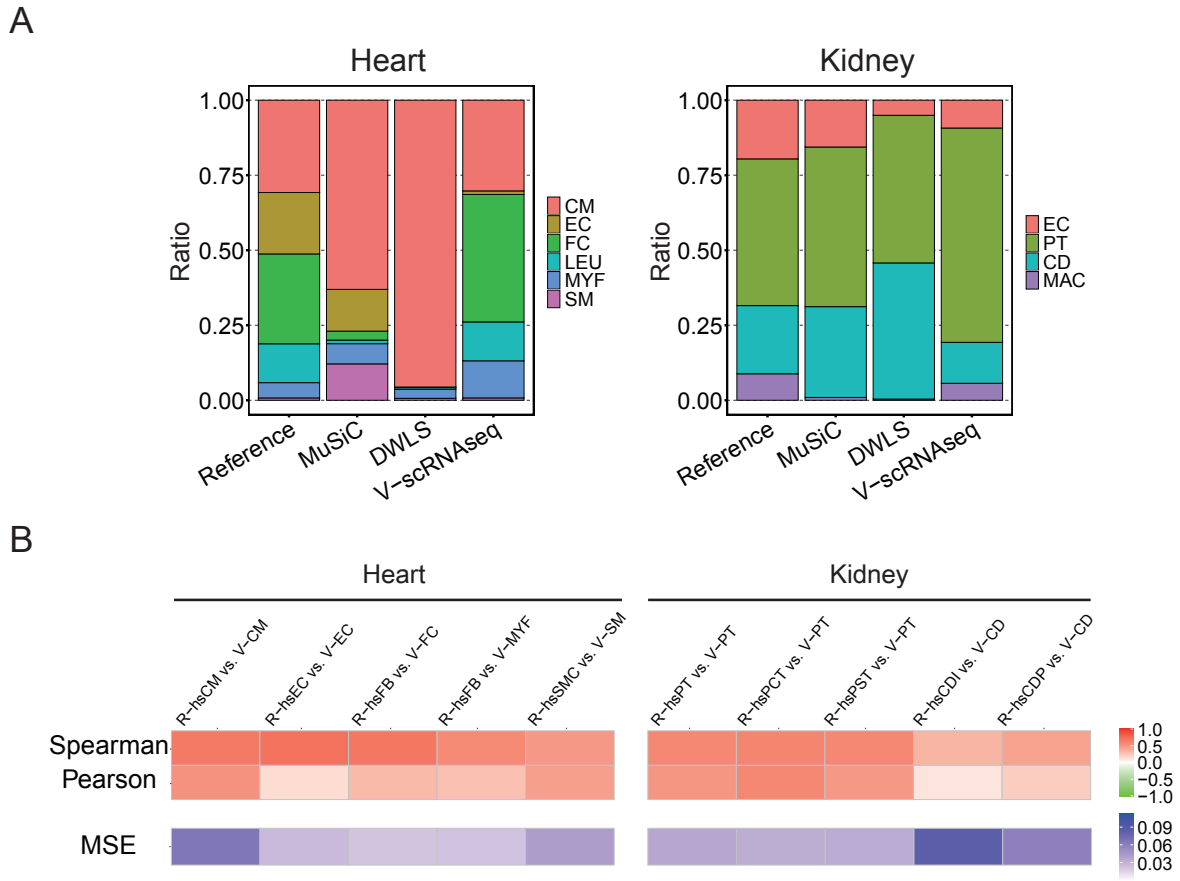


**Figure 6. Cell-type and transcriptome deconvolution of the human heart and kidney.** **(A)** The estimated cell type ratios of the human heart and kidney. Mouse weighting-factors are used to calculate the cell-type ratios using human bulk-tissue RNA-seq data. The cell type annotations are indicated by abbreviations at the bottom for each organ and they are the same as those of the mouse. **(B)** The heatmap showing the Spearman correlation coefficients, Pearson correlation coefficients and MSE (Mean Squared Errors) between virtual and real transcriptome of each cell type.

**Figure 6 – Source data 1.** Cell-type ratios computed by the reference and all methods, MSE/Pearson/Spearman of the cell-type ratio and the transcriptome comparisons between the reference and our method for each organ, the mean and variance of the counts of all genes for each cell-type, Spearman/Pearson/MSE of the transcriptome comparisons between the related cell-types, virtual vs. real counts of all genes for each cell type comparison.

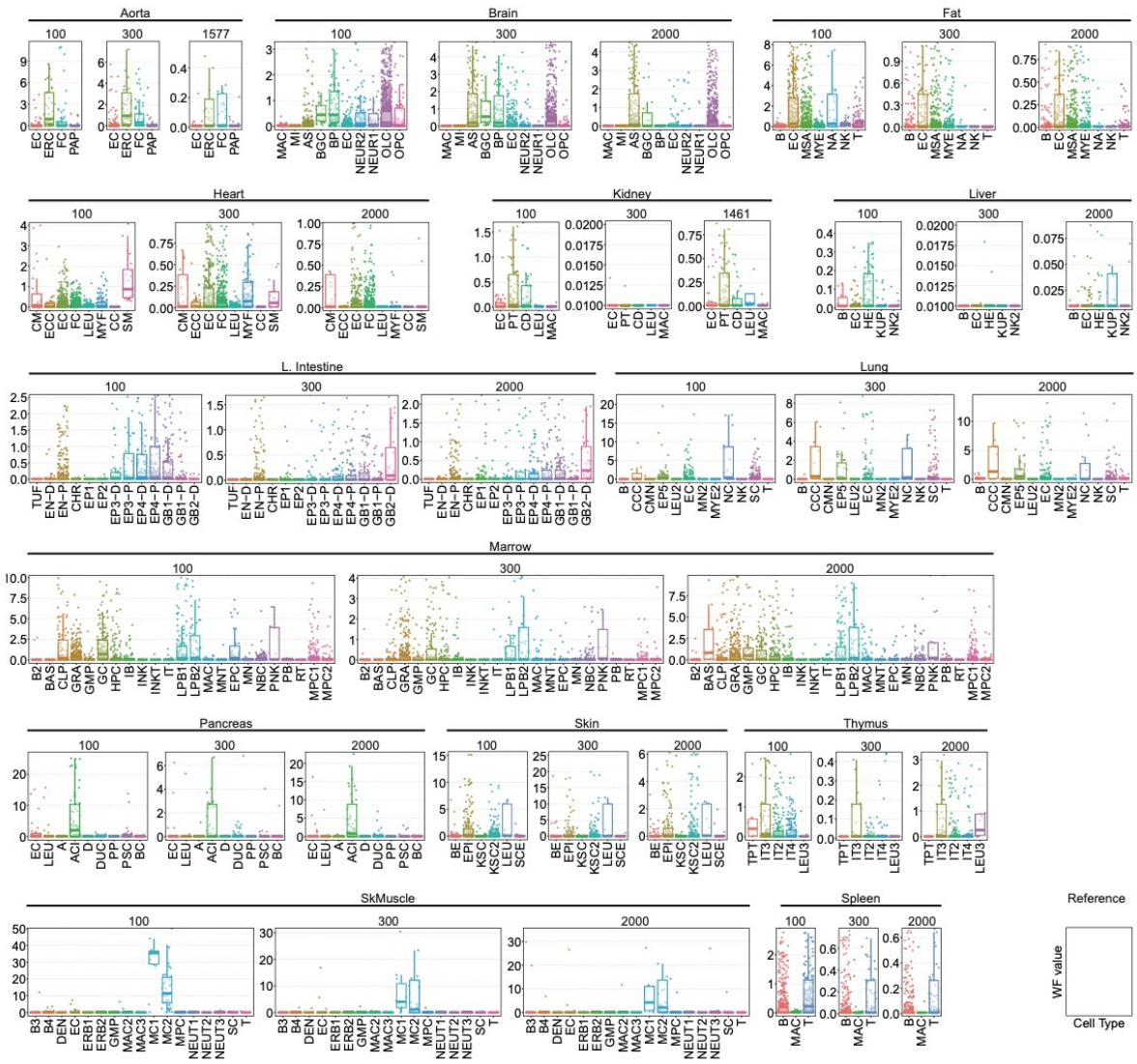


**Figure Supplement 1. The computed cell type-specific weighting-factors**

Box and scatter plots showing the weighting-factors for each cell-type in each-organ. Shown are the means of individual virtual-cells/cell-subjects for each cell-type in each organ. The results with 100, 300, 2,000 (1,577 for the aorta, 1,461 for the kidney) cell type-specific signature gene sets (indicated at the top of each box-plot) are shown. The scale of the y-axis is selected to show all boxplots for each cell-type. The labels on x and y axes are described at the reference plot at the right bottom. WF: cell-type specific weighting-factors.

**Figure 2 – Figure supplement 1 – source data 1.** The number of all signature genes, the list of all signature genes and their Mean Decrease in Gini, the weighting-factors for each cell-type and their mean and variance.