bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 **Article type: Original Research**

2 # Scalable batch-correction approach for integrating large-scale single-cell
3 # transcriptomes

4

5 Xilin Shen[1], Hongru Shen[1], Dan Wu[1], Mengyao Feng[1], Jiani Hu[1], Jilei Liu[1],
6 Yichen Yang[1], Meng Yang[1], Yang Li[1], Lei Shi[3*], Kexin Chen[2*], Xiangchun Li[1*]

7

8 [1]Tianjin Cancer Institute, National Clinical Research Center for Cancer, Key
9 Laboratory of Cancer Prevention and Therapy, Tianjin Medical University
10 Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China;
11 [2]Department of Epidemiology and Biostatistics, National Clinical Research
12 Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Key
13 Laboratory of Molecular Cancer Epidemiology of Tianjin, Tianjin Medical
14 University Cancer Institute and Hospital, Tianjin Medical University, Tianjin,
15 China; [3]State Key Laboratory of Experimental Hematology, The Province and
16 Ministry Co-sponsored Collaborative Innovation Center for Medical
17 Epigenetics, Key Laboratory of Breast Cancer Prevention and Therapy
18 (Ministry of Education), Key Laboratory of Immune Microenvironment and
19 Disease (Ministry of Education), Department of Biochemistry and Molecular
20 Biology, School of Basic Medical Sciences, Tianjin Medical University Cancer
21 Institute and Hospital, Tianjin Medical University, Tianjin 300070, China

22

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

23 *Correspondence to: Prof. Xiangchun Li, Tianjin Cancer Institute, Tianjin
24 Medical University Cancer Institute and Hospital, Huanhu Xi Road, Tiyuan Bei,
25 Hexi District, Tianjin, 300060, China. Tel (Fax): (86)22- 23372231; Email:
26 lixiangchun@tmu.edu.cn

27 Prof. Kexin Chen, Department of Epidemiology and Biostatistics, Tianjin
28 Medical University Cancer Institute and Hospital, Huanhu Xi Road, Tiyuan Bei,
29 Hexi District, Tianjin, 300060, China. Tel: (86) 2237 2231; Email:
30 chenkexin@tmu.edu.cn.

31 Prof. Lei Shi, Department of Biochemistry and Molecular Biology, School of
32 Basic Medical Sciences, Tianjin Medical University, Qixiangtai Road, Heping
33 District, Tianjin, 300070, China. Tel: (86)22- 83336998; Email:
34 shilei@tmu.edu.cn.

35

36 **Abstract**

37 Integration of the evolving large-scale single-cell transcriptomes requires
38 scalable batch-correction approaches. Here we propose a simple
39 batch-correction method that is scalable for integrating super large-scale
40 single-cell transcriptomes from diverse sources. The core idea of the method is
41 encoding batch information of each cell as a trainable parameter and added to
42 its expression profile; subsequently, a contrastive learning approach is used to
43 learn feature representation of the additive expression profile. We demonstrate
44 the scalability of the proposed method by integrating 18 million cells obtained
45 from the Human Cell Atlas. Our benchmark comparisons with current
46 state-of-the-art single-cell integration methods demonstrated that our method

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

47 could achieve comparable data alignment and cluster preservation. Our study
48 would facilitate the integration of super large-scale single-cell transcriptomes.
49 The source code is available at https://github.com/xilinshen/Fugue.

50

51 Background

52 Single-cell sequencing offers tremendous opportunities for biomedical
53 research to explore the cellular ecosystem and molecular mechanisms [1].
54 Advances in single-cell technologies have spurred the establishment of
55 several public repositories of single-cell data, including the Human Cell Atlas
56 (HCA), the Single-cell Expression Atlas and the Mouse Cell Atlas [2, 3]. The
57 HCA project is committed to curate millions to trillions of single-cells for
58 constructing a comprehensive reference map of all human cells. As can be
59 foreseen, integration of super large-scale single-cells across heterogeneous
60 tissues from diverse sources will be a leading wave for deep exploration of
61 biology [4, 5]. Therefore, scalable computational methods are crucial for
62 integration of single-cell transcriptomes and subsequently their translation into
63 biological significance.

64 Batch effects are fundamental issues to be addressed for integration of
65 single-cell transcriptomes. Batch effects are inevitable as single-cell data were
66 generated by various groups with diverse experimental protocols and
67 sequencing platforms [6]. Considerable progress has been made on
68 batch-correction of single-cell expression. For instance, MNN [7], Scanorama
69 [8] and BBKNN [9] are all based on mutual nearest neighbors (MNNs)
70 identification were successfully applied to guide single-cell integration. The

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

71 Seurat integration [10] utilizes canonical correlation analysis to identify
72 correlations across datasets and computes MNNs to correct data. Harmony
73 [11] integrates datasets by clustering similar cells from different batches while
74 maximizing the diversity of batches within each cluster. scVI [12] applies a
75 deep learning model to learn shared embedding space among datasets for the
76 elimination of batch effects. However, these methods are not designed for the
77 integration of super large-scale single-cells.

78 To satisfy this need, we present Fugue, a simple yet efficient solution for
79 batch-correction of super large-scale single-cell transcriptomes. The method
80 extended the deep learning method at the heart of our recently published
81 Miscell approach [13]. Miscell learns representations of single-cell expression
82 profiles through contrastive learning and achieves high performance on
83 canonical single-cell analysis tasks including cell clustering and cell-specific
84 markers inferring. In this study, we expand Miscell through encoding batch
85 information as trainable parameters and adding them into expression profiles.
86 In concept, the gene expression profiles of same cell from different batches
87 could be seen as superposition of the same biological information and different
88 batch information. Fugue incorporates addictive batch information as learnable
89 parameters into gene expression matrix. The batch information can be
90 properly represented after training. By taking batch information as trainable
91 variable, Fugue is scalable in atlasing-scale data integration with fixed memory
92 usage.

93 We demonstrated the scalability and efficiency of Fugue by applying it to
94 analyze 18 million single-cells obtained from HCA and benchmarked its
95 performance on diverse datasets along with current state-of-the-art methods.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

96 We showed that Fugue achieved favorably performance as compared with
97 current state-of-the-art methods. The reference map of HCA dissected by
98 Fugue demonstrated that it can learn smooth embedding for time course
99 trajectory and joint embedding space for immune cells from heterogeneous
100 tissues.

101

102 **Results**

103 **Overview of Fugue**

104 Fugue integrates single-cells through adding batch information into expression
105 profile and learns batch information by contrastive learning. Specifically, we
106 construct Fugue as a deep learning-based feature encoder to learn dimension
107 reduction representation of expression profile. Given a set of uncorrected
108 single-cells (**Figure 1A**), Fugue embeds their batch information as a learnable
109 matrix (i.e. batch embedding matrix) and adds them to the corresponding
110 expression profile (**Figure 1B**). A DenseNet of 21 layers [14] is used as feature
111 encoder to learn the additive expression profiles. The feature encoder is
112 trained in a self-supervised manner through contrastive learning (**Figure 1C**)
113 [15]. Contrastive loss minimizes the distance between the cell and its
114 noise-added view, and maximizes the distance between different cells. The
115 trained feature encoder is used to extract feature representations of
116 single-cells (**Figure 1D**). We remove the batch embedding matrix from the
117 input. As a result, only biological signals are retained in the embedding space.
118 The representation could be utilized for downstream analysis such as
119 single-cell cluster delineation (**Figure 1E**). Details are described in **Methods**

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

120 section.

121

122 **Benchmark evaluations**

123 On the *simulation dataset* of 30,000 cells of 3 cell types among 5 batches,
124 each cell type was divided by batches (**Figure 2A**) before batch correction.
125 After integration with Fugue, cells of the same types were well-mixed and cells
126 of different types were dispersed across batches (**Figure 2B**). In addition, we
127 ran Fugue on this simulation dataset after removing a specific cell type from
128 four batches (See **Methods and Supplementary Figure 1**). The result
129 showed that Fugue could maintain batch-specific cell types (**Figure 2C**).

130 We used this simulation dataset to search for three hyperparameters that are
131 adjusted for contrastive learning, including size of memory bank and
132 momentum coefficient. The kBET [16] and ARI scores were applied to evaluate
133 its performance (**see Methods**). Fugue was insensitive to variation of these
134 hyperparameters in terms of ARI and kBET scores (**Supplementary Figure**
135 **1A and 1B**). Data augmentations include random dropout and position
136 shuffling. We set the dropout rate to 30% and random shuffle rate to 10%
137 based on the value of ARI and kBET (**Supplementary Figure 1C**).

138 We compared Fugue to 8 single-cell integration methods on the *cell line* (n =
139 9,531) and *PBMC* (n = 28,541) *datasets* (See Methods, **Figure 2D, G**). Fugue
140 yielded similar result as these 8 methods on UMAP plots (**Figure 2E, H and**
141 **Supplementary Figure 2,3**). Quantitatively, Fugue achieved comparable
142 kBET and ARI scores (**Figure 2F, I**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

143

144 **Fugue could accurately remove batch effects**

145 We applied Fugue to integrate all available data from HCA repository (75
146 cohorts totaling 18,056,192 cells) (**Supplementary Table 1**). The batch effect
147 removing efficiency of Fugue was evaluated on three datasets included in HCA,
148 including the *census of immune project,* the *lung and* the *brain dataset*.

149 Common cell types of the *census of immune project* (cord blood, n = 133,264;
150 bone marrow, n = 176,571) revealed a minimal overlap before integration
151 (**Supplementary Figure 4A**). Fugue clustered cells into biologically coherent
152 groups and removed batch-specific variations (**Figure 3A**), and UMAP plot
153 was similar to the aforementioned benchmark methods (**Supplementary**
154 **Figure 4B-H**). Fugue achieved comparable kBET and ARI scores as
155 compared with these methods (**Figure 3B**).

156 The C30.1 (n = 75,387) and C47 (n = 2,532) from *lung dataset* showed
157 minimal overlap before batch correction (**Figure 3C**). After correction with
158 Fugue, cells from different datasets were mapped into corresponding area
159 (**Figure 3D**). The UMAP plot was consistent with the aforementioned
160 benchmark methods (**Supplementary Figure 5**). Quantitatively, Fugue
161 achieved comparable kBET and ARI scores with these methods (**Figure 3F**).
162 Unsupervised clustering and cell types annotation revealed 11 cell types in the
163 *lung dataset*, including monocytes, mast cells and ciliated cells (**Figure 3E**).
164 Conventional cell markers [17, 18] were expressed uniquely in each cell
165 cluster (**Figure 3G**), and invariant across batches (**Figure 3H**).

166 We evaluated the batch removing efficiency on the *brain dataset* with the same

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

167 process applied for the *lung dataset*. The result also demonstrated that Fugue
168 can robustly integrate cells from multiple studies (**Supplementary Figure 6**).
169

170 Fugue captures the real batch information

171 We hypothesize that sequencing samples of the same cohort are subjected to
172 lower batch variation. Therefore, the batch embeddings of samples from the
173 same cohort should be more similar than those from different cohorts.

174 We extracted the batch embeddings of 373 samples from 75 cohorts in HCA.
175 The result showed that samples from the same cohort had higher similarity of
176 batch embeddings as compared with samples from different cohorts
177 (**Supplementary Figure 7A)**. We found that batch embeddings of 4 patients
178 from the C2 cohort are almost identical (**Supplementary Figure 7B**), which
179 was consistent with the previous report that there was no batch effect among
180 these four patients [19]. For the *census of immune project*, we observed higher
181 similarity of batch embeddings within the same batch than between batches
182 (**Supplementary Figure 7C**). For the PBMC and tonsil tissue from C39
183 subjected to the same sequencing protocol [20], we also observed high
184 similarity among them, especially among samples from the same tissue
185 (**Supplementary Figure 7D**).
186

187 Fugue aligned precise immune cell subtypes in HCA

188 Immune cells are highly homogeneous across tissues [21]. Therefore, Fugue
189 should be able to map the same immune cell types together across HCA.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

190 Forty-six clusters were inferred from HCA (n = 3,424,607) (**Supplementary**
191 **Figure 8A and Supplementary Table 2**). Most clusters consist of multiple
192 cohorts, while some come from specific organs (**Supplementary Figure 8B**).
193 For example, 26 projects had over 100 cells in endothelial cell_1 cluster; C2
194 was the only project associated with lymphatic tissue [19] and made up the
195 majority of lymphatic endothelial cell cluster (endothelial cell_7)
196 (**Supplementary Figure 8B**).

197 We reclustered the immune cells **(Supplementary Figure 9)** corresponding to
198 17 subtypes (**Figure 4A**). Different cell types were readily separable from each
199 other, and dataset specific cell types were retained, such as *in-vivo* stimulated
200 NKT cells (**Figure 4A**). Canonical markers were expressed in corresponding
201 cell types (**Figure 4B**). For example, Pan-B cell markers *CD79A* and *CD79B*
202 were expressed in B cell clusters. B cell precursor specific markers *VPREB1*
203 and *IGLL1* were expressed in the relevant cell type. We observed stable
204 expression of marker genes among batches (**Figure 4C, D and**
205 **Supplementary Figure 10**). For example, natural killer cell markers *PRF1* and
206 *KLRD1* were expressed in all of the 27 cohorts (**Figure 4D**).

207

208 **Fugue integrates time course development trajectories**

209 On the *embryonic mouse cardiac dataset,* batch effects were observed among
210 embryo development stage before correction (**Figure 5A**). Fugue integrated
211 cells from different embryo periods (**Figure 5B**). We classified the single-cells
212 into 5 cell types based on the specific cell markers (**Figure 5C, D and**
213 **Supplementary Figure 11**). The FLE dimension reduction showed that

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

214 expression representation extracted from Fugue captured the embryonic
215 developmental trajectories for each cell type (**Figure 5E**). Cells from
216 embryonic (E) day 10.5, 13.5 and 16.5 were orderly arranged according to
217 pseudo-time trajectory (**Figure 5E**). The expression patterns of canonical cell
218 differentiation markers were consistent with developmental stages
219 (**Supplementary Figure 12**). For instance, early erythrocyte markers *GYPA*
220 and *TFRC* expressed highly in E10.5 erythrocytes and negatively correlated
221 with pseudo-time (**Supplementary Figure 12**), which was consistent with the
222 previous studies [22, 23].

223 We next recovered cell development trajectories during hematopoiesis from
224 the *census of immune project*. FLE plot indicated clear overlap of the identified
225 cell types from cord blood and bone marrow across pseudo-time trajectory
226 (**Supplementary Figure 13**). Clear trajectories that quadrifurcate from
227 hematopoietic stem cells (HSCs) into B cell, T cell, mono-dendritic and
228 megakaryocyte-erythroid series were constructed (**Figure 5F**). The trajectories
229 were ordered by cell development stages and branching by cell differentiation
230 types (**Figure 5G-J**).

231

232 Discussion and conclusions

233 In this study, we attempt to tackle the batch effect removal issue in the rapidly
234 developing single-cell transcriptomic with a simple yet effective solution. Fugue
235 could be deployed as a scalable deep learning model to integrate single-cells
236 of any magnitude with fixed memory. We provide evidence that Fugue
237 showcases superior performance in terms of integrating millions of single-cells

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

238 from various sources. Fugue is expected to assemble all human cells to
239 construct a comprehensive single-cell atlas.

240 In application, we showcase the robustness of Fugue in super large-scale
241 datasets integration. Specifically, Fugue was applied to integrate all available
242 single-cells among HCA repository. Three datasets included in HCA were
243 utilized to represent the data integrated effectiveness of Fugue, for that most of
244 the benchmark methods cannot handle atlasing-scale datasets due to memory
245 overflow. Moreover, there are currently no suitable indicators to assess the
246 batch-correction performance of complex datasets with multiple distinct or
247 dataset-specific cell types spanning dozens of batches. Fugue performed on
248 par with current state-of-the-art single-cell integration methods in terms of
249 batch-correction and cluster preservation performance. Fugue therefore offers
250 better trade-offs between data integration performance and scalability, and it is
251 a key advantage of Fugue to integrate super large-scale datasets.
252 Furthermore, we show that Fugue could integrate millions of immune cells to
253 reflect delicate cell functional status while retaining distinct cell subtypes.
254 Additional analysis demonstrates that time course trajectories could be
255 correctly constructed and ordered after single-cell integration by Fugue. The
256 algorithm can thus facilitates the exploration of subtle biological differences
257 among atlasing-scale datasets.

258 A great deal of batch-correction methods learn batch information based on
259 prior assumptions. For example, Combat assumes batches as a function of
260 gene expressions [24]. Methods based on MNN learn batch information
261 through paired cells between batches, and highly depend on the qualities of
262 MNNs [7, 8]. Fugue is a hypothesis-free deep learning network. It simply

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

263 learns batch information through contrastive learning and does not require
264 domain specific knowledge. The flexibility of this approach could be
265 demonstrated through the integration of single-cells from HCA. For that explicit
266 batch information are not always available from researchers, we employed
267 sample labels as batch information for HCA projects. We demonstrated the
268 compatibility of this configuration through benchmark the performance of
269 Fugue with current state-of-the-art methods, for which accurate batch labels
270 were set. The batch information learned by Fugue also show little variation
271 within the same batch as compared with that between batches
272 (**Supplementary Figure 7**). Therefore, the simple batch correction approach
273 is flexible and can be a good candidate for multi-millions of single-cells where
274 explicit batch information are not always available.

275 Although immune cell markers have been studied extensively, the knowledge
276 might be limited by their definition via a restricted set of organs or cell types.
277 The integrated analysis of atlasing-scale single-cells enabled cross-organ
278 comparisons and provide new perspectives for the understanding of marker
279 genes. Based on the reference map of HCA, we found many conventional
280 immune cell markers are expressed in nonimmune cell types. For example,
281 conventional monocyte marker *S100A9* was expressed in esophageal
282 squamous epithelium cells (Epithelial cell_4) (**Supplementary Figure 9**),
283 which was confirmed by previous studies [25, 26]. Canonical HSC marker
284 *SPINK2* was expressed higher in epididymal epithelial cells (Epithelial cell_7)
285 than HSCs (Stem cell_1) (**Supplementary Figure 9**). The enrichment of
286 *SPINK2* in epididymal tissue was confirmed in the previous report [27].

287 Fugue could be improved in several aspects. First, as an artificial intelligence

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

288 model, black-box nature of the approach is a limitation that should be resolved
289 [28, 29]. We explored the batch embedding matrix and found that similar
290 batches have more similar batch embeddings than dissimilar batches. It brings
291 insights into the interpretability of batch information learned by Fugue. Second,
292 as an unsupervised learning model, hyperparameters tuning might to some
293 extent influence the performance of Fugue [30]. In our analysis, we proved the
294 stability of Fugue to hyperparameters tuning (**Supplementary Figure 1**). We
295 also used the same hyperparameters of model structure throughout the study
296 to ensure the generalization of the result.

297 In summary, we present Fugue, a simple yet efficient deep learning model for
298 super large-scale single-cell transcriptomes integration. We anticipate Fugue
299 will be helpful for researchers to transform growing scale of single-cell
300 transcriptomes into the understanding of biology and disease, driving new
301 ways for disease diagnosis and treatment.

302

303 **Methods**

304 **Batch embedding**

305 The key idea of batch-effect removal is decoupling biological signals from
306 nuisance factors of batch effects. We explicitly encoded batch information as a
307 learnable batch embedding matrix (BE) and added them to expression matrix
308 (E) to obtain expression matrix with batch embedding information ($X = BE + E$),
309 subsequently performing feature representation learning on X. The batch
310 embedding matrix BE was randomly initialized and updated during training.
311 For the purpose of point-wise addition between BE and E, the dimension of

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

312 matrices BE and E must be identical.

313

314 **Network architecture and training**

315 We used DenseNet architecture [14] as feature encoder to learn expression
316 embedding of single-cells. The DenseNet has 21 layers that are consisted of 4
317 dense blocks. The DenseNet architecture is featured by concatenating all the
318 outputs from preceding layers as input for the next layer to make feature
319 transmission more efficient. We replaced convolutional layer of the DenseNet
320 with linear layer to make it able to process gene expression matrix.
321 Self-supervised learning with momentum contrast [15] was adopted to train the
322 feature encoder. We applied multi-layer perceptron (MLP) as project head,
323 which was demonstrated to be beneficial for contrastive learning [31].

324

325 Here we adapt contrastive learning for feature encoder development, through
326 which the model was trained by constructing positive and negative pairs [32].
327 For a given integrated input I_{cell}, a feature encoder represents it as $C_q = f_q (I_{cell})$,
328 where f_q *is a* query encoder network and C_q is a query sample. A key encoder
329 network f_k encode the noise-adding view of the input I_{cell+} *as* C_{k+} (likewise, $C_{k+} =$
330 $f_q (I_{cell+})$). One cell C_q and its noise-adding view C_{k+} form a positive pair, and
331 assemble with a different cell C_{k-} *to* form a negative pair. The contrastive loss is
332 optimized through learning the same representation of the positive pairs and
333 dissimilar representation of negative pairs:

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

334 $$L_{C_q, C_{k+}, \{C_{k-}\}} = -\log \frac{\exp(C_q \times C_{k+} / \tau)}{\exp(C_q \times C_{k+} / \tau) + \sum_{k-} \exp(C_q \times C_{k-} / \tau)},$$

335 where C_{k-} denotes a dictionary of the negative samples. The dictionary was
336 built as a queue C_{k1-}, C_{k2-}, ..., C_{kn-}. The current mini-batch en queue and the
337 oldest mini-batch de queue. We set the queue size to 10% of the training data.
338 τ is a temperature hyper-parameter and was set to 0.2. We performed data
339 augmentations through random zero out to 30% and shuffling to 10% of genes.
340 These hyperparameters were determined through grid search (see
341 **Supplementary Figure 1**).

342

343 The parameters of query encoder θ_q were updated by back-propagation; the
344 parameters of key encoder θ_k were updated according to θ_q:

345 $$\theta_k \leftarrow m\theta_k + (1-m)\theta_q,$$

346 where m stands a momentum coefficient, which was set to 0.999. We trained
347 the network at a learning rate of 0.01. The training was ended until loss did not
348 improve over a specified number of epochs (see **Supplementary Table 3**).

349

350 The network was trained through mini-batch stochastic gradient descent
351 algorithm [33] with a weight decay of 1e-4. The convergence speed of deep
352 learning model is affected by batch-size [34]. For that we integrated thousands
353 to multi-millions single-cells, we set the size of mini-batch from 16 to 256,
354 which was dependent on the volume of training data (**Supplementary Table**

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

355 **3**).

356

357 At the stage of feature extraction, we applied the developed feature encoder f_q
358 as feature extractor. Only expression matrix E_{cell} was provided to f_q:

359
$$F_{\text{cell}} = f_q(E_{cell}),$$

360 where F_{cell} is the feature representation of the single-cell transcriptome.

361

362 **Data sources**

363 *Simulation dataset*

364 We simulated a total of 30,000 single-cell read counts using Splatter
365 package[35]. The resultant *simulation dataset* contains 3 cell types; each cell
366 type consists of 5 batches (**Figure 2A**). Each batch contains 2000 genes with
367 a differential expression factor of 0.4. To estimate the performance of Fugue
368 on batch-specific cell types detection, we manually removed cell type 1 from
369 batches 2-5 and maintained them in batch 1 (**Supplementary Figure 14**). We
370 named the resultant dataset as *simulation_rm dataset*.

371

372 *Cell line dataset*

373 This dataset consists of the cell lines of “Jurkat”, “293 T” and the 50/50 mixture
374 of both cell lines [36]. The dataset is composed of 9,531 single-cells generated
375 by 10x 3’ protocol. For mixture cell lines, cells were clustered with Louvain

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

376 algorithm based on Scanpy pipeline. Cell clusters with high expression of *XIST*
377 were annotated as "293 T" while others as "Jurkat".
378

379 *Human peripheral blood mononuclear cell (PBMC) dataset*

380 The data included two batches of PBMC from five samples [37]. One sample
381 was excluded from the analysis because it was stimulated *in vitro*. This dataset
382 contains 28,541 single-cells, which could be grouped into B cells, CD4+ T cell,
383 CD8+ T cell, NK cells, monocytes, megakaryocytes and dendritic cells. The
384 cell labels were provided by the original publication [37].
385

386 *Human cell atlas*

387 We downloaded single-cell data from HCA portal [2] on 16 July 2021.
388 Fifty-three projects (C1-C53) following HCA data processing pipeline were
389 collected (**Supplementary Table 1**). These projects consist of 75 cohorts. We
390 filtered out samples with available cell numbers less than 1000. A total number
391 of 373 samples were maintained for downstream analysis. This dataset
392 contains 18,056,192 cells from multiple organs, including blood, lung, brain
393 and cardiac (**Supplementary Table 1**). We used the sample labels as batch
394 information given that explicit batch information is not always available for
395 every dataset. All of that 18,056,192 cells were used by Fugue for batch
396 information learning. A total of 3,424,607 cells with more than 500 expressed
397 genes were utilized to construct the reference map of HCA.
398

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

399 The following projects were selected for the assessment of dataset alignment
400 and biological significance preservation performance of the reference map.
401 The first dataset was the c*ensus of immune project* (C1). The c*ensus of*
402 *immune project* consists of two batches that can be referred to as cord blood
403 (C1.0) and bone marrow (C1.1). The two batches contain immune cells from
404 diverse development statuses. We downloaded cell type labels from HCA
405 repository on 28 August 2020. The *lung dataset* consists of C30.1 and C47.
406 Both C30.1 and C47 came from lung tissue and have similar cell types [18, 38].
407 The *brain dataset* consists of C19, C28 and C32, which contain cells from
408 different subsections of brain tissue with overlapping cell types among each
409 other [17, 39, 40]. The *embryonic mouse cardiac dataset* consists of C18 and
410 C20. C18 contains mouse cardiac cells from embryonic state of E10.5 and
411 E13.5. C20 includes mouse cardiac cells from embryonic state of E16.5 . Only
412 healthy embryos were taken into account in this analysis. Since the original
413 author of C18 denotes batch effect exists between cells from E10.5 and E13.5
414 mouse [41], the embryonic periods were employed as batch labels for the
415 benchmarking methods.

416

417 **Data prepossessing**

418 We applied Scanpy (**version 1.7.0**) for data preprocessing. We used
419 “highly_variable_genes” function with the default parameters to identify highly
420 variable genes. A total of 1,959 and 2,085 HVGs were selected from the *cell*
421 *line* and *PBMC* datasets, respectively. For HCA project, 14550 genes shared
422 among datasets were selected.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

423

424 For all of the aforementioned datasets, we normalized the count matrix to
425 counts per million normalization (CPM) and took logarithmic transformation (i.e.
426 log2(CPM+1)). Subsequently, the expression of each gene was scaled by
427 subtracting its average expression then divided by its standard deviation. The
428 scaled expression matrix was applied as inputs for the model.

429

430 **Benchmark methods**

431 We benchmarked the performance of Fugue with eight state-of-the-art batch
432 correction methods, including Seurat V3, ComBat, Harmony, BBKNN,
433 Scanorama, scVI, Pegasus L/S adjustment and INSCT. All methods were
434 performed with the default parameters (see **Supplementary Table 4** for
435 detailed information) throughout the study. Seurat V3 ran out of memory on our
436 server (maximum memory: 256 Gb) for dataset with more than 100,000 cells
437 and therefore it was not evaluated on dataset >100,000 cells. For the c*ensus*
438 *of immune project*, cord blood and bone marrow were utilized as batch
439 information. We employed different cohorts as batch information for the *lung*
440 and *brain datasets* and embryonic development periods as batch information
441 for the *embryonic mouse cardiac dataset*. We provided these methods with
442 explicit batch information because it's the general configuration and suitable
443 for these methods [7-12].

444

445 **Evaluation functions**

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

446 We employed kBET acceptance rate [16] for the assessment of batch effect
447 through *Pegasus* package [42]. The kBET acceptance rate measures whether
448 batches are well-mixed in the local neighborhood of each cell. The resulting
449 score ranges from 0 to 1, where a higher score means a better mix. We
450 computed kBET scores based on each cell type and used the average score to
451 evaluate the degree of batch mixing. The adjusted rand index (ARI) score was
452 applied to evaluate batch correction method in terms of cell type mixing. The
453 ARI score measures the percentage of matches between two label lists. The
454 resulting score ranges from -1 to 1, where a high score denotes that the data
455 point fits well in the current cluster. We used the Louvain community detection
456 algorithm implemented in "tl.louvain" of Scanpy package (**version 1.7.0**) for
457 cell clustering. In our study, Louvain algorithm would generate much more cell
458 clusters than real cell types when the resolution was 0.5 and far fewer when
459 the resolution was 0.01. Thus, we set the resolution parameter range from 0.5
460 to 0.01 with a step of 0.01 and computed ARI score with *sklearn* package for
461 each step. The maximum ARI score was employed as the final evaluation
462 index. On account of BBKNN cannot give the corrected feature representation,
463 we calculated the evaluation indexes in UMAP embedding space. The
464 embedding was computed with the default parameters based on the same
465 random seed through *umap-learn* package (**version 0.4.6**). For the *census of*
466 *immune project*, we assessed the performance based on 20, 000 random
467 sampled cells and averaged the scores of 10 replications.

468

469 **Cell marker inferring and cell type identification**

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

470 The marker genes of cell clusters were calculated as mentioned in Miscell [13].
471 Specifically, we constructed a new deep neural network (denoted as $F: R^n \rightarrow$
472 $[0,1]$) by freezing the parameters of the trained encoder and adding a single
473 linear classifier at the end of it. The classifier was trained for cell cluster
474 prediction. We used the importance score calculated by integrated gradient
475 algorithm [43] as the surrogate metric for the impact of each gene on
476 classification output. In specificity, the integrating gradient algorithm calculates
477 the important score of the ith gene as the gradient of $F(x)$ along the ith
478 dimension, which is defined as:

$$\mathrm{IntegratedGrad}_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

479

480 The x and x' are the actual and baseline expression levels respectively. We set
481 x' to 0. A higher importance score represents a more significant impact of gene
482 for the specific cell cluster. We manually annotated cell types according to
483 genes with the highest importance scores.

484

485 **External software**

486 Louvain community detection algorithm implemented in *Scanpy* package
487 (**version 1.7.0**) was applied for cell clustering. We applied UMAP algorithm to
488 visualize cells in a two-dimensional space if unspecified. UMAP failed on
489 3,424,607 cells after 72 hours; thus t-Distributed Stochastic Neighbor
490 Embedding (t-SNE) algorithm from *FIt-SNE* package was utilized to construct
491 a global view of HCA embedding space. Force-directed layout embedding
492 (FLE) from Pegasus package was applied for trajectories inferring.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

493

494 **Availability of data and materials**

495 The source code of Fugue is available at https://github.com/xilinshen/Fugue.
496 The datasets supporting the conclusions of this article are publicly available
497 through onlines sources. The simulation dataset was available at
498 https://github.com/xilinshen/Fugue/tree/master/data; the cell line dataset was
499 available at
500 https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/j
501 urkat,
502 https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/2
503 93t and
504 https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/j
505 urkat:293t_50:50; the PBMC dataset was downloaded from
506 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96583; all available
507 single cells of HCA repository was downloaded from
508 https://www.humancellatlas.org/.

509

510 **Author contribution**

511 X.L., K.C. and L.S. designed and supervised the study. X.L. and X.S. wrote the
512 manuscript. X.L., K.C., L.S. and X.S. revised the manuscript. H.S., D.W., M.F.,
513 J.H., J.L. collected the data. X.S., Y.Y., M.Y., Y.L. processed the data. X.S., X.L.,
514 K.C., L.S., Y.Y., M.Y. and Y.L. interpreted the results. All authors reviewed and
515 approved the submission of this manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

516

517 ## Acknowledgment

518 We want to thank all the researchers for their generosity to make their data
519 publicly available.

520

521 ## Funding

522 This work was supported by the National Natural Science Foundation of China
523 [31801117]; the Program for Changjiang Scholars and Innovative Research
524 Team in University in China [IRT_14R40]; the Tianjin Science and Technology
525 Committee Foundation [17JCYBJC25300]; and the Chinese National Key
526 Research and Development Project [2018YFC1315600].

527

528 ## Disclosure

529 The authors declare that they have no conflict of interest.

530

531 ## Reference

532 1. Paik DT, Tian L, Williams IM, Rhee S, Zhang H, Liu C, Mishra R, Wu SM,
533 Red-Horse K, Wu JC: **Single-Cell RNA Sequencing Unveils Unique**
534 **Transcriptomic Signatures of Organ-Specific Endothelial Cells.**
535 *Circulation* 2020, **142:**1848-1862.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

536 2. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E,
537 Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al: **The Human**
538 **Cell Atlas.** *Elife* 2017, **6**.

539 3. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z,
540 Chen H, Ye F, et al: **Mapping the Mouse Cell Atlas by Microwell-Seq.**
541 *Cell* 2018, **172:**1091-1107 e1017.

542 4. Li Y, Ren P, Dawson A, Vasquez HG, Ageedi W, Zhang C, Luo W, Chen
543 R, Li Y, Kim S, et al: **Single-Cell Transcriptome Analysis Reveals**
544 **Dynamic Cell Populations and Differential Gene Expression**
545 **Patterns in Control and Aneurysmal Human Aortic Tissue.**
546 *Circulation* 2020, **142:**1374-1388.

547 5. Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, Kang B, Liu Z,
548 Jin L, Xing R, et al: **Global characterization of T cells in**
549 **non-small-cell lung cancer by single-cell sequencing.** *Nat Med*
550 2018, **24:**978-985.

551 6. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE,
552 Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and**
553 **critical impact of batch effects in high-throughput data.** *Nat Rev*
554 *Genet* 2010, **11:**733-739.

555 7. Haghverdi L, Lun ATL, Morgan MD, Marioni JC: **Batch effects in**
556 **single-cell RNA-sequencing data are corrected by matching**

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

557 **mutual nearest neighbors.** *Nat Biotechnol* 2018, **36:**421-427.

558 8. Hie B, Bryson B, Berger B: **Efficient integration of heterogeneous**
559 **single-cell transcriptomes using Scanorama.** *Nat Biotechnol* 2019,
560 **37:**685-691.

561 9. Polanski K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE:
562 **BBKNN: fast batch alignment of single cell transcriptomes.**
563 *Bioinformatics* 2020, **36:**964-965.

564 10. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: **Integrating**
565 **single-cell transcriptomic data across different conditions,**
566 **technologies, and species.** *Nat Biotechnol* 2018, **36:**411-420.

567 11. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K,
568 Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S: **Fast, sensitive and**
569 **accurate integration of single-cell data with Harmony.** *Nat Methods*
570 2019, **16:**1289-1296.

571 12. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N: **Deep generative**
572 **modeling for single-cell transcriptomics.** *Nat Methods* 2018,
573 **15:**1053-1058.

574 13. Shen H, Li Y, Feng M, Shen X, Wu D, Zhang C, Yang Y, Yang M, Hu J,
575 Liu J, et al: **Miscell: An efficient self-supervised learning approach**
576 **for dissecting single-cell transcriptome.** *iScience* 2021, **24:**103200.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

577 14. Huang G, Liu Z, van der Maaten L, Weinberger KQ: **Densely**
578 **Connected Convolutional Networks.** pp. arXiv:1608.06993;
579 2016:arXiv:1608.06993.

580 15. Chen X, Fan H, Girshick R, He K: **Improved Baselines with**
581 **Momentum Contrastive Learning.** *arXiv* 2020.

582 16. Buttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ: **A test metric for**
583 **assessing single-cell RNA-seq batch correction.** *Nat Methods* 2019,
584 **16:**43-49.

585 17. Welch J, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko E:
586 **Integrative inference of brain cell similarities and differences from**
587 **single-cell genomics.** *ArXiv* 2018.

588 18. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL,
589 Peter L, Chung MI, Taylor CJ, Jetter C, et al: **Single-cell RNA**
590 **sequencing reveals profibrotic roles of distinct epithelial and**
591 **mesenchymal lineages in pulmonary fibrosis.** *Sci Adv* 2020,
592 **6:**eaba1972.

593 19. Kinchen J, Chen HH, Parikh K, Antanaviciute A, Jagielowicz M,
594 Fawkner-Corbett D, Ashley N, Cubitt L, Mellado-Gomez E, Attar M, et al:
595 **Structural Remodeling of the Human Colonic Mesenchyme in**
596 **Inflammatory Bowel Disease.** *Cell* 2018, **175:**372-386 e317.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

597 20. Cillo AR, Kurten CHL, Tabib T, Qi Z, Onkar S, Wang T, Liu A, Duvvuri U,
598 Kim S, Soose RJ, et al: **Immune Landscape of Viral- and**
599 **Carcinogen-Driven Head and Neck Cancer.** *Immunity* 2020,
600 **52:**183-199 e189.

601 21. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager
602 MA, Aldinger KA, Blecher-Gonen R, Zhang F, et al: **A human cell atlas**
603 **of fetal gene expression.** *Science* 2020, **370**.

604 22. Andersson LC, Gahmberg CG, Teerenhovi L, Vuopio P: **Glycophorin A**
605 **as a cell surface marker of early erythroid differentiation in acute**
606 **leukemia.** *Int J Cancer* 1979, **24:**717-720.

607 23. Levy JE, Jin O, Fujiwara Y, Kuo F, Andrews NC: **Transferrin receptor**
608 **is necessary for development of erythrocytes and the nervous**
609 **system.** *Nat Genet* 1999, **21:**396-399.

610 24. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva**
611 **package for removing batch effects and other unwanted variation**
612 **in high-throughput experiments.** *Bioinformatics* 2012, **28:**882-883.

613 25. Chi ZL, Hayasaka Y, Zhang XY, Cui HS, Hayasaka S: **S100A9-positive**
614 **granulocytes and monocytes in lipopolysaccharide-induced**
615 **anterior ocular inflammation.** *Exp Eye Res* 2007, **84:**254-265.

616 26. Pawar H, Srikanth SM, Kashyap MK, Sathe G, Chavan S, Singal M,

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

617 Manju HC, Kumar KV, Vijayakumar M, Sirdeshmukh R, et al:
618 **Downregulation of S100 Calcium Binding Protein A9 in**
619 **Esophageal Squamous Cell Carcinoma.** *ScientificWorldJournal* 2015,
620 **2015:**325721.

621 27. Bui FQ, Almeida-da-Silva CLC, Huynh B, Trinh A, Liu J, Woodward J,
622 Asadi H, Ojcius DM: **Association between periodontal pathogens**
623 **and systemic disease.** *Biomed J* 2019, **42:**27-35.

624 28. Ghosh A, Kandasamy D: **Interpretable Artificial Intelligence: Why**
625 **and When.** *AJR Am J Roentgenol* 2020, **214:**1137-1138.

626 29. Moore JH, Boland MR, Camara PG, Chervitz H, Gonzalez G, Himes BE,
627 Kim D, Mowery DL, Ritchie MD, Shen L, et al: **Preparing**
628 **next-generation scientists for biomedical big data: artificial**
629 **intelligence approaches.** *Per Med* 2019, **16:**247-257.

630 30. Xinlei Chen, Fan. H, Ross Girshick, He K: **Improved Baselines with**
631 **Momentum Contrastive Learning.** *arXiv* 2020.

632 31. Chen X, Fan H, Girshick R, He K: **Improved Baselines with**
633 **Momentum Contrastive Learning.** pp. arXiv:2003.04297;
634 2020:arXiv:2003.04297.

635 32. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Girshick R: **Momentum**
636 **Contrast for Unsupervised Visual Representation Learning.** *arXiv*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

637 2020.

638 33. Mu Li, Tong Zhang, Yuqiang Chen, Smola AJ: **Efficient mini-batch**
639 **training for stochastic optimization.** *Association for Computing*
640 *Machinery* 2014, **2014.**

641 34. Byrd RH, Chin GM, Nocedal J, Wu Y: **Sample size selection in**
642 **optimization methods for machine learning.** *Mathematical*
643 *Programming* 2012, **134:**127-155.

644 35. Zappia L, Phipson B, Oshlack A: **Splatter: simulation of single-cell**
645 **RNA sequencing data.** *Genome Biol* 2017, **18:**174.

646 36. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo
647 SB, Wheeler TD, McDermott GP, Zhu J, et al: **Massively parallel**
648 **digital transcriptional profiling of single cells.** *Nat Commun* 2017,
649 **8:**14049.

650 37. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy
651 E, Wan E, Wong S, Byrnes L, Lanata CM, et al: **Multiplexed droplet**
652 **single-cell RNA-sequencing using natural genetic variation.** *Nat*
653 *Biotechnol* 2018, **36:**89-94.

654 38. Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani
655 KT, Georgakopoulos N, Harding P, Polanski K, Huang N,
656 Nowicki-Osuch K, et al: **scRNA-seq assessment of the human lung,**

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

657 **spleen, and esophagus tissue stability after cold preservation.**
658 *Genome Biol* 2019, **21:**1.

659 39. Agarwal D, Sandor C, Volpato V, Caffrey TM, Monzon-Sandoval J,
660 Bowden R, Alegre-Abarrategui J, Wade-Martins R, Webber C: **A**
661 **single-cell atlas of the human substantia nigra reveals cell-specific**
662 **pathways associated with neurological disorders.** *Nat Commun*
663 2020, **11:**4183.

664 40. Jakel S, Agirre E, Mendanha Falcao A, van Bruggen D, Lee KW,
665 Knuesel I, Malhotra D, Ffrench-Constant C, Williams A, Castelo-Branco
666 G: **Altered human oligodendrocyte heterogeneity in multiple**
667 **sclerosis.** *Nature* 2019, **566:**543-547.

668 41. Hill MC, Kadow ZA, Li L, Tran TT, Wythe JD, Martin JF: **A cellular atlas**
669 **of Pitx2-dependent cardiac development.** *Development* 2019, **146**.

670 42. Li B, Gould J, Yang Y, Sarkizova S, Tabaka M, Ashenberg O, Rosen Y,
671 Slyper M, Kowalczyk MS, Villani AC, et al: **Cumulus provides**
672 **cloud-based data analysis for large-scale single-cell and**
673 **single-nucleus RNA-seq.** *Nat Methods* 2020, **17:**793-798.

674 43. Mukund Sundararajan AT, Qiqi Yan: **Axiomatic Attribution for Deep**
675 **Networks.** pp. arXiv:1703.01365; 2017:arXiv:1703.01365.

676

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

677 **Figures**

678 **Figure 1. Overview of Fugue.** (A) Given a set of uncorrected single-cells, (B)
679 Fugue embedded their batch information as a learnable matrix and added
680 them to the expression profile for feature encoder training. (C) The feature
681 encoder was trained with contrastive loss. (D) At the feature extraction stage,
682 single-cell expression profiles were provided to the feature encoder to extract
683 embedding representation. (E) The embedding representation could be
684 utilized for downstream analysis such as visualization and cell clustering.

685 **Figure 2. Benchmark of batch-correction performance of Fugue across**
686 **the *simulation*, *cell line* and *PBMC datasets*.** (A) UMAP plot of cells from
687 simulation dataset, which consists of 5 different batches and 3 cell types. (B)
688 UMAP visualization of Fugue batch effect removing performance on the
689 *simulation dataset*. (C) UMAP plot of Fugue batch effect removing
690 performance on the *simulation_rm* dataset. (D) UMAP plot displays cells from
691 the *cell line dataset*, which consists of 3 different batches and 2 cell types. (E)
692 UMAP plot of Fugue batch effect removing performance on *cell line dataset*. (F)
693 Quantitative assessments of different batch effect removal methods on *cell line*
694 *dataset*. (G) UMAP plot displaying cells from *PBMC* dataset, which consists of
695 2 different batches and 8 cell types. (H) UMAP plot of Fugue batch effect
696 removing performance on *PBMC* dataset. (I) Quantitative assessments of
697 different batch effect removing methods on *PBMC* dataset. For (A-E, G-H),
698 cells are colored by batch (left panel) and cell type (right panel).

699 **Figure 3. Assessment of the batch-correction performance of Fugue.** (A)
700 UMAP plot of Fugue batch effect removing performance on the *census of*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

701 *immune project*. Cells are colored by batch in the left panel and cell ontology
702 label provided in the original publication in the right panel. (B) Quantitative
703 assessments of different batch effect removal methods on the *census of*
704 *immune project*. (C-E) UMAP plot depicting cells in the *lung dataset* before (C)
705 and after (D-E) Fugue integration. Cells are colored by batch in (C-D) and cell
706 cluster label in (E). (F) Bar plot depicting kBET scores of different batch effect
707 removing methods on the *lung dataset*. (G) Expression of cell type markers
708 across the feature embedding space. Dark and light colors represent low and
709 high relative expression values, respectively. (H) Dot plot representing cell
710 markers across batches. The size of each circle reflects the percentage of
711 cells in a cluster where the gene is detected, and the color intensity reflects the
712 average expression level within each cluster.

713 **Figure 4. Joint analysis of all immune cells across HCA repository with**
714 **Fugue.** (A) UMAP plot of the 17 immune cell types inferred from Fugue. Cells
715 are colored by cell type labels. (B) Dot plot showing cell type markers across
716 cell clusters. The size of each circle reflects the percentage of cells in a cluster
717 where the gene is detected, and the color intensity reflects the average
718 expression level within each cluster. (C-D) Violin plot deciphering expression
719 levels of cell type markers for hematopoietic stem cells (C) and natural killer
720 cells (D) across HCA cohorts. Cohorts with more than 1000 cells in each
721 cluster were displayed.

722 **Figure 5. Joint analysis of Fugue on batch-correction and gene**
723 **expression trajectory recovering during cell development.** (A-B) UMAP
724 plot of cells from the *embryonic mouse cardiac dataset* before (A) and after (B)
725 Fugue integration. Cells are colored by batch. (C) UMAP plot of the *embryonic*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

726 *mouse cardiac dataset* integrated by Fugue, colored by cell clusters. The
727 surrounding circle plot from inner to outer shows the cell types, batch labels
728 and pseudo-time scores of the 1% randomly downsampled cells. (D) Violin plot
729 deciphering expression levels of cell type markers across cell clusters. The
730 color intensity reflects the average expression level within each cluster. (E)
731 FLE plot revealing time course trajectories of cardiac development across
732 different cell types. Arrows indicate inferred cell state transition directions from
733 early to late pseudo time. (F) FLE plot revealing cell state transition directions
734 from HSC to all main blood lineages. (G-J) FLE plots of the main development
735 trajectory from HSC to B cell, T cell, monocyte and erythrocyte, respectively. B
736 cell series (G) were separation from HSCs towards B cell progenitors,
737 precursors of B cells and matured naïve B cells. B cells also differentiate into
738 mature B cells, plasma cells and memory B cells. T cell series trajectory (H)
739 was started from HSCs, followed by naïve T cells and finally mature T cells and
740 NK cells. Monocytes series trajectory (I) was started from HCA, and
741 transferred into DCs and CD14+ and CD16+ mature monocytes. Erythrocyte
742 series (J) differentiates from HSCs to megakaryocytes and erythroid cells. Pro,
743 progenitor; Pre, precursor; HSC, hematopoietic stem progenitor cell; DC,
744 dendritic cell; cDC, canonical dendritic cell; NK cells, natural killer cell; MSC,
745 multipotent progenitor cell.

746

747 **Supplementary figure and table legends**

748 **Supplementary Figure 1. Evaluation of Fugue's robustness over changes**
749 **of hyperparameters based on the *simulation dataset*.** (A) Effect of

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

750 momentum and queue size on the performance of Fugue. (B) The changes of
751 loss values versus epochs. Error bands are standard deviations determined
752 across 10 runs. (C) The performance of Fugue over the choice of data
753 augmentation ratios.

754 **Supplementary Figure 2.** UMAP plot of batch effect removing performance
755 on the *cell line dataset* across Seurat V3, ComBat, Harmony, BBKNN,
756 Scanorama, scVI, Pegasus L/S adjustment and INSCT. Cells are colored by
757 batch and cell type respectively.

758 **Supplementary Figure 3.** UMAP plot of batch effect removing performance
759 on the *PBMC* dataset across Seurat V3, ComBat, Harmony, BBKNN,
760 Scanorama, scVI, Pegasus L/S adjustment and INSCT. Cells are colored by
761 batch and cell label.

762 **Supplementary Figure 4.** UMAP plot of the *census of immune project* before
763 (A) and after (B-H) batch correction using ComBat, Harmony, BBKNN,
764 Scanorama, scVI, Pegasus L/S adjustment and INSCT. Cells are colored by
765 batch and cell type respectively.

766 **Supplementary Figure 5.** UMAP plot of batch effect removing performance
767 on the *lung dataset* across Seurat V3, Harmony, ComBat, Scanorama,
768 Scanorama, Pegasus L/S adjustment, scVI, BBKNN and INSCT. Cells are
769 colored by batch.

770 **Supplementary Figure 6. Assessment of the performance of Fugue on**
771 **the *brain dataset*.** (A-b) UMAP plot showing cells in the *brain dataset* before
772 (A) and after (B-C) Fugue integration. Cells are colored by batch in (A-B) and
773 cell cluster label in (C). (D) Bar plot depicting kBET scores of different batch

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

774 effect removing methods on HCA brain cohorts. (E) Expression of cell type
775 markers across the integrated embedding space. Dark and light colors
776 represent low and high relative expression values, respectively. (F) Dot plot of
777 cell type markers across batches. The size of each circle reflects the
778 percentage of cells in a cluster where the gene is detected, and the color
779 intensity reflects the average expression level within each cluster.

780 **Supplementary Figure 7. The similarity across batch embedding**
781 **representation of all samples in HCA repository.** (A) Heatmap of cosine
782 similarity of dimension reduction representations of the batch embedding
783 matrix across all samples. Each red frame represents samples from one cohort.
784 (B-D) show 3 representative projects from (A), namely C2, C1 and C39.

785 **Supplementary Figure 8. Fugue inferred cell clusters from HCA**
786 **embedding space.** (A) Importance scores of the top 5 marker genes for each
787 cell cluster. Representative markers are displayed on the right side. (B) Bar
788 plot displaying the cohort composition of cell clusters.

789 **Supplementary Figure 9.** TSNE plot of all quality-controlled cells from HCA.
790 TSNE plot in the top left corner is labeled by cell cluster labels. The others are
791 colored by the expression level of marker genes of immune cells. Light and
792 deep red represent low and high relative expression values, respectively.

793 **Supplementary Figure 10.** Violin plot deciphering expression levels of cell
794 type markers across immune cell subtype in HCA repository. Violins were
795 colored by cohorts. Cohorts with more than 1000 cells were displayed.

796 **Supplementary Figure 11.** Dot plot of cell type markers of cardiac cells
797 across batches. The size of each circle reflects the percentage of cells in a

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.12.472307; this version posted December 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

798 cluster where the gene is detected, and the color intensity reflects the average
799 expression level within each cluster.

800 **Supplementary Figure 12.** Correlation of pseudo-time and expression level of
801 cell differentiation markers across cardiac cell types. The curves representing
802 polynomial fits for each batch.

803 **Supplementary Figure 13. FLE embedding space of the *census of***
804 ***immune project* integrated by Fugue.** Cells are colored by batch (A) and cell
805 type (B-C). (B) and (C) displaying the major cell types in cord blood (B) and
806 bone marrow (C), respectively.

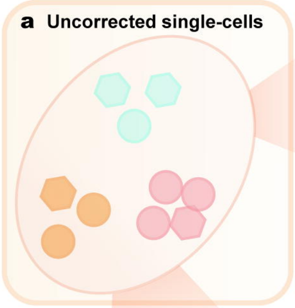
807 **Supplementary Figure 14. UMAP plot of the simulated cells.** (A-B)
808 deciphering the *simulation dataset* and (C-D) deciphering the *simulation_rm*
809 *dataset,* which was obtained by manually removing cell type 1 from batches
810 2-5 and retaining them in batch 1.

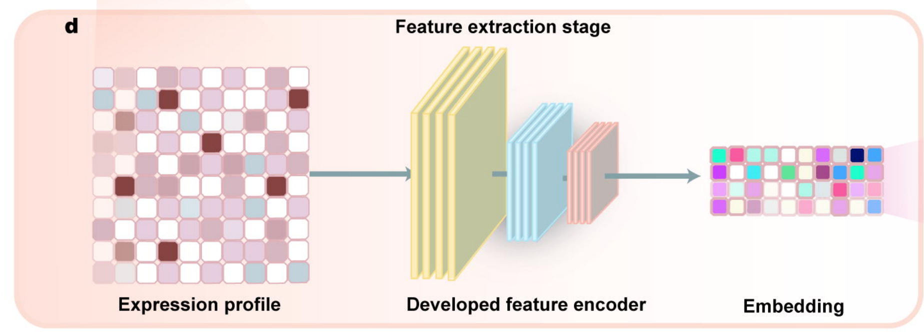
811 **Supplementary Table 1.** Detailed information of datasets from HCA
812 repository.

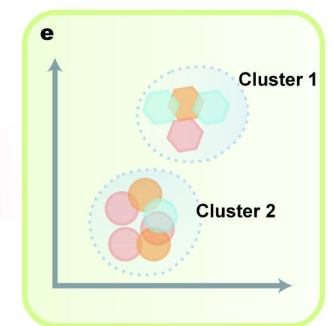
813 **Supplementary Table 2.** The 250 genes with the highest importance scores of
814 HCA cell clusters were inferred from Fugue. Marker genes of each cluster
815 were colored in blue.

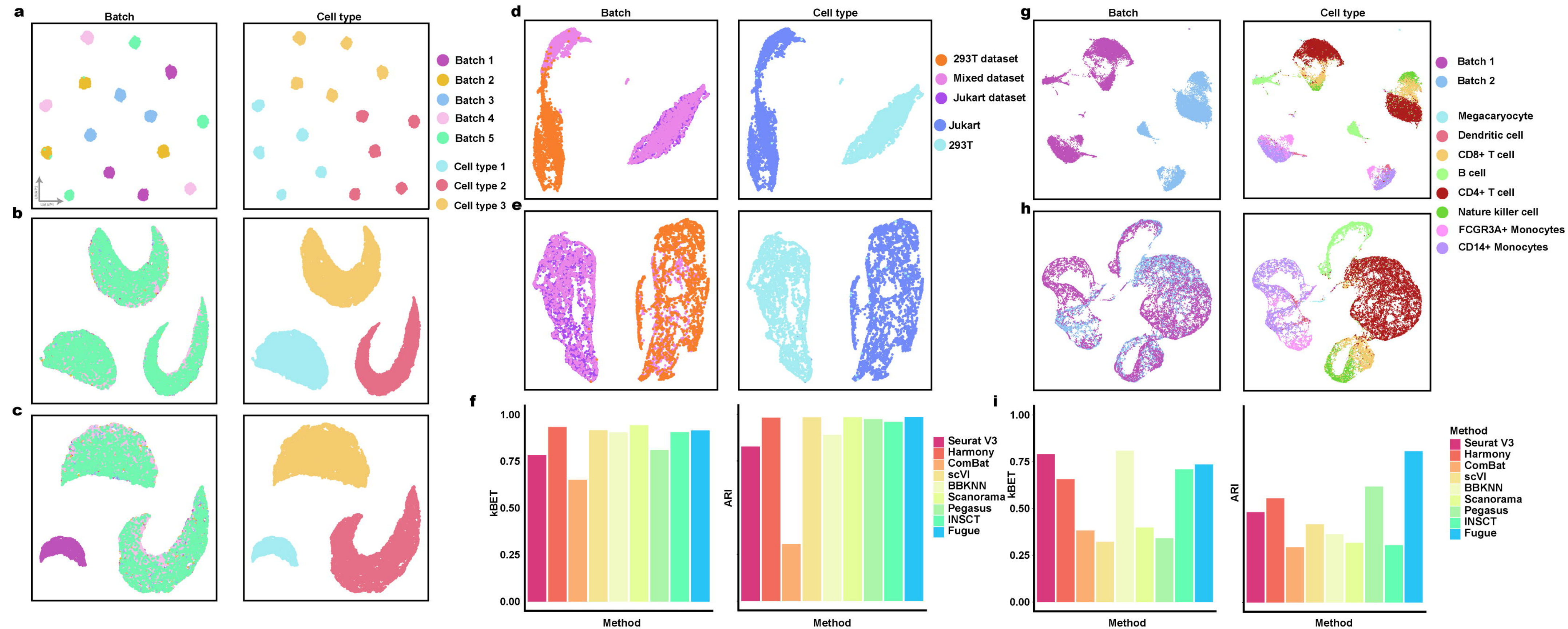
816 **Supplementary Table 3.** Detailed information of benchmark datasets, their
817 gene filtering and hyperparameter settings of Fugue.

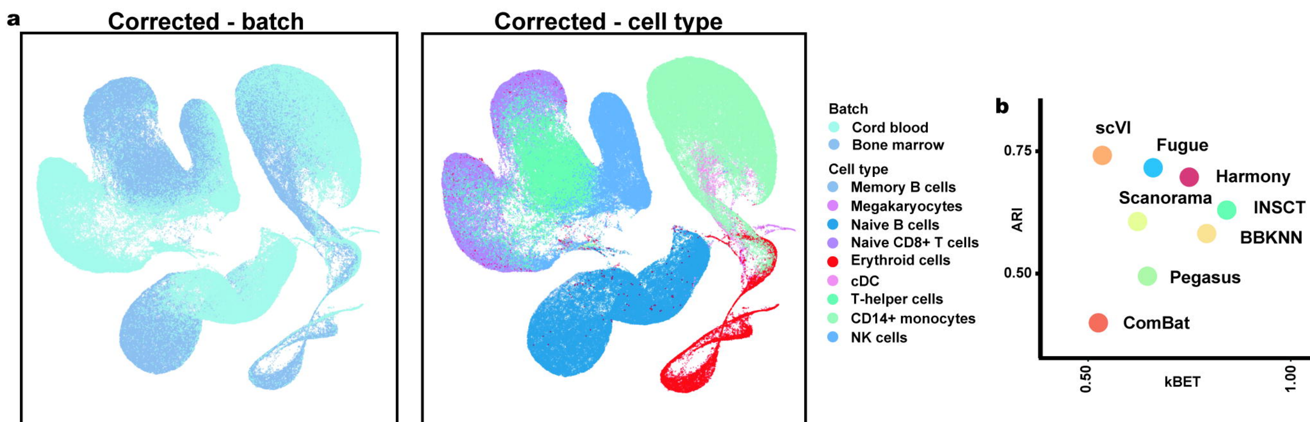
818 **Supplementary Table 4.** Detailed information of the benchmark methods.

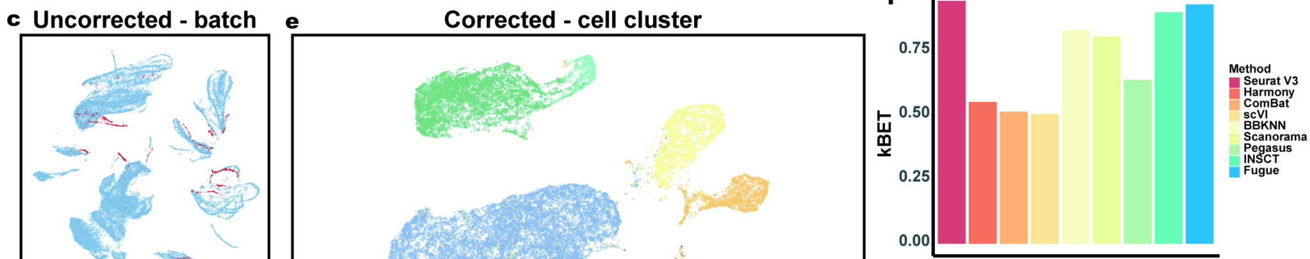

Batch
Cell type
a Uncorrected single-cells
Training stage
b
Expression profile
Batch embedding
Gene
Single-cell
Concatenation
Input
c Self-supervised learning
Contrastive loss
Feature encoder

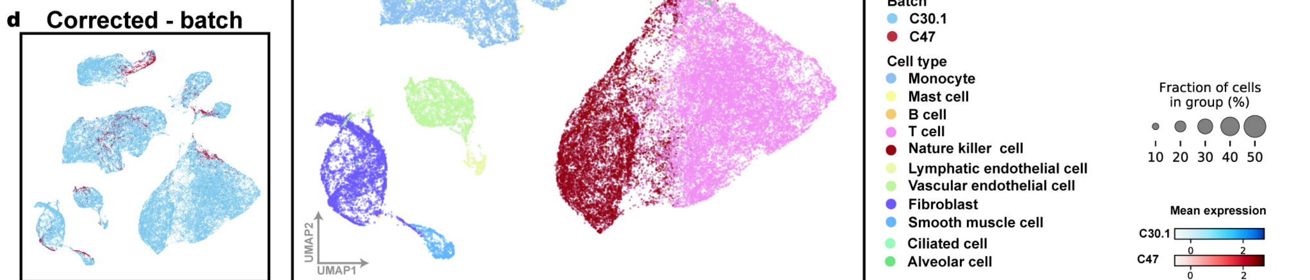

d
Feature extraction stage
Expression profile
Developed feature encoder
Embedding

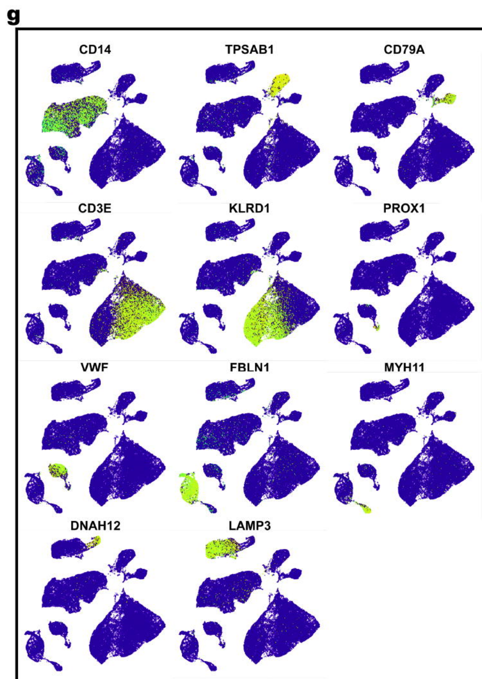

e
Cluster 1
Cluster 2

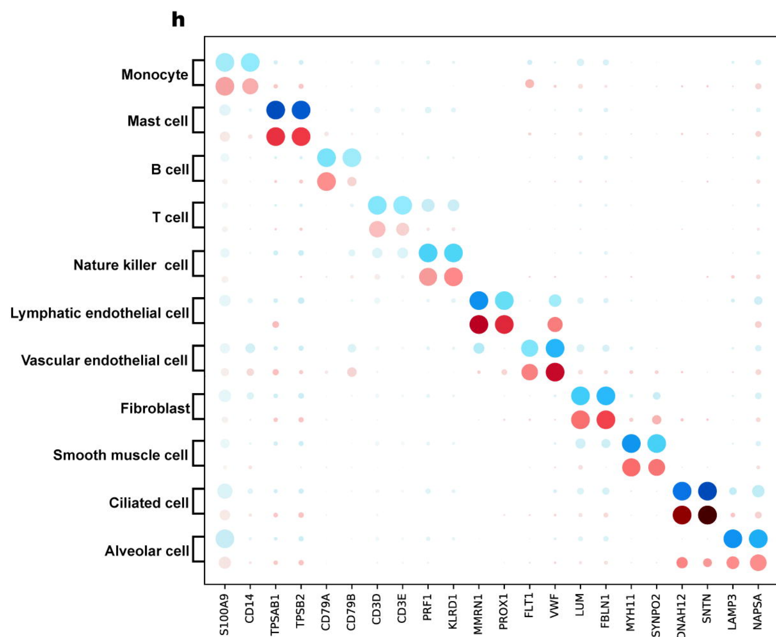

a
Batch
Cell type
Batch 1
Batch 2
Batch 3
Batch 4
Batch 5
Cell type 1
Cell type 2
Cell type 3
b
c
d
Batch
Cell type
293T dataset
Mixed dataset
Jukart dataset
Jukart
293T
e
f
kBET
ARI
Method
Seurat V3
Harmony
ComBat
scVI
BBKNN
Scanorama
Pegasus
INSCT
Fugue
g
Batch
Cell type
Batch 1
Batch 2
Megacaryocyte
Dendritic cell
CD8+ T cell
B cell
CD4+ T cell
Nature killer cell
FCGR3A+ Monocytes
CD14+ Monocytes
h
i
kBET
ARI
Method
1.00
0.75
0.50
0.25
0.00

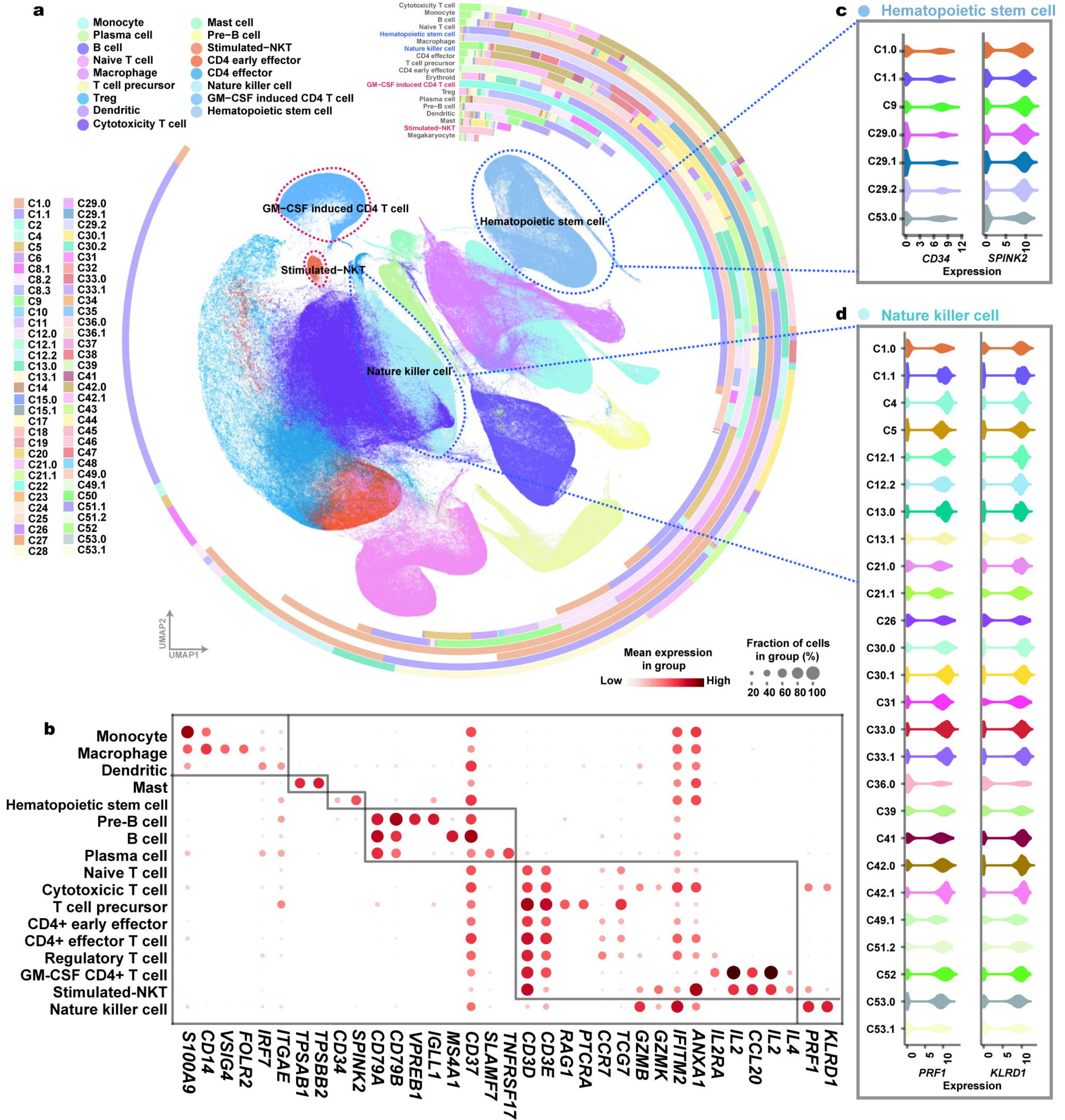

a
Corrected - batch
Corrected - cell type
Batch
Cord blood
Bone marrow
Cell type
Memory B cells
Megakaryocytes
Naive B cells
Naive CD8+ T cells
Erythroid cells
cDC
T-helper cells
CD14+ monocytes
NK cells
b
ARI
kBET
scVI
Fugue
Harmony
Scanorama
INSCT
BBKNN
Pegasus
ComBat
0.75
0.50
0.50
1.00

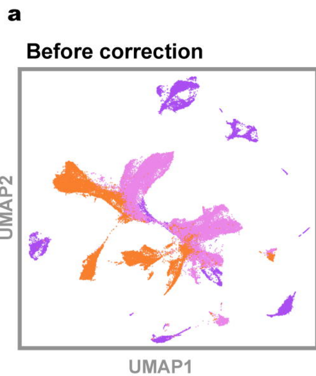

c Uncorrected - batch
e Corrected - cell cluster
f
kBET
Method
Seurat V3
Harmony
ComBat
scVI
BBKNN
Scanorama
Pegasus
INSCT
Fugue
0.75
0.50
0.25
0.00

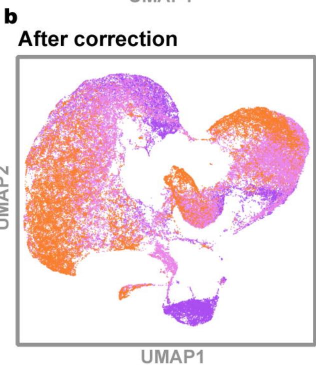

d Corrected - batch
UMAP2
UMAP1
Batch
C30.1
C47
Cell type
Monocyte
Mast cell
B cell
T cell
Nature killer cell
Lymphatic endothelial cell
Vascular endothelial cell
Fibroblast
Smooth muscle cell
Ciliated cell
Alveolar cell
Fraction of cells in group (%)
10 20 30 40 50
Mean expression
C30.1
C47
0 2

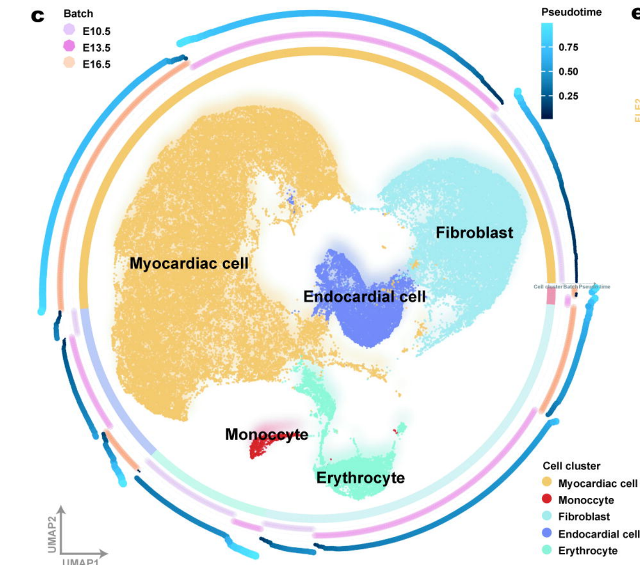

g
CD14
TPSAB1
CD79A
CD3E
KLRD1
PROX1
VWF
FBLN1
MYH11
DNAH12
LAMP3

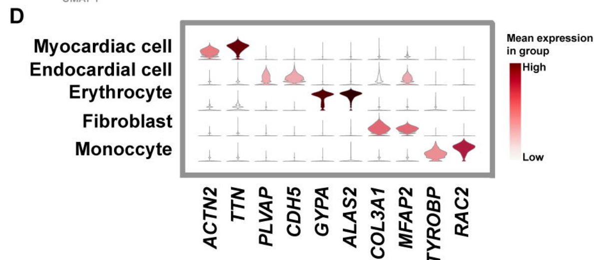

h
Monocyte
Mast cell
B cell
T cell
Nature killer cell
Lymphatic endothelial cell
Vascular endothelial cell
Fibroblast
Smooth muscle cell
Ciliated cell
Alveolar cell
S100A9
CD14
TPSAB1
TPSB2
CD79A
CD79B
CD3D
CD3E
PRF1
KLRD1
MMRN1
PROX1
FLT1
VWF
LUM
FBLN1
MYH11
SYNPO2
DNAH12
SNTN
LAMP3
NAPSA

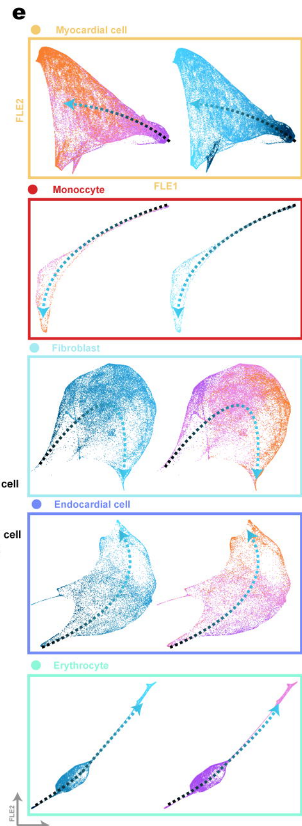

a
Monocyte
Plasma cell
B cell
Naive T cell
Macrophage
T cell precursor
Treg
Dendritic
Cytotoxicity T cell
Mast cell
Pre−B cell
Stimulated−NKT
CD4 early effector
CD4 effector
Nature killer cell
GM−CSF induced CD4 T cell
Hematopoietic stem cell
GM−CSF induced CD4 T cell
Stimulated−NKT
Hematopoietic stem cell
Nature killer cell
UMAP1
UMAP2
c Hematopoietic stem cell
CD34
SPINK2
Expression
d Nature killer cell
PRF1
KLRD1
Expression
Mean expression in group
Low
High
Fraction of cells in group (%)
20 40 60 80 100
b
Monocyte
Macrophage
Dendritic
Mast
Hematopoietic stem cell
Pre-B cell
B cell
Plasma cell
Naive T cell
Cytotoxicic T cell
T cell precursor
CD4+ early effector
CD4+ effector T cell
Regulatory T cell
GM-CSF CD4+ T cell
Stimulated-NKT
Nature killer cell
S100A9 CD14 VSIG4 FOLR2 IRF7 ITGAE TPSAB1 TPSBB2 CD34 SPINK2 CD79A CD79B VPREB1 IGLL1 MS4A1 CD37 SLAMF7 TNFRSF17 CD3D CD3E RAG1 PTCRA CCR7 TCG7 GZMB GZMK IFITM2 ANXA1 IL2RA IL2 CCL20 IL2 IL4 PRF1 KLRD1

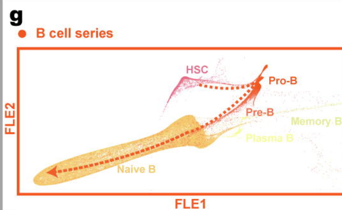

a
Before correction
UMAP2
UMAP1

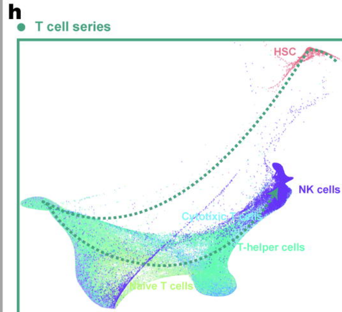

b
After correction
UMAP2
UMAP1

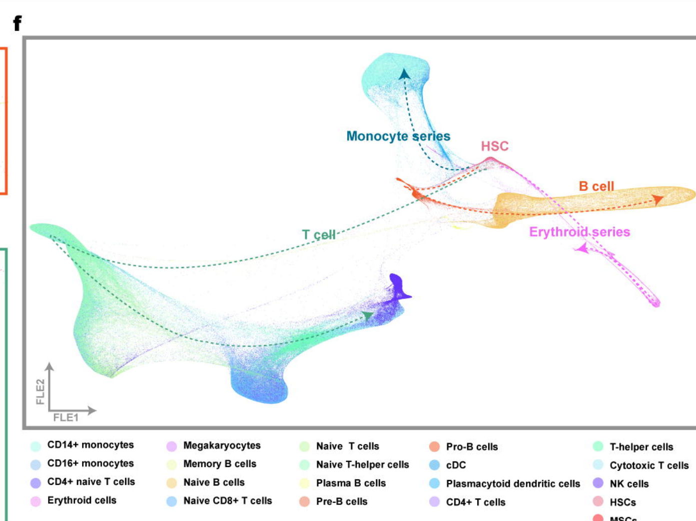

c
Batch
E10.5
E13.5
E16.5
Pseudotime
0.75
0.50
0.25
Myocardiac cell
Fibroblast
Endocardial cell
Monoccyte
Erythrocyte
Cell cluster
Myocardiac cell
Monoccyte
Fibroblast
Endocardial cell
Erythrocyte
UMAP2
UMAP1

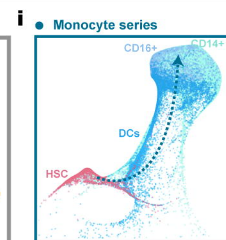

D
Myocardiac cell
Endocardial cell
Erythrocyte
Fibroblast
Monoccyte
Mean expression in group
High
Low
ACTN2
TTN
PLVAP
CDH5
GYPA
ALAS2
COL3A1
MFAP2
TYROBP
RAC2

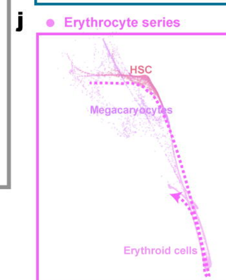

e
Myocardial cell
FLE2
FLE1
Monocyte
Fibroblast
Endocardial cell
Erythrocyte


g
B cell series
HSC
Pro-B
Pre-B
Memory B
Plasma B
Naive B
FLE2
FLE1


h
T cell series
HSC
NK cells
T-helper cells
Naive T cells


f
Monocyte series
HSC
B cell
T cell
Erythroid series
FLE2
FLE1
CD14+ monocytes
CD16+ monocytes
CD4+ naive T cells
Erythroid cells
Megakaryocytes
Memory B cells
Naive B cells
Naive CD8+ T cells
Naive T cells
Naive T-helper cells
Plasma B cells
Pre-B cells
Pro-B cells
cDC
Plasmacytoid dendritic cells
CD4+ T cells
T-helper cells
Cytotoxic T cells
NK cells
HSCs
MSCs


i
Monocyte series
CD16+
CD14+
DCs
HSC


j
Erythrocyte series
HSC
Megacaryocytes
Erythroid cells