Cell type-specific weighting-factors to solve solid organs-specific limitations of single cell RNA-sequencing

Short title: Cell type-specific weighting-factors for solid organ scRNA-seq

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

While single-cell RNA-sequencing (scRNA-seq) is a popular method to analyze gene expression and cellular composition at single-cell resolution, it harbors shortcomings: The failure to account for cell-to-cell variations of transcriptome-size (i.e., the total number of transcripts per cell) and also cell dissociation/processing-induced cryptic gene expression. This is particularly a problem when analyzing highly heterogeneous solid tissues/organs, which requires cell dissociation for the analysis. As a result, there exists a discrepancy between bulk RNA-seq result and virtually reconstituted bulk RNA-seq result using its composite scRNA-seq data. To fix this problem, we propose a computationally calculated coefficient, "cell type-specific weighting-factor (cWF)". Here, we introduce a concept and a method of its computation and report cWFs for 76 cell-types across 10 solid organs. Their fidelity is validated by more accurate reconstitution and deconvolution of bulk RNA-seq data of diverse solid organs using the scRNA-seq data and the cWFs of their composite cells. Furthermore, we also show that cWFs effectively predict aging-progression, implicating their diagnostic applications and also their association with aging mechanism. Our study provides an important method to solve critical limitations of scRNA-seq analysis of complex solid tissues/organs. Furthermore, our findings suggest a diagnostic utility and biological significance of cWFs.

Abbreviations: cWF, cell type-specific weighting factor; r-woRNAseq, real whole-organ RNA-seq; syn-woRNAseq, synthetic whole-organ RNA-seq; V-scRNAseq, virtual single cell RNA-seq.

Introduction

Since the advent of single-cell RNA-sequencing (scRNA-seq) technologies early this century, this method is becoming one of the essential tools in biomedical fields [1]. While scRNA-seq provides useful information of cellular composition and gene expression at single cell resolution, it has limitations.

It is well-known that the total number of RNA molecules per cell, also known as transcriptome-size, varies from-cell-to-cell [2-4]. However, in scRNA-seq analyses, the transcripts are sequenced in individual cells and their transcript-counts are normalized, allowing the comparison of the relative abundance of each transcript-species across individual cells/cell-types [5-9]. This internal normalization process of the analyses cancels out the putative transcriptome size-variations among different cells, concealing potentially important functional differences among the cells.

This differential transcriptome-size can be determined by counting the library depth of the cells and/or by including normalizing spike-in RNA in the samples [2, 3, 10, 11]. Alternative, it can be quantified by PCR-based methods [12, 13]. However, these approaches are applicable only to culture cells, blood/immune cells, and cancer cells which are relatively uniform cell types and do not require cell dissociation. In contrast, they are not applicable to solid tissues/organs composed of highly heterogeneous cell types. Individual cells can be dissociated from solid tissues/organs and then analyzed by scRNA-seq. However, the cell-dissociation and subsequent harvesting and/or purification steps induce cryptic gene expression [14]. Due to such limitations, it is difficult to measure and compare the absolute transcript-counts of each gene in individual cells of the complex solid tissues/organs such as the heart, brain, kidney, etc.

Furthermore, the differential transcriptome-size has biological importance. Several studies indicate that the cell-size influences the overall cellular transcription [15-18]. It is also

69 reported that c-myc amplifies the global transcription in tumor cells [2, 19], lymphocytes and
70 embryonic stem cells [20]. Another study shows that the global transcription is repressed by
71 the loss of MECP2, a methyl CpG binding protein, in embryonic stem cell-derived neurons
72 [21]. The genome structure, such as ploidy/gene-dosage and nucleosome state, affects the
73 overall transcription [12, 22, 23]. In additions, it is shown that the total mRNA-contents of
74 tumor cells are associated with the extent of the cancer progression [24]. These results illustrate
75 biological importance of accounting for the transcriptome size differences among cells.

76 In addition to the transcriptome-size, cell dissociation/processing-induced cryptic gene
77 expression is another problem with scRNA-seq, in particular with solid tissues/organs [14].
78 The scRNA-seq analyses of the solid tissues and organs require enzymatic digesting and
79 mechanical dissociation of individual cells. Spatial scRNA-seq requires tissue processing such
80 as sectioning and laser-mediated harvesting [8, 9]. Such technical procedures induce cryptic
81 gene expression [14]. The currently available scRNA-seq largely disregard this problem.

82 In this study, we address and overcome these limitations/problems associated with solid
83 tissues/organs by developing a method to compute coefficients, cell type-specific weighting-
84 factors (cWFs), to offset the transcriptome size- and cell dissociation/processing-induced
85 problem. Furthermore, we also calculate cWFs for 76 cell-types across 10 solid organs and use
86 them to illustrate their diagnostic and biological utilities.

87

88 Results

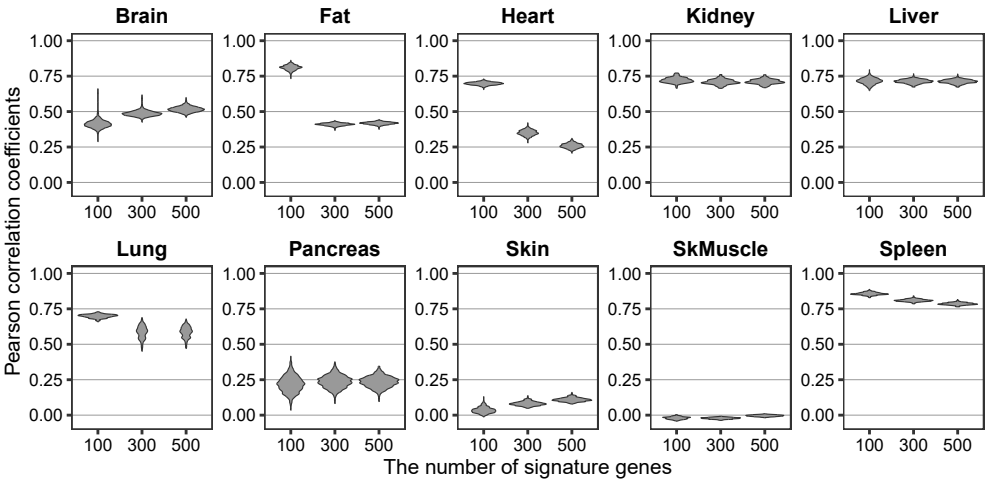
89 The cell type-specific weighting factor (cWF)

90 First, the major premise is follows:

91 If the transcriptome size differences and the cell dissociation/processing-induced cryptic gene
92 expression (abbreviated as “cryptic gene expression” hereafter for convenience) are corrected

93 in the conventional scRNA-seq, the bulk (i.e., tissue/organ) RNA-seq should be the sum of
94 gene expression of each composite cell-type weighted by cell-type ratios.

95 We evaluated this premise. The synthetic whole-organ data were constructed from
96 single-cell transcript counts combined according to the experimentally determined fraction of
97 each cell-type in the organ (see Materials and methods). Briefly, the synthetic whole-organ
98 RNA-seq data are calculated as the sum of the normalized transcript counts weighted by the
99 known ratios of the composite cell-types of the organs, using the signature genes for
100 distinguishing the cell-types for each organ. The signature genes were determined by the
101 random forest classifier. These synthetic data for 10 organs (brain, fat, heart, kidney, liver, lung,
102 pancreas, skin, skeletal muscle, spleen) were then compared to the corresponding real whole-
103 organ RNA-seq data (Fig 1 and S1 Table). The comparisons were performed by calculating
104 their Pearson correlation coefficients.



105

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

110

111 The result shows low Pearson correlation coefficients (< 0.75) for most of the organs
112 (brain, heart, kidney, liver, lung, pancreas, skin, skeletal muscle) regardless of the number (100,
113 300, 500) of the top-ranked signature genes used. Particularly low coefficients are found with
114 the brain (< 0.6 for 100, 300, 500 signature genes), the fat (< 0.5 for 300, 500 signature genes),
115 the heart (< 0.4 for 300, 500 signature genes), the pancreas (< 0.3 for 100, 300, 500 signature
116 genes), the skin (< 0.2 for 100, 300, 500 signature genes), and the skeletal muscle (< 0.1 for
117 100, 300, 500 signature genes). These low Pearson correlation coefficients suggest that the
118 transcriptome-size of individual cell-types varies, resulting in the large gaps.

119 These results fail to satisfy the above-described major premise; hence, the differential
120 transcriptome size-variations and/or the cryptic gene expression are unrepresented in the
121 scRNA-seq data of the solid organs. To solve this problem, we developed an algorithm to
122 compute such unrepresented factors (Fig 2 and see Materials and methods). The concept of the
123 algorithm is schematically described in Fig 2 and as follows: In the real scRNA-seq data
124 analyses, the total transcript counts per cell (i.e., the transcriptome-size) of the real scRNA-seq
125 data are normalized (i.e., the total counts per cell are equal among all composite cell-types of
126 the organs). This results in the loss of a factor representing the putative differences of the
127 transcriptome-size and/or the cryptic gene expression across cell-types. This factor is computed
128 as the cell type-specific weighting-factor (cWF) (indicated as w_1, w_2, w_3, etc. in the Fig 2).

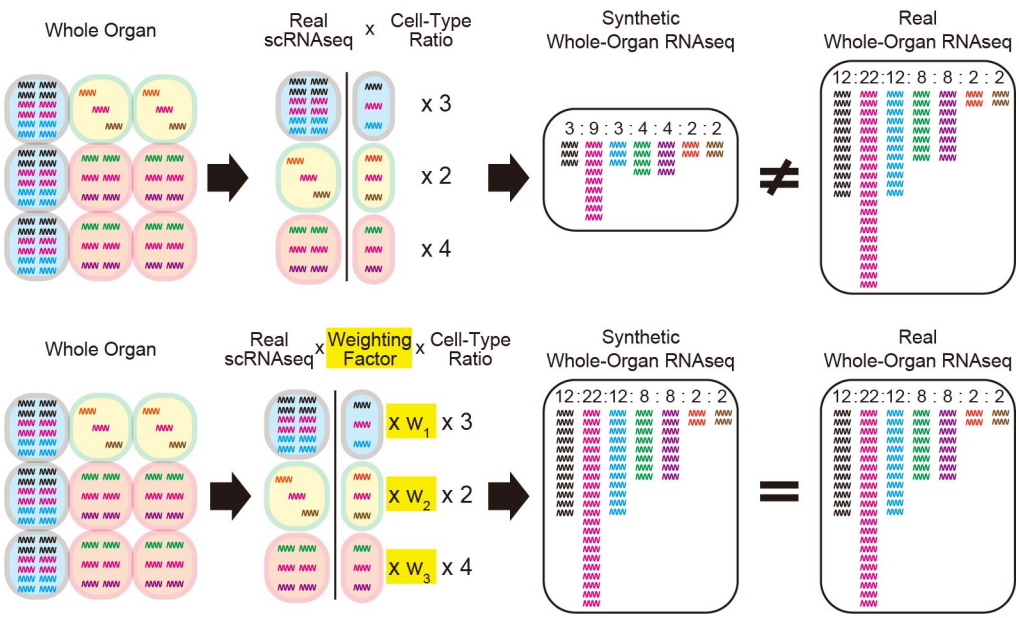


Fig 2. Graphical description of the overall concept of the cell type-specific weighting factors (cWFs). The top row describes the case without the cWFs, failing to reconstitute the real whole-organ RNA-seq data by the synthetic whole-organ RNA-seq data generated by assembling the composite scRNA-seq data without the cWFs. The bottom describes the case with the cWFs, accurately reconstitute the real whole-organ RNA-seq data by the synthetic whole-organ RNA-seq 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 RNA-seq are indicated by the numbers at the top of the colored transcripts.

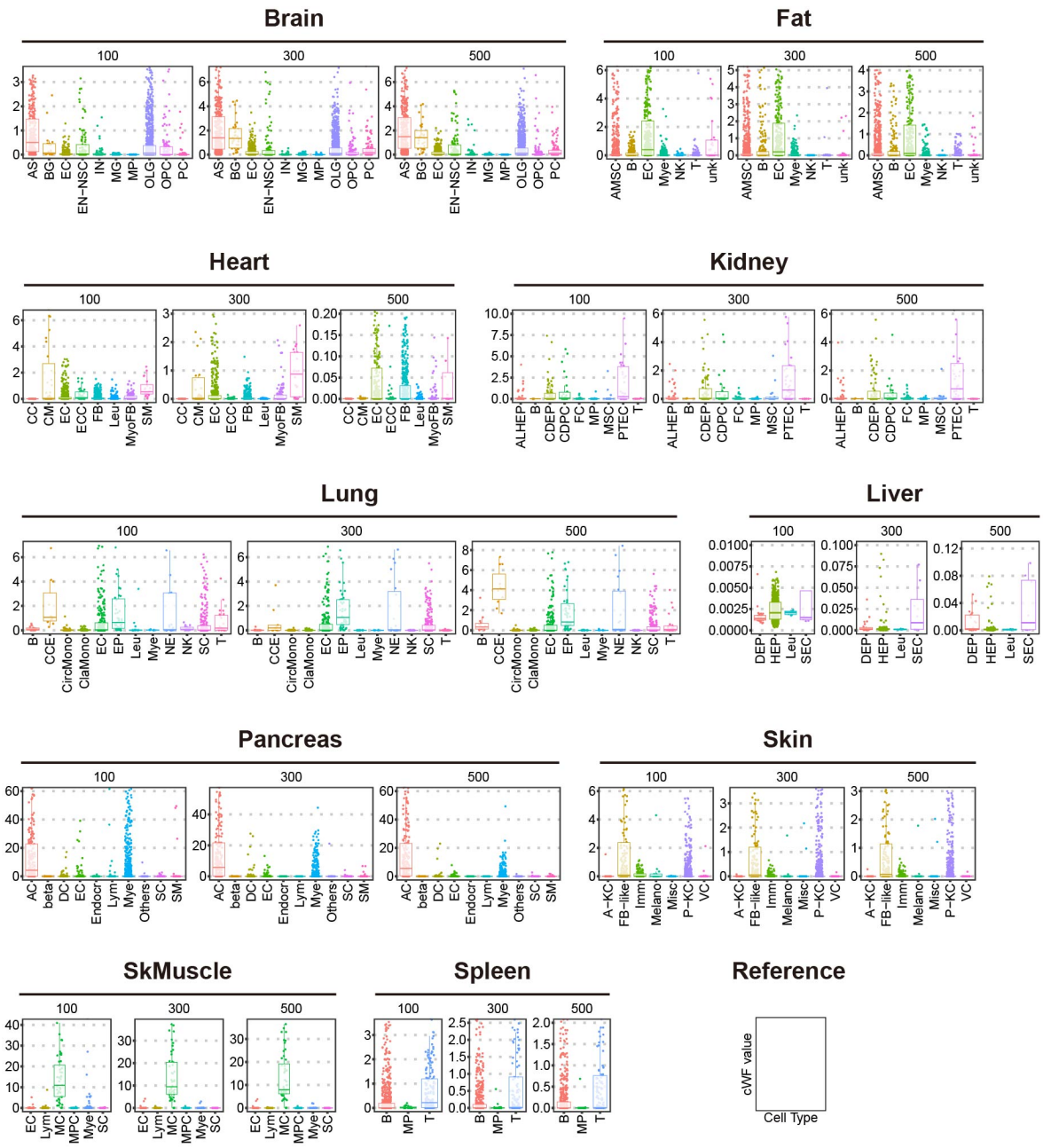


Fig 3. The cWFs for each cell-type in each organ. The values of the cWFs for each cell-type (bottom of each plot) for each number of the signature genes (indicated above each plot) for each organ (indicated at the top of the corresponding plots) are shown as box plots. The corresponding raw data are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

Using this algorithm, we computed cWFs for 76 cell-types across 10 organs (Fig 3 and S2 Table). The result shows significant variations of the cWFs among different cell-types in

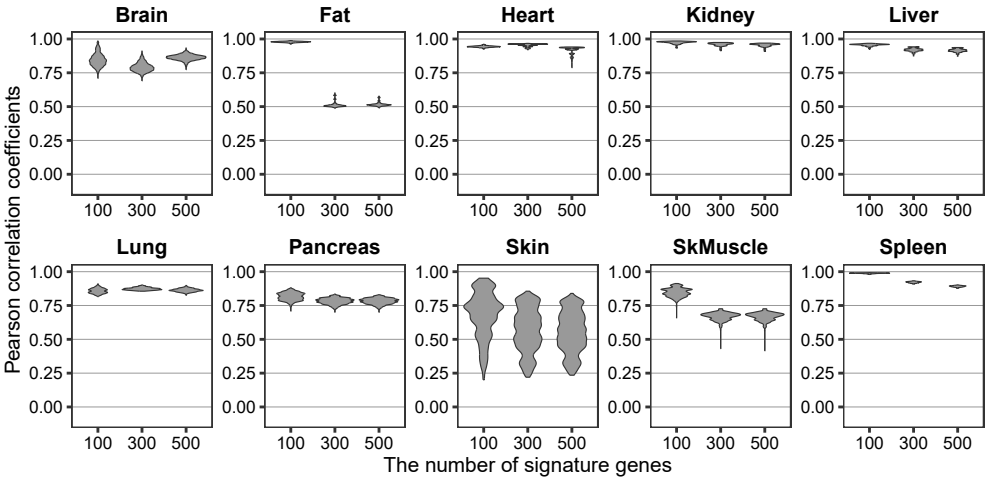
147 many of the organs, uncovering the body-wide degree of such variations. In particular, the
148 following cell-types show significantly higher cWFs than the others within the same organ:
149 astrocyte (AS) in the brain, endothelial cell (EC) in the fat, muscle cell-types in the heart and
150 the skeletal muscle (CM in the heart, MC in the skeletal muscle), epithelial cell of proximal
151 tubule (PTEC) in the kidney, epithelial cell (EP) and neuroendocrine cell (NE) in the lung,
152 fibroblast-like cell (FB-like) in the skin, and T-cell (T) in the spleen. This result suggests that
153 these cell-types contain larger amounts of transcripts (i.e., the larger transcriptome-size) than
154 the others in the corresponding organs. Furthermore, the results show that such variable degrees
155 of differences in the transcriptome-size are widespread across diverse cell-types and organs.
156

157 # Validation of cWFs

158 We validate the computed cWFs by three methods: Reconstitution (Fig 4), deconvolution (Figs
159 5 and 6) of the whole-organ RNAseq data using their composite scRNA-seq data, and the
160 comparison of the cWFs to the experimentally measured transcript-contents (Fig 7).

161 First, we validated the cWFs by the reconstitution (Fig 4 and S3 Table, and see
162 Materials and methods). Using the mean value of cWFs for each cell-type, we weighted the
163 transcript counts for each cell-type. We then summed these weighted transcript counts for all
164 composite cell-types of each organ according to their ratios to construct the synthetic whole-
165 organ RNA-seq (syn-woRNAseq(+cWF)), data. Then, they are compared to the corresponding
166 real whole-organ RNA-seq (r-woRNAseq) data. The result shows the Pearson correlation
167 coefficients of 0.8 – 1.0 between the syn-woRNAseq(+cWF) and r-woRNAseq for all 10
168 organs with all or at least one of the 100, 300, 500 signature genes (Fig 4). These are significant
169 improvements from the 0 – 0.75 of the Pearson correlation coefficients without the cWFs (Fig
170 1). The heart shows the Pearson correlation coefficients of almost 1.0 with the cWFs (Fig 4),
171 as compared to those of 0.25 – 0.7 without the cWFs (Fig 1). Even for the spleen where the

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



178

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

183

184 Next, the validation by the deconvolution (Fig 5) was performed. We developed a
185 deconvolution method that integrates cWFs in a Bayesian framework, as schematically
186 described in Fig 5A.

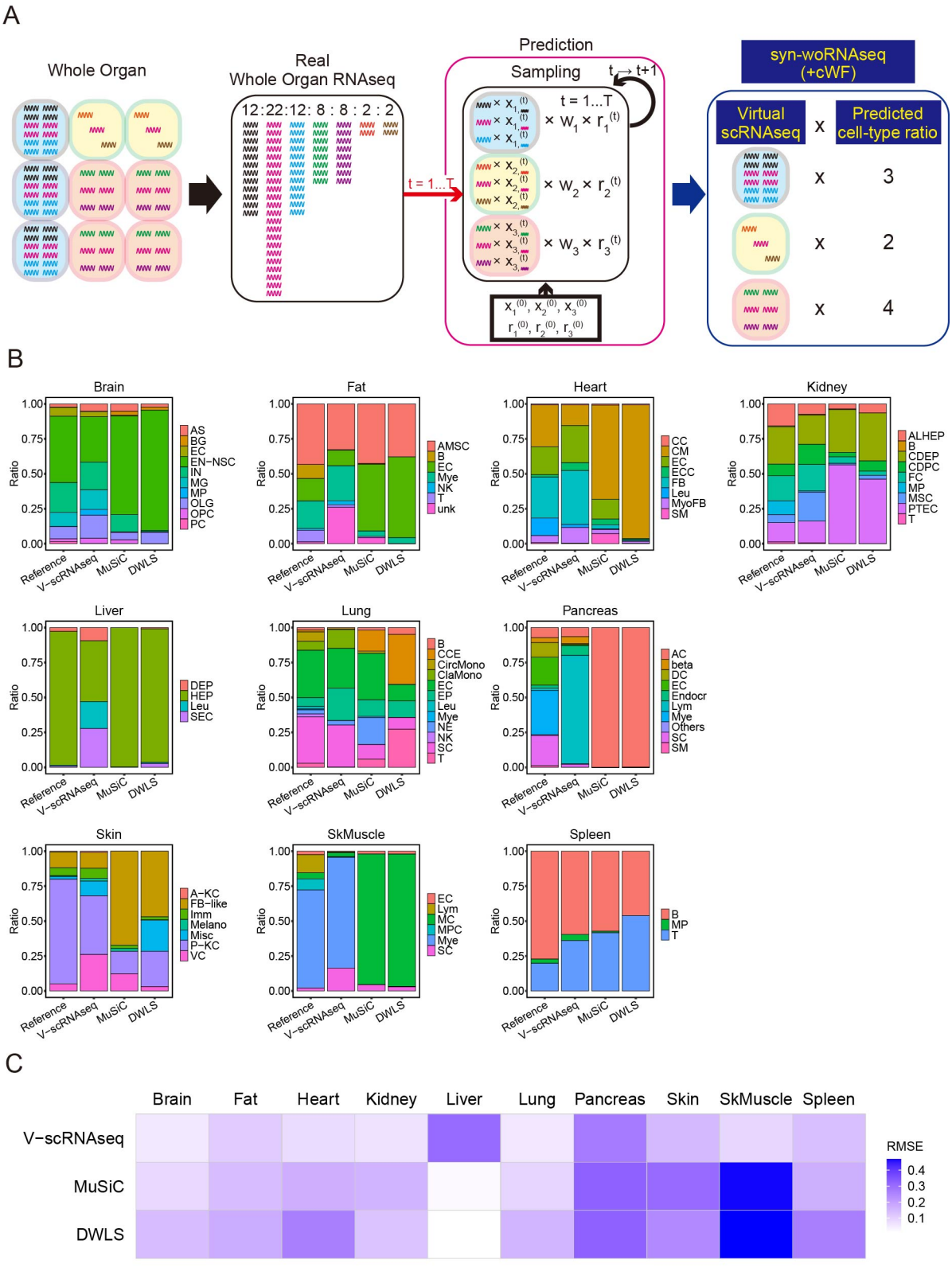


Fig 5. Accurate deconvolution of the whole organ RNA-seq by cWFs. (A) Graphical description of the overall concept of the deconvolution method with the cWFs. Shown is the case of an organ consisting of 3 cell-types (1, 2, 3) with 7 different genes (wiggling lines of different colors). The mean of the 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 (see 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 S4 Table). 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 S4 Table. SkMuscle: skeletal muscle.

Briefly, it works as follows (see also Materials and methods): The mean of the cWF (w_1, w_2, w_3 in Fig 5A, the lower case 1, 2, 3 indicate cell-type 1, 2, 3) is used to compute virtual transcript counts ("x" in Fig 5A) for each gene (indicated as wiggling lines of different colors in Fig. 5A) and the ratio (indicated as "r" in Fig 5A) of each cell-type in each organ (painted in different colors in Fig 5A). Using the signature-gene set, the computation is iteratively performed by the sampling method of No-U-Turn Sampler (NUTS), in the Bayesian framework with two hyperparameters, α and β, to account for the combinatorial influence of the cell-type ratios and the gene expression patterns of each cell-type on the whole-organ-level transcriptome, respectively. For the iteration (t = 1, …, T), the initial condition is set as $x_1^{(0)}$ and $r_1^{(0)}$ for the transcript counts and the ratio of cell-type 1 (see the diagram labels as "Prediction" in Fig 5A). Those for cell-types 2 and 3 are indicated accordingly (see the diagram

219 labels as "Prediction" in Fig 5A). Such initial conditions are the cWF-weighted transcript count
220 for each gene and the reference cell-type ratio, respectively. The iteration is repeated until all
221 sampling variables converge with the signature-gene set. The counts of the non-signature gene
222 set are computed by an analytic approach in the Bayesian framework with a hyperparameter,
223 γ, which has the similar nature of β as above. The "r" is fixed as the above-estimated value,
224 and only "x" in the non-signature gene set is computed.

225 The deconvolution results for 10 organs using this method were compared to the real
226 cell-composition data of the corresponding organ (Fig 5, B and C). The results were also
227 compared to those of two conventional deconvolution methods without the cWF, MuSiC and
228 DWLS [25, 26]. These two were chosen as they have been applied to 1 – 4 organs and appear
229 to outperform the other published methods. The prediction of the cell-type ratios show that our
230 method (V-scRNAseq) outperforms the methods without the cWFs (MuSiC, DWLS) for 9
231 organs (brain, fat, heart, kidney, lung, pancreas, skin, skeletal muscle, spleen). For these organs,
232 the cell-type ratios predicted by our method are more similar to those of the real data than those
233 predicted by MuSiC or DWLS (Fig 5B and S4 Table). Notably, our method corrected the
234 abnormally large ratios of cardiac muscle and muscles in the heart and the skeletal muscle,
235 respectively, estimated by MuSiC and DWLS (Fig 5B and S4 Table). The lower root-mean-
236 squared error (RMSE) for these 9 organs with our method (Fig 5C and S4 Table) further
237 confirms the better performance of our method (Fig 5C and S4 Table). For the liver, our method
238 failed (Fig 5, B and C, and S4 Table), despite the high performance in the reconstitution (Fig
239 4). This may be due to the predominantly hepatocyte composition of the liver. Our method
240 assumes the symmetric Dirichlet distribution as prior distribution which influences the
241 posterior distribution – hence, the method is less suitable for the organ, such as the liver,
242 consisting of a single dominant cell-type.

In addition to the cell-type ratios, we also computed the virtual transcriptome of 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 in each of the "virtual" cell-type and compared them to those of the real cell-type for the 10 organs (Fig 6 and see Materials and methods). Our approach computes a posterior distribution and its Expected *A Posteriori* (EAP) for counts of each transcript for each cell-type. Hence, we calculated Pearson correlation coefficients for each cell-type, using the EAPs of all genes in virtual and real cell-types for each organ (Fig 6 and S5 Table).

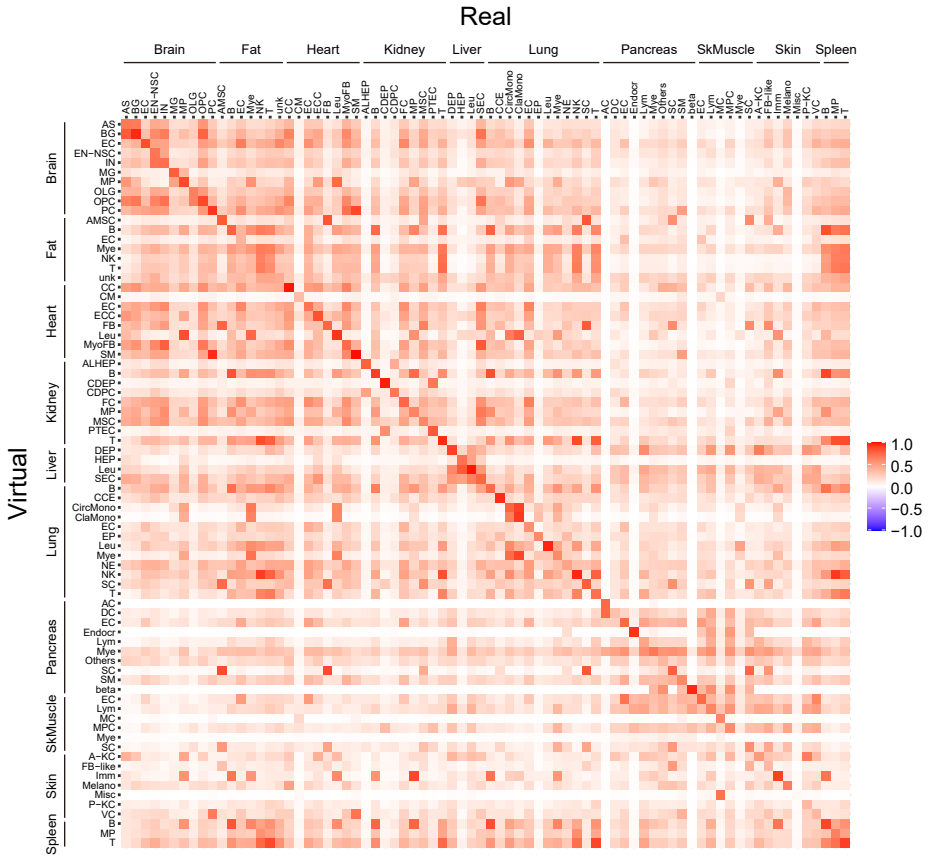
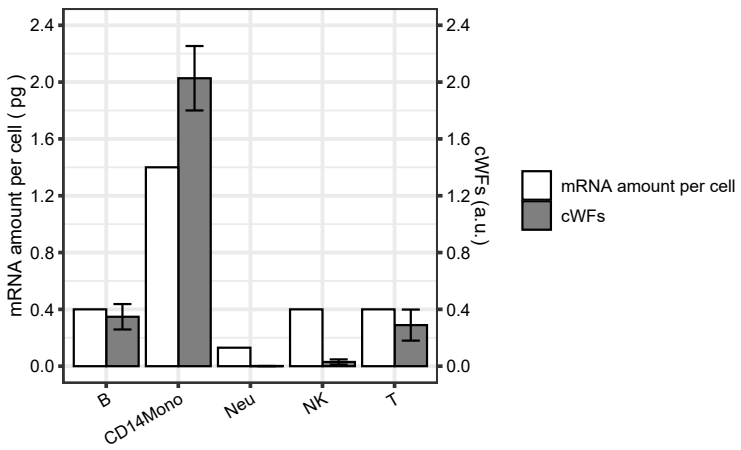


Fig 6. Computation of the complete transcriptome of the composite cell-types of 10 mouse organs. The heatmap of Pearson correlation coefficients for the transcriptome of all virtual and real cell-types across all 10 organs. The Pearson correlation coefficients are calculated for 23,131 (Brain, Fat, Heart, 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 S5 Table. SkMuscle: skeletal muscle.

257

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

264 The reconstitution and deconvolution results demonstrate that cWFs enable the
265 accurate representation of individual cells in the context of the whole-organs of highly
266 heterogenous compositions, presumably by weighting the variations of the transcriptome-size
267 of different cell-types. We further validate this account by directly comparing the cWFs
268 variations to those of the experimentally-determined total mRNA-contents of the
269 corresponding cell-types.



270

271 **Fig 7. Comparison of the cWFs to the experimentally-determined total mRNA-contents of human**
272 **PBMC cell-types.** The comparison is shown as bar graphs of the experimentally-determined mRNA-
273 amounts per cell (mRNA amount per cell: empty bar) and the computed cWFs (cWFs: filled-bar) for each
274 cell-type. The bars are shown as mean ± S.E. (standard error). B: B-cell, CD14Mono: $CD14^+$ monocytes,
275 Neu: Neutrophil, NK: Natural killer cell, T: T-cell. The validation of the cWFs by reconstitution and

276 deconvolution are found in S1 Fig (raw data are in S6 – S8 Tables). The corresponding raw data for Fig 7
277 are found in S9 Table.
278
279 The measurement of such variations for individual single-cells of solid organs are
280 difficult and the methods to measure such variations are limited to cultured cells and
281 systemic/circulating cells such as immune cells, and tumor cells [2, 10, 13, 24]. Therefore, we
282 used the experimentally-determined total mRNA-contents of the cell-types composing human
283 peripheral blood mononuclear cells (PBMCs) and compared them to the computed cWFs of
284 the corresponding cell-types (Fig 7).
285 We used bulk and single-cell RNA-seq data of human PBMCs [27-30] and computed
286 the cWFs of B-cells, $CD14^+$ monocytes, $CD16^+$ monocytes, dendritic cells, neutrophils, natural
287 killer (NK) cells, plasmablasts (PB), T-cells, and the others (e.g., red-blood-cells, platelets,
288 etc.) (see Materials and methods). The reconstitution and deconvolution analyses confirmed
289 that they correct the biased gene-expression in the scRNA-seq data caused by the normalization
290 of the transcriptome-size, as indicated by the higher Pearson correlation coefficients (S1A Fig,
291 and, S6 and S7 Tables) and lower RMSE/higher Person correlation coefficient (S1B Fig, and,
292 S8 Table) with the cWFs in the reconstitution and deconvolution analyses, respectively.
293 The total mRNA-amounts are experimentally measured and determined for B-cells,
294 $CD14^+$ monocytes, neutrophils, NK cells, and T-cells of PBMCs [11]. We made side-by-side
295 comparison of the computed cWFs and their experimentally-measured transcriptome-size of
296 the corresponding cell-types (Fig 7 and S9 Table). The result shows that the computed cWFs
297 accurately recapitulate the 3.5 - 10.8-fold higher transcriptome-size of the $CD14^+$ monocytes
298 compared to the B-, T-, NK cells, or neutrophils, respectively (i.e., $CD14^+$ monocytes: 1.4
299 pg/cell vs. B-/T-/NK-cells: 0.4 pg/cell, Neutrophils: 0.13 pg/cell). It also recapitulates the 3-
300 fold higher values of the B- and T-cells than that of neutrophils (i.e., 0.4 vs. 0.13 pg/cell). The
301 experimental values of the mRNA-amounts of B- and T-cells are approximately the same,

302 which is also recapitulated by virtually the same cWFs of these two cell-types. These side-by-
303 side comparisons provide another layer of evidence supporting the account that cWFs indeed
304 represent the transcriptome size-variations among different cell-types.
305

306 # Utilities of cWFs

307 Previously, biological and diagnostic utilities of differential transcriptome size are reported [2,
308 12, 15-24]. Hence, we explored such utilities for cWFs. We hypothesized that the
309 transcriptome-size of each cell-type differentially change over aging and that such a phenotype
310 could underlie the aging mechanism and could also be exploited as a new type of aging-
311 biomarker. To test this hypothesis, we examined whether the patterns of the cWFs change over
312 the course of aging. Furthermore, we examined whether such putative changes of the cWFs
313 actually indicate the progression of aging.

314 For this purpose, we used the *Tabula Muris Senis*, the atlas of the multiple-organ single-
315 cell transcriptomics across multiple aging stages of mouse [29, 31]. This atlas provides both
316 the whole-organ RNA-seq data and their corresponding composite scRNA-seq data. Using
317 these data, we computed cWFs of the composite cell-types of 5 organs (brain, heart, kidney,
318 lung, spleen) across 3 aging-stages (3, 18, and 24 months-old) of mice. The 3, 18, 24 months-
319 old (mos.) in mouse correspond to 25 - 26 years-old, 60 - 80 years-old, and ≥ 80 years-old in
320 human, respectively [32]. These 5 organs were chosen for the following two reasons: 1) they
321 performed well in both the reconstitution and deconvolution validations (Figs 4 and 5); 2) their
322 equivalent whole-organ/single-cell data are also present in our 10 organs-analysis described
323 above.

324 Using these datasets, we computed the cWFs of all composite cell-types for each organ.
325 The reconstitution (S2 Fig, and, S10 and S11 Tables) and deconvolution (S12 Table) analyses
326 confirmed that they correct the biased representations of the gene expression and cell-type

327 compositions, respectively, in the scRNA-seq data with one or more of the signature gene set(s)
328 for each aging-stage. Based on these results, we selected the best performing numbers of the
329 signature genes for each organ and aging-stage by omitting the conditions that failed to
330 significantly improve the reconstitution (highlighted in pink in S12 Table) and selecting the
331 ones with the lowest RMSE and the highest Pearson correlation coefficients in the
332 deconvolution analysis (highlighted in light-green in S12 Table). This screening selected the
333 cWFs computed with the following conditions: Brain (3 mos. – 500 signature genes, 18 mos.
334 – 300 signature genes, 24 mos. – 300 signature genes), Heart (3 mos. – 100 signature genes,
335 18 mos. – 500 signature genes, 24 mos. – 300 signature genes), Kidney (3 mos. – 300 signature
336 genes, 18 mos. – 500 signature genes, 24 mos. – 500 signature genes), Lung (3 mos. – 500
337 signature genes, 18 mos. – 300 signature genes, 24 mos. – 300 signature genes), Spleen (3 mos.
338 – 300 signature genes, 18 mos. – 100 signature genes, 24 mos. – 500 signature genes).

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

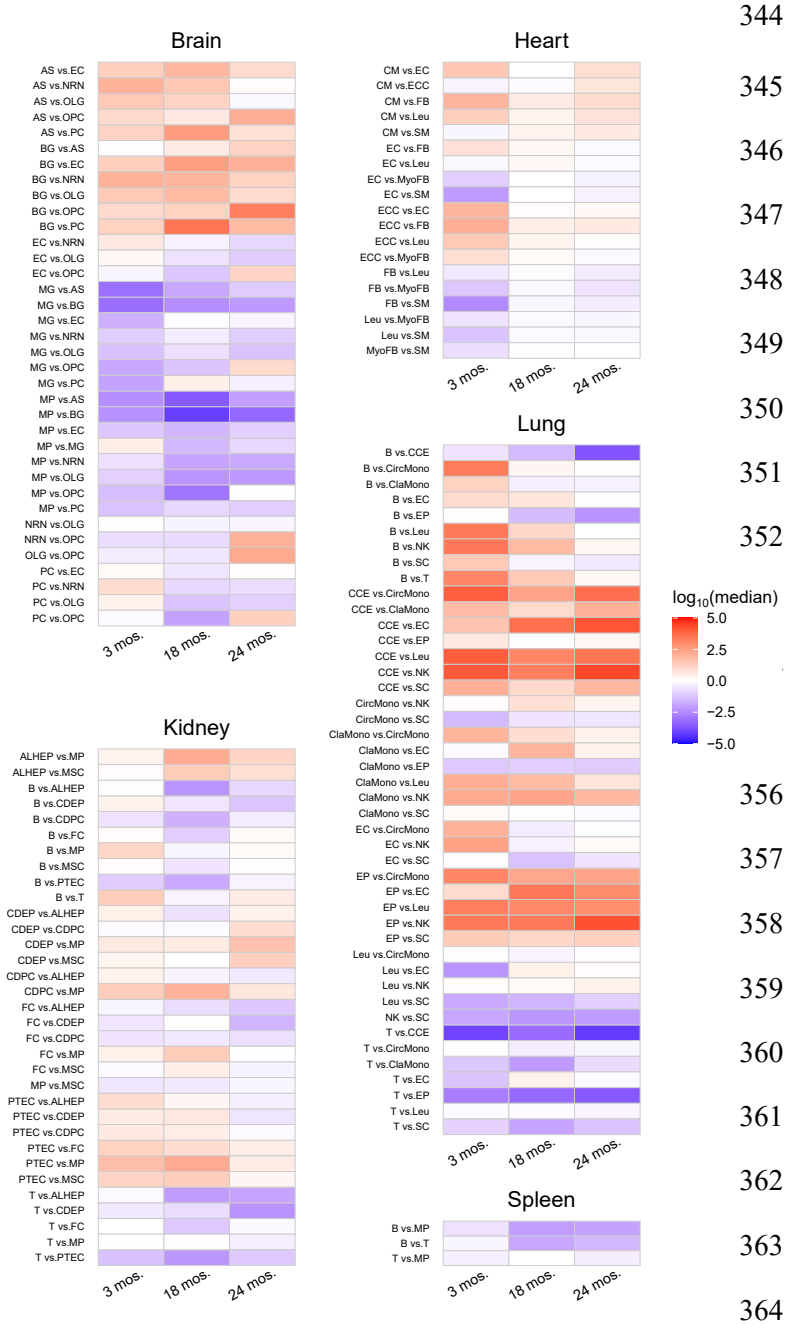


Fig 8. Differential cWFs during aging of mouse. The changes of the ratios of cWFs of cell type-pairs across three aging-stages (3 months-old: 3 mos., 18 months-old: 18 mos., 24 months-old: 24 mos.) are shown as heatmap for each organ (Brain, Heart, Kidney, Lung, Spleen). The cell type-pairs are indicated 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 (S2 Fig, and, S10 and S11 Tables) and the best deconvolution results (S12 Table) for each organ. The full names of the cell type abbreviations and the raw data of the heatmap are found in S13 Table.

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 [33], to the dataset of the pairwise cWF ratios between the cell type-pairs listed above (see 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).

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

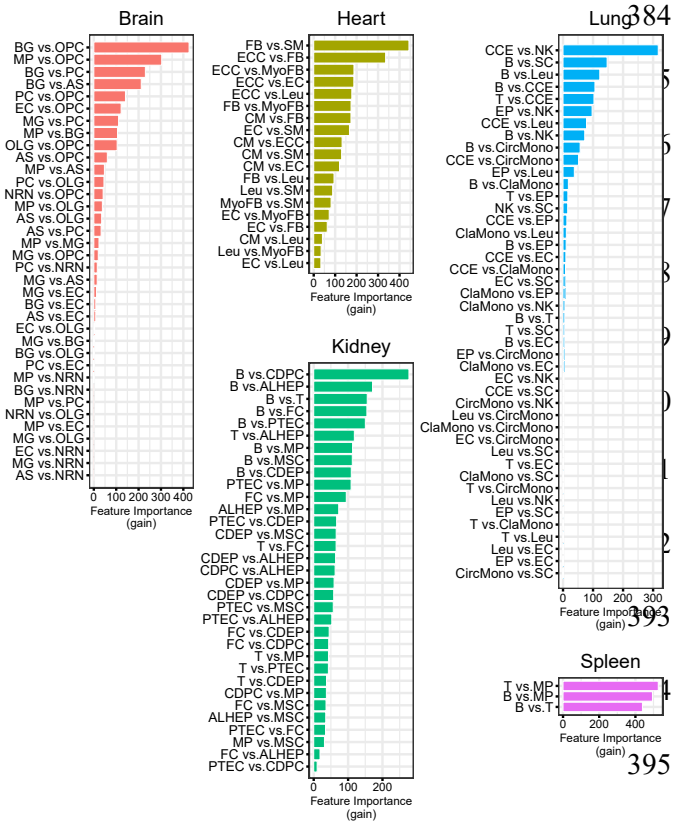


Fig 9. Prediction of the aging-stage by the differential cWFs of each organ. (A) The prediction scores (accuracy, macro-precision, macro-recall, macro-F1 score) for each organ (Brain, Heart, Kidney, Lung, Spleen) are shown. **(B)** The top-ranked features (i.e., pairwise cell-types) important for the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

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

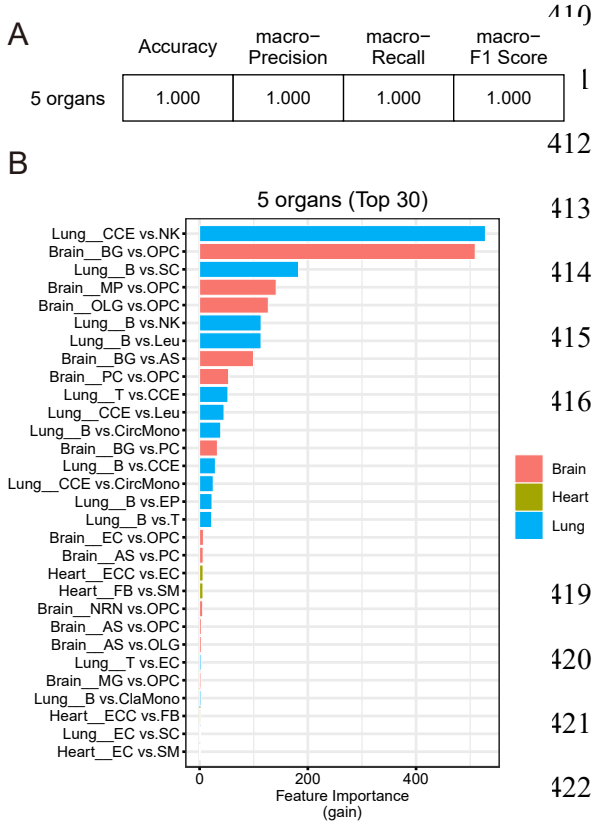


Fig 10. Prediction of the aging-stage by the differential cWFs of all 5 organs combined. (A) The prediction scores (accuracy, macro-precision, macro-recall, macro-F1) by combining the cWF of all pairwise cell-types across all organs are shown. **(B)** The top 30 important features (i.e., the pairwise 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 S14 Table.

425 # Discussion

426 In this study, we report a computational method to calculate cWFs, coefficients to offset the
427 lack of the representation of transcriptome size variations and differential cryptic gene
428 expression among different cells of solid organs (Fig 2). Using this method, we computed and
429 describe cWFs for 76 cell-types across 10 organs (Fig 3). We show that they indeed account
430 for the relative variations of the transcriptome-size of the cells and their differential cryptic
431 gene expression by demonstrating their requisite role for accurately reconstituting and
432 deconvolving the whole-organ RNA-seq data using their composite scRNA-seq data weighted
433 by the corresponding cWFs (Figs 4 – 6). We also show that they recapitulate the
434 experimentally-determined total mRNA content-differences, using PBMC data (Fig 7).

435 Furthermore, we show that cWFs differentially change among diverse cell-types across
436 various organs over aging and can be used to effectively predict aging progression in the aging
437 mouse model (Figs 8 – 10). These results suggest a possible role of differential transcriptome
438 size-change over lifetime in the aging mechanism. The results also indicate their possible
439 diagnostic applications as biomarkers for aging progression and/or aging associated diseases.

440 Despite the usefulness of cWFs, some limitations must be noted. One limitation is the
441 requirement of both bulk and single-cell RNA-seq data of the organs and the tissues of interest.
442 Our method requires both of these datasets for the target organs/tissues to infer the cWFs of
443 the composing cell-types. However, this may become less of a limiting factor as the availability
444 of such data is exponentially growing. Once we determine the cWFs for the specific cell-types
445 of organs, the same cWFs can be used for the same cell-types of the same organs regardless of
446 their sources. Therefore, the cWFs reported in this study can be used for the same 76 cell-types
447 and 10 organs for other studies. Another requirement is that the method needs at least a rough
448 a priori knowledge about the cell-type compositions (i.e., the ratio of cell-types) of the organs
449 and the tissue of interest -i.e., the method works only in a supervised framework. This limitation

450 could also become less limiting as a growing body of scRNA-seq data provides information on
451 the cell-type compositions of previously less characterized organs/tissues.
452

453 Materials and methods

454 Datasets

455 For the whole-organ RNA-seq data, two murine datasets are used: One from 11 weeks-old male
456 C57BL6/N Jcl mice and derived from 'iOrgans Atlas' (http://i-organs.atr.jp) [34, 35] and the
457 other from 3, 18, and 24 months-old male C57BL6/N mice and derived from the *Tabula Muris*
458 *Senis* [29]. For the scRNA-seq datasets, used are the brain, the fat, the heart, the liver, the lung,
459 and the spleen data from 11 weeks-old male C57BL6/JN mice described in *Tabula Muris* [36].
460 The kidney data are from 3 months-old male C57BL6/JN mice in *Tabula Muris Senis* [31].
461 The skeletal muscle and the pancreas data are from 8-10 weeks-old female C57BL6/J reported
462 in Mouse Cell Atlas [37]. The skin data are from 9 weeks-old female C57BL6/J mice [38]. The
463 aging mouse scRNA-seq datasets in the brain, the heart, the kidney, the lung and the spleen are
464 from 3, 18, 24 months-old male C57BL6/JN mice in *Tabula Muris Senis* [29, 31]. The human
465 PBMC datasets are GSM2871599 in GSE107572 (bulk RNA-seq) [43, 44] and GSM4557334
466 in GSE150728 (scRNA-seq) [30], both are from healthy donors.
467

468 Data preprocessing

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

474 characters). For the skeletal muscle, 'B cell_***', 'Dendritic cell', and 'T cell' into 'Lymphatic
475 cell', 'Erythroblast_***', 'Granulocyte monocyte progenitor cell', 'Macrophage_***' and
476 'Neutrophil_***' into 'Myeloid cell', 'Muscle cell_***' into 'Muscle cell'. For the pancreas,
477 'B cell' and 'T cell' into 'Lymphatic cell', 'Dendritic cell', 'Erythroblast_***', 'Granulocyte',
478 and 'Macrophage_***' into 'Myeloid cell', 'Endothelial cell_***' into 'Endothelial cell',
479 'Smooth muscle cell_***' into 'Smooth muscle cell', 'Stromal cell_***' into 'Stromal cell',
480 'Dividing cell' and 'Glial cell' into 'Others'. For the heart dataset in *Tabula Muris Senis* [31],
481 the labels were replaced with the ones in *Tabula Muris* [20] using single-cell reference mapping
482 by '*Seurat*' v4 [39], as follows. First, the reference scRNA-seq data in *Tabula Muris* were
483 preprocessed by '*Seurat*' analyses with the log-normalization (where the scale factor is 10^6),
484 the highly variable genes selection with default parameters, the scaling for all genes, and
485 principal component analysis (PCA) acquiring 200 principal components (PCs). The number
486 of PCs used for the downstream analysis was determined at p-value below 0.0001 in JackStraw
487 analysis, resulting in 33 PCs. Using these 33 PCs reference data, the query heart scRNA-seq
488 datasets from *Tabula Muris Senis* [31] were annotated by the functions of
489 'FindTransferAnchors()' and 'TransferData()' in '*Seurat*'. For the human PBMC, we
490 combined cell-type labels as follows: 'Class-switched B' and 'B' into 'B', 'CD8m T', 'CD4m
491 T', 'CD4n T', 'CD4 T', 'gd T' and 'CD8eff T' into 'T', 'IgG PB' and 'IgA PB' into 'PB',
492 'pDC' and 'DC' into 'DC', 'RBC', 'Platelet', 'Activated Granulocyte' and 'SC & Eosinophil'
493 into 'Others'. The gene symbol-matching between the scRNA-seq and the whole-organ RNA-
494 seq data was conducted by using entrez gene IDs derived from 'org.Mm.egALIAS2EG' in an
495 R package, '*org.Mm.eg.db*'. ERCC-labelled genes were removed as they are spike-in genes. In
496 addition, the counts of the three genes (*Rn45s, Akap5, Lrrc17*) were removed as they are non-
497 mRNA artifacts significantly influencing the total counts. The normalization was performed

498 by adjusting the total counts of each cell in the scRNA-seq datasets to a million. The same
499 normalization process was also applied to each of the whole-organ RNA-seq data.
500

501 Calculation of reference cell-type ratios

502 The reference cell-type ratios were as described in each original single-cell study, except for
503 the heart and the brain. For the heart, the cell-type ratios described in *Tabula Muris* [36] were
504 based on the separate analysis of the cardiac muscles and the non-muscle cells, resulting in a
505 large under-estimation of the cardiac muscle ratio – 3.1% [36]. In contrast, the ratio determined
506 by multiple other methods and considered as the gold-standard in the cardiovascular fields is
507 30 - 40% [40-46]. According to this gold-standard ratios, we made the following modifications:
508 We set the ratio of the cardiac-muscle cells at 30% and the remaining at 70%. The latter, then,
509 was divided by the non-muscle cell types by maintaining their ratios the same to those in the
510 *Tabula Muris* data. For the brain, the cell-type ratio reported in the NIH database
511 (https://www.nervenet.org/papers/brainrev99.html#Numbers) was used to modify those in the
512 *Tabula Muris* scRNA-seq data [36], as follows. First, we divided the brain cell-type classes
513 into four classes: 'neurons', 'glial cells', 'endothelial cells', 'others', and then set the ratio of
514 these classes at 75 : 23 : 7 : 4, respectively, accordingly to their estimated ratios in the NIH
515 murine brain database (https://www.nervenet.org/papers/brainrev99.html#Numbers). Second,
516 we further divided 'neurons', 'glial cells', and 'others' into more cell-type classes according to
517 *Tabula Muris* [36]. The 'neurons' were further divided into 'neuron-excitatory neurons and
518 some neuronal stem cells' and 'neuron-inhibitory neurons'. The 'others' class was divided into
519 'brain pericyte-NA' and 'oligodendrocyte precursor cell-NA'. For the 'glial cells' class, we
520 introduced three assumptions: 1) the 'glial cells' can be classified into four cell-types,
521 'microglial cell-NA', 'astrocyte-NA', 'Bergmann glial cells-NA', and 'oligodendrocyte-NA',
522 according to *Tabula Muris* [36]; 2) the ratios of these four glial cell types follow those of

523 *Tabula Muris* [36] among them, and 3) the ratio of 'microglial cells-NA' in the whole brain is
524 0.1, as it is reported that the 'microglial cells' account for 10 – 15 % of the whole brain cells.
525 On the basis of these assumptions, the ratios of each brain cell type was estimated as
526 'macrophage-NA' (ca. 0.2%), 'microglial cell-NA' (10.0%), 'astrocyte-NA' (ca. 2.2%),
527 'Bergmann glial cell-NA' (ca. 2.1%), 'brain pericyte-NA' (ca. 1.5%), 'endothelial cell-NA'
528 (ca. 6.4%), 'neuron-excitatory neurons and some neuronal stem cells' (ca. 47.5%), 'neuron-
529 inhibitory neurons' (ca. 21.3%), oligodendrocyte-NA' (ca. 8.7%) and 'oligodendrocyte
530 precursor cell-NA' (ca. 1.9%) and used as the reference. For the human PBMC, the reference
531 cell-type ratios are as described [27, 28], except for those of "$CD16^+$ monocytes" and "PB and
532 Others". The ratios of these cell-types are derived as follows: The ratios of the cell-type label
533 "the others" was divided to '$CD16^+$ monocytes' (ca. 7.21%), 'PB' (ca. 0.24%), 'Others' (ca.
534 3.77%), according to their RNA-seq data-based ratios.
535

536 ## Selection of signature gene sets

537 Random Forest (RF) was performed to extract signature genes for distinguishing cell-types
538 from one another in each organ using the scRNA-seq count data derived from the cell-type
539 reference datasets described in the above section. In this study, the '*randomForest*' package in
540 R was used for tuning and producing a classifier by RF. The scRNA-seq data were first divided
541 by 8:2 into the training and the test data. Using the training data, we determined two parameters,
542 'mtry' and 'ntree' for producing an RF-model. The parameter 'mtry' means the number of
543 features used for an RF-model and was tuned by a function 'tuneRF()', while the parameter
544 'ntree' means the number of trees generated in an RF-model and was set as 500, which is
545 sufficient to converge error rate in cell-type classification. The produced RF-model was
546 validated by using the test data and F1-score calculated by a function 'F1_Score()' in an R
547 package '*MLmetrics*'. Then we confirmed the validity of the model by F1-score over 0.8.

548 Following the classifier production, the important features in the classifier were extracted as
549 the signature genes for each cell-type, where we used 'Mean Decrease in Gini' values as the
550 importance indicator for each gene.
551

552 # Computation of the cell type-specific weighting-factors (cWFs)

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

560 $$m\mathbf{y} = \sum_j w_j \mathbf{x}_j \ \ (j \in C^m) \quad (1)$$

561 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
562 vector, the cWF for each cell subject j, and the normalized scRNA-seq counts vector for each
563 cell subject j, respectively. The n is the total number of the signature genes for the organs. In
564 this study, 100, 300, and 500 signature genes were selected according to the ranking on the
565 basis of 'Mean Decrease in Gini' values calculated by the RF analyses. The numbers, 100, 300,
566 and 500 were arbitrarily selected but the smaller number of signature genes is computationally
567 more cost-effective (see the details also in the section below). The combination of cell subjects,
568 C^m was selected at random while maintaining the reference cell-type ratio described above
569 with the total number of cell subjects, m, which is arbitrarily set to 100. In addition, we set C^m
570 containing at least 1 cell subject in each cell-type. With this model, w_j was calculated by
571 solving a quadratic problem under the constraint that the resulting value is non-negative as
572 follows:

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

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:

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

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:

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

where, k, C^k, and N_k denote a cell-type, a 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:

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

$$\mathrm{P}(\mathbf{r}) = Dirichlet(\mathbf{r}|\alpha\mathbf{1}_K) \tag{9}$$

where, α and $\mathbf{1}_K$ denote a hyperparameter and the K-dimensional all-ones vector. In prior distribution $\mathrm{P}(\mathbf{X})$, $\mathcal{N}(\bar{\mathbf{x}}_k|\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}_k^{-1})$ 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 $\boldsymbol{\Sigma}_k^{-1} = \frac{1}{N_k}\sum_j \sigma_{w_j}^2 \mathrm{diag}(\mathbf{x}_j \odot \mathbf{x}_j)$. 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 sampled elements of $\mathbf{X}$. The 'stan' sampling parameters, such as the number of iterations, max_treedepth, adapt_delta, thin, etc., were appropriately chosen to converge well the sampled distributions. The convergence was evaluated by R-hat values (Gelman-Rubin statistics) in all sampling variables below 1.1, which

is a generally accepted value for convergence evaluation. For the expression pattern of non-signature genes, we set another noise parameter γ in Equation (6), instead of β and estimated $\bar{\mathbf{x}}_k$ (the expression patterns in non-signature genes) by the following probability equation:

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

where, $\bar{\mathbf{r}}$ denotes the mean values of $\mathbf{r}$ determined by the sampled distribution above.

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 the following equations:

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

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

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

Selection of the optimal number of the signature genes

The optimal number (100, 300, or 500) of the signature genes were determined based on the three criteria as follows: 1) Improvement of the Pearson correlation coefficient between the r-woRNAseq and the syn-woRNAseq(+cWF), 2) Convergence in the deconvolution, 3) Higher similarity to the reference cell type composition (i.e., lower RMSE). The results with the selected optimal number of the signature genes were used in the comparisons with those

643 obtained by MuSiC and DWLS, and also for the analyses with the aging data and for the cWFs
644 analysis of the human PBMC data described below.
645

646 MuSiC and DWLS

647 MuSiC was conducted as described [26]. DWLS was performed as described [25], except that
648 a function, 'solve.QP()' (R package: '*quadprog*') was substituted by 'solve_osqp()' (R
649 package: '*osqp*') for quadratic problem solver.
650

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

652 The mean counts of each cell-type in each organ were compared by Pearson correlation
653 coefficients. For the virtual scRNA-seq, the estimated mean count for each cell-type was
654 normalized to a million. For the real scRNA-seq, the mean count for each individual cell was
655 normalized to a million and averaged for each cell-type.
656

657 Analysis of differential cWFs in aging progression

658 The distribution of the pairwise cWF ratios was calculated for each aging-stage. The ratios of
659 the pairwise cWFs in each cell-type, instead of the cWFs themselves, were used as the cWFs
660 are relative values calculated from the normalized count datasets of the real whole-organ RNA-
661 seq and thus the cWFs values themselves can be used only for their comparison among the
662 cell-types within each organ at each aging-stage (i.e., the cWFs cannot be compared across the
663 aging-stages). Hence, their pairwise ratios were used for the comparisons across the aging-
664 stages. For each pairwise cell-type within an organ, the distribution of the cWF ratios was
665 calculated for all pair combinations of the consisting cells. For example, assuming that cell-
666 type X has 1,000 cells and cell-type Y has 500 cells with one cWF per cell, respectively, the

667 distribution of the ratio of cWFs of cell-type X vs. Y consisting of 500,000 values of the ratios
668 was calculated. When a cell-type contains zero value of cWF, we substituted the zero value for
669 the second minimum cWF value in the cell-type to avoid the division by zero. Next, for each
670 cell-types pair, we performed Mann-Whitney U-test to make statistical evaluations of the
671 differences of their ratio-distributions between the aging-stages (i.e., 18 mos. vs. 3 mos., 24
672 mos. vs. 3 mos., 24 mos. vs. 18 mos.). Zeros were generated if the resulting values were below
673 the lower bound of R program, 1e-308.
674

675 Machine learning prediction of the aging-stages by the differential 676 cWF ratios

677 For this analysis, we omitted the data of the cell-types pairs of which statistical analysis
678 described above resulted in $p \geq 0.01$ across all pairs of the aging-stages. Using the remaining
679 data, we performed LightGBM classifier analysis [33]. We first generated input data for
680 LightGBM classifier analysis using the cWF ratio distributions obtained as described in the
681 above section. The 100 cWF ratio distribution-values were randomly sampled for each cell-
682 types pair for each aging-stage. The sampled values were at random combined with each other,
683 resulting in the matrix of "the 100 cWFs ratio distribution values" x "the number of the cell-
684 type pairs" for each aging-stage. Using these matrices for all three aging-stages as the input
685 data, we generated a LightGBM classifier of aging-stages on Python, using the '*lightgbm*'
686 package [33]. The input data was divided by 8:2 into training data and test data. The training
687 data was used for tuning the parameters of LightGBM. The tuning was performed by the
688 '*optuna*' package [47], searching for the best parameters to maximize macro-F1 score.
689 Specifically, the searched parameters are max_depth, num_leaves, subsample, subsample_freq,
690 colsample_bytree, min_child_samples, ranging from 3 to 12, 2 to 256, 0.1 to 1.0, 1 to 7, 0.1 to
691 1.0, and 5 to 100, respectively. The model was generated using the best parameters and

evaluated using test data by accuracy, macro-Precision, macro-Recall, and macro-F1_score. These evaluations were performed by a Python package, 's*cikit-learn*' [48]. In addition, we also extracted feature importance based on 'gain' from the model and confirmed the high-ranked contributors in the model (i.e., aging-stage classification).

Supporting information

S1 Fig. Reconstitution and deconvolution of the bulk human PBMC RNA-seq by the composite scRNA-seq. (**A)** Reconstitution results with and without (no) cWFs are compared. The results with the 100, 300, 500 signature genes are shown. The similarity is shown as violin plot of the Pearson correlation coefficients. The corresponding raw data are found in tables S6 and S7. **(B)** Bar graph showing the cell type-ratios computed by the 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 S8). The bar graphs are composed of the cell-types computed to be present for each organ by our method. The similarity scores (RMSE: Root Mean Squared Errors, Pearson correlation coefficient) are indicated at the top of the bar. The corresponding raw data of S1 Fig. are found in S6, S7, S8 Tables.

S2 Fig. Reconstitution of the whole organ RNA-seq of the aging model mouse by the composite scRNA-seq. The results with and without (no) cWFs are compared for each aging-stage (3 mos., 18 mos. 24 mos.) for each number of the signature gens (100, 300, 500) and for each organ (Brain, Heart, Kidney, Lung, Spleen). The similarity is shown as violin plots of the Pearson correlation coefficients. The corresponding raw data are found in S10 Table (no cWFs) and S11 (with cWFs). The corresponding raw data of S2 Fig. are found in S10 and S11 Tables. All supplementary tables below are available at https://github.com/k-teji/VscRNAseq.

S1 Table (Separate file)

The raw data of Fig. 1.

S2 Table (Separate file)

The raw data of Fig. 3.

S3 Table (Separate file)

The raw data of Fig. 4.

S4 Table (Separate file)

The raw data of Fig. 5, B and C.

S5 Table (Separate file)

The raw data of Fig. 6.

S6 Table (Separate file)

The raw data of S1 Fig. (Pearson correlation coefficients for the result with no cWFs).

S7 Table (Separate file)

The raw data of S1A Fig. (Pearson correlation coefficients for the result with cWFs).

S8 Table (Separate file)

The raw data of S1B Fig. The best performing result (the lowest RMSE and the highest Pearson correlation coefficient) shown in S1B Fig. is highlighted in light green.

S9 Table (Separate file)

The raw data of Fig. 7.

S10 Table (Separate file)

The raw data of S2 Fig. (Pearson correlation coefficients for the result with no cWFs).

S11 Table (Separate file)

The raw data of S2 Fig. (Pearson correlation coefficients for the result with cWFs).

S12 Table (Separate file)

Deconvolution of the aging-mouse whole-organ RNA-seq data. The results omitted due to the lack of the significant improvement by the cWFs in the reconstitution are highlighted in pink. The results used in the analyses shown in Figs. 7, 8, 9 (the pairwise comparisons between the two aging-stages show $p<0.01$ by U-test) are highlighted in light-green.

S13 Table (Separate file)

The raw data of Fig. 8.

S14 Table (Separate file)

The raw data of Fig. 9 (sheet 1) & Fig. 10 (sheet 2).

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.).

773 **Author contributions:** Conceptualization: TNS, Methodology: KT, SK, Investigation: TNS,
774 KT, SK, Visualization: KT, Supervision: TNS, Writing-original draft: TNS, KT, Writing-
775 review & editing: TNS, KT, SK.
776
777 **Completing interests:** T.N.S. is the inventor of the patents filed on the cell type-specific
778 weighting-factors and holds shares of Karydo TherapeutiX, Inc., which owns the right to this
779 patent. K.T. and S.K. are also employed by Karydo TherapeutiX, Inc. Correspondence and
780 requests for materials should be addressed to T.N.S. (island1005@gmail.com).
781
782 **Data and materials availability:** The public datasets are as indicated and the raw data as
783 supplementary tables and the code for our algorithm are available at https://github.com/k-
784 teji/VscRNAseq.
785

786 # References

787 1. Aldridge S, Teichmann SA. Single cell transcriptomics comes of age. Nat Commun.
788 2020;11(1):4307. Epub 20200827. doi: 10.1038/s41467-020-18158-5. PubMed PMID:
789 32855414; PubMed Central PMCID: PMCPMC7453005.
790 2. Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting global
791 gene expression analysis. Cell. 2012;151(3):476-82. doi: 10.1016/j.cell.2012.10.012. PubMed
792 PMID: 23101621; PubMed Central PMCID: PMCPMC3505597.
793 3. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From
794 single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing.
795 Genome Res. 2014;24(3):496-510. Epub 2013/12/05. doi: 10.1101/gr.161034.113. PubMed
796 PMID: 24299736; PubMed Central PMCID: PMCPMC3941114.

797 4. Coate JE, Doyle JJ. Variation in transcriptome size: are we getting the message?
798 Chromosoma. 2015;124(1):27-43. Epub 20141126. doi: 10.1007/s00412-014-0496-3. PubMed
799 PMID: 25421950.
800 5. Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-
801 sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):75.
802 Epub 20170818. doi: 10.1186/s13073-017-0467-4. PubMed PMID: 28821273; PubMed
803 Central PMCID: PMCPMC5561556.
804 6. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and
805 bioinformatics pipelines. Exp Mol Med. 2018;50(8):1-14. Epub 20180807. doi:
806 10.1038/s12276-018-0071-8. PubMed PMID: 30089861; PubMed Central PMCID:
807 PMCPMC6082860.
808 7. Adil A, Kumar V, Jan AT, Asger M. Single-Cell Transcriptomics: Current Methods
809 and Challenges in Data Acquisition and Analysis. Front Neurosci. 2021;15:591122. Epub
810 20210422. doi: 10.3389/fnins.2021.591122. PubMed PMID: 33967674; PubMed Central
811 PMCID: PMCPMC8100238.
812 8. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to
813 spatial transcriptomics for biomedical research. Genome Med. 2022;14(1):68. Epub 20220627.
814 doi: 10.1186/s13073-022-01075-1. PubMed PMID: 35761361; PubMed Central PMCID:
815 PMCPMC9238181.
816 9. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-
817 cell and spatial multi-omics. Nature reviews Genetics. 2023:1-22. Epub 20230302. doi:
818 10.1038/s41576-023-00580-2. PubMed PMID: 36864178; PubMed Central PMCID:
819 PMCPMC9979144.
820 10. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of
821 cellular mixtures based on linearity of transcriptional signatures. Nat Commun.

822 2019;10(1):2209. Epub 2019/05/19. doi: 10.1038/s41467-019-09990-5. PubMed PMID:
823 31101809; PubMed Central PMCID: PMCPMC6525259.
824 11. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous
825 enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife.
826 2017;6. Epub 20171113. doi: 10.7554/eLife.26476. PubMed PMID: 29130882; PubMed
827 Central PMCID: PMCPMC5718706.
828 12. Coate JE, Doyle JJ. Quantifying whole transcriptome size, a prerequisite for
829 understanding transcriptome evolution across species: an example from a plant allopolyploid.
830 Genome Biol Evol. 2010;2:534-46. Epub 20100705. doi: 10.1093/gbe/evq038. PubMed
831 PMID: 20671102; PubMed Central PMCID: PMCPMC2997557.
832 13. Jonasson E, Andersson L, Dolatabadi S, Ghannoum S, Aman P, Stahlberg A. Total
833 mRNA Quantification in Single Cells: Sarcoma Cell Heterogeneity. Cells. 2020;9(3). Epub
834 20200319. doi: 10.3390/cells9030759. PubMed PMID: 32204559; PubMed Central PMCID:
835 PMCPMC7140709.
836 14. Raj A, Rifkin SA, Andersen E, van Oudenaarden A. Variability in gene expression
837 underlies incomplete penetrance. Nature. 2010;463(7283):913-8. doi: 10.1038/nature08781.
838 PubMed PMID: 20164922; PubMed Central PMCID: PMC2836165.
839 15. Schmidt EE, Schibler U. Cell size regulation, a mechanism that controls cellular RNA
840 accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-
841 Y and the liver-enriched transcription factor DBP. The Journal of cell biology.
842 1995;128(4):467-83. Epub 1995/02/01. doi: 10.1083/jcb.128.4.467. PubMed PMID: 7532171;
843 PubMed Central PMCID: PMCPMC2199888.
844 16. Zhurinsky J, Leonhard K, Watt S, Marguerat S, Bahler J, Nurse P. A coordinated global
845 control over cellular transcription. Current biology : CB. 2010;20(22):2010-5. Epub 20101021.
846 doi: 10.1016/j.cub.2010.10.002. PubMed PMID: 20970341.

847 17. Marguerat S, Bahler J. Coordinating genome expression with cell size. Trends in
848 genetics : TIG. 2012;28(11):560-5. Epub 20120802. doi: 10.1016/j.tig.2012.07.003. PubMed
849 PMID: 22863032.
850 18. Miettinen TP, Pessa HK, Caldez MJ, Fuhrer T, Diril MK, Sauer U, et al. Identification
851 of transcriptional and metabolic programs related to mammalian cell size. Current biology :
852 CB. 2014;24(6):598-608. Epub 20140306. doi: 10.1016/j.cub.2014.01.071. PubMed PMID:
853 24613310; PubMed Central PMCID: PMCPMC3991852.
854 19. Lin CY, Loven J, Rahl PB, Paranal RM, Burge CB, Bradner JE, et al. Transcriptional
855 amplification in tumor cells with elevated c-Myc. Cell. 2012;151(1):56-67. doi:
856 10.1016/j.cell.2012.08.026. PubMed PMID: 23021215; PubMed Central PMCID:
857 PMCPMC3462372.
858 20. Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, et al. c-Myc is a universal amplifier
859 of expressed genes in lymphocytes and embryonic stem cells. Cell. 2012;151(1):68-79. doi:
860 10.1016/j.cell.2012.08.033. PubMed PMID: 23021216; PubMed Central PMCID:
861 PMCPMC3471363.
862 21. Li Y, Wang H, Muffat J, Cheng AW, Orlando DA, Loven J, et al. Global transcriptional
863 and translational repression in human-embryonic-stem-cell-derived Rett syndrome neurons.
864 Cell stem cell. 2013;13(4):446-58. doi: 10.1016/j.stem.2013.09.001. PubMed PMID:
865 24094325; PubMed Central PMCID: PMCPMC4053296.
866 22. Chandler MG, Pritchard RH. The effect of gene concentration and relative gene dosage
867 on gene output in Escherichia coli. Mol Gen Genet. 1975;138(2):127-41. doi:
868 10.1007/BF02428117. PubMed PMID: 1105148.
869 23. Hu Z, Chen K, Xia Z, Chavez M, Pal S, Seol JH, et al. Nucleosome loss leads to global
870 transcriptional up-regulation and genomic instability during yeast aging. Genes Dev.

871 2014;28(4):396-408. doi: 10.1101/gad.233221.113. PubMed PMID: 24532716; PubMed
872 Central PMCID: PMCPMC3937517.
873 24. Cao S, Wang JR, Ji S, Yang P, Dai Y, Guo S, et al. Estimation of tumor cell total mRNA
874 expression in 15 cancer types predicts disease progression. Nature biotechnology.
875 2022;40(11):1624-33. Epub 20220613. doi: 10.1038/s41587-022-01342-x. PubMed PMID:
876 35697807; PubMed Central PMCID: PMCPMC9646498.
877 25. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan GC. Accurate estimation of cell-
878 type composition from gene expression data. Nat Commun. 2019;10(1):2975. Epub
879 2019/07/07. doi: 10.1038/s41467-019-10802-z. PubMed PMID: 31278265; PubMed Central
880 PMCID: PMCPMC6611906.
881 26. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with
882 multi-subject single-cell expression reference. Nat Commun. 2019;10(1):380. Epub
883 2019/01/24. doi: 10.1038/s41467-018-08023-x. PubMed PMID: 30670690; PubMed Central
884 PMCID: PMCPMC6342984.
885 27. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Correction to:
886 Molecular and pharmacological modulators of the tumor immune contexture revealed by
887 deconvolution of RNA-seq data. Genome Med. 2019;11(1):50. Epub 20190729. doi:
888 10.1186/s13073-019-0655-5. PubMed PMID: 31358023; PubMed Central PMCID:
889 PMCPMC6661746.
890 28. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and
891 pharmacological modulators of the tumor immune contexture revealed by deconvolution of
892 RNA-seq data. Genome Med. 2019;11(1):34. Epub 20190524. doi: 10.1186/s13073-019-0638-
893 6. PubMed PMID: 31126321; PubMed Central PMCID: PMCPMC6534875.
894 29. Schaum N, Lehallier B, Hahn O, Palovics R, Hosseinzadeh S, Lee SE, et al. Ageing
895 hallmarks exhibit organ-specific temporal signatures. Nature. 2020;583(7817):596-602. Epub

896 20200715. doi: 10.1038/s41586-020-2499-y. PubMed PMID: 32669715; PubMed Central
897 PMCID: PMCPMC7757734.
898 30. Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martinez-Colon GJ, McKechnie JL, et al. A
899 single-cell atlas of the peripheral immune response to severe COVID-19. medRxiv. 2020. Epub
900 20200423. doi: 10.1101/2020.04.17.20069930. PubMed PMID: 32511639; PubMed Central
901 PMCID: PMCPMC7276995.
902 31. Tabula Muris C. A single-cell transcriptomic atlas characterizes ageing tissues in the
903 mouse. Nature. 2020;583(7817):590-5. Epub 20200715. doi: 10.1038/s41586-020-2496-1.
904 PubMed PMID: 32669714; PubMed Central PMCID: PMCPMC8240505.
905 32. Dutta S, Sengupta P. Men and mice: Relating their ages. Life Sci. 2016;152:244-8.
906 Epub 20151024. doi: 10.1016/j.lfs.2015.10.025. PubMed PMID: 26596563.
907 33. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient
908 Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems.
909 2017;30.
910 34. Kozawa S, Sagawa F, Endo S, De Almeida GM, Mitsuishi Y, Sato TN. Predicting
911 Human Clinical Outcomes Using Mouse Multi-Organ Transcriptome. iScience.
912 2020;23(2):100791. Epub 20200109. doi: 10.1016/j.isci.2019.100791. PubMed PMID:
913 31928967; PubMed Central PMCID: PMCPMC7033637.
914 35. Kozawa S, Ueda R, Urayama K, Sagawa F, Endo S, Shiizaki K, et al. The Body-wide
915 Transcriptome Landscape of Disease Models. iScience. 2018;2:238-68. Epub 2018/11/15. doi:
916 10.1016/j.isci.2018.03.014. PubMed PMID: 30428375; PubMed Central PMCID:
917 PMCPMC6135982.
918 36. Tabula Muris C, Overall c, Logistical c, Organ c, processing, Library p, et al. Single-
919 cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727):367-

920 72. Epub 2018/10/05. doi: 10.1038/s41586-018-0590-4. PubMed PMID: 30283141; PubMed
921 Central PMCID: PMCPMC6642641.
922 37. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the Mouse Cell Atlas by
923 Microwell-Seq. Cell. 2018;172(5):1091-107 e17. Epub 2018/02/24. doi:
924 10.1016/j.cell.2018.02.001. PubMed PMID: 29474909.
925 38. Joost S, Annusver K, Jacob T, Sun X, Dalessandri T, Sivan U, et al. The Molecular
926 Anatomy of Mouse Skin during Hair Growth and Rest. Cell stem cell. 2020;26(3):441-57 e7.
927 Epub 20200227. doi: 10.1016/j.stem.2020.01.012. PubMed PMID: 32109378.
928 39. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al.
929 Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87 e29. Epub
930 20210531. doi: 10.1016/j.cell.2021.04.048. PubMed PMID: 34062119; PubMed Central
931 PMCID: PMCPMC8238499.
932 40. Banerjee I, Fuseler JW, Price RL, Borg TK, Baudino TA. Determination of cell types
933 and numbers during cardiac development in the neonatal and adult rat and mouse. American
934 journal of physiology Heart and circulatory physiology. 2007;293(3):H1883-91. Epub
935 2007/07/03. doi: 10.1152/ajpheart.00514.2007. PubMed PMID: 17604329.
936 41. Bergmann O, Zdunek S, Felker A, Salehpour M, Alkass K, Bernard S, et al. Dynamics
937 of Cell Generation and Turnover in the Human Heart. Cell. 2015;161(7):1566-75. Epub
938 2015/06/16. doi: 10.1016/j.cell.2015.05.026. PubMed PMID: 26073943.
939 42. Nag AC. Study of non-muscle cells of the adult mammalian heart: a fine structural
940 analysis and distribution. Cytobios. 1980;28(109):41-61. Epub 1980/01/01. PubMed PMID:
941 7428441.
942 43. Pinto AR, Ilinykh A, Ivey MJ, Kuwabara JT, D'Antoni ML, Debuque R, et al.
943 Revisiting Cardiac Cellular Composition. Circulation research. 2016;118(3):400-9. Epub

944 2015/12/05. doi: 10.1161/CIRCRESAHA.115.307778. PubMed PMID: 26635390; PubMed
945 Central PMCID: PMCPMC4744092.
946 44. Raulf A, Horder H, Tarnawski L, Geisen C, Ottersbach A, Roll W, et al. Transgenic
947 systems for unequivocal identification of cardiac myocyte nuclei and analysis of
948 cardiomyocyte cell cycle status. Basic Res Cardiol. 2015;110(3):33. Epub 2015/05/01. doi:
949 10.1007/s00395-015-0489-2. PubMed PMID: 25925989; PubMed Central PMCID:
950 PMCPMC4414935.
951 45. Walsh S, Ponten A, Fleischmann BK, Jovinge S. Cardiomyocyte cell cycle control and
952 growth estimation in vivo--an analysis based on cardiomyocyte nuclei. Cardiovascular research.
953 2010;86(3):365-73. Epub 2010/01/15. doi: 10.1093/cvr/cvq005. PubMed PMID: 20071355.
954 46. Zhou P, Pu WT. Recounting Cardiac Cellular Composition. Circulation research.
955 2016;118(3):368-70. Epub 2016/02/06. doi: 10.1161/CIRCRESAHA.116.308139. PubMed
956 PMID: 26846633; PubMed Central PMCID: PMCPMC4755297.
957 47. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation
958 Hyperparameter Optimization Framework. Proceedings of the ACM SIGKDD International
959 Conference on Knowledge Discovery and Data Mining. 2019:2623-31. doi:
960 10.1145/3292500.3330701.
961 48. Pedregosa Fabianpedregosa F, Michel V, Grisel Oliviergrisel O, Blondel M,
962 Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. Journal of Machine
963 Learning Research. 2011;12(85):2825-30.
964