Body-wide landscape of cell type-specific total cellular transcriptome-size and its physiological significance

Short Title:

Cell type-specific transcriptome-size

Kengo Tejima*, Satoshi Kozawa*, Thomas N. Sato¶

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

¶Corresponding Authors:

Thomas N. Sato

Karydo TherapeutiX, Inc.

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

Email: island1005@gmail.com

*These two authors contributed equally to this work

Abstract

Total cellular transcriptome-size varies from cell-to-cell. However, the degree of such variations among diverse cell-types and their function in the body remains elusive. Moreover, such differential transcriptome-size is mostly ignored in the widely used single-cell analyses, veiling its functional importance. To overcome such limitations, currently available experimental approaches are insufficient. Hence, we developed and report a computational method to determine its body-wide variations. Using this method, we compute weighting-factors that account for the differences of the total cellular transcriptome-size for 76 cell-types across 10 organs. We demonstrate their validity by accurate reconstitution and deconvolution of the whole-organ transcriptomics using their composite single-cell data weighted by these factors. Importantly, we also show such weighting-factors indeed recapitulate experimentally-determined differential total cellular transcriptome-size. Furthermore, we show that such body-wide variations accurately predict aging-progression. These results unveil the body-wide landscape of the cell type-specific total cellular transcriptome-size and their physiological importance.

Teaser

Differential mRNA contents of 76 cell-types across 10 organs and their importance in aging are computationally determined.

40 **Introduction**

41 The total cellular transcript-contents, also known as transcriptome-size, vary from cell-to-cell (*1-3*).
42 Several studies indicate that the cell-size influences the overall cellular transcription (*4-7*). It is also
43 reported that c-myc amplifies the global transcription in tumor cells (*2, 8*), lymphocytes and embryonic
44 stem cells (*9*). Another study shows that the global transcription is repressed by the loss of MECP2, a
45 methyl CpG binding protein, in embryonic stem cell-derived neurons (*10*). The genome structure, such
46 as ploidy/gene-dosage and nucleosome state, affects the overall transcription (*11-13*). Failure to
47 integrate such variations in the single-cell analyses could hide a potentially important cell-to-cell gene
48 expression variation or could lead to finding a false cell-to-cell gene expression heterogeneity (*1, 2, 14*).
49 Furthermore, it is shown that the total cellular transcriptome-size of tumor cells is shown to estimate
50 the extent of the cancer progression (*15*), indicating its potential medical significance.

51

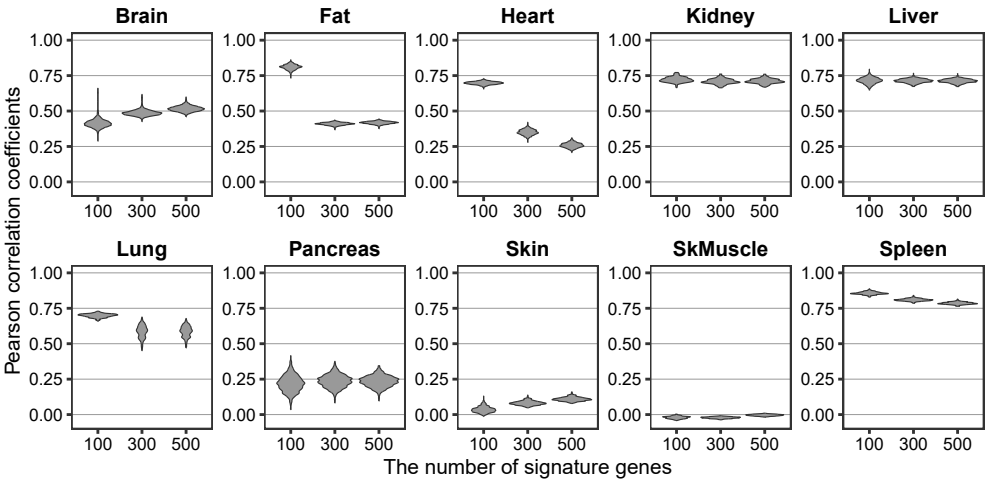
52 Despite such importance, little is known about the extent to which the total cellular transcriptome-size
53 varies among different cell-types across diverse types of organs in the body. Furthermore, such apparent
54 variations are mostly ignored in the current single-cell analyses, such as single-cell RNA sequencing
55 (scRNA-seq) and other types of single cell transcriptomics methods (*16-20*). In such methods, the
56 transcripts are sequenced in individual cells and their transcript-counts are normalized, allowing the
57 comparison of the relative abundance of each transcript-species across individual cells/cell-types (*21-*
58 *25*). This internal normalization process of the methods cancels out the putative total cellular
59 transcriptome size-variations among different cells, hiding potentially important functional differences
60 among the cells.

61

62 Such shortfalls are primarily caused by the lack of experimental methods to measure the total
63 transcriptome-size of individual cells across diverse cell-types and organs in the body. The utilities of
64 the currently available methods are limited to culture cells, immune cells and tumor cells (*2, 3, 14, 15,*
65 *26, 27*), and not applicable to diverse cell-types and organs of the body. This limitation is imposed by
66 two factors: First, these experimental methods require dissociation and purification (e.g., cell-sorting)
67 of each individual cell-types from the solid organs and the accurate quantification of the transcript-

68 counts per cell. This is a challenge for the solid organs consisting of highly heterogenous cell-types.
69 Second factor is an apparent cryptic gene expression induced by the cell-dissociation and purification
70 (*28*).
71
72 In this study, we overcome these challenges by developing a computational method, by exploiting a
73 growing body of multi-organ bulk RNA-seq and the corresponding composite scRNA-seq data, to
74 body-widely infer the relative differences of the total cellular transcriptome-size of individual single-
75 cells across diverse organs. Using this method, we determined and report cell type-specific "weighting-
76 factors" for 76 cell-types cells across 10 mouse organs that account for the quantitative differences of
77 their total cellular transcriptome-size. Furthermore, we show that such body-wide variations of the total
78 cellular transcriptome-size can predict aging-stages in the mouse model. These results together unveil
79 the landscape of the body-wide variations of the total cellular transcriptome-size and their possible
80 functional importance in physiological processes such as aging.
81
82 **Results**
83 **The cell type-specific weighting factor (cWF)**
84 First, we challenge the following conventional notion that the total cellular transcriptome-size is equal
85 across different cell-types:
86 Let all cells be assumed to contain the equal amounts of transcripts. Then, the whole-organ RNA-seq
87 should be the sum of gene expression of each composite cell-type weighted by cell-type ratios.
88
89 We evaluated this assumption as follows:
90 The synthetic whole-organ data were constructed from single-cell transcript counts combined according
91 to the experimentally determined fraction of each cell-type in the organ (Materials and Methods).
92 Briefly, the synthetic whole-organ RNAseq data are calculated as the sum of the normalized transcript
93 counts weighted by the known ratios of the composite cell-types of the organs, using the signature genes
94 for distinguishing the cell-types for each organ. These synthetic data for 10 organs (brain, fat, heart,
95 kidney, liver, lung, pancreas, skin, skeletal muscle, spleen) were then compared to the corresponding

96 real whole-organ RNA-seq data (Fig. 1, table S1). The comparisons were performed by calculating their
97 Pearson correlation coefficients.



98

Fig. 1. Incomplete reconstitution of the whole organ RNA-seq by the composite scRNA-seq. The similarity is shown as violin plot of the Pearson correlation coefficient for each number of the signature genes (100, 300, 500) for each organ (indicated above each plot). The corresponding raw data are found in table S1. SkMuscle: skeletal muscle.

99 The result shows low Pearson correlation coefficients (< 0.75) for most of the organs (brain, heart,
100 kidney, liver, lung, pancreas, skin, skeletal muscle) regardless of the number (100, 300, 500) of the top-
101 ranked signature genes used. Particularly low coefficients are found with the brain (< 0.6 for 100, 300,
102 500 signature genes), the fat (< 0.5 for 300, 500 signature genes), the heart (< 0.4 for 300, 500 signature
103 genes), the pancreas (< 0.3 for 100, 300, 500 signature genes), the skin (< 0.2 for 100, 300, 500
104 signature genes), and the skeletal muscle (< 0.1 for 100, 300, 500 signature genes). These low Pearson
105 correlation coefficients suggest that the total cellular transcriptome-size of individual cell-types varies,
106 resulting in the large gaps between the real and synthetic whole-organ RNA-seq data.
107
108 To solve this apparent problem, we developed an algorithm to compute such variations and examined
109 whether they can correct the disagreement between the real and synthetic whole-organ RNA-seq data
110 (Fig. 2 and Materials and Methods). The concept of the algorithm is schematically described in Fig. 2

and as follows: In the real scRNA-seq data analyses, the total transcript counts per cell (i.e., the total cellular transcriptome-size) of the real scRNA-seq data are normalized (i.e., the total counts per cell are equal among all composite cell-types of the organs). This results in the loss of a factor representing the putative differences of the total cellular transcriptome-size across various cell-types. This factor is computed as the cell type-specific weighting-factor (cWF) (indicated as w_1, w_2, w_3, etc. in the Fig. 2).

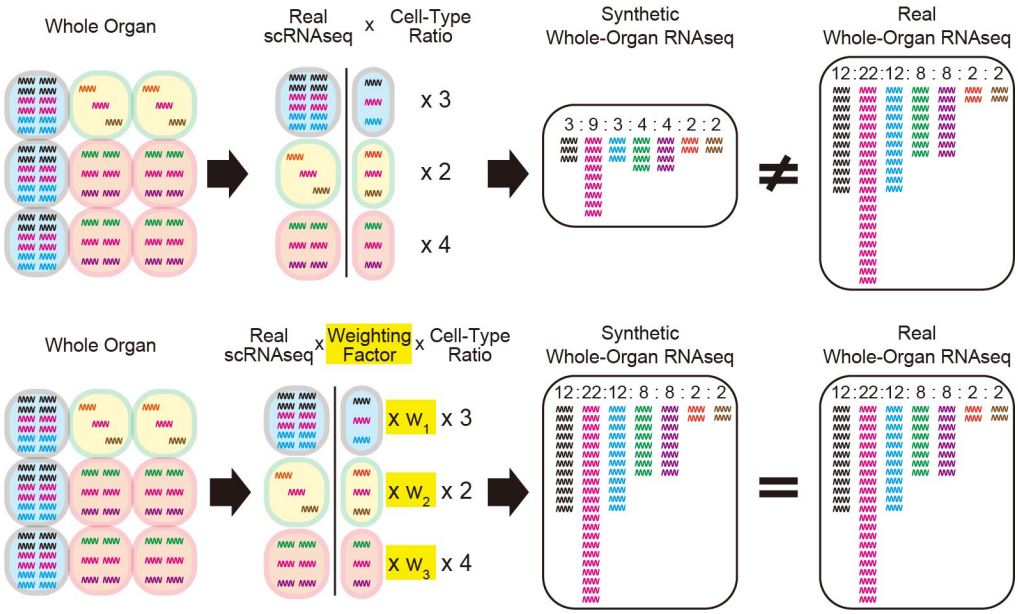
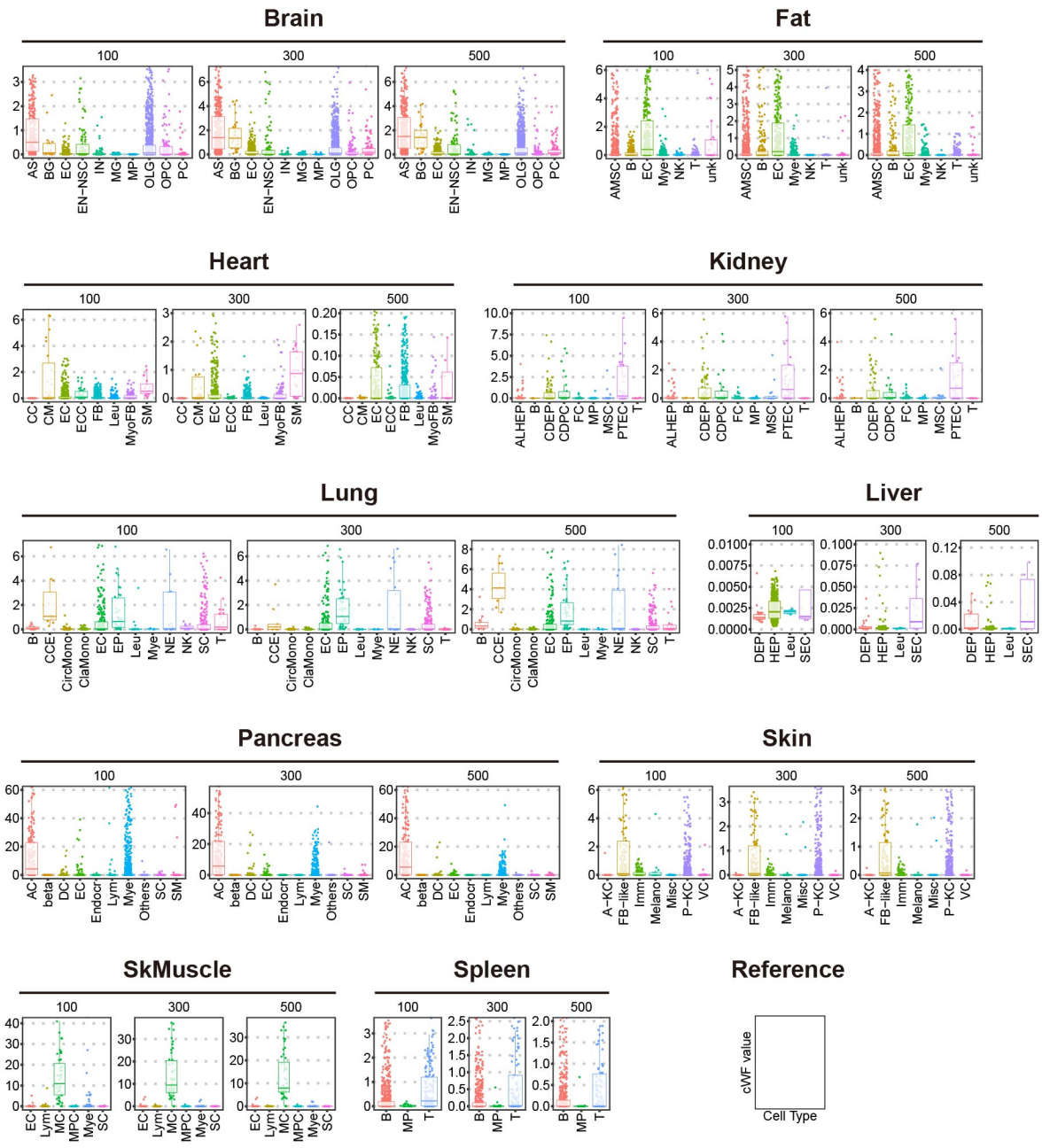


Fig. 2. Graphical description of the overall concept of the cell type-specific weighting factors (cWFs). The top row describes the case without the cWFs, failing to reconstitute the real whole-organ RNAseq data by the synthetic whole-organ RNAseq data generated by assembling the composite scRNA-seq data without the cWFs. The bottom describes the case with the cWFs, accurately reconstitute the real whole-organ RNAseq data by the synthetic whole-organ RNAseq data generated by assembling the composite scRNA-seq data with the cWFs (w_1, w_2, w_3). The wiggling lines are cellular transcripts where different gene transcripts are in different colors. The ratios of the cellular transcript-counts in the synthetic and real whole-organ RNAseq are indicated by the numbers at the top of the colored transcripts.



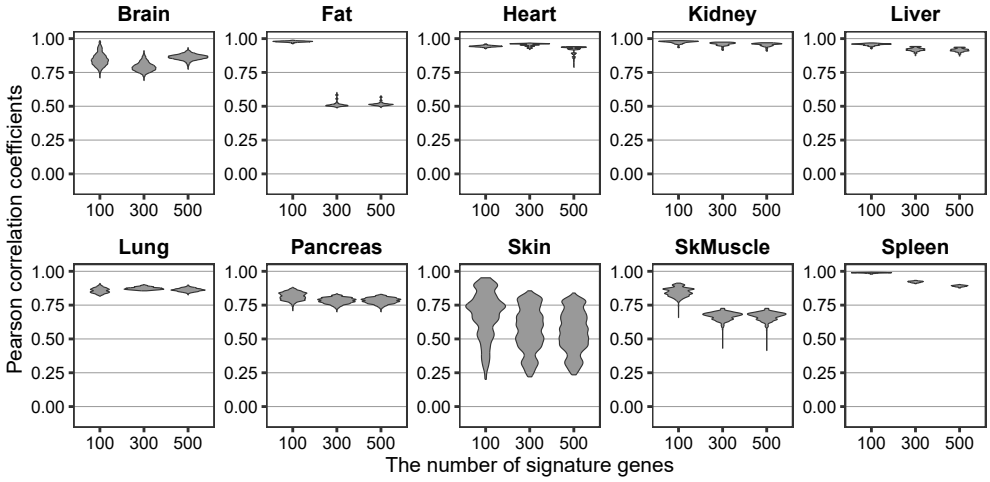
119

Fig. 3. The cWFs for each cell-type in each organ. The values of the cWFs for each cell-type (bottom of each plot) for each number of the signature genes (indicated above each plot) for each organ (indicated at the top of the corresponding plots) are shown as box plots. The corresponding raw data are found table S2. The cell-types are indicated as abbreviations and their full names are found in table S2. SkMuscle: skeletal muscle.

120

121 Using this algorithm, we computed cWFs for 76 cell-types across 10 organs (Fig. 3, table S2). The
122 result shows significant variations of the cWFs among different cell-types in many of the organs,
123 uncovering the body-wide degree of such variations. In particular, the following cell-types show
124 significantly higher cWFs than the other cell-types within the same organ: astrocyte (AS) in the brain,
125 endothelial cell (EC) in the fat, muscle cell-types in the heart and the skeletal muscle (CM in the heart,
126 MC in the skeletal muscle), epithelial cell of proximal tubule (PTEC) in the kidney, epithelial cell (EP)
127 and neuroendocrine cell (NE) in the lung, fibroblast-like cell (FB-like) in the skin, and T-cell (T) in the
128 spleen. This result suggests that these cell-types contain larger amounts of transcripts (i.e., the larger
129 total cellular transcriptome-size) than the others in the corresponding organs. Furthermore, the results
130 show that such variable degrees of differences in the total cellular transcriptome-size are widespread
131 across diverse cell-types and organs in the body.
132
133 **Validation of the cWFs**
134 We demonstrate that the cWFs indeed account for the variations of the total cellular transcriptome-size
135 by three methods: Reconstitution (Fig. 4) and deconvolution (Figs. 5 and 6) of the whole-organ RNAseq
136 data using their composite scRNA-seq data, and their comparison to the experimentally measured
137 values (Fig. 7).
138
139 First, we validated the cWFs by the reconstitution (Fig. 4, table S3, Materials and Methods). Using the
140 mean value of cWFs for each cell-type, we weighted the transcript counts for each cell-type. We then
141 summed these weighted transcript counts for all composite cell-types of each organ according to their
142 ratios to construct the synthetic whole-organ RNAseq (syn-woRNAseq(+cWF)), data. Then, they are
143 compared to the corresponding real whole-organ RNAseq (r-woRNAseq) data. The result shows the
144 Pearson correlation coefficients of 0.8 – 1.0 between the syn-woRNAseq(+cWF) and r-woRNAseq for
145 all 10 organs with all or at least one of the 100, 300, 500 signature genes (Fig. 4). These are significant
146 improvements from the 0 – 0.75 of the Pearson correlation coefficients without the cWFs (Fig. 1). The
147 heart shows the Pearson correlation coefficients of almost 1.0 with the cWFs (Fig. 4), as compared to
148 those of 0.25 – 0.7 without the cWFs (Fig. 1). Even for the spleen where the Pearson correlation

149 coefficients are 0.75 – 0.8 without the cWF (Fig. 1), the cWF (i.e., syn-woRNAseq(+cWF)) improves
150 the Pearson correlation coefficients up to 0.9 – 1.0 (Fig. 4). The results demonstrate that the cWFs
151 correct the large gap between the synthetic whole-organ RNAseq without the cWFs (syn-woRNAseq(-
152 cWF)) and r-woRNAseq, supporting the notion that the cWFs reflect the total cellular transcriptome
153 size-variations of the cells and their requirement for the accurate single-cell analyses.



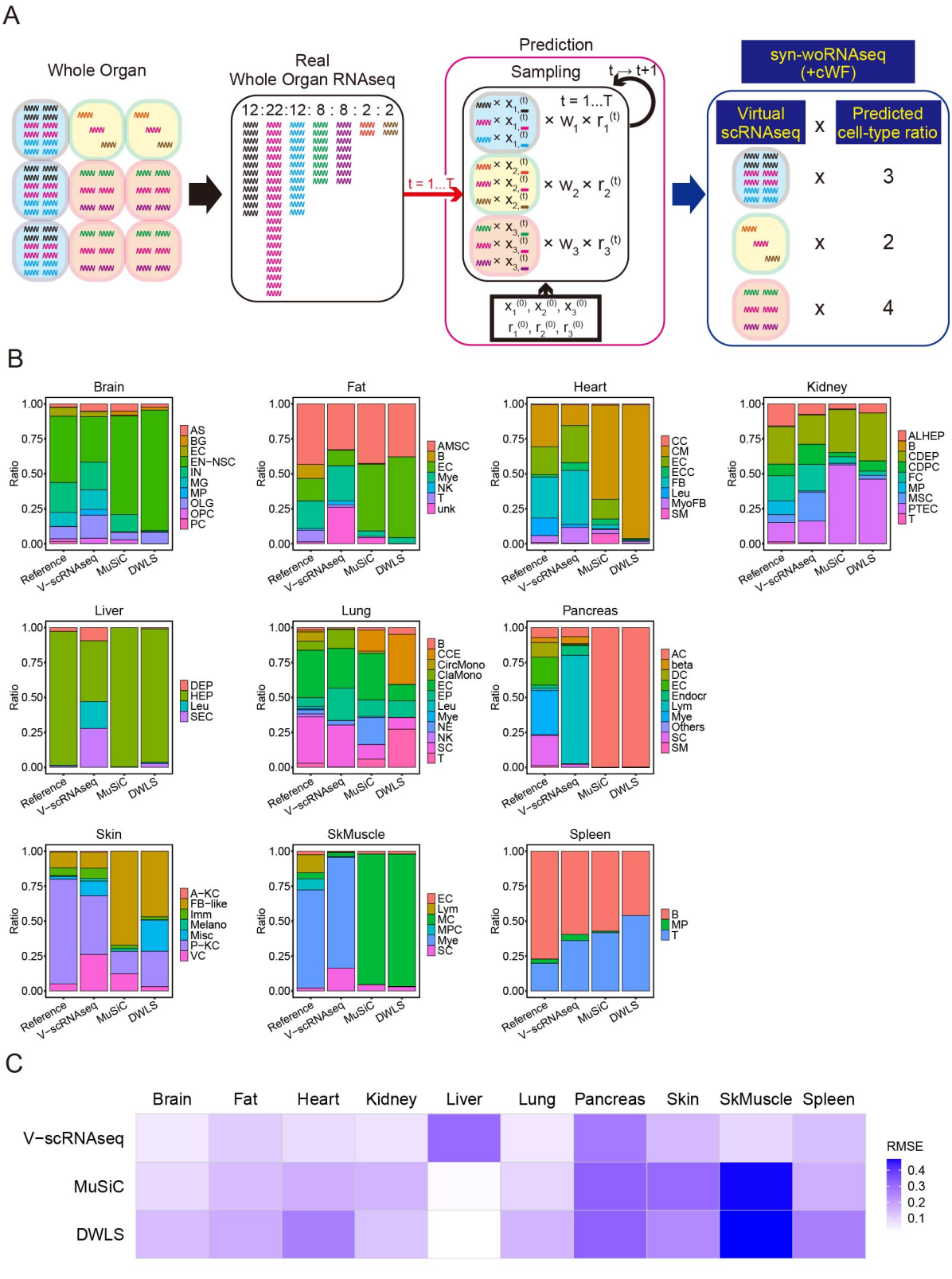
154

Fig. 4. Accurate reconstitution of the whole organ RNA-seq by the composite scRNA-seq and cWFs. The similarity is shown as violin plot of the Pearson correlation coefficient for each number of the signature genes (100, 300, 500) for each organ (indicated above each plot). The corresponding raw data are found in table S3. SkMuscle: skeletal muscle.

155

156 Next, the validation by the deconvolution (Fig. 5) was performed. Several deconvolution methods were
157 previously reported (*29-42*). However, they fail to account for the variations of the total cellular
158 transcriptome-size. Hence, we developed a deconvolution method that integrates the cWFs in a
159 Bayesian framework, which is schematically described in Fig. 5A. Briefly, it works as follows
160 (Materials and Methods): The mean of the cWF (w_1, w_2, w_3 in Fig. 5A, the lower case 1, 2, 3 indicate
161 cell-type 1, 2, 3) is used to compute virtual transcript counts ("x" in Fig. 5A) for each gene (indicated
162 as wiggling lines of different colors in Fig. 5A) and the ratio (indicated as "r" in Fig. 5A) of each cell-
163 type in each organ (painted in different colors in Fig. 5A). Using the signature-gene set, the computation
164 is iteratively performed by the sampling method of No-U-Turn Sampler (NUTS), in the Bayesian

165 framework with two hyperparameters, α and β, to account for the combinatorial influence of the cell-
166 type ratios and the gene expression patterns of each cell-type on the whole-organ-level transcriptome,
167 respectively. For the iteration (t = 1, …, T), the initial condition is set as $x_1^{(0)}$ and $r_1^{(0)}$ for the transcript
168 counts and the ratio of cell-type 1 (see the diagram labels as "Prediction" in Fig. 5A). Those for cell-
169 types 2 and 3 are indicated accordingly (see the diagram labels as "Prediction" in Fig. 5A). Such initial
170 conditions are the cWF-weighted transcript count for each gene and the reference cell-type ratio,
171 respectively. The iteration is repeated until all sampling variables converge with the signature-gene set.
172 The counts of the non-signature gene set are computed by an analytic approach in the Bayesian
173 framework with a hyperparameter, γ, which has the similar nature of β as above. The "r" is fixed as the
174 above-estimated value, and only "x" in the non-signature gene set is computed.
175
176 The deconvolution results for 10 organs using this method were compared to the real cell-composition
177 data of the corresponding organ (Fig. 5, B and C). The results were also compared to those of two
178 conventional deconvolution methods without the cWF, MuSiC and DWLS (*40, 41*). These two were
179 chosen as they have been applied to 1 – 4 organs and appear to outperform the other published methods.
180 The prediction of the cell-type ratios show that our method (V-scRNAseq) outperforms the methods
181 without the cWFs (MuSiC, DWLS) for 9 organs (brain, fat, heart, kidney, lung, pancreas, skin, skeletal
182 muscle, spleen). For these organs, the cell-type ratios predicted by our method are more similar to those
183 of the real data than those predicted by MuSiC or DWLS (Fig. 5B, table S4). Notably, our method
184 corrected the abnormally large ratios of cardiac muscle and muscles in the heart and the skeletal muscle,
185 respectively, estimated by MuSiC and DWLS (Fig. 5B, table S4). The lower root-mean-square error
186 (RMSE) for these 9 organs with our method (Fig. 5C, table S4) further confirms the better performance
187 of our method (Fig. 5C, table S4). For the liver, our method failed (Fig. 5, B and C, table S4), despite
188 the high performance in the reconstitution (Fig. 4). This may be due to the predominantly hepatocyte
189 composition of the liver. Our method assumes the symmetric Dirichlet distribution as prior distribution
190 which influences the posterior distribution – hence, the method is less suitable for the organ, such as
191 the liver, consisting of a single dominant cell-type.
192


A
Whole Organ
Real
Whole Organ RNAseq
12:22:12: 8 : 8 : 2 : 2
t = 1...T
Prediction
Sampling
t → t+1
t = 1...T
syn-woRNAseq
(+cWF)
Virtual
scRNAseq
x
Predicted
cell-type ratio
x 3
x 2
x 4
B
Brain
Fat
Heart
Kidney
Liver
Lung
Pancreas
Skin
SkMuscle
Spleen
Ratio
Reference
V-scRNAseq
MuSiC
DWLS
AS
BG
EC
EN-NSC
IN
MG
MP
OLG
OPC
PC
AMSC
B
EC
Mye
NK
T
unk
CC
CM
EC
ECC
FB
Leu
MyoFB
SM
ALHEP
B
CDEP
CDPC
FC
MP
MSC
PTEC
T
DEP
HEP
Leu
SEC
B
CCE
CircMono
ClaMono
EC
EP
Leu
Mye
NE
NK
SC
T
AC
beta
DC
EC
Endocr
Lym
Mye
Others
SC
SM
A-KC
FB-like
Imm
Melano
Misc
P-KC
VC
EC
Lym
MC
MPC
Mye
SC
B
MP
T
C
Brain
Fat
Heart
Kidney
Liver
Lung
Pancreas
Skin
SkMuscle
Spleen
V-scRNAseq
MuSiC
DWLS
RMSE
0.4
0.3
0.2
0.1

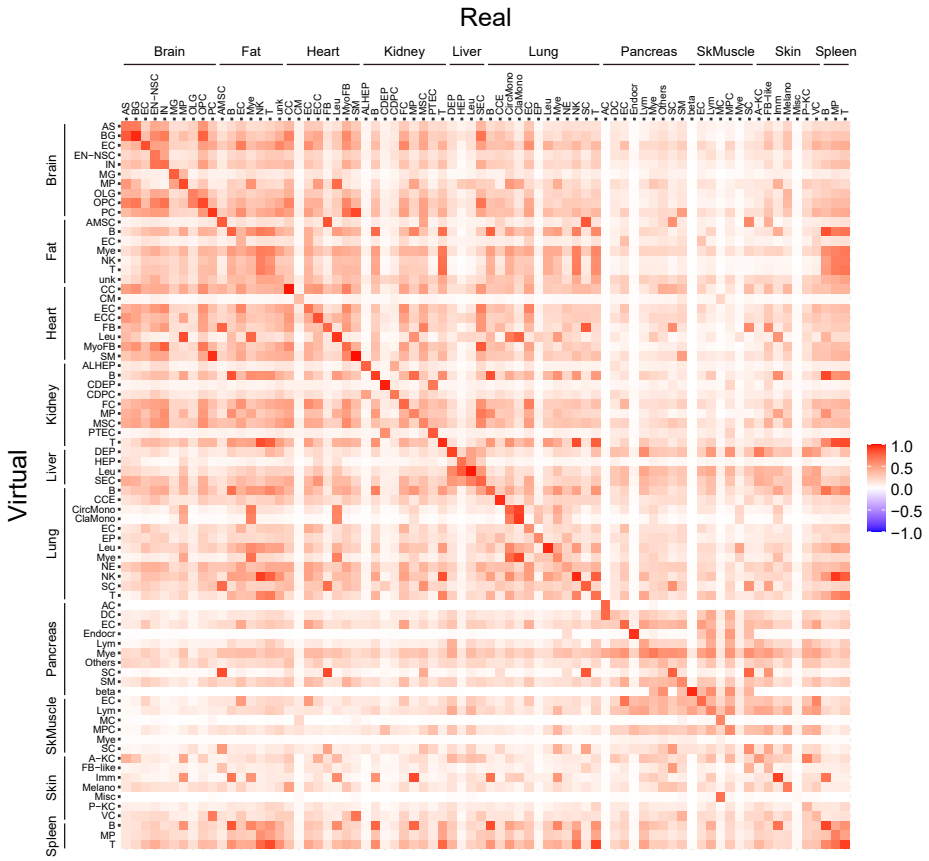
196

Fig. 5. Accurate deconvolution of the whole organ RNA-seq by cWFs. (A) Graphical description of the overall concept of the deconvolution method with the cWFs. 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 cWFs: w_1, w_2, w_3 for each cell-type 1, 2 and 3, respectively. The virtual transcript counts ($x_1^{(t)}$, $x_2^{(t)}$, $x_3^{(t)}$ where the different genes are indicated by the 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 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 iteration (t= 1, …, T) is performed until both **X** and **r** converge (Materials and Methods). Upon completing the iterations, the sum of the virtual transcript counts weighted by the cWFs (indicated as "Virtual scRNAseq") weighted by the virtual cell-type ratios (indicated as "Predicted cell-type ratio) generates the synthetic whole-organ transcriptome (indicated as "syn-woRNAseq (+cWF)"), as shown at the far right of this panel. **(B)** Bar graph showing the cell type-ratios computed by our deconvolution method (V-scRNAseq) for each organ. The deconvolution was performed with the cWFs computed using the optimal number of the signature genes for each organ (indicated in the accompanying table S4). The reference cell-type ratios and the ratios estimated by the conventional deconvolution methods without cWFs, MuSiC and DWLS, are shown as comparisons. The bar graphs are composed of the cell-types computed to be present for each organ by our method. **(C)** The quantitative similarity of the computed ratios to the reference ratios is shown by heatmap of RMSE for our method with cWFs (V-scRNAseq), MuSiC, and DWLS. RMSE: Root Mean Squared Errors. The corresponding raw data are found in table S4. SkMuscle: skeletal muscle.

197

198 In addition to the cell-type ratios, we also computed the virtual transcriptome of 23,131 (Brain, Fat,
199 Heart, Lung, Spleen), 22,742 (Kidney), 23,104 (Liver), 15,682 (Pancreas), 21,233 (Skin) and 14,129
200 (the skeletal muscle) genes in each of the "virtual" cell-type and compared them to those of the real
201 cell-type for the 10 organs (Fig. 6 and Materials and Methods). Our approach computes a posterior
202 distribution and its Expected A Posteriori (EAP) for counts of each transcript for each cell-type. Hence,
203 we calculated Pearson correlation coefficients for each cell-type, using the EAPs of all genes in virtual

204 and real cell-types for each organ (Fig. 6, table S5). Their Pearson correlation coefficients indicate that
205 the virtual transcriptomes are comparable to the real ones across all cell-types and organs. They also
206 recapitulate the transcriptomics-relatedness of the same/related cell types across different organs (Fig.
207 6). For example, the similarity of endothelial cells in multiple organ such as brain, fat, heart, liver and
208 lung is recapitulated in the virtual transcriptomes among these organs (Fig. 6). The relatedness of each
209 of the immune system cells across multiple organs is also represented (Fig. 6).



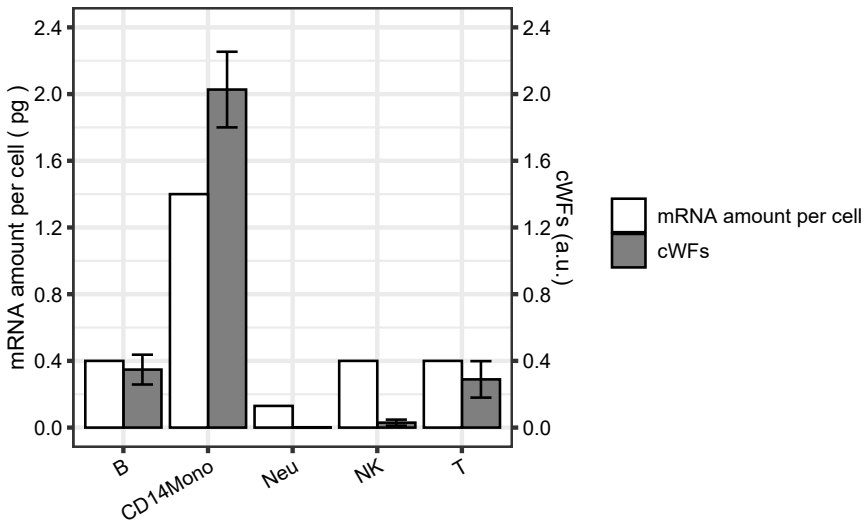
210

Fig. 6. Computation of the complete transcriptome of the composite cell-types of 10 mouse organs. 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 (Brain, Fat, Heart, Lung, Spleen), 22,742 (Kidney), 23,104 (Liver), 15,682 (Pancreas), 21,233 (Skin) and 14,129 (the skeletal muscle) genes. The corresponding raw data are found in table S5. SkMuscle: skeletal muscle.

211

212

213 The reconstitution and deconvolution results demonstrate that the cWFs enable the accurate
214 representation of individual cells in the context of the whole-organs of highly heterogenous
215 compositions, presumably by weighting the variations of the total cellular transcriptome-size of
216 different cell-types. We further validate this account by directly comparing the cWFs variations to those
217 of the experimentally-determined total transcriptome-contents of the corresponding cell-types.
218 Currently, as described in Introduction, the accurate measurement of such variations for individual
219 single-cells of the solid organs are difficult and the methods to measure such variations are limited to
220 cultured cells and systemic/circulating cells such as immune cells, and tumor cells (*2, 14, 15, 26*).
221 Therefore, we used the experimentally-determined total cellular transcript-contents of the cell-types
222 composing human peripheral blood mononuclear cells (PBMCs) and compared them to the computed
223 cWFs of the corresponding cell-types (Fig. 7).



224

Fig. 7. Comparison of the cWFs to the experimentally-determined total cellular transcript-contents of human PBMC cell-types. The comparison is shown as bar graphs of the experimentally-determined transcript-contents per cell (mRNA amount per cell: empty bar) and the computed cWFs (cWFs: filled-bar) for each cell-type. The bars are shown as mean ± S.E. (standard error). B: B-cell, CD14Mono: $CD14^+$ monocytes, Neu: Neutrophil, NK: Natural killer cell, T: T-cell. The validation of the cWFs by reconstitution and deconvolution are found in fig. S1 (raw data are in tables S6 – S8). The corresponding raw data for Fig. 7 are found in table S9.

225

226 We used bulk and single-cell RNA-seq data of human PBMCs (*43-45*) and computed the cWFs of B-
227 cells, $CD14^+$ monocytes, $CD16^+$ monocytes, dendritic cells, neutrophils, natural killer (NK) cells,
228 plasmablasts (PB), T-cells, and the others (e.g., red-blood-cells, platelets, etc.) (Materials and Methods).
229 The reconstitution and deconvolution analyses confirmed that they correct the biased gene-expression
230 in the scRNA-seq data caused by the total cellular transcriptome-size normalization, as indicated by the
231 higher Pearson correlation coefficients (fig. S1A, tables S6 and S7) and lower RMSE/higher Person
232 correlation coefficient (fig. S1B, table S8) with the cWFs in the reconstitution and deconvolution
233 analyses, respectively.

234

235 The total cellular RNA-contents are experimentally measured and determined for the B-cells, $CD14^+$
236 monocytes, neutrophils, NK cells, and T-cells of PBMCs (*27*). We made the side-by-side comparison
237 of the computed cWFs and their experimentally-measured total cellular transcriptome-size of the
238 corresponding cell-types (Fig. 7, table S9). The result shows that the computed cWFs accurately
239 recapitulate the 3.5 - 10.8-fold higher transcriptome-size of the $CD14^+$ monocytes compared to the B-,
240 T-, NK cells, or neutrophils, respectively (i.e., $CD14^+$ monocytes: 1.4 pg/cell vs. B-/T-/NK-cells: 0.4
241 pg/cell, Neutrophils: 0.13 pg/cell). It also recapitulates the 3-fold higher values of the B- and T-cells
242 than that of neutrophils (i.e., 0.4 vs. 0.13 pg/cell). The experimental values of the transcript-contents of
243 the B- and T-cells are approximately the same, which is also recapitulated by virtually the same cWFs
244 of these two cell-types. These side-by-side comparisons provide another layer of evidence supporting
245 the account that the cWFs indeed represent the total cellular transcriptome size-variations among
246 different cell-types.

247

248 **Physiological significance of cWFs**

249 The analyses described above demonstrate that the total cellular transcriptome-size significantly varies
250 across different cell-types and the cWFs account for such variations. Furthermore, we unveil their body-
251 wide landscape. We next addressed their possible physiological significance. In this regard, we set a
252 hypothesis that the total cellular transcriptome-size of each cell-type differentially change over aging

253 and that such a phenotype could underlie the aging mechanism and could also be exploited as a new
254 type of aging-indicator. To test this hypothesis, we examined whether the patterns of the cWFs change
255 over the course of aging. Furthermore, we examined whether such putative changes of the cWFs
256 actually indicate the progression of aging.
257
258 For this purpose, we used the *Tabula Muris Senis*, the atlas of the multiple-organ single-cell
259 transcriptomics across multiple aging stages of mouse (*46, 47*). This atlas provides both the whole-
260 organ RNA-seq data and their corresponding composite scRNA-seq data. Using these data, we
261 computed cWFs of the composite cell-types of 5 organs (brain, heart, kidney, lung, spleen) across 3
262 aging-stages (3, 18, and 24 months-old) of mice. The 3, 18, 24 months-old (mos.) in mouse correspond
263 to 25 - 26 years-old, 60 - 80 years-old, and ≥ 80 years-old in human, respectively (*48*). These 5 organs
264 were chosen for the following two reasons: 1) they performed well in both the reconstitution and
265 deconvolution validations (Fig. 4 and 5); 2) their equivalent whole-organ/single-cell data are also
266 present in our 10 organs-analysis described above.
267
268 Using these datasets, we computed the cWFs of all composite cell-types for each organ. The
269 reconstitution (fig. S2, tables S10 and S11) and deconvolution (table S12) analyses confirmed that they
270 correct the biased representations of the gene expression and cell-type compositions, respectively, in
271 the scRNA-seq data with one or more of the signature gene set(s) for each aging-stage. Based on these
272 results, we selected the best performing numbers of the signature genes for each organ and aging-stage
273 by omitting the conditions that failed to significantly improve the reconstitution (highlighted in pink in
274 table S12) and selecting the ones with the lowest RMSE and the highest Pearson correlation coefficients
275 in the deconvolution analysis (highlighted in light-green in table S12). This screening selected the cWFs
276 computed with the following conditions: Brain (3 mos. – 500 signature genes, 18 mos. – 300 signature
277 genes, 24 mos. – 300 signature genes), Heart (3 mos. – 100 signature genes, 18 mos. – 500 signature
278 genes, 24 mos. – 300 signature genes), Kidney (3 mos. – 300 signature genes, 18 mos. – 500 signature
279 genes, 24 mos. – 500 signature genes), Lung (3 mos. – 500 signature genes, 18 mos. – 300 signature

genes, 24 mos. – 300 signature genes), Spleen (3 mos. – 300 signature genes, 18 mos. – 100 signature genes, 24 mos. – 500 signature genes).

With these selected cWFs, we examined whether their relative differences change over the course of aging. For this purpose, we compared the ratios of the cWFs of a pair of two distinct cell-types in each organ across three aging-stages (Fig. 8, table S13). The analysis detected their statistically significant ($p < 0.01$ by U-test) changes of many of the cell type-pairs in all organs over the course of aging.

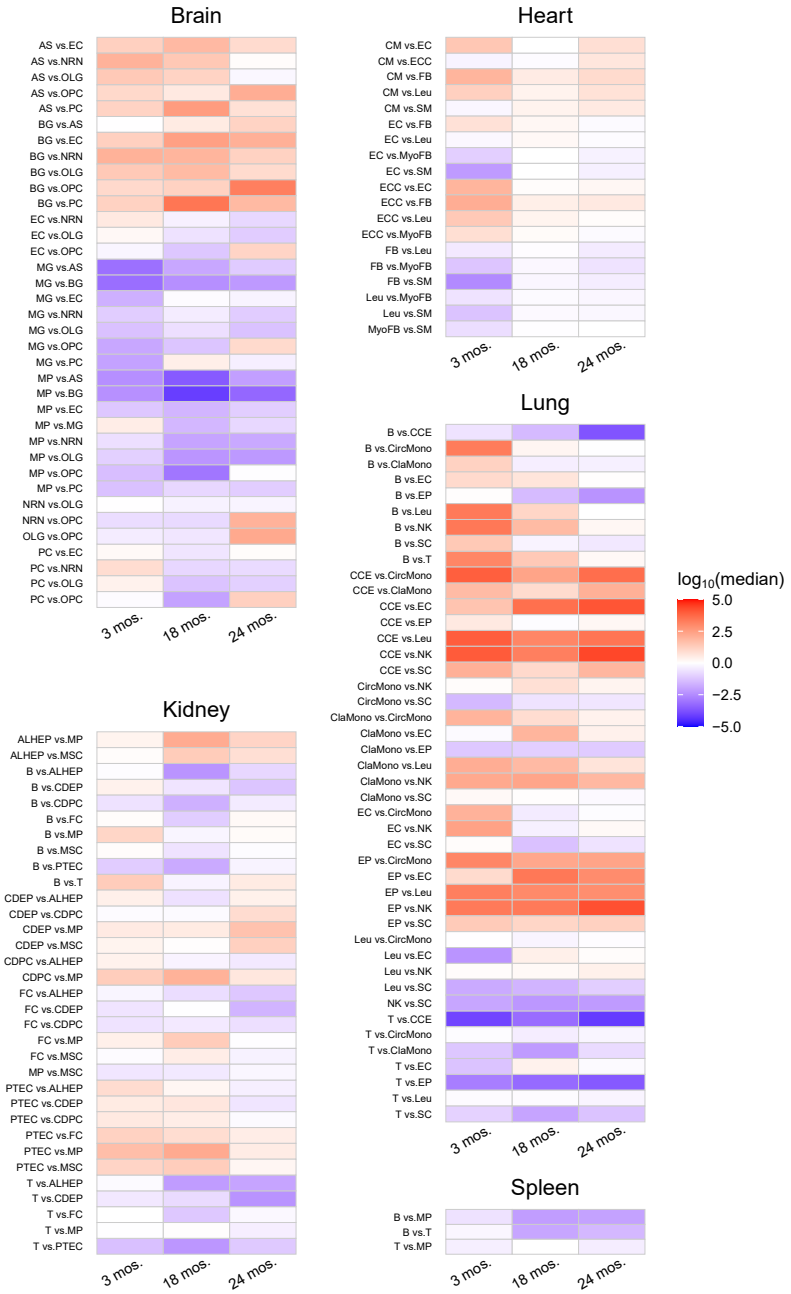


Fig. 8. Differential cWFs during aging of mouse. The changes of the ratios of cWFs of cell type-pairs across three aging-stages (3 months-old: 3 mos., 18 months-old: 18 mos., 24 months-old: 24 mos.) are shown as heatmap for each organ (Brain, Heart, Kidney, Lung, Spleen). The cell type-pairs are indicated at the left of each heatmap. The ratios are the pairwise ratios of the corresponding cell-types. For example, AS vs. EC indicates the ratio of the cWF for AS (numerator) and EC (denominator). The ratios are indicated as $\log_{10}$ of the medians. Shown are the pairs resulted in $p < 0.01$ by U-test in one or more of the three pairwise aging stage-comparisons (18 mos. vs. 3 mos., 24 mos. vs. 3 mos., 24mos. vs. 18 mos.). The analysis was conducted with those showing the significant improvements in the reconstitution (fig. S2, tables S10 and S11) and the best deconvolution results (table S12) for each organ. The full names of the cell type abbreviations and the raw data of the heatmap are found in table S13.

We next investigated whether such differential pairwise ratio-changes of the cWFs predict the progression of aging. For this purpose, we applied a machine learning algorithm, LightGBM (*49*), to the dataset of the pairwise cWF ratios between the cell type-pairs listed above (Materials and Methods). First, each organ dataset was independently evaluated (Fig. 9). The result shows high predictability with the pairwise cWF ratios between the brain, the lung, the heart, and the kidney cells, as indicated by their high prediction index scores (i.e., accuracy, macro-recall, macro-precision, macro-F1 score) of 0.88 – 0.98 (Fig. 9A).

The feature importance analysis identified important cell type-pairs for the high prediction performance in each organ (Fig. 9B, table S14). The following pairs are found especially important (the top 5 in each organ): Brain - BG/OPC, MP/OPC, BG/PC, BG/AS, PC/OPC; Lung - CCE/NK, B/SC, B/Leu, B/CCE, T/CCE; Heart - FB/SM, ECC/FB, ECC/MyoFB, ECC/EC, ECC/Leu; Kidney - B/CDPC, B/ALHEP, B/T, B/FC, B/PTEC. This result suggests that the cWFs of the following cell-types are more important contributors to the prediction, as indicated by their more frequent appearance in the top 5 ranked pairs: Brain – Bergmann glial cells (BG) and oligodendrocyte precursor cells (OPC), Lung – B cells (B) and

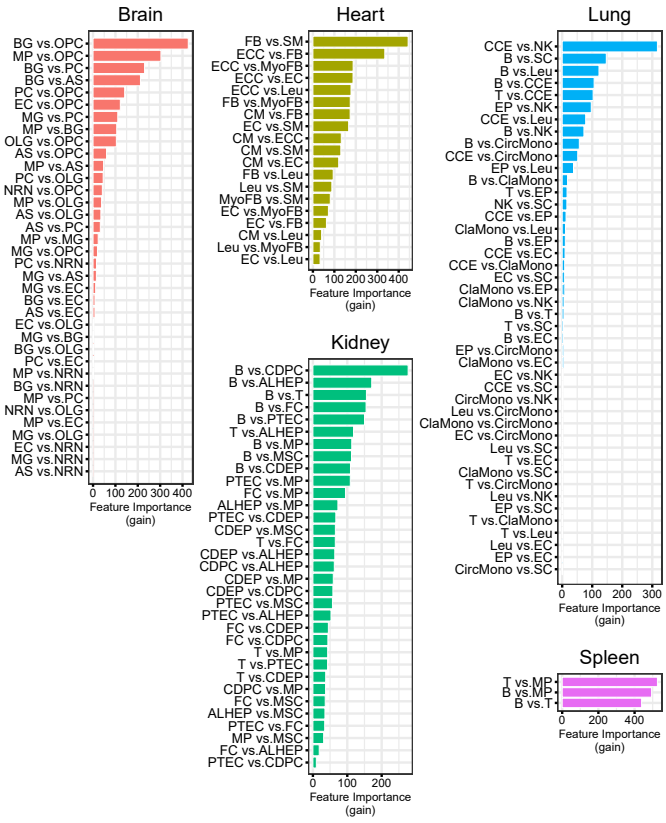
ciliated columnar cells of tracheobronchial tree multiciliated cells (CCE), Heart – endocardial cells (ECC), Kidney – B cells (B).

Fig. 9. Prediction of the aging-stage by the differential cWFs of each organ. (A) The prediction scores (accuracy, macro-precision, macro-recall, macro-F1 score) for each organ (Brain, Heart, Kidney, Lung, Spleen) are shown. **(B)** The top-ranked features (i.e., pairwise cell-types) important for the prediction are shown as bar graphs for each organ. The feature importance is indicated by gain. The corresponding raw data are found in table S14.

A

	Accuracy	macro−Precision	macro−Recall	macro−F1 Score
Brain	0.983	0.984	0.983	0.983
Heart	0.883	0.893	0.883	0.882
Kidney	0.883	0.883	0.883	0.881
Lung	0.967	0.967	0.967	0.967
Spleen	0.517	0.514	0.517	0.514

B



The LightGBM prediction with all organs combined also shows high performance as indicated by its prediction index scores of 1 (Fig. 10A). The feature importance analysis shows the top-ranked cell type-pairs are those of the lung and the brain, supporting the higher performance results with the cell-types of these organs when they are independently analyzed (Fig. 9A). The cell-types indicated as important contributors to the prediction listed above for the individual organ analyses also ranked higher in this

314 all-organ-combined analysis (Fig. 10B, table S14). Taken together, the results demonstrate that the
315 differential cWF ratios of the brain, the lung, the heart and the kidney accurately predict the progression
316 of aging, suggesting a possible role of the differential total cellular transcriptome-size among the cells
317 in these organs in the aging mechanism. The results also indicate a possibility of using these indices to
318 determine the progression of aging.
319

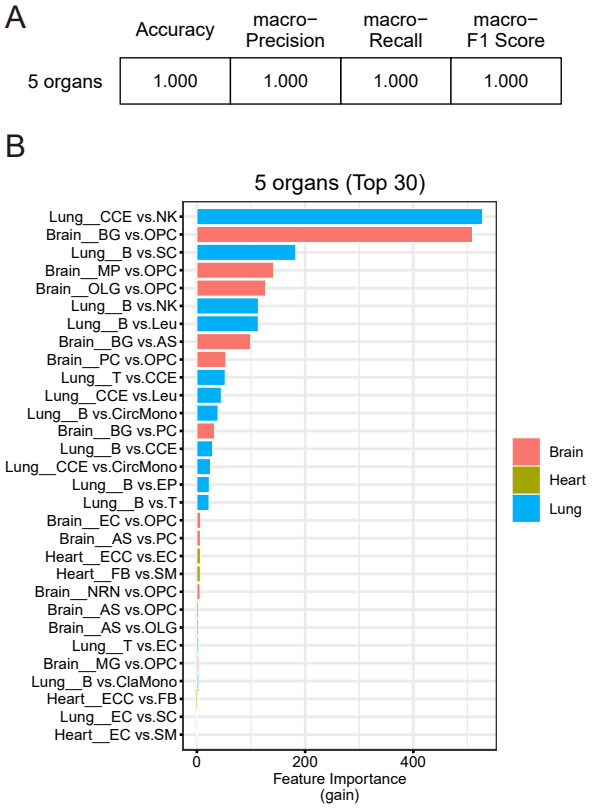


Fig. 10. Prediction of the aging-stage by the differential cWFs of all 5 organs combined. (A) The prediction scores (accuracy, macro-precision, macro-recall, macro-F1) by combining the cWF of all pairwise cell-types across all organs are shown. **(B)** The top 30 important features (i.e., the pairwise ratios of the cWFs) for the prediction are shown as bar graphs for each organ. The feature importance is indicated by gain. The cell type-pairs and their organs are indicated as organ-name_cell-type pair (e.g., Lung_CCE vs. NK: the CCE vs. NK pair of the lung). The corresponding raw data are found in table S14.

323

324 **Discussion**

325 In this study, we report a computational method to determine the relative variations of the total cellular
326 transcriptome-size of the cells in the solid organs (Fig. 2). Using this method, we computed and describe
327 such variations among 76 cell-types across 10 organs as cell type-specific weighting factors (cWFs)
328 (Fig. 3). We show that they indeed account for the relative variations of the total cellular transcriptome-

329 size of the cells by demonstrating their requisite role in accurately reconstituting and deconvolving the
330 whole-organ RNA-seq data using their composite scRNA-seq data weighted by the corresponding
331 cWFs (Figs. 4 – 6). We also show that they indeed recapitulate the real experimentally-determined total
332 cellular transcript content-differences, using PBMC data as a model (Fig. 7). Furthermore, we show
333 that they change over aging and can be used to effectively predict aging progression in the aging mouse
334 model (Figs. 8 – 10), suggesting their possible role in the aging mechanism and also their medical
335 applications such as diagnosis of aging progression and/or aging associated diseases.
336
337 We propose that the cWFs described here reflect the relative variations of the total cellular
338 transcriptome-size among the different cell-types in each organ. This is supported by not only the
339 reconstitution and deconvolution results (Figs. 4 – 6), but importantly also by the recapitulation of the
340 experimentally-measured total cellular transcript-contents of five cell-types of PBMCs (i.e., B-cells,
341 $CD14^+$ monocytes, neutrophils, NK cells, and T-cells) (Fig. 7). These results together indicate that the
342 cWFs indeed account for the relative variations of the total cellular transcriptome-size among the
343 different cell-types of the organs. Are there any additional factors contributing to the differential cWFs?
344 One such factor could be cell dissociation-associated aberrant gene expression (*50*). The scRNA-seq
345 analyses of the solid tissues and organs require enzymatic digesting and mechanical dissociation of
346 individual cells and such technical procedures induce cryptic gene expression. Another possible factor
347 is a bias introduced by differential RNA stability during the cell-dissociation and RNA-extraction
348 procedures in scRNA-seq. It is possible that distinct cell-types are differentially susceptible to overall
349 RNA-degradation, which could contribute to the biased scRNA-seq data. In these regards, the cWFs
350 may account for the total transcriptome size-variations of both biological and procedural origins.
351
352 What is the mechanism of the cell type-specific differential total transcriptome-size? Two molecular
353 pathways, c-myc and MECP2 pathways, are implicated for their roles in regulating the total cellular
354 transcriptome-size (*2, 8-10*). A physicochemical alteration of the genome appears to also contribute to
355 the overall cellular transcription(*13*). Cell-size is another regulator of the total cellular transcriptome-
356 size (*4, 5, 7, 14*). In this regard, we find larger cells such as cardiac muscle (CM) and skeletal muscle

357 (MC) cells exhibit significantly larger cWFs values as compared to those of the other smaller-sized cell-
358 types in the heart and the skeletal muscle, respectively (Fig. 3). In the heart, the volume of the cardiac
359 muscle cell is 20 – 25 times of that of non-muscle cells such as endothelial cells and fibroblasts (*51-*
360 *55*), supporting the notion that the larger cWF of the cardiac muscle cells reflects their larger cell-size.
361 In the brain, the astrocytes (AS), a larger cell-type than the microglial cells (MG), also show larger
362 cWFs than those of MG (Fig. 3). These results indicate that larger cell size is at least a contributor to
363 the larger cWFs.
364
365 Does the total cellular transcriptome-size regulate any physiological and/or disease processes? The
366 MECP2 mutation in Rett syndrome causes the global transcriptional repression via a defect in the
367 AKT/mTOR pathway activity (*10*). Recently, the utility of the tumor cell type-specific total cellular
368 transcriptome-size for predicting tumor phenotypes and clinical outcomes is reported (*15*). In our study,
369 we show the differential cWFs in organs can accurately predict aging progression in the mouse model
370 (Figs. 9A and 10A). Furthermore, we identified specific cell-types of which cWFs are major
371 contributors to the prediction (Figs. 9B and 10B). We also show that the cWFs of the lung and brain
372 cells are dominant determinants in the prediction (Figs. 9B and 10B), suggesting their differential organ
373 contributions to this phenotyping. Taken together, our results indicate a possibility that the cWFs, i.e.,
374 the total cellular transcriptome-size variations, are a part of the aging mechanism. The results also
375 demonstrate a utility of the cWFs to predict the aging-stages. Hence, our study provides evidence for
376 an important role of the total cellular transcriptome-size in regulating physiological processes such as
377 aging, and also indicate its possible utilities for the diagnosis of age-associated diseases.
378
379 Some limitations, in particular to the method itself, must be noted. One limitation is the requirement of
380 both bulk and single-cell RNA-seq data of the organs and the tissues of interest. Our method requires
381 both of these datasets for the target organs/tissues to infer the cWFs of the composing cell-types. As the
382 availability of such data is exponentially growing, this may become less of a limiting factor. Another
383 requirement is that the method needs at least a rough a priori knowledge about the cell-type
384 compositions (i.e., the ratio of cell-types) of the organs and the tissue of interest -i.e., the method works

385 only in a supervised framework. This limitation could also become less limiting as a growing body of
386 scRNA-seq data provides information on the cell-type compositions of previously less characterized
387 organs/tissues. There is also a computational limitation. When too many cell type-labels (e.g., hundreds
388 of distinct cell-types) are present in the data, the deconvolution fails to converge or the computational
389 cost becomes a rate-limiting factor. Despite such limitations, our method provides an important tool to
390 unveil previously less characterized cellular and molecular phenotypes, i.e., the differential cell type-
391 specific transcriptome-size, that appear to have significant physiological impacts.
392

393 **Materials and Methods**

394 **Datasets**

395 For the whole-organ RNA-seq data, two murine datasets are used: One from 11 weeks-old male
396 C57BL6/N Jcl mice and derived from 'iOrgans Atlas' (http://i-organs.atr.jp) (*56, 57*) and the other from
397 3, 18, and 24 months-old male C57BL6/N mice and derived from the *Tabula Muris Senis* (*46*). For the
398 scRNA-seq datasets, used are the brain, the fat, the heart, the liver, the lung, and the spleen data from
399 11 weeks-old male C57BL6/JN mice described in *Tabula Muris* (*20*). The kidney data are from 3
400 months-old male C57BL6/JN mice in *Tabula Muris Senis* (*47*). The skeletal muscle and the pancreas
401 data are from 8-10 weeks-old female C57BL6/J reported in Mouse Cell Atlas (*17*). The skin data are
402 from 9 weeks-old female C57BL6/J mice (*58*). The aging mouse scRNA-seq datasets in the brain, the
403 heart, the kidney, the lung and the spleen are from 3, 18, 24 months-old male C57BL6/JN mice in
404 *Tabula Muris Senis (46, 47)*. The human PBMC datasets are GSM2871599 in GSE107572 (bulk RNA-
405 seq) (*43, 45*) and GSM4557334 in GSE150728 (scRNA-seq) (*44*), both are from healthy donors.
406

407 **Data preprocessing**

408 Cell-type labels of the organs used in this study are as described in each corresponding original single-
409 cell study, except for those of the skeletal muscle, the pancreas, and the heart (the *Tabula Muris Senis,*
410 but not the *Tabula Muris,* data) and the human PBMC. For the skeletal muscle and the pancreas, the
411 multiple labels of the cell-types of the similar gene expression patterns in the original study are
412 combined into the single label as follows (*** indicating arbitrary characters). For the skeletal muscle,

413 ‘B cell_***’, ‘Dendritic cell’, and ‘T cell’ into ‘Lymphatic cell’, ‘Erythroblast_***’, ‘Granulocyte
414 monocyte progenitor cell’, ‘Macrophage_***’ and ‘Neutrophil_***’ into ‘Myeloid cell’, ‘Muscle
415 cell_***’ into ‘Muscle cell’. For the pancreas, ‘B cell’ and ‘T cell’ into ‘Lymphatic cell’, ‘Dendritic
416 cell’, ‘Erythroblast_***’, ‘Granulocyte’, and ‘Macrophage_***’ into ‘Myeloid cell’, ‘Endothelial
417 cell_***’ into ‘Endothelial cell’, ‘Smooth muscle cell_***’ into ‘Smooth muscle cell’, ‘Stromal
418 cell_***’ into ‘Stromal cell’, ‘Dividing cell’ and ‘Glial cell’ into ‘Others’. For the heart dataset in
419 *Tabula Muris Senis* (*47*), the labels were replaced with the ones in *Tabula Muris* (*20*) using single-cell
420 reference mapping by ‘*Seurat*’ v4 (*59*), as follows. First, the reference scRNA-seq data in *Tabula Muris*
421 were preprocessed by ‘*Seurat*’ analyses with the log-normalization (where the scale factor is 10^6), the
422 highly variable genes selection with default parameters, the scaling for all genes, and principal
423 component analysis (PCA) acquiring 200 principal components (PCs). The number of PCs used for the
424 downstream analysis was determined at p-value below 0.0001 in JackStraw analysis, resulting in 33
425 PCs. Using these 33 PCs reference data, the query heart scRNA-seq datasets from *Tabula Muris Senis*
426 *(47)* were annotated by the functions of ‘FindTransferAnchors()’ and ‘TransferData()’ in ‘*Seurat*’. For
427 the human PBMC, we combined cell-type labels as follows: ‘Class-switched B’ and ‘B’ into ‘B’,
428 ‘CD8m T’, ‘CD4m T’, ‘CD4n T’, ‘CD4 T’, ‘gd T’ and ‘CD8eff T’ into ‘T’, ‘IgG PB’ and ‘IgA PB’
429 into ‘PB’, ‘pDC’ and ‘DC’ into ‘DC’, ‘RBC’, ‘Platelet’, ‘Activated Granulocyte’ and ‘SC & Eosinophil’
430 into ‘Others’. The gene symbol-matching between the scRNA-seq and the whole-organ RNA-seq data
431 was conducted by using entrez gene IDs derived from ‘org.Mm.egALIAS2EG’ in an R package,
432 ‘*org.Mm.eg.db*’. ERCC-labelled genes were removed as they are spike-in genes. In addition, the counts
433 of the three genes (*Rn45s, Akap5, Lrrc17*) were removed as they are non-mRNA artifacts significantly
434 influencing the total counts. The normalization was performed by adjusting the total counts of each cell
435 in the scRNA-seq datasets to a million. The same normalization process was also applied to each of the
436 whole-organ RNA-seq data.
437
438 **Calculation of reference cell-type ratios**
439 The reference cell-type ratios were as described in each original single-cell study, except for the heart
440 and the brain. For the heart, the cell-type ratios described in *Tabula Muris* (*20*) were based on the

441 separate analysis of the cardiac muscles and the non-muscle cells, resulting in a large under-estimation
442 of the cardiac muscle ratio – 3.1% (*20*). In contrast, the ratio determined by multiple other methods and
443 considered as the gold-standard in the cardiovascular fields is 30 - 40% (*55, 60-65*). According to this
444 gold-standard ratios, we made the following modifications: We set the ratio of the cardiac-muscle cells
445 at 30% and the remaining at 70%. The latter, then, was divided by the non-muscle cell types by
446 maintaining their ratios the same to those in the *Tabula Muris* data. For the brain, the cell-type ratio
447 reported in the NIH database (https://www.nervenet.org/papers/brainrev99.html#Numbers) was used to
448 modify those in the *Tabula Muris* scRNA-seq data (*20*), as follows. First, we divided the brain cell-type
449 classes into four classes: 'neurons', 'glial cells', 'endothelial cells', 'others', and then set the ratio of
450 these classes at 75 : 23 : 7 : 4, respectively, accordingly to their estimated ratios in the NIH murine brain
451 database (https://www.nervenet.org/papers/brainrev99.html#Numbers). Second, we further divided
452 'neurons', 'glial cells', and 'others' into more cell-type classes according to *Tabula Muris* (*20*). The
453 'neurons' were further divided into 'neuron-excitatory neurons and some neuronal stem cells' and
454 'neuron-inhibitory neurons'. The 'others' class was divided into 'brain pericyte-NA' and
455 'oligodendrocyte precursor cell-NA'. For the 'glial cells' class, we introduced three assumptions: 1) the
456 'glial cells' can be classified into four cell-types, 'microglial cell-NA', 'astrocyte-NA', 'Bergmann glial
457 cells-NA', and 'oligodendrocyte-NA', according to *Tabula Muris* (*20*); 2) the ratios of these four glial
458 cell types follow those of *Tabula Muris (20)* among them, and 3) the ratio of 'microglial cells-NA' in
459 the whole brain is 0.1, as it is reported that the 'microglial cells' account for 10 - 15 % of the whole
460 brain cells. On the basis of these assumptions, the ratios of each brain cell type was estimated as
461 'macrophage-NA' (ca. 0.2%), 'microglial cell-NA' (10.0%), 'astrocyte-NA' (ca. 2.2%), 'Bergmann
462 glial cell-NA' (ca. 2.1%), 'brain pericyte-NA' (ca. 1.5%), 'endothelial cell-NA' (ca. 6.4%), 'neuron-
463 excitatory neurons and some neuronal stem cells' (ca. 47.5%), 'neuron-inhibitory neurons' (ca. 21.3%),
464 oligodendrocyte-NA' (ca. 8.7%) and 'oligodendrocyte precursor cell-NA' (ca. 1.9%) and used as the
465 reference. For the human PBMC, the reference cell-type ratios are as described (*43, 45*), except for
466 those of "$CD16^+$ monocytes" and "PB and Others". The ratios of these cell-types are derived as follows:
467 The ratios of the cell-type label "the others" was divided to '$CD16^+$ monocytes' (ca. 7.21%), 'PB' (ca.
468 0.24%), 'Others' (ca. 3.77%), according to their RNA-seq data-based ratios.

469

470 **Selection of signature gene sets**

471 Random Forest (RF) was performed to extract signature genes for distinguishing cell-types from one
472 another in each organ using the scRNA-seq count data derived from the cell-type reference datasets
473 described in the above section. In this study, the '*randomForest*' package in R was used for tuning and
474 producing a classifier by RF. The scRNA-seq data were first divided by 8:2 into the training and the
475 test data. Using the training data, we determined two parameters, 'mtry' and 'ntree' for producing an
476 RF-model. The parameter 'mtry' means the number of features used for an RF-model and was tuned
477 by a function 'tuneRF()', while the parameter 'ntree' means the number of trees generated in an RF-
478 model and was set as 500, which is sufficient to converge error rate in cell-type classification. The
479 produced RF-model was validated by using the test data and F1-score calculated by a function
480 'F1_Score()' in an R package '*MLmetrics*'. Then we confirmed the validity of the model by F1-score
481 over 0.8. Following the classifier production, the important features in the classifier were extracted as
482 the signature genes for each cell-type, where we used 'Mean Decrease in Gini' values as the importance
483 indicator for each gene.

484

485 **Computation of the cell type-specific weighting-factors (cWFs)**

486 The overall concept for computing the cWFs is graphically described in Fig. 2 and is as follows: The
487 observed scRNA-seq counts are normalized (i.e., the total counts per cell is the same for all individual
488 cells). Hence, the putative differential total cellular transcriptome-size among various cell-types is lost
489 from the count data. The cWFs (w_1, w_2, w_3, etc.) offset such loss. In their computation, we calculated
490 at least 100 cWFs for each virtual-cell/cell-subject of each cell-type and made them follow Gaussian
491 distribution, instead of one cWF per cell-type. Based on this concept, we developed a model as follows:

492 $$m\mathbf{y} = \sum_j w_j \mathbf{x}_j \ \ (j \in C^m) \quad (1)$$

493 where $\mathbf{y} \in \mathbb{R}^n$, $w_j \geq 0$ and $\mathbf{x}_j \in \mathbb{R}^n$ denote the normalized whole-organ RNA-seq counts vector, the
494 cWF for each cell subject j, and the normalized scRNA-seq counts vector for each cell subject j,
495 respectively. The n is the total number of the signature genes for the organs. In this study, 100, 300,

and 500 signature genes were selected according to the ranking on the basis of 'Mean Decrease in Gini' values calculated by the RF analyses. The numbers, 100, 300, and 500 were arbitrarily selected but the smaller number of signature genes is computationally more cost-effective (see the details also in the section below). The combination of cell subjects, C^m was selected at random while maintaining the reference cell-type ratio described above with the total number of cell subjects, m, which is arbitrarily set to 100. In addition, we set C^m containing at least 1 cell subject in each cell-type. With this model, w_j was calculated by solving a quadratic problem under the constraint that the resulting value is non-negative as follows:

$$\widetilde{w}_j = \mathrm{argmin}_{w_j} \sum_i^S \left| m\mathbf{y}_i - \sum_j w_j \mathbf{x}_j \right|^2 \; s.t. \; w_j \geq 0 \tag{2}$$

Here, S is the number of whole-organ RNA-seq count data. This quadratic problem was solved by using '*osqp*' 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 subjects. In every recursion, Pearson correlation coefficients were calculated between the real whole-organ RNA-seq and the corresponding synthetic whole-organ RNA-seq generated with or without the weighting-factors. 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:

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

where, $\tilde{\mathbf{x}}_j$, μ_{w_j}, and $\sigma_{w_j}^2$ denote the weighted count vector for the cell subject j, the mean of the cWFs for the cell subject j, and the variance of the cWFs for the cell subject j, respectively. The 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:

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

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

Computing the cell-type ratios & gene expression patterns

The overall concept is graphically described as Fig. 5A and the model is as follows:

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

where, $\mathbf{y}$, $\mathbf{X} = [\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_k, \dots, \bar{\mathbf{x}}_K]$, 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, Gaussian noise $\mathcal{N}(\mathbf{0}, \beta\mathbf{I})$ was added to Equation (5) and a probabilistic model was developed as follows:

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

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

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

$$\mathrm{P}(\mathbf{r}) = Dirichlet(\mathbf{r}|\alpha\mathbf{I}) \tag{9}$$

where, α and $\mathbf{1}_K$ denote a hyperparameter and the K-dimensional all-ones vector. In prior distribution $\mathrm{P}(\mathbf{X})$, $\mathcal{N}\left(\bar{\mathbf{x}}_k \middle| \boldsymbol{\mu}'_k, \boldsymbol{\Sigma}_k^{-1}\right)$ 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 $\boldsymbol{\Sigma}_k^{-1} = \frac{1}{N_k}\sum_j \sigma_{w_j}^2 \mathrm{diag}\left(\mathbf{x}_j \odot \mathbf{x}_j\right)$. To maximize the posterior distribution, we employed an extension of Hamiltonian Monte Carlo (HMC) sampler, the No-U-Turn Sampler (NUTS) by a software, 'stan', which was performed on R program by an R package, '*rstan*', and used only signature gene sets for sampling, as the running time is reduced. The probabilities of hyperparameters, α and β were set as zero-truncated standard normal distribution. In addition, we set zero for the negative values in the

546 sampled elements of $\mathbf{X}$. The 'stan' sampling parameters, such as the number of iterations,
547 max_treedepth, adapt_delta, thin, etc., were appropriately chosen to converge well the sampled
548 distributions. The convergence was evaluated by R-hat values (Gelman-Rubin statistics) in all sampling
549 variables below 1.1, which is a generally accepted value for convergence evaluation. For the expression
550 pattern of non-signature genes, we set another noise parameter γ in Equation (6), instead of β and
551 estimated $\bar{\mathbf{x}}_k$ (the expression patterns in non-signature genes) by the following probability equation:

552 $$\mathrm{P}(\bar{\mathbf{x}}_k|\mathbf{y}, \bar{\mathbf{r}}, \{\bar{\mathbf{x}}_l\}_{l\neq k}) \propto \mathrm{P}(\mathbf{y}|\mathbf{X}, \bar{\mathbf{r}})\mathrm{P}(\bar{\mathbf{x}}_k) \quad (10)$$

553 where, $\bar{\mathbf{r}}$ denotes the mean values of $\mathbf{r}$ determined by the sampled distribution above.
554 Then, $\mathrm{P}(\bar{\mathbf{x}}_k|\mathbf{y}, \bar{\mathbf{r}}, \{\bar{\mathbf{x}}_l\}_{l\neq k})$ follows Gaussian distribution and its mean and variance can be calculated by
555 the following equations:

556 $$\mathbb{E}[\bar{\mathbf{x}}_k] = \left(\gamma \mathrm{r}_k^2\mathbf{I} + \boldsymbol{\Sigma}_k^{-1}\right)^{-1}\left(\gamma \mathrm{r}_k\mathbf{y} + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k' - \gamma\sum\nolimits_{l\neq k}\mathrm{r}_k\mathrm{r}_l\bar{\mathbf{x}}_l\right) \quad (11)$$

557 $$\mathrm{Var}[\bar{\mathbf{x}}_k] = \gamma \mathrm{r}_k^2\mathbf{I} + \boldsymbol{\Sigma}_k^{-1} \quad (12)$$

558 In the resulting $\mathbb{E}[\bar{\mathbf{x}}_k]$, all negative values were set to zero. The hyperparameter γ was set as either of
559 $10^{-5}, 10^{-4}, \ldots, 10^{5}$. The estimating calculation was recursively performed until the root mean square
560 error (RMSE) of $\bar{\mathbf{x}}$ between iterations gets below 1 or until at most 100 times enough to converge, where
561 one iteration is the estimating calculations of Equation (11) and (12) for all k (i.e., all cell-types) and
562 for each iteration, $\bar{\mathbf{x}}_l$ is from the prior iteration. The optimal value of the hyperparameter γ was
563 determined by the least RMSE between the original whole-organ RNA-seq (i.e., $\mathbf{y}$) and the estimated
564 one (i.e., $\bar{\mathbf{y}}$).
565

566 **Selection of the optimal number of the signature genes**

567 The optimal number (100, 300, or 500) of the signature genes were determined based on the three
568 criteria as follows: 1) Improvement of the Pearson correlation coefficient between the r-woRNAseq and
569 the syn-woRNAseq(+cWF), 2) Convergence in the deconvolution, 3) Higher similarity to the reference
570 cell type composition (i.e., lower RMSE). The results with the selected optimal number of the signature
571 genes were used in the comparisons with those obtained by MuSiC and DWLS, and also for the analyses
572 with the aging data and for the cWFs analysis of the human PBMC data described below.

573

574 **MuSiC and DWLS**

575 MuSiC was conducted as described (*41*). DWLS was performed as described (*40*), except that a function,
576 'solve.QP()' (R package: '*quadprog*') was substituted by 'solve_osqp()' (R package: '*osqp*') for
577 quadratic problem solver.

578

579 **Comparison of the real and virtual scRNA-seq counts of 10 organs**

580 The mean counts of each cell-type in each organ were compared by Pearson correlation coefficients.
581 For the virtual scRNA-seq, the estimated mean count for each cell-type was normalized to a million.
582 For the real scRNA-seq, the mean count for each individual cell was normalized to a million and
583 averaged for each cell-type.

584

585 **Analysis of differential cWFs in aging progression**

586 The distribution of the pairwise cWF ratios was calculated for each aging-stage. The ratios of the
587 pairwise cWFs in each cell-type, instead of the cWFs themselves, were used as the cWFs are relative
588 values calculated from the normalized count datasets of the real whole-organ RNA-seq and thus the
589 cWFs values themselves can be used only for their comparison among the cell-types within each organ
590 at each aging-stage (i.e., the cWFs cannot be compared across the aging-stages). Hence, their pairwise
591 ratios were used for the comparisons across the aging-stages. For each pairwise cell-type within an
592 organ, the distribution of the cWF ratios was calculated for all pair combinations of the consisting cells.
593 For example, assuming that cell-type X has 1,000 cells and cell-type Y has 500 cells with one cWF per
594 cell, respectively, the distribution of the ratio of cWFs of cell-type X vs. Y consisting of 500,000 values
595 of the ratios was calculated. When a cell-type contains zero value of cWF, we substituted the zero value
596 for the second minimum cWF value in the cell-type to avoid the division by zero. Next, for each cell-
597 types pair, we performed Mann-Whitney U-test to make statistical evaluations of the differences of
598 their ratio-distributions between the aging-stages (i.e., 18 mos. vs. 3 mos., 24 mos. vs. 3 mos., 24 mos.
599 vs. 18 mos.). Zeros were generated if the resulting values were below the lower bound of R program,
600 1e-308.

601

602 **Machine learning prediction of the aging-stages by the differential cWF ratios**

603 For this analysis, we omitted the data of the cell-types pairs of which statistical analysis described above
604 resulted in $p \geq 0.01$ across all pairs of the aging-stages. Using the remaining data, we performed
605 LightGBM classifier analysis (*49*). We first generated input data for LightGBM classifier analysis using
606 the cWF ratio distributions obtained as described in the above section. The 100 cWF ratio distribution-
607 values were randomly sampled for each cell-types pair for each aging-stage. The sampled values were
608 at random combined with each other, resulting in the matrix of "the 100 cWFs ratio distribution values"
609 x "the number of the cell-type pairs" for each aging-stage. Using these matrices for all three aging-
610 stages as the input data, we generated a LightGBM classifier of aging-stages on Python, using the
611 '*lightgbm*' package (*49*). The input data was divided by 8:2 into training data and test data. The training
612 data was used for tuning the parameters of LightGBM. The tuning was performed by the '*optuna*'
613 package (*66*), searching for the best parameters to maximize macro-F1 score. Specifically, the searched
614 parameters are max_depth, num_leaves, subsample, subsample_freq, colsample_bytree,
615 min_child_samples, ranging from 3 to 12, 2 to 256, 0.1 to 1.0, 1 to 7, 0.1 to 1.0, and 5 to 100,
616 respectively. The model was generated using the best parameters and evaluated using test data by
617 accuracy, macro-Precision, macro-Recall, and macro-F1_score. These evaluations were performed by
618 a Python package, 's*cikit-learn*' (*67*). In addition, we also extracted feature importance based on 'gain'
619 from the model and confirmed the high-ranked contributors in the model (i.e., aging-stage
620 classification).

621

622 **References**

623 1. J. E. Coate, J. J. Doyle, Variation in transcriptome size: are we getting the message?
624 *Chromosoma* **124**, 27-43 (2015).

625 2. J. Loven *et al.*, Revisiting global gene expression analysis. *Cell* **151**, 476-482 (2012).

626 3. G. K. Marinov *et al.*, From single-cell to cell-pool transcriptomes: stochasticity in gene
627 expression and RNA splicing. *Genome Res* **24**, 496-510 (2014).

628 4. S. Marguerat, J. Bahler, Coordinating genome expression with cell size. *Trends in genetics :*
629 *TIG* **28**, 560-565 (2012).
630 5. T. P. Miettinen *et al.*, Identification of transcriptional and metabolic programs related to
631 mammalian cell size. *Current biology : CB* **24**, 598-608 (2014).
632 6. J. Zhurinsky *et al.*, A coordinated global control over cellular transcription. *Current biology :*
633 *CB* **20**, 2010-2015 (2010).
634 7. E. E. Schmidt, U. Schibler, Cell size regulation, a mechanism that controls cellular RNA
635 accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-
636 Y and the liver-enriched transcription factor DBP. *The Journal of cell biology* **128**, 467-483
637 (1995).
638 8. C. Y. Lin *et al.*, Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56-
639 67 (2012).
640 9. Z. Nie *et al.*, c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic
641 stem cells. *Cell* **151**, 68-79 (2012).
642 10. Y. Li *et al.*, Global transcriptional and translational repression in human-embryonic-stem-cell-
643 derived Rett syndrome neurons. *Cell stem cell* **13**, 446-458 (2013).
644 11. M. G. Chandler, R. H. Pritchard, The effect of gene concentration and relative gene dosage on
645 gene output in Escherichia coli. *Mol Gen Genet* **138**, 127-141 (1975).
646 12. J. E. Coate, J. J. Doyle, Quantifying whole transcriptome size, a prerequisite for understanding
647 transcriptome evolution across species: an example from a plant allopolyploid. *Genome Biol*
648 *Evol* **2**, 534-546 (2010).
649 13. Z. Hu *et al.*, Nucleosome loss leads to global transcriptional up-regulation and genomic
650 instability during yeast aging. *Genes Dev* **28**, 396-408 (2014).
651 14. K. Zaitsev, M. Bambouskova, A. Swain, M. N. Artyomov, Complete deconvolution of cellular
652 mixtures based on linearity of transcriptional signatures. *Nat Commun* **10**, 2209 (2019).
653 15. S. Cao *et al.*, Estimation of tumor cell total mRNA expression in 15 cancer types predicts
654 disease progression. *Nature biotechnology* **40**, 1624-1633 (2022).

655 16. Q. Deng, D. Ramskold, B. Reinius, R. Sandberg, Single-cell RNA-seq reveals dynamic,
656 random monoallelic gene expression in mammalian cells. *Science* **343**, 193-196 (2014).
657 17. X. Han *et al.*, Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107 e1017
658 (2018).
659 18. A. Regev *et al.*, Science Forum: The Human Cell Atlas. *Elife* **6**, (2017).
660 19. R. Sandberg, Entering the era of single-cell transcriptomics in biology and medicine. *Nature*
661 *methods* **11**, 22-24 (2014).
662 20. C. Tabula Muris *et al.*, Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris.
663 *Nature* **562**, 367-372 (2018).
664 21. A. Adil, V. Kumar, A. T. Jan, M. Asger, Single-Cell Transcriptomics: Current Methods and
665 Challenges in Data Acquisition and Analysis. *Front Neurosci* **15**, 591122 (2021).
666 22. A. Haque, J. Engel, S. A. Teichmann, T. Lonnberg, A practical guide to single-cell RNA-
667 sequencing for biomedical research and clinical applications. *Genome Med* **9**, 75 (2017).
668 23. B. Hwang, J. H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics
669 pipelines. *Exp Mol Med* **50**, 1-14 (2018).
670 24. K. Vandereyken, A. Sifrim, B. Thienpont, T. Voet, Methods and applications for single-cell
671 and spatial multi-omics. *Nature reviews. Genetics*, 1-22 (2023).
672 25. C. G. Williams, H. J. Lee, T. Asatsuma, R. Vento-Tormo, A. Haque, An introduction to spatial
673 transcriptomics for biomedical research. *Genome Med* **14**, 68 (2022).
674 26. E. Jonasson *et al.*, Total mRNA Quantification in Single Cells: Sarcoma Cell Heterogeneity.
675 *Cells* **9**, (2020).
676 27. J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, D. Gfeller, Simultaneous enumeration
677 of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6**, (2017).
678 28. A. Raj, S. A. Rifkin, E. Andersen, A. van Oudenaarden, Variability in gene expression underlies
679 incomplete penetrance. *Nature* **463**, 913-918 (2010).
680 29. A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, H. F. Clark, Deconvolution of blood
681 microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS*
682 *one* **4**, e6098 (2009).

683 30. F. Avila Cobos, J. Vandesompele, P. Mestdagh, K. De Preter, Computational deconvolution of
684 transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969-1979 (2018).
685 31. R. Gaujoux, C. Seoighe, Semi-supervised Nonnegative Matrix Factorization for gene
686 expression deconvolution: a case study. *Infect Genet Evol* **12**, 913-921 (2012).
687 32. T. Gong *et al.*, Optimal deconvolution of transcriptional profiling data using quadratic
688 programming with application to complex clinical blood samples. *PloS one* **6**, e27156 (2011).
689 33. T. Gong, J. D. Szustakowski, DeconRNASeq: a statistical framework for deconvolution of
690 heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083-1085 (2013).
691 34. B. Li *et al.*, Comprehensive analyses of tumor immunity: implications for cancer
692 immunotherapy. *Genome biology* **17**, 174 (2016).
693 35. A. M. Newman *et al.*, Robust enumeration of cell subsets from tissue expression profiles.
694 *Nature methods* **12**, 453-457 (2015).
695 36. D. Repsilber *et al.*, Biomarker discovery in heterogeneous tissue samples -taking the in-silico
696 deconfounding approach. *BMC bioinformatics* **11**, 27 (2010).
697 37. S. S. Shen-Orr, R. Gaujoux, Computational deconvolution: extracting cell type-specific
698 information from heterogeneous samples. *Curr Opin Immunol* **25**, 571-578 (2013).
699 38. N. Wang *et al.*, UNDO: a Bioconductor R package for unsupervised deconvolution of mixed
700 gene expressions in tumor samples. *Bioinformatics* **31**, 137-139 (2015).
701 39. Y. Zhong, Y. W. Wan, K. Pang, L. M. Chow, Z. Liu, Digital sorting of complex tissues for cell
702 type-specific gene expression profiles. *BMC bioinformatics* **14**, 89 (2013).
703 40. D. Tsoucas *et al.*, Accurate estimation of cell-type composition from gene expression data. *Nat*
704 *Commun* **10**, 2975 (2019).
705 41. X. Wang, J. Park, K. Susztak, N. R. Zhang, M. Li, Bulk tissue cell type deconvolution with
706 multi-subject single-cell expression reference. *Nat Commun* **10**, 380 (2019).
707 42. K. Kang *et al.*, CDSeq: A novel complete deconvolution method for dissecting heterogeneous
708 samples using gene expression data. *PLoS computational biology* **15**, e1007510 (2019).
709 43. F. Finotello *et al.*, Molecular and pharmacological modulators of the tumor immune contexture
710 revealed by deconvolution of RNA-seq data. *Genome Med* **11**, 34 (2019).

711 44. A. J. Wilk *et al.*, A single-cell atlas of the peripheral immune response to severe COVID-19.
712 *medRxiv*, (2020).
713 45. F. Finotello *et al.*, Correction to: Molecular and pharmacological modulators of the tumor
714 immune contexture revealed by deconvolution of RNA-seq data. *Genome Med* **11**, 50 (2019).
715 46. N. Schaum *et al.*, Ageing hallmarks exhibit organ-specific temporal signatures. *Nature* **583**,
716 596-602 (2020).
717 47. C. Tabula Muris, A single-cell transcriptomic atlas characterizes ageing tissues in the mouse.
718 *Nature* **583**, 590-595 (2020).
719 48. S. Dutta, P. Sengupta, Men and mice: Relating their ages. *Life Sci* **152**, 244-248 (2016).
720 49. G. Ke *et al.*, LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in*
721 *Neural Information Processing Systems* **30**, (2017).
722 50. S. C. van den Brink *et al.*, Single-cell sequencing reveals dissociation-induced gene expression
723 in tissue subpopulations. *Nature methods* **14**, 935-936 (2017).
724 51. P. Anversa, G. Olivetti, M. Melissari, A. V. Loud, Stereological measurement of cellular and
725 subcellular hypertrophy and hyperplasia in the papillary muscle of adult rat. *Journal of*
726 *molecular and cellular cardiology* **12**, 781-795 (1980).
727 52. J. Dammrich, U. Pfeifer, Cardiac hypertrophy in rats after supravalvular aortic constriction. I.
728 Size and number of cardiomyocytes, endothelial and interstitial cells. *Virchows Arch B Cell*
729 *Pathol Incl Mol Pathol* **43**, 265-286 (1983).
730 53. T. Mattfeldt, K. L. Kramer, R. Zeitz, G. Mall, Stereology of myocardial hypertrophy induced
731 by physical exercise. *Virchows Arch A Pathol Anat Histopathol* **409**, 473-484 (1986).
732 54. Y. Tang, J. R. Nyengaard, J. B. Andersen, U. Baandrup, H. J. Gundersen, The application of
733 stereological methods for estimating structural parameters in the human heart. *Anat Rec*
734 *(Hoboken)* **292**, 1630-1647 (2009).
735 55. P. Zhou, W. T. Pu, Recounting Cardiac Cellular Composition. *Circulation research* **118**, 368-
736 370 (2016).
737 56. S. Kozawa *et al.*, Predicting Human Clinical Outcomes using Mouse Multi-Organ
738 Transcriptome. *iScience* **In press**, (2019).

739 57. S. Kozawa *et al.*, The Body-wide Transcriptome Landscape of Disease Models. *iScience* **2**,
740 238-268 (2018).
741 58. S. Joost *et al.*, The Molecular Anatomy of Mouse Skin during Hair Growth and Rest. *Cell stem*
742 *cell* **26**, 441-457 e447 (2020).
743 59. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529
744 (2021).
745 60. I. Banerjee, J. W. Fuseler, R. L. Price, T. K. Borg, T. A. Baudino, Determination of cell types
746 and numbers during cardiac development in the neonatal and adult rat and mouse. *American*
747 *journal of physiology. Heart and circulatory physiology* **293**, H1883-1891 (2007).
748 61. O. Bergmann *et al.*, Dynamics of Cell Generation and Turnover in the Human Heart. *Cell* **161**,
749 1566-1575 (2015).
750 62. A. C. Nag, Study of non-muscle cells of the adult mammalian heart: a fine structural analysis
751 and distribution. *Cytobios* **28**, 41-61 (1980).
752 63. A. R. Pinto *et al.*, Revisiting Cardiac Cellular Composition. *Circulation research* **118**, 400-409
753 (2016).
754 64. A. Raulf *et al.*, Transgenic systems for unequivocal identification of cardiac myocyte nuclei
755 and analysis of cardiomyocyte cell cycle status. *Basic Res Cardiol* **110**, 33 (2015).
756 65. S. Walsh, A. Ponten, B. K. Fleischmann, S. Jovinge, Cardiomyocyte cell cycle control and
757 growth estimation in vivo--an analysis based on cardiomyocyte nuclei. *Cardiovascular*
758 *research* **86**, 365-373 (2010).
759 66. T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation
760 Hyperparameter Optimization Framework. *Proceedings of the ACM SIGKDD International*
761 *Conference on Knowledge Discovery and Data Mining*, 2623-2631 (2019).
762 67. F. Pedregosa Fabianpedregosa *et al.*, Scikit-learn: Machine Learning in Python. *Journal of*
763 *Machine Learning Research* **12**, 2825-2830 (2011).
764
765
766

Acknowledgements

We thank K. Sugisaka, R. Takahashi, T. Ninomiya, R. Kitaura, R. Ishikawa, S. Taniyama for their administrative and laboratory management assistance. We are also grateful to the members of Karydo TherapeutiX, Inc. and Sato laboratory at ATR for advice and discussion throughout the course of this work. **Funding:** This work was funded by Innovative Science and Technology Initiative for Security Grant Number JPJ004596 ATLA Japan (T.N.S.), JST ERATO Grant Number JPMJER1303 Japan (T.N.S.), Nakatani Foundation (T.N.S.). **Author contributions:** Conceptualization: TNS, Methodology: KT, SK, Investigation: TNS, KT, SK, Visualization: KT, Supervision: TNS, Writing-original draft: TNS, KT, Writing-review & editing: TNS, KT, SK. **Completing interests:** T.N.S. is the inventor of the patents filed on the cell type-specific weighting-factors and holds shares of Karydo TherapeutiX, Inc., which owns the right to this patent. K.T. and S.K. are also employed by Karydo TherapeutiX, Inc. Correspondence and requests for materials should be addressed to T.N.S. (island1005@gmail.com). **Data and materials availability:** The public datasets are as indicated and the raw data are available as supplementary tables in this paper. The code for our algorithm is freely available (https://github.com/k-teji/VscRNAseq).