Unbiased integration of single cell multi-omics data

Jinzhuang Dou[1], jdou1@mdanderson.org

Shaoheng Liang[1], sliang3@mdanderson.org

Vakul Mohanty[1], vmohanty@mdanderson.org

Xuesen Cheng[2], xuesenc@bcm.edu

Sangbae Kim[2], Sangbae.Kim@bcm.edu

Jongsu Choi[2], Jongsu.Choi@bcm.edu

Yumei Li[2], yumeil@bcm.edu

Katayoun Rezvani[4], krezvani@mdanderson.org

Rui Chen[2,3], ruichen@bcm.edu

Ken Chen[1,*]. kchen3@mdanderson.org

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center

[2]HGSC, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA

[3]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX, 77030, USA

[4]Department of Stem Cell Transplantation and Cellular Therapy, The University of Texas MD Anderson Cancer Center, Houston, Texas

* Correspondence: kchen3@mdanderson.org

1 Abstract

2 Acquiring accurate single-cell multiomics profiles often requires performing unbiased *in silico*

3 integration of data matrices generated by different single-cell technologies from the same

4 biological sample. However, both the rows and the columns can represent different entities in

5 different data matrices, making such integration a computational challenge that has only been

6 solved approximately by existing approaches. Here, we present bindSC, a single-cell data

7 integration tool that realizes simultaneous alignment of the rows and the columns between data

8 matrices without making approximations. Using datasets produced by multiomics technologies as

9 gold standard, we show that bindSC generates accurate multimodal co-embeddings that are

10 substantially more accurate than those generated by existing approaches. Particularly, bindSC

11 effectively integrated single cell RNA sequencing (scRNA-seq) and single cell chromatin

12 accessibility sequencing (scATAC-seq) data towards discovering key regulatory elements in

13 cancer cell-lines and mouse cells. It achieved accurate integration of both common and rare cell

14 types (<0.25% abundance) in a novel mouse retina cell atlas generated using the 10x Genomics

15 Multiome ATAC+RNA kit. Further, it achieves unbiased integration of scRNA-seq and 10x

16 Visium spatial transcriptomics data derived from mouse brain cortex samples. Lastly, it

17 demonstrated efficacy in delineating immune cell types via integrating single-cell RNA and

18 protein data. Thus, bindSC, available at https://github.com/KChen-lab/bindSC, can be applied in

19 a broad variety of context to accelerate discovery of complex cellular and biological identities and

20 associated molecular underpinnings in diseases and developing organisms.

Introductions

21

22 Advances in high-throughput single-cell technology such as single-cell RNA-sequencing (scRNA-

23 seq) [1] and mass cytometry [2] have enabled systematic delineation of cell types based on thousands

24 to millions of cells sampled from developing organisms or patient biopsies. For example, recent

25 application of combinatorial indexing based technology has generated the transcriptomic and

26 chromatin accessibility profiles of millions of cells in developing human fetus samples [3]. Rare cell

27 types and complex cellular states, however, remain challenging to discover, which necessitates the

28 development of multiomics technologies to simultaneously measure other cellular features,

29 including DNA methylation [4,5], chromatin accessibility [6-8] and spatial positions [9,10] in the same

30 cells. Although available single-cell multiomics technologies [8,11-14] can profile thousands to

31 millions of cells per experiment, the cost of the experiments is still quite high [15]; and the data

32 generated are often of lower throughput than those generated by unimodal technologies. These

33 restrictions necessitate the development of computational approaches that can accurately integrate

34 multiple data matrices generated by different technologies from the same biological samples to

35 acquire an accurate characterization of cellular identity and function.

36

37 However, different technologies create data matrices of different rows and columns, which

38 correspond to different sets of cells and different types of features. How to align cells and features

39 simultaneously across matrices is a core computational challenge. When the two sets of cells are

40 sampled uniformly from the same biological sample, it is safe to assume that there exists an optimal

41 way to align together cells of similar identities and features associated with these identities. This

42 is mathematically challenging, however, as there are many possible ways to simultaneously align

43 a large number of cells and features. To address this challenge, existing computational approaches

3

44 followed two directions: 1) aligning features empirically before aligning cells [16-19]; 2) obtaining

45 separate embeddings for each modality, followed by performing unsupervised manifold alignment

46 [20-22]. Taking integration of scRNA-seq and singe cell assay for transposase accessible chromatin

47 sequencing (scATAC-seq) as an example, the first category of methods require constructing a

48 "gene activity matrix" from scATAC-seq data by counting DNA reads aligned near and within

49 each gene [23]. This strategy considers only the basic cis-regulatory relations and ignores long-range,

50 trans-regulatory relationship established via other regulatory elements such as enhancers [6], which

51 are often critical to decipher cell identities. It also substantially simplifies (or loses) multifactorial

52 relations between transcription factors (TF) and target genes [24]. Based on pre-aligned features

53 generated by such empirical rules, Seurat applies canonical correlation analysis (CCA) and mutual

54 nearest neighbors (MNNs) to identify cells anchoring the two data matrices [17]; LIGER uses an

55 integrative non-negative matrix factorization (iNMF) to delineate shared and dataset-specific

56 features [19]; Harmony projects cells onto a shared embedding using principle components analysis

57 (PCA) and removes batch effects iteratively [18]. All these programs suffer from the aforementioned

58 limitations and thereby cannot yield a comprehensive, unbiased gene regulatory network,

59 particularly when chromatin changes are asynchronous from RNA transcriptions in cells

60 undergoing state transitions [25]. The second category of methods [20-22] do not require prior feature

61 alignment and are fully unsupervised. However, they depend heavily on the assumption that

62 feature variation across cells is driven by a few latent variables in both modalities [22]. This

63 assumption can get violated easily in datasets of complex biology involving dynamic processes

64 such as differentiation, reprogramming and transdifferentiation [22].

65

4

66   In this study, we develop a novel computational tool called bindSC (bi-order integration of single-

67   cell data). The key algorithm implemented in bindSC is called bi-CCA (bi-order canonical

68   correlation analysis). Bi-CCA learns the optimal alignment among rows and columns from two

69   data matrices generated by two different experiments. The alignment matrix derived from bi-CCA

70   can thereby be utilized to derive *in silico* multiomics profiles from aligned cells.

71

72   We assess our method on several challenging multimodality integration tasks between 1)

73   transcriptomic and chromatin accessibility data, 2) transcriptomic and spatial transcriptomic data,

74   and 3) transcriptomic and proteomic data. We validate scRNA-seq and scATAC-seq integration

75   accuracy using datasets obtained directly from multiomics technologies, including a novel mouse

76   retina cell atlas created by the 10x Genomics Multiome ATAC+RNA kit. We show that bindSC

77   enables comprehensive characterization of epigenetic regulatory states in a lung adenocarcinoma

78   cell-line A549 in response to dexamethasone treatment. And bindSC can align mouse retina cell

79   types accurately, for multi-subtype bipolar cells and rare horizontal cells. Moreover, bindSC

80   enables unbiased integration of spatial transcriptomics data with scRNA-seq data on mouse brain

81   cortex samples, as well as single-cell RNA data with protein data from peripheral blood

82   mononuclear cells. BindSC is implemented as an open-source R package available at

83   https://github.com/KChen-lab/bindSC.

84

85   **Results**

86

87   **Bi-order integration of multi-omics data**

88   BindSC takes as input two single-cell data matrices (X and Y) generated uniformly from the same

89   cell population by two different technologies (**Fig. 1a**). In most single-cell multi-omics integration

90   tasks, neither the alignment between the cells in X and those in Y, nor the alignment between the

91   features in X and those in Y is known. BindSC employs a bi-CCA algorithm developed in this

92   study to address this challenge (**Fig. 1b**). Briefly, bi-CCA introduces a gene score matrix Z to link

93   X and Y. The gene score matrix has the same rows as does X and the same columns as does Y. To

94   reduce computational cost, Z can be initialized based on prior knowledge. Taking integration of

95   scRNA-seq and scATAC-seq as an example, the gene score matrix can be initialized using the

96   "gene activity matrix" estimated by other programs such as Seurat. Bi-CCA then iteratively

97   updates Z to find an optimal solution which maximizes the correlation between X and Z and

98   between Y and Z in the latent space simultaneously. Details about this iterative procedure can be

99   found in **Methods** and **Supplementary Fig. 1a**.

100

101   Bi-CCA outputs canonical correlation vectors (CCVs), which project cells from two datasets onto

102   a shared latent space (referring below as "co-embedding"). A K-nearest neighbor (KNN) graph is

103   constructed based on Euclidean distances observed in the latent space, followed by modularity

104   optimization techniques to partition the KNN into highly interconnected subgraphs, each of which

105   corresponds to a putative cell type or state (**Fig. 1c**). Within each cluster, sub-clustering using

106   similar strategies is further performed to derive what we call pseudo-cells (**Methods**). Each

107   pseudo-cell encloses tens of cells from both datasets and thus has a consensus multiomic profile

108   summarized from constituting cells (**Fig. 1c-d**). The joint multiomic profiles thus enable 1)

109   characterizing gene and chromatin-accessibility relations from aligned scRNA-seq and scATAC-

110   seq data; 2) associating transcriptomic profiles with spatial locations from aligned scRNA-seq and

111 spatial transcriptomic data; 3) associating transcriptomic profiles with proteomic profiles from

112 aligned scRNA-seq and CyTOF data, and so on (**Fig. 1e**).

113

114 **Benchmarking bindSC performance on simulation datasets**

115 Existing integration methods such as Seurat, LIGER, and Harmony require pre-aligning features

116 across modalities, i.e., compressing cell-peak matrices obtained from scATAC-seq onto cell-gene-

117 activity matrices based on reference genome annotations. BindSC overcomes that restriction: its

118 generic mathematical formulations allow free alignment amongst features to be established from

119 data.

120

121 Under our formulation (**Methods**), $\mathbf{Z}$ has features (rows) aligned with $\mathbf{X}$ and cells (columns)

122 aligned with $\mathbf{Y}$. The introduction of $\mathbf{Z}$ enables bi-order alignment of the cells and the features,

123 respectively.

124

125 To quantify how much this step matters to overall integration accuracy, we performed a set of

126 simulation experiments. We started by creating a dataset $\mathbf{X}$ consisting of 3 cell clusters (types),

127 each having 333 cells and 1,000 genes using Splatter [26] (**Supplementary Fig. 2a**). We created a

128 second dataset $\mathbf{Y}$ and made it identical to $\mathbf{X}$: $\mathbf{X} = \mathbf{Y}$. We then constructed a gene score matrix $\mathbf{Z}$

129 from $\mathbf{Y}$ by permuting a fraction of features (rows), termed misalignment rate (MR), into different

130 orders. The features between $\mathbf{Z}$ and $\mathbf{Y}$ are perfectly aligned if MR equals 1 and are independent if

131 MR equals 0. We further added white noise on all the entries of $\mathbf{Z}$ at a given signal-noise-ratio

132 (SNR) level.

133

7

134 We then provide (**X**, **Z**) as input to the other methods (**Supplementary Fig. 2b**), mimicking how

135 they perform integration, while provide both (**X**, **Z**) and (**Y**, **Z**) to bindSC (**Supplementary Fig.**

136 **2c**). As described, rather than taking **Z** as it is from the input, bindSC will iteratively update **Z**

137 until reaching convergence.

138

139 Since we know the true cell type and dataset origin of the cells in these experiments, we can assess

140 the integration performance in terms of cell type classification accuracy and dataset alignment

141 accuracy in the co-embeddings. It is necessary to measure both types of accuracy, as a high cell

142 type classification accuracy can be achieved by simply projecting cells onto local clusters without

143 achieving uniform mixing of the two datasets. Similarly, a high dataset alignment accuracy can be

144 achieved by uniformly mixing cells from the datasets, regardless of their cellular identity. We used

145 Silhouette score for measure cell type classification accuracy and alignment mixing score to

146 measure the dataset alignment accuracy (**Methods**). We compared bindSC, CCA, Seurat, LIGER

147 and Harmony under default settings (**Supplementary Note 1**).

148

149 We obtained results from a range of MRs under SNR = 0.25 (**Fig. 2**). When there was no feature

150 misalignment (MR = 0), all methods achieved good performance. Even under this ideal scenario,

151 bindSC achieved the highest Silhouette score (> 0.75) (**Fig. 2a**). The worse performance of other

152 methods can be explained by the noise introduced to distort the manifold structures between **X** and

153 **Z**. CCA showed better performance than Seurat, which may be due partly to label transferring

154 errors introduced by Seurat's empirical anchor-based alignment approach. As MR increased from

155 0 to 0.9, the Silhouette score for bindSC remained stable (> 0.7), while all the other methods

156 showed a decreasing trend, especially for LIGER and Harmony. Harmony worked well when MR

8

157 ≤ 0.15 (**Fig. 2a-b**) but had a substantial drop on Silhouette score (< 0.1) when MR > 0.15. In

158 addition, its alignment mixing score dropped to 0 when MR > 0.2, with no mixing of cells from **X**

159 and **Z** in the co-embedding UMAP (**Fig. 2b**; MR = 0.5). Harmony takes cell coordinates from a

160 reduced dimensional PCA space and runs an iterative algorithm to adjust for dataset-specific

161 effects. When MR > 0.15, cells from **X** and **Z** already formed two dis-joint groups, which made

162 the downstream integration impossible for Harmony. The Silhouette score of LIGER showed

163 fluctuations but was always lower than 0.4. LIGER utilizes an integrated nonnegative matrix

164 factorization (iNMF) method to identify shared and dataset-specific metagenes across two datasets.

165 If it worked as designed, the errors caused by feature misalignment should be contained within

166 dataset-specific modules. However, variance explained by the data-specific modules appeared to

167 be small (< 1%). When MR ≥ 0.95, all methods including bindSC failed to achieve reasonable

168 integration. That was expected as **X** and **Z** (as well as **Y** and **Z**) became nearly independent.

169

170 As expected, increasing SNR level worsened the integration performance for most of the methods

171 except bindSC. For example, both CCA and Seurat had acceptable performance under MR = 0.5

172 and SNR = 0 (**Supplementary Fig. 3a**), but Seurat failed to separate cell type 2 and 3 accurately

173 when SNR = 0.25 or 0.5 (**Supplemental Fig. 4a**; **Fig. 2**). For SNR = 0.5, Harmony failed in both

174 alignment mixing (< 0.2) and classification (= 0) accuracy, even when MR was as low as 0.1

175 (**Supplementary Fig. 4**).

176

177 We repeated the above experiments by increasing the number of cells to 5,000 and 10,000,

178 respectively. Similarly, bindSC showed robust performance regardless of MR and SNR levels,

179 which was not achieved by other methods (**Supplementary Tables S2-3**). Overall, the simulation

180   results demonstrated that bindSC is robust to bias introduced by noise in the data and via pre-

181   aligning features, thanks to its ability to align both cells and features simultaneously.

182

183   **Integrating single cell epigenomic data with single cell transcriptomic data**

184   Integrating single cell epigenomic data with single cell transcriptomic data obtained from

185   unimodal technologies provides an opportunity to decipher epigenetic regulatory mechanisms

186   underpinning cell transcriptomic identity. We examined the performance of bindSC in integrating

187   the scRNA-seq and scATAC-seq data derived from lung adenocarcinoma (A549) cells after 0, 1,

188   and 3 hours of dexamethasone (DEX) treatment [6]. This dataset was generated using a

189   combinatorial indexing-based coassay (sci-CAR), which enabled jointly measurement of

190   chromatin accessibility and transcriptome in the same cells. In this dataset, 6,005 cells have sci-

191   RNA-seq profiles and 3,628 cells have sci-ATAC-seq profiles. Among them, 1,429 cells have both

192   RNA-seq and ATAC-seq profiles, which can be used as a gold standard for evaluating integration

193   accuracy of various methods (**Methods**).

194

195   For comparison, we ran the 4 methods on the same data and derived *in silico* co-embeddings. There

196   was relatively clear separation between cells acquired at 0 hour and those at 1 or 3 hours in the co-

197   embeddings (**Fig. 3a**). In terms of classifying cells by time, bindSC achieved the highest Silhouette

198   score and Harmony the second, whereas Seurat had the lowest score with many sub-clusters in its

199   co-embedding (**Fig. 3a-b**). As to alignment accuracy, bindSC and Harmony had similar scores,

200   whereas Seurat received a relatively low score (**Fig. 3b**). Similar trends were observed in a

201   previous study analyzing the same dataset [27]. As suggested by simulation, the low alignment

202   mixing score of Seurat was likely attributable to bias introduced in its anchor-based integration.

203

204   A perfect integration method would place the two instances of the 1,429 co-assayed cells onto

205   identical locations in the co-embeddings. We leveraged this expectation to compare the accuracy

206   of various methods. We defined a metric, called anchoring distance that measures the normalized

207   Euclidean distance between the two instances of a co-assayed cell in the co-embeddings

208   (**Methods**). BindSC achieved substantially shorter anchoring distances than the other methods (p

209     $< 2.2e\text{-}16$; Student t-test; **Fig. 3b**).

210

211   We further compared how accurately TF (or peak) -gene correlations can be inferred from the co-

212   embeddings produced by each method. For a fair comparison, we applied the same bindSC

213   workflow to derive pseudo-cells for the 4 methods (**Methods; Supplementary Note 2**).

214

215   For each TF-gene (and peak-gene) pair, we calculated a Spearman rank correlation coefficient

216   (SRCC) between the TF activity (and normalized peak) level and the gene expression level in the

217   pseudo-cells (**Methods**). We repeated the same calculation in the co-assayed cells to create a gold

218   standard. For each of the 4 methods in 3 types of relations: TF-gene, cis- peak-gene and trans-

219   peak-gene, we calculated a summary SRCC between the SRCCs obtained from the pseudo-cells

220   and the SRCCs obtained from the co-assayed cells. The summary SRCCs resulting from bindSC

221   were consistently higher than those obtained from Seurat, LIGER and Harmony in all the

222   categories of comparison, indicating that the bindSC multiomic profile had the highest accuracy.

223

224   We further examined the peak-gene association identified from the co-assayed cell profiles and

225   found 585 trans- peak-gene pairs being supported by isogenic Hi-C data generated in an

11

226   independent study [28]. Compared with other approaches, bindSC derived peak-gene SRCCs of the

227   highest level of agreement with those observed in the co-assayed cells (**Supplementary Fig. 5**).

228   Among the 585 trans- peaks, 470 appeared more strongly correlated with the corresponding gene

229   expression levels than did the corresponding cis- peaks. One example was the gene *CFLAR* and a

230   trans- peak at chr2:201,770,437-201,770,992, which is 200-kb upstream of *CFLAR* transcription

231   start site, spanning over three genes (**Fig. 3e**). The SRCC of this pair was 0.32 in the co-assayed

232   cells. It was lower but comparable (0.23) in the bindSC pseudo-cells, however, became

233   substantially lower (< 0.11) in the pseudo-cells generated by the other methods (**Supplementary**

234   **Fig. 5**).

235

236   The DEX treatment specifically targets the glucocorticoid receptor encoded by *NR3C1*, a TF that

237   activates the mRNA transcription of a handful of downstream genes. BindSC accurately

238   reconstructed the gene expression and TF activity kinetics of *NR3C1* (**Fig. 3d**), consistent with

239   what was depicted in the original study [6] using the co-assayed cells: the *NR3C1* expression level

240   decreased over time while its activity level increased; Even the slowing down trend of *NR3C1*

241   activity was captured.

242

243   We further evaluated the performance of bindSC in integrating scRNA-seq and scATAC-seq on

244   another available multi-omics dataset generated recently by SHARE-seq technology[14]. There were

245   a total of 37,774 cells from mouse skin tissues that had paired RNA and ATAC profiles. Compared

246   with other methods, bindSC again achieved significantly shorter anchoring distances

247   (**Supplementary Fig. 7**; **Supplementary Note 4**).

248

Comprehensive evaluation using a novel mouse retinal cell atlas

For comprehensive evaluation and comparison, we generated a novel multi-omics dataset from single nuclei of wild type mouse retina. Mouse retina is heterogeneous, composed of multiple neuronal and non-neuronal cell types, including five major neuron classes: photoreceptors (rods and cones), retinal ganglion cells (RGC), horizontal cells (HC), bipolar cells (BC), amacrine cells (AC), and a non-neuronal Müller glial cell (MG) [4,29,30]. While we [29] and others [31-33] have provided high-resolution single cell transcriptomic profiles of whole retina or specially sorted cell types on mouse and human retina tissue, little is known on the single-cell chromatin landscape of mouse retina tissue. Numerous studies [34-36] demonstrate the importance of transcription factors (TFs) on establishing or maintaining the chromatin landscapes that define retina cell identity. Therefore, integration of ATAC and RNA profiles at single cell resolution provides an exciting opportunity to comprehensively characterize cell types and rare cell subtypes in mouse retina.

We applied the newly released 10x Genomics Multiome ATAC+RNA kit on nuclei suspension acquired from adult mice retina samples. After performing standard quality control, we obtained an atlas of 9,383 nuclei of high-quality ATAC+RNA profiles. To define cell types, we first clustered the RNA and the ATAC data individually. Nineteen (19) clusters were identified from the RNA data alone, which included all the known major cell types with some subtypes identified: rod, BC (BC1~BC10), AC, RGC, cone, HC, MG and retina progenitor cells (RPC) (**Fig 4a and Supplementary Fig. 8**). Nineteen (19) clusters were also identified from the peak files of the ATAC data alone (**Fig. 4b**). Although known cell types appeared to be well separated in both modalities, there were some noticeable differences. For example, RGC cells and rod cells were separated clearly in the RNA data but partly blended together in the ATAC data, whereas ACs and

13

272   RGC cells were blended in the RNA data but well separated in the ATAC data. Interestingly, all

273   the 10 BC cell subtypes, defined based on RNA expression levels, were well separated in the

274   ATAC data except for BC1 and BC6. However, after reducing ATAC data to gene level in a gene

275   activity matrix, the cell types became considerably harder to delineate (**Fig. 4c**).

276

277   To obtain *in silico* multiomics profiles, we ran bindSC together with three other methods on the

278   data without using the known cell correspondence. As shown in the co-embedding UMAP

279   (**Fig. 4d-e**), bindSC successfully aligned cell types across modalities, with most cell types well

280   separated out (**Fig. 4d-f**). Interestingly, bindSC successfully aligned the HCs, which is quite rare

281   in the dataset (23 cells, <0.25% abundance). None of the other methods aligned the HCs correctly

282   as it was already difficult to separate the HCs from the ACs in the gene-level chromatin profiles

283   (**Fig. 4c and Supplementary Fig. 9**), the input to the other methods. Overall, the anchoring

284   distances in the co-embeddings generated by bindSC were considerably smaller than those

285   generated by the other methods in all the cell types assessed (**Fig. 4f**).

286

287   Note that bindSC aligned the 10 BC subtypes reasonably well (**Fig. 4g**), although separations in

288   the ATAC modality were not as clean as they were in the RNA modality. In comparison, Seurat

289   and LIGER failed to generate meaningful alignments among the BC subtypes (**Fig. 4h-j** and

290   **Supplementary Fig. 9**) while Harmony aligned a few subtypes successfully. These were due

291   partly to the fact that these methods used the low precision gene-level chromatin accessibility

292   profiles as the input (**Fig. 4c**).

293

14

294   Overall, our study demonstrated the power of multiomics in delineating rare cell types and proves

295   that bindSC can generate *in silico* multiomics profiles that are considerably more accurate than do

296   existing tools.

297

298   **Integrating scRNA-seq data with spatial transcriptomics (ST) data**

299   BindSC can integrate scRNA-seq data with spatial transcriptomics data to 1) assign spatial

300   locations to cells in the scRNA-seq data and 2) associate additional RNA features to the spatial

301   data for higher resolution delineation. For demonstration, we applied bindSC to integrate the

302   SMART-Seq2 data with the *in situ* spatial transcriptomics data generated by 10x Visium from the

303   same mouse frontal cortex tissue. These two datasets differ widely in number of cells: 1,072 spots

304   in the ST data versus 14,249 cells in the scRNA-seq data **(Supplementary Fig. 10a)**. The spots

305   on the Visium assay are at ~50 um resolution and each spot can contain tens of cells. There were

306   6 clusters identified from the ST data alone, which linked to distinct layers in the corresponding

307   histology images (**Supplementary Fig. 10b-c**) and 23 cell types from the scRNA-seq data alone

308   (**Supplementary Fig. 10d**).

309

310   We used bindSC and other programs to derive co-embeddings containing datapoints from both

311   datasets (**Fig. 5a**). BindSC achieved evidently higher alignment mixing scores than the other

312   programs (**Supplementary Fig. 11c**) while the Silhouette scores were similar (**Supplementary**

313   **Fig. 11b**). For each pseudo-cell in the scRNA-seq data, we calculated its probability to map to a

314   spatial location in the histology image. We then overlaid these cells on the histology image

315   coloring by their probability scores (**Methods**). Noticeably, several cell types in the scRNA-seq

316   data mapped to distinct spatial layers in the histology image, which is consistent with the known

317   cellular anatomy of mouse cortex, particularly for the laminar excitatory neuron cell types such as

318   L2.3 IT, L4, L5.IT, L5.PT, L6.IT, L6.CT, L6B and NP (**Fig. 5b**). Consistent with previous

319   observations, the oligodendrocyte-rich white matter (oligo cells) was mapped below the cortex.

320   BindSC and Seurat were also able to map inhibitory clear cell types such as Lamp5, Vip, Pvalb

321   and Sst in the scRNA-seq data to the histology image, but these cell types did not form distinct

322   spatial patterns. LIGER and Harmony, which had worse alignment mixing scores (**Supplementary**

323   **Fig. 11c**), failed to map these cells (**Supplementary Figs. 13-14),** especially the Vip cells. The

324   poor mapping of the inhibitory cells may also be attributable to the limited resolution of the Visium

325   technology.

326

327   Given that each spot in the ST data may encompass multiple cells from multiple cell types, we

328   hypothesized that the probability scores calculated from the co-embeddings can reveal the

329   composition of the cell types at each spot. **Fig. 5c** showed the relationship between cell type

330   abundance in the scRNA-seq data and abundance estimated based on probability scores in the ST

331   data. Results from bindSC achieved the best correlation (Pearson's R = 0.9). L6.IT, Sst and Vip

332   cell types were the top 3 most abundant cell types in both the scRNA-seq data and the ST data.

333   Seurat also performed reasonably well (Pearson's R = 0.83) while LIGER and Harmony performed

334   worse. Note that Lamp5 was the cell type that showed the largest discordance in the bindSC result.

335   In examining the spatial distributions of Lamp5 specific gene expressions such as *Lsp1*, *Npy2r*,

336   and *Dock5*, we could not find any spatial patterns (**Supplementary Fig. 11 d-e).** This finding may

337   indicate that Lamp5 does not have a characteristic spatial distribution.

338

339   **Integrating single-cell RNA with protein data**

340   Complex interplay exists between mRNAs and proteins [37]. Single-cell proteomic methods such as

341   mass cytometry (CyTOF) [2,38] measure abundance of a small set of (often 10-50) surface proteins

342   (epitopes) and provide functional quantification of various cell populations. Integrating single-cell

343   RNA and protein data from the same sample can potentially achieve higher resolution

344   characterization and enable discovery of novel cellular states and associated features. BindSC can

345   be applied for such a task. Notice that this task cannot be achieved by any of the existing tools

346   because the mRNA and protein expression levels derived from the same genes are not well

347   correlated, due to complex post-transcriptional modifications and technological limitations [39].

348   CITE-seq [40] performs jointly profiling of epitope and mRNA levels in the same cells and can be

349   used to evaluate the results of *in silico* integration.

350

351   We used a CITE-seq dataset consisting of 30,672 human bone marrow cells with a panel of 25

352   antibodies [17]. We split the data into an RNA matrix and a protein matrix. Unsupervised clustering

353   of the RNA matrix revealed cell types largely consistent with those in the protein matrix, except

354   for some noticeable differences (**Fig. 6a-b**). CD8+ and CD4+ T cells were partly blended together

355   in the RNA data but separated clearly in the protein data. On the other hand, conventional dendritic

356   cells (cDC2) were separated from other clusters in the RNA data but were intermixed with other

357   cell types in the protein data. In contrast, unsupervised clustering of the gene expression levels of

358   the 25 protein-homologous RNAs could not yield meaningful classification (**Fig. 6c**).

359   Consequently, Seurat, LIGER and Harmony, which work with only data matrix of 25 homologous

360   features, failed to produce meaningful co-embeddings (**Supplementary Fig. 15**): the cells from

361   the protein data were well clustered, but those from the RNA data were not meaningfully

362   distributed in the co-embeddings.

363

364   We then tested bindSC on this task. The matrix **X** was set as the protein matrix, **Y** the RNA matrix

365   of 3,000 highly variable genes, and **Z** the RNA matrix containing only the 25 protein-homologous

366   genes. Remarkably, the majority of the cells from the two modalities became well aligned in the

367   co-embedding (**Fig. 6d-e**), as they are expected to be. Similar to our previous experiments, we

368   calculated the anchoring distance between the protein and the RNA cells deriving from the same

369   original cells in the co-embeddings. The overall anchoring distance for bindSC was significantly

370   lower than those obtained by Seurat, LIGER, Harmony, or random guesses (p-value < 2.2e-16;

371   Student t-test; **Fig. 6f**). Notably, the bulk of CD4+ and CD8+ T cells in the RNA data became well

372   separated in the co-embedding (**Fig. 6d-e**), thanks to the power of integration. Moreover, the

373   anchoring distances revealed the extent of differences between the levels of the RNAs and those

374   of the homologous proteins in individual cell types (**Fig. 6g**). Interestingly, relatively rare cell

375   types such as HSC, Prog/NK, LMPP, and CD16+ Mono appeared relatively well anchored,

376   whereas relatively common cell types such as CD8 naïve, CD8 memory, B progenitor, Treg, etc.

377   appeared less well anchored. This indicates that there are higher degrees of post-transcriptional

378   heterogeneity in cell types conducting adaptive immune surveillance [41].

379

Discussion

381   Despite the ground-breaking advances in single-cell technologies, including multiomics

382   technologies, there always exists a need to computationally integrate multiple data matrices of

383   different modalities from the same biological samples to derive a more comprehensive

384   characterization of cellular identities and functions.

385

386   Our method bi-CCA and tool bindSC appeared to have addressed this important analytical

387   challenge without compromising biological complexity in the data. In our experiments, bindSC

388   successfully integrated data obtained from a wide variety of vastly different technologies covering

389   transcriptomes, epigenomes, spatial-transcriptomes and proteomes, and clearly outperformed

390   existing tools such as Seurat, LIGER and Harmony, when being evaluated objectively using true

391   single-cell multiomics data derived from the same cells. In particular, Seurat, LIGER, and

392   Harmony are essentially first-order solutions that can be applied to only rows or columns but not

393   both simultaneously. That approach introduced biases in the results and restricted the utility of

394   those tools in discovering complex cell-type relations and molecular interactions. For instance,

395   they consider only the basic cis-regulatory relations and ignores trans-regulatory relations [6]

396   established via distal enhancers, as exemplified in the interaction between *CFLAR* and a 200 kbps

397   upstream putative enhancer site discovered by bindSC and validated by Hi-C in the DEX-treated

398   A549 data. Other scATAC-seq analysis pipelines such as MAESTRO [16] and ArchR [42] have similar

399   restrictions.

400

401   Similarly, bindSC was able to meaningfully associate the expression levels of mRNAs with those

402   of the surface proteins, a very challenging task due to complexity in post-transcriptional

403   modification. The resulting co-embedding offered deeper biological insights than embeddings

404   derived from single modality or by using other existing approaches. For example, CD4+ T cells

405   became evidently separated from CD8+ T cells and so did pDC cells from other cell types.

406

407   BindSC also achieved meaningful mapping of scRNA-seq data to spatial locations in the brain

408   cortex samples, after integrating with the ST data. Even though the two datasets were not both at

409   single-cell resolution, bindSC was still able to achieve a meaningful integration.

410

411   Bi-CCA made two assumptions: 1) the two sets of cells are sampled uniformly from the same

412   biological sample; 2) the features of the two datasets are linearly correlated. These two

413   assumptions are met under many scenarios of current investigations, however, could be violated

414   when there are insufficient number of cells obtained from a rapidly developing cell population.

415   Consequently, the accuracy of the co-embedding could vary, depending on the sampling density

416   and the complexity of the population. We measured accuracy with respect to data complexity in

417   the simulation experiments, however, accuracy on a real dataset could be complex to gauge *a*

418   *priori* and will require case by case investigation in the context of a specific study, followed by

419   necessary experimental validation. Nonetheless, in this study we clearly proved based on objective

420   ground truth data that bi-CCA substantially avoided bias introduced by existing methods and that

421   bindSC is a robust implementation that can be applied to derive meaningful results on most recent

422   datasets containing thousands to tens of thousands cells (**Supplementary Table 1**).

423

424   BindSC is efficiently implemented in R. The major computational cost for bindSC is from

425   calculating cell/feature co-embedding coordinates using singular value decomposition (SVD)

426   (**Methods**); It typically requires $O(MNL)$ floating-point operations to construct MN cell-cell

427   distance matrix as input to SVD decomposition, where M and N are cell number of the two

428   modalities, respectively. To address this computational challenge, bindSC implements the "divide-

429   and-conquer eigenvalue algorithm". The divide part first splits cells into different blocks specified

20

430   by users, which can be solved in parallel with lower memory usage (**Supplementary Fig. 1b**). The

431   conquer part then merges results from each block recursively. Therefore, the maximal memory

432   usage of bindSC is independent of the total cell number.

433

434   Taken together, we believe that bindSC is likely the first tool that has achieved unbiased integration

435   of data matrices generated by different technologies and can be applied in broad settings. In the

436   single-cell domain, bindSC can clearly be applied to align cells and features simultaneously, which

437   are important for ongoing investigations in the Human Cell Atlas [43], the NIH HubMap [44], the

438   Human Tumor Cell Network [45] and on remodeling of tumor microenvironment [46]. Further, bindSC

439   can potentially be applied to other domains, such as integrating patient sample mRNA profiles

440   with cell-line drug-sensitivity data [47].

441 ## Methods

442 **BindSC workflow**

443 BindSC workflow for creating *in silico* single cell multi-omics embeddings consists of five steps:

444   1. individual dataset preprocessing including variable feature selection and cell clustering,

445   2. initializing feature matching across modalities (i.e., constructing gene score matrix),

446   3. identifying cell correspondence using the bi-CCA algorithm,

447   4. jointly clustering cells between two modalities in the co-embedding latent space and

448   constructing pseudo-cell level multi-omics profiles, and

449   5. downstream analysis for various integration tasks.

450 We formulate our method for the case of two modalities. Let $X \in \mathbb{R}^{M \times K}$ be a single-cell dataset of

451 features $g_1, g_1, \cdots, g_M$ by cells $c_1, c_1, \cdots, c_K$ and $Y \in \mathbb{R}^{N \times L}$ be a single-cell dataset of feature

452 $p_1, p_2, \cdots, p_N$ by cells $d_1, d_1, \cdots, d_L$. M and N are the numbers of features (e.g., gene expression,

453 chromatin accessibility, protein abundance level) in the two datasets. K and L are the number of

454 cells in the two datasets. Without loss of generality, we assume that features

455 $g_1, g_1, \cdots, g_M$ represent the gene expression levels and $M \leq N$. The important component of each

456 step is described as follows.

457

458   **1. Individual modality preprocessing**

459 For each modality, we follow standard processing pipeline, which includes variable feature

460 selection and unsupervised cell clustering. The cluster information derived from each modality is

461 used for downstream parameter optimization.

462

463   **2. Initializing feature matching across modalities**

464   Because features in the two datasets are generally different, bindSC requires one additional

465   transition matrix $Z \in \mathbb{R}^{M \times L}$ as input for bridging the integration of X and Y. The transition matrix

466   Z can be understood as the projection of Y to the feature space of the first dataset X. Taking the

467   integration of scRNA-seq and scATAC-seq as an example, the matrix Z can be derived from

468   scATAC-seq profiles by summing reads in gene bodies [17,19,23]. This can also be input from the

469   regulatory potential (RP) model in MAESTRO [16]. In a simpler case where X and Y have matched

470   features, the integration tasks fall into two categories: 1) batch correction for scRNA-seq data

471   across individuals, species, or technologies; 2) integration of scRNA-seq with spatial

472   transcriptome data. In those cases, the transition matrix Z is initialized as Y. In bi-CCA, Z is

473   updated iteratively. In the following text, the initial value of Z is denoted by $Z^{(0)}$.

474

475   **3. Bi-order canonical correlation analysis (Bi-CCA)**

476   The key algorithm implemented in bindSC is Bi-CCA, the concept of which extends traditional

477   CCA[17,24,48] to both rows and columns to enable capturing of correlated variables in cells and

478   features simultaneously. Bi-CCA introduces two cell-level projection matrices $U \in \mathbb{R}^{K \times E}$, $S \in$

479     $\mathbb{R}^{L \times E}$ such that the correlations between indices XU and ZS are maximized, and two feature-level

480   projection matrices $T \in \mathbb{R}^{M \times E}$, $V \in \mathbb{R}^{N \times E}$ such that the correlations between indices $Z'T$ and $Y'V$

481   are maximized. The optimization framework can be formulated as:

482 $$\underset{U,S,T,V,Z}{\arg\max} \, tr\{(XU)'ZS + (Z'T)' \, Y'V\} \tag{1}$$

483   subject to $(XU)'XU = I$, $(ZS)' \, ZS = I$, $(Z'T)' \, Z'T = I$, $(Y'V)' \, Y'V = I$.

484

485   If the transition matrix Z is known, the objective (1) can be divided into two disjoint traditional

486   canonical correlation analysis (CCA) problems. The left term is performed to identify cells of

487   similar (aligned) features, while the right term is performed to identify features shared by the

488   (aligned) cells, each of which can be solved in the CCA framework. However, it is difficult to

489   update transition matrix Z in equation (1) even when matrices U, S, T, V are available. This is

490   because: a) left optimization problem requires Z as input and the right optimization problem

491   requires Z' as input, leading (1) to a non-linear optimization problem; b) transition matrix Z shows

492   up in constraints.

493

494   Therefore, we modify equation (1) in a much more practical way. First, we standardize X to let it

495   have $X'X = I$, and standardize Y so that $YY' = I$. The standardization process can be seen in

496   **Algorithm 1**. Thus, equation (1) could be simplified as

497
$$\underset{U,S,T,V,Z}{\operatorname{argmax}} \, tr\{(XU)'ZS + (Z'T)'Y'V\} \tag{2}$$

498   subject to $U'U = I$, $(ZS)'\,ZS = I$, $(Z'T)'\,Z'T = I$, $V'V = I$.

499

500   To eliminate transition matrix Z from constraints, we introduce two transition matrices $Z_l \in \mathbb{R}^{M \times L}$

501   and $Z_r \in \mathbb{R}^{M \times L}$ and optimize the following problem:

502
$$\underset{U,S,T,V,Z_l,Z_r}{\operatorname{argmax}} \, tr\{(XU)'Z_l S + (Z_r'T)'\,Y'V\} + \|Z_l - Z_r\|_2 \tag{3}$$

503   subject to $U'U = I$, $S'S = I$, $T'T = I$, $V'V = I$, $Z_l Z_l' = I$, $Z_r' Z_r = I$.

504

505   To solve equation (3), we also standardize $Z^{(0)}$ to let $Z^{(0)'}Z^{(0)} = I$, and initialized with $Z_l := Z^{(0)}$.

506   The standard singular value decomposition (SVD) can be implemented to obtain the canonical

507   correlation vectors (CCVs) at cell levels. We used a user-defined number (E) of singular vectors

508   that approximate the CCVs (**Algorithm 2**). Here we term E to represent the cell-level

24

509   "dimensionality" in the latent space, which is a parameter required to be optimized (Details seen

510   in **Parameter optimization**).

511
$$(U, S) := \underset{U,S}{\operatorname{argmax}}\ tr(U'\ X'Z_l S) \text{ subject to } U'U = I,\ S'S = I. \tag{4}$$

512

513   Having CCV pair (U, S) obtained, we have cell correspondence in the latent space between two

514   datasets. The left transition matrix Z_l can be updated by:

515
$$(Z_l) := \underset{Z_l}{\operatorname{argmax}}\ tr(U'\ X'\ Z_l S) \text{ subject to } Z_l Z_l' = I. \tag{5}$$

516

517   The details of solving optimization problem (5) is in **Algorithm 2**.

518

519   We then set

520
$$(Z_r) := Z_l. \tag{6}$$

521   The similar SVD algorithm (**Algorithm 2**) is used to approximate CCVs:

522

523
$$(T, V) := \underset{T,V}{\operatorname{argmax}}\ tr(T'\ Z_r Y'V) \text{ subject to } T'T = I,\ V'V = I. \tag{7}$$

524

525   Once CCV pairs (T, V) are obtained, the features are matched in the latent space between two

526   datasets. The right transition matrix Z_r could be updated as:

527

528
$$(Z_r) := \underset{Z_r}{\operatorname{argmax}} \boldsymbol{tr}\{(Z_r'T)'Y'V\} \text{ subset to } Z_r'Z_r = I. \tag{8}$$

529   Next, we set

530
$$(Z_l) := Z_r, \tag{9}$$

25

531 The update process (4) ~ (9) are repeated until convergence. Because each of the subproblems is

532 convex with respect to the block variables being optimized, the algorithm is guaranteed to

533 converge to a fixed point (local minimum).

534

535 In the above framework, the transition matrix $\boldsymbol{Z}$ (represented by $\boldsymbol{Z}_l$ and $\boldsymbol{Z}_r$) is updated based on

536 original observed matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. In practice, we introduce the couple coefficient α ($0 \leq \alpha \leq 1$)

537 to assign weights on initialized matrix $\boldsymbol{Z}^{(0)}$ on transition process (6) and (9).

538
$$(\boldsymbol{Z}_r) := (1 - \alpha)\boldsymbol{Z}^{(0)} + \alpha\boldsymbol{Z}_l \tag{10}$$

539 and

540
$$(\boldsymbol{Z}_l) := (1 - \alpha)\boldsymbol{Z}^{(0)} + \alpha\boldsymbol{Z}_r \tag{11}$$

541 The couple coefficient α can reflect the contribution of initial $\boldsymbol{Z}^{(0)}$ on linking two modalities.

542 Equations (10) and (11) will be reduced to Equations (6) and (9) if $\alpha = 1$. The bi-CCA algorithm

543 will be reduced to traditional CCA if $\alpha = 0$. Selection of coefficient α can be seen in **Parameter**

544 **optimization**. Notably, the final $\boldsymbol{Z}_r$ and $\boldsymbol{Z}_l$ will be converged to different matrices if $\alpha < 1$. The

545 workflow of the iterative process is shown in **Supplementary Fig. 1a**.

546

547 **Jointly clustering cells across datasets in shared latent space and constructing pseudo-cell**

548 **level multi-omics profiles**

549 Equation (4) projects cells of two datasets into a correlated E-dimensional space with cell

550 coordinates $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_K)$ and $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_L)$, respectively. L2-normalization is

551 performed to remove global differences in scale, therefore

552
$$\hat{\boldsymbol{u}}_i = \boldsymbol{u}_i / \|\boldsymbol{u}_i\|_2 , i = 1,2, ..., K, \tag{12}$$

553
$$\hat{\boldsymbol{s}}_i = \boldsymbol{s}_i / \|\boldsymbol{s}_i\|_2 , i = 1,2, ..., L.$$

554

555   The shared nearest neighbor (SNN) graph is constructed by calculating the l-nearest neighbors (20

556   by default) based on the Euclidean distance of L2-normlized latent space. The fraction of shared

557   nearest neighbors between the cell and its neighbors is used as weights of the SNN graph. The

558   modularity optimization technique Leiden algorithm [49] is used to group cells into interconnected

559   clusters (termed meta-cluster) based on constructed SNN graph with a resolution parameter setting

560   by users (default 0.5).

561

562   To understand the molecular-level interaction among modalities, we construct the pseudo-cell

563   level multi-omics profiles. Briefly, for cells in each meta-cluster identified, the Leiden algorithm

564   is further performed based on SNN graph with a higher resolution (default = 2). In this way, cells

565   in each meta-cluster are further grouped into highly interconnected sub-clusters. We term such

566   sub-clusters as pseudo-cells. Only pseudo-cells that consist of at least n cells (default = 10) are

567   kept for downstream analysis, while the others are considered data-specific and discarded. Profiles

568   of the pseudo-cells are constructed by aggregating the cells included. We denote by $X^{pseudo} \in$

569      $\mathbb{R}^{M \times P}$ be pseudo-cell profiles of feature $g_1, g_1, \cdots, g_M$ and $Y^{pseudo} \in \mathbb{R}^{N \times P}$ be pseudo-cell profiles

570   of feature $p_1, p_2, \cdots, p_N$. P is the number of pseudo-cells.

571

572   **Algorithm 1. Standardizing inputs**

573   For input matrix X, we denote $\Sigma_{X'X}^{-1}$ as the generalized inverse of matrix $X'X$, and redefine $X :=$

574      $\Sigma_{X'X}^{-1/2} X$. For input matrix Y, we denote $\Sigma_{YY'}^{-1}$ as the generalized inverse of matrix YY', and redefine

575      $Y := Y\Sigma_{YY}^{-1/2}$. The standardization of Z_r and Z_l is the same as above.

576

27

577 **Algorithm 2. Calculating CCVs using SVD**

578 Take subproblem from the Equation (4) as an example, the goal of this module is to find projection

579 matrix $U \in \mathbb{R}^{K \times E}$ and $S \in \mathbb{R}^{L \times E}$ such that the correlations between two indices XU and $Z_l S$ are

580 maximized.

$$\underset{U,S}{\text{argmax}} \; tr(U' \, X' \, Z_l S) \; \text{ subject to } \; U'U = I \, , \, S'S = I. \tag{A1}$$

581

582 We define $\Sigma_{X'Z_l} := X'Z_l$. Let $U \in \mathbb{R}^{K \times D}$ and $S \in \mathbb{R}^{K \times D}$ be the matrices of the first E left- and

583 right singular vectors of $\Sigma_{X'Z_l}$. Then the optimum in Equation (A1) is solved.

584

585 **Algorithm 3. Updating transition matrix with orthogonality constraints**

586 Take subproblem from the Equation (5) as an example, the goal of this module is to optimize Z_l.

$$(Z_l) := \underset{Z_l}{\text{argmax}} \; tr(U' \, X' \, Z_l S), \text{ subject to } \; S'Z_l'Z_l S = I \tag{A2}$$

587

588 Equation (A2) is maximized when $Z_l S = XU$. Therefore, we can update Z_l as

$$(Z_l) := XU \, S' \Sigma_{SS'}^{-1}, \tag{A3}$$

589

590 where $\Sigma_{SS'}^{-1}$ denotes the generalized inverse of matrix SS'.

591

592 **Parameter optimization**

593 There are two key hyperparameters when running bindSC for integration. The first one is the

594 dimensionality E in the latent space and the second one is the couple coefficient α. Similar with

595 previous integration methods, the number of dimension E is very important on cell type

596 classification. We provide heuristics to guide the selection of E based on integration metrics

597 defined below, though sometimes helpful, are not substitute for biological insights. As a general

598 suggestion, we recommend starting with a value of E the same with the minimal number of

599 principle components (PCs) used on single modality clustering. The selection of couple coefficient

600 α depends on whether initialized $Z^{(0)}$ can represent the gene score of Y. We devise two metrics to

601 aid in selecting α, which measure integration performance on accuracy (no mixing of cell type)

602 and alignment (mixing of datasets) as defined below.

603

604 **1) Silhouette score**

605 To measure integration accuracy, we use the Silhouette score. Cluster for each cell is defined

606 using the cell type labels assigned from single dataset clustering. The Silhouette score assesses the

607 separation of cell types, where a high score suggests that cells of the same cell type are close

608 together and far from cells of a different type. The Silhouette score $s(i)$ for each cell is calculated

609 as following. Let $a(i)$ be the average distance of cell i to all other cells within i's cluster and $b(i)$

610 the average distance of i to all cells in the nearest cluster, to which cell i does not belong. Cell-cell

611 distance is computed in the L2-normalized co-embeddings (**Equation 12**). $s(i)$ can be computed

612 as:

613
$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & if\ a(i) < b(i) \\ 0 & if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & if\ a(i) > b(i) \end{cases}$$

614 Notably, given accurate correspondence between two modalities unknown, calculating $s(i)$ for cell

615 i in above equation only includes cells from the same dataset. We average values across all cells

616 to obtain an overall silhouette score for integration task.

617

618 **2) Alignment mixing score**

29

619   To measure integration mixing level, we use an alignment mixing score similar to those of previous

620   studies [50]. We first build a 20-nearest neighbor graph for each cell from L2-normalized co-

621   embeddings (**Equation 12**). For cell i, assuming proportions of cells from two modalities are p_{1i}

622   and p_{2i}, respectively, the alignment mixing score is calculated as

623     $$H(i) = -p_{1i} \log_2 p_{1i} - p_{2i} \log_2 p_{2i}$$

624   This corresponds to a mixing metric per cell, and we average values across all cells to obtain an

625   overall mixing metric.

626

627   We run bindSC by ranging α from 0 to 1 (with step size 0.1). Silhouette score and alignment

628   mixing score is calculated for each scenario. We select appropriate α that generally has best

629   performance in Silhouette score and alignment mixing score. Parameter values used in this study

630   can be seen in **Supplementary Table S1**.

631

632   **Performance and benchmarking**

633   In our evaluation, in addition to Silhouette score and alignment mixing score, we also consider

634   anchoring distance for evaluation datasets from multi-omics technologies, in which each cell has

635   paired profiles. For cell i from the first data, we calculate its distance (Euclidean distance) with all

636   cells in the second data as $\boldsymbol{D}_i$, and its distance with cell i in the second data as d_i. The anchoring

637   distance for cell i is calculated as $2d_i/max(\boldsymbol{D}_i)$. We then average anchoring distance across all

638   cells to obtain an overall anchor distance metric. The anchoring distance of cell i is 0 when it is

639   anchored correctly. The overall anchoring distance is 1 if we randomly layout cells on co-

640   embeddings.

641

642 **Simulation dataset**

643 We generated simulation dataset to evaluate method performance in integrating two modalities

644 assuming accurate feature matching unknown. We used Splatter tool [26] to simulate dataset X with

645 1,000 genes and cells with different sizes (from 1,000 to 10,000). The whole population is

646 consisted of three batches (cell types). To mimic the feature unmatching case, we first generated

647 dataset X_1 by randomly permutating genes of X under specified misalignment rates (MR). MR

648 ranges from 0 to 1 with step size being 0.05 in this study. X_1 is the same as X if MR = 0. Half of

649 genes are matched between X_1 and X if MR = 0.5. No genes are matched between X_1 and X if MR

650 = 1. Then we generated matrix Z by adding X_1 with white noise at certain level (i.e., Signal-Noise-

651 Ratio; SNR). SNR is set to be three levels (0, 0.25 and 0.5).

652

653 For method comparison, previous methods including traditional CCA, Seurat [17], Liger [19], and

654 Harmony tools [18] take X and Z as input assuming that cell correspondence between them is

655 unknown. bindSC takes two parts as input: 1) X and Z with cell correspondence unknown; 2) X

656 and Z with feature-level matching unknown (**Supplementary Fig. 2**).

657

658 **Preparation of dexamethasone (DEX) treated A549 cell dataset**

659 To investigate the ability of bindSC in integrating scRNA-seq and scATAC-seq profiles, we

660 explored the DEX-treated A549 dataset generated from sci-CAR technology, which uses

661 combinatorial indexing-based assay to jointly profile chromatin accessibility and mRNA on same

662 cell [6]. In the A549 dataset, DEX is a synthetic corticosteroid which activates glucocorticoid

663 receptor (GR), binds to thousands of locations, and alternates the expression of hundreds of genes

664 [51]. The human lung adenocarcinoma derived A549 cells after 0, 1, or 3 hours of 100nM DEX

665   treatment are assayed. The sci-RNA-seq dataset was from GSE117089

666   (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117089) and sci-ATAC-seq data was

667   from GSM3271041 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3271041). The

668   original A549 data includes sci-RNA-seq profiles for 6,150 cells and sci-ATAC-seq profiles for

669   6,260 cells. There are 1,429 cells co-assayed. Following Cao et al., pre-processing pipeline

670   (https://github.com/KChen-

671   lab/bindSC/blob/master/vignettes/A549/A549_preprocess.ATAC.Rmd), we binarized peak count

672   matrix for cells from both ATAC-seq only and co-assay. Loci present in less than 5 cells and cells

673   with less than 300 accessible loci were removed. Peaks within 1kb were merged and reads in

674   merged peaks were aggregated to generate a merged peak matrix, leading to 3,628 cells with

675   32,791 loci. Each locus' accessibility in each cell was calculated by dividing the cell's raw read

676   count by cell specific size factor using *estimateSizeFactors* function in *Monocle 2* [52]. For RNA-

677   seq data, cells with expression counts less than 500 and more than 9100 were removed. The gene

678   expression in each cell was also calculated by dividing the cell's raw read count by cell specific

679   size factor, followed by *log2* normalization. Genes with no variation in expression across cells

680   were further removed. The gene activity matrix was collapsed from the peak matrix by summing

681   all counts with the gene body plus 2kb upstream using *CreateGeneActivityMatrix* function in

682   Seurat3 [17]. We then picked top 10,000 variable genes in both sci-RNA-seq data and gene activity

683   data and used the overlapped 4,759 genes between them for integration. Finally, the sci-RNA-seq

684   matrix was composed of 6,005 cells with 4,759 genes, the gene activity matrix was composed of

685   3,628 cells with 4,758 genes, and the sci-ATAC-seq matrix was composed of 3,628 cells with

686   24,953 loci. There were 1,429 cells co-assayed.

687

Preparation of the mouse skin cell data

We examined the performance of bindSC in integrating the scRNA-seq and scATAC-seq data derived from mouse skin tissue. This dataset was generated using SHARE-seq (3) which included 34,774 cells that have joint profiles of RNA and ATAC profiles. The RNA data was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4156608. The ATAC data was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4156597. The final ATAC-seq matrix includes 25,594 cells on 74,161 peaks after quality control (including removing cells with less than 350 genes expressed; peaks that exist in less than 500 cells). In addition, 4,894 genes were identified that were highly variable in both gene expression and gene activity profiles. For this evaluation, we only focused on the third metric (e.g., anchoring distance) that represents the chance for the two instances of a co-assayed cell to appear in the co-embeddings.

Preparation of the mouse retina 10x Genomics Multiome ATAC+RNA data

One mouse retina was dissociated by papain-based enzymatic digestion as described previously [53] with slight modifications. Briefly, 45 U of activated papain solution (with 1.2 mg L-cysteine (Sigma) and 1200U of DNase I (Affymetrix) in 5ml of HBSS buffer) was added to the tissue and incubated at 37 °C for 20 minutes to release live cells. Post-incubation, papain solution was replaced and deactivated with ovomucoid solution (15 mg ovomucoid (Worthington biochemical) and 15 mg BSA (Thermo Fisher Scientific) in 10 ml of MEM (Thermo Fisher Scientific)). The remaining tissue clumps were further triturated in the ovomucoid solution and filtered through a 20nm nylon mesh. After centrifugation at 300g 10min at 4C, the singe cells were resuspended PBS with 0.04% BSA and checked for viability and cell count. About 1 million cells were pelleted and resuspend in chilled lysis buffer (10x Genomics), incubate for 2 minutes on ice while monitored

33

711   under microscope. 1ml of chilled wash buffer (10x Genomics) was added and sample was spun

712   down at 500g 5min at 4C and washed before resuspended in Diluted Nuclei Buffer (10x Genomics).

713   Nuclei concentration was determined using countess and proceed with transposition according to

714   manufacturer's recommendation (10x Genomics). After incubation for one hour at 37C, the

715   transposed nuclei were combined with barcoded gel beads, RT mix and partition oil on Chromium

716   to generate gel beads in Emulsion (GEMs). Single cell ATACseq library and 3'RNAseq library

717   were subsequently generated following recommended protocol from 10x Genomics. Libraries

718   were quantified and loaded on Novaseq 6000 and run with the following parameter: 151, 8, 8,

719   151bp. Data was analyzed using bcl2fastq (to generate fastq files) and cellranger pipeline (10x

720   Genomics).

721

722   **Preparation of the mouse frontal cortex cell data**

723   We investigate bindSC ability in integrating spatially resolved transcriptomic (ST) with

724   dissociated scRNA-seq. For the ST dataset, we used sagittal mouse brain slices generated from the

725   Visium v1 chemistry. The dataset was downloaded from https://support.10xgenomics.com/spatial-

726   gene-expression/datasets. The pre-processing workflow was guided by the Seurat3

727   (https://satijalab.org/seurat/v3.2/spatial_vignette.html). Briefly, cells were subset from anterior

728   region, followed by *sctransform* [54]. We then proceed to run dimensionality reduction and clustering

729   using standard workflow as did for scRNA-seq. Cluster ID 1,2,3,5,6,7 was extracted, followed by

730   segment based on exact position (Details in **Subset out anatomical regions** part in Seurat3

731   tutorial), leading to 1,072 cortical cells left for the ST data. One cortical scRNA-seq data composed

732   of ~14,000 adult mouse cortical cell taxonomy from the Allen Institute was collected

733   (https://www.dropbox.com/s/cuowvm4vrf65pvq/allen_cortex.rds?dl=1). This dataset was

734   generated using the SMART-Seq2 protocol [55]. The *sctransform* normalization was performed

735   based on 3,000 variable genes. We used the cell type annotation provided by published meta data

736   available. There was a total of 14,294 cortical cells with 34,617 genes for the scRNA-seq data.

737   Integration of scRNA-seq and ST is based on 2,316 variable genes overlapped between two

738   datasets.

739

740   To predict locations of each cell type from scRNA-seq in the histological images, we built a

741   support vector machine (SVM) that trained on cell profiles from scRNA-seq data. In the training

742   model, features were identified as cell coordinates in co-embeddings and labels were

743   corresponding cell types. The trained SVM was applied to ST data and output predicted probability

744   of each cell type at each spot. The *SpatialFeaturePlot* function in Seurat3 was used to overlay

745   predicted probabilities for each cell type on top of tissue histology.

746

747   **Preparation of human bone marrow cell dataset**

748   We examined the performance of bindSC in integrating the single-cell RNA and protein data

749   derived from human bone marrow tissue. This dataset was generated using the CITE-seq

750   technology [40], which included 30,672 cells that have joint profiles of RNA and a panel of 25

751   antibodies.   The   dataset   was   downloaded   from

752   https://satijalab.org/seurat/v4.0/weighted_nearest_neighbor_analysis.html. We extracted the 25

753   protein-homologous gene expression profile from the RNA data and kept cells that have total

754   expression count > 2. The final protein matrix includes 28,609 cells with 25 protein abundance

755   levels. The gene expression matrix includes 28,609 cells with 3,000 genes. The protein-

756   homologous RNA matrix includes 28,609 cells with the RNA levels of the 25 genes homologous

35

757   to the 25 proteins. To measure anchoring accuracy for each cell type, we used the third metric,

758   anchoring distance, which measures the distance of protein and gene expression for each cell in

759   co-embeddings.

760

761   **Motif-based Transcription Factors (TFs) activity estimation**

762   To estimate transcription factor activity from scATAC-seq data, we used default settings in

763   chromVAR [56] package. This approach quantifies accessibility variation across single cells by

764   aggregating accessible regions containing a specific TF motif. It calculated motif-based TF activity

765   by comparing the observed accessibility of all the peaks containing a TF motif to a background set

766   of peaks normalizing against known technical confounders.

767

768   **Acknowledgements**

779

Author contributions

780

781 K.C. conceptualized and supervised the project. J.D. designed the bindSC tool, implemented the

782 software and performed analysis. R.C., Y. L., X. C., S.K., J.C., contributed to mouse retina 10x

783 Genomics ATAC+RNA data generation, curation. V.M. contributed to data interpretation. J.D.,

784 S.L. and K.C. drafted the manuscript. All authors reviewed, edited, and approved the manuscript.

785

Competing interests

786

787 The authors declare no competing interests.

Reference

789   1   Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).

791   2   Spitzer, M. H. & Nolan, G. P. Mass cytometry: single cells, many features. *Cell* **165**, 780-791 (2016).

793   3   Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370** (2020).

794   4   Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600-604 (2017).

796   5   Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nature biotechnology* **36**, 428-431 (2018).

798   6   Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385 (2018).

800   7   Cusanovich, D. A. *et al.* A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309-1324. e1318 (2018).

802   8   Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature biotechnology* **36**, 70-80 (2018).

804   9   Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362** (2018).

806   10   Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361** (2018).

808   11   Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & Van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nature biotechnology* **33**, 285-289 (2015).

811   12   Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods* **12**, 519-522 (2015).

813   13   Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487-491 (2019).

815   14   Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* (2020).

817   15   Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nature methods* **17**, 11-14 (2020).

819   16   Wang, C. *et al.* Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome biology* **21**, 1-28 (2020).

821   17   Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).

823   18   Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*, 1-8 (2019).

825   19   Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-182 (2018).

827   20   Singh, R. *et al.* Unsupervised manifold alignment for single-cell multi-omics data. *BioRxiv* (2020).

829   21   Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised Topological Alignment for Single-Cell Multi-Omics Integration. *bioRxiv* (2020).

831 22 Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals
832   correspondence between single cell transcriptome and epigenome dynamics. *Genome*
833   *biology* **18**, 1-19 (2017).
834 23 Pliner, H. A. *et al.* Cicero predicts cis-regulatory DNA interactions from single-cell
835   chromatin accessibility data. *Molecular cell* **71**, 858-871. e858 (2018).
836 24 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals
837   folding principles of the human genome. *science* **326**, 289-293 (2009).
838 25 Lara-Astiaso, D. *et al.* Chromatin state dynamics during blood formation. *science* **345**, 943-
839   949 (2014).
840 26 Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing
841   data. *Genome biology* **18**, 1-15 (2017).
842 27 Jin, S., Zhang, L. & Nie, Q. scAI: an unsupervised approach for the integrative analysis of
843   parallel single-cell transcriptomic and epigenomic profiles. *Genome biology* **21**, 1-19
844   (2020).
845 28 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome.
846   *Nature* **489**, 57-74 (2012).
847 29 Liang, Q. *et al.* Single-nuclei RNA-seq on human retinal tissue provides improved
848   transcriptome profiling. *Nature communications* **10**, 1-12 (2019).
849 30 Masland, R. H. The neuronal organization of the retina. *Neuron* **76**, 266-280 (2012).
850 31 Menon, M. *et al.* Single-cell transcriptomic atlas of the human retina identifies cell types
851   associated with age-related macular degeneration. *Nature communications* **10**, 1-9 (2019).
852 32 Clark, B. S. *et al.* Single-cell RNA-seq analysis of retinal development identifies NFI
853   factors as regulating mitotic exit and late-born cell specification. *Neuron* **102**, 1111-1126.
854   e1115 (2019).
855 33 Shekhar, K. *et al.* Comprehensive classification of retinal bipolar neurons by single-cell
856   transcriptomics. *Cell* **166**, 1308-1323. e1330 (2016).
857 34 Brunet, I. *et al.* The transcription factor Engrailed-2 guides retinal axons. *Nature* **438**, 94-
858   98 (2005).
859 35 Nishida, A. *et al.* Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal
860   gland development. *Nature neuroscience* **6**, 1255-1263 (2003).
861 36 Marquardt, T. *et al.* Pax6 is required for the multipotent state of retinal progenitor cells.
862   *Cell* **105**, 43-55 (2001).
863 37 Ramanathan, M., Porter, D. F. & Khavari, P. A. Methods to study RNA–protein
864   interactions. *Nature methods* **16**, 225-234 (2019).
865 38 Krishnaswamy, S. *et al.* Conditional density-based analysis of T cell signaling in single-
866   cell data. *Science* **346** (2014).
867 39 Efremova, M. & Teichmann, S. Computational methods for single-cell omics across
868   modalities. *Nature Methods* **17**, 14-17 (2020).
869 40 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells.
870   *Nature methods* **14**, 865 (2017).
871 41 Doyle, H. A. & Mamula, M. J. Post-translational protein modifications in antigen
872   recognition and autoimmunity. *Trends in immunology* **22**, 443-449 (2001).
873 42 Granja, J. M. *et al.* ArchR: An integrative and scalable software package for single-cell
874   chromatin accessibility analysis. *bioRxiv* (2020).
875 43 Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A. & Teichmann, S. A. The Human Cell
876   Atlas: from vision to reality. *Nature News* **550**, 451 (2017).

877 44   Consortium, H. The human body at cellular resolution: the NIH Human Biomolecular Atlas
878   Program. *Nature* **574**, 187 (2019).
879 45   Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: charting tumor transitions
880   across space and time at single-cell resolution. *Cell* **181**, 236-249 (2020).
881 46   Sharma, A. *et al.* Onco-fetal Reprogramming of Endothelial Cells Drives
882   Immunosuppressive Macrophages in Hepatocellular Carcinoma. *Cell* **183**, 377-394. e321
883   (2020).
884 47   Warren, A. *et al.* Global computational alignment of tumor and cell line transcriptional
885   profiles. *BioRxiv* (2020).
886 48   Hardoon, D. R., Szedmak, S. & Shawe-Taylor, J. Canonical correlation analysis: An
887   overview with application to learning methods. *Neural computation* **16**, 2639-2664 (2004).
888 49   Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-
889   connected communities. *Scientific reports* **9**, 1-12 (2019).
890 50   Welch, J. D. *et al.* Single-cell multi-omic integration compares and contrasts features of
891   brain cell identity. *Cell* **177**, 1873-1887. e1817 (2019).
892 51   Vockley, C. M. *et al.* Direct GR binding sites potentiate clusters of TF binding across the
893   human genome. *Cell* **166**, 1269-1281. e1219 (2016).
894 52   Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature
895   methods* **14**, 979 (2017).
896 53   Siegert, S. *et al.* Transcriptional code and disease map for adult retinal cell types. *Nature
897   neuroscience* **15**, 487-495 (2012).
898 54   Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-
899   seq data using regularized negative binomial regression. *Genome biology* **20**, 1-15 (2019).
900 55   Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols
901   **9**, 171-181 (2014).
902 56   Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
903   transcription-factor-associated accessibility from single-cell epigenomic data. *Nature
904   methods* **14**, 975-978 (2017).
905 57   Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics.
906   *Nature neuroscience* **19**, 335-346 (2016).
907

908 **Figure Legends**

909 **Fig. 1 Overview of bindSC.** BindSC takes as input two data matrices produced by different modalities

910 from the same cell population (**a**). The modalities may include transcriptomes, epigenomes, spatial

911 transcriptomes and proteomes. Bi-order integration of two modalities (**X** and **Y**) with unpaired cells and

912 unmatched features using Bi-CCA algorithm **(b)**. In the data matrices, each row represents one gene/locus,

913 and each column represents one cell. The gene score matrix **Z** that links the first modality with the second

914 one is initialized by prior gene activity modeling (see **Methods**). Bi-CCA algorithm aims to update gene

915 score matrix **Z** iteratively by maximizing the correlation of between **X** and **Z** and between **Y** and **Z**

916 simultaneously. Based on canonical correlation vectors (CCVs) in the derived latent space, K-nearest

917 neighbor (KNNs) clustering is performed to define cell types in both modalities **(c)**. Within each cell-type

918 cluster, KNN clustering is further performed at a higher resolution to define pseudo-cells consisting of 10s

919 cells from both modalities. *In silico* multimodal profiles are constructed from cells assigned to the same

920 pseudo-cell **(d)**. The color in each box indicates the relative level of each feature, with white corresponding

921 to missing values. The multiomics feature profiles enable us to 1) link genes to regulatory elements, 2) map

922 RNA expressions to spatial locations and 3) delineate cells by both RNA and protein signatures **(e)**.

923

924 **Fig. 2 Benchmarking bindSC performance on simulation datasets.** Comparison of bindSC to CCA,

925 Seurat, LIGER, and Harmony based on Silhouette score and alignment mixing score **(a)**. The dataset

926 contains 1,000 genes and 1,000 cells equally distributed in 3 cell types. Signal-to-noise ratio (SNR) was set

927 at 0.25. X-axes denote the misalignment rates (MR) between features in the two datasets, which ranges

928 from 0 to 1. The features between two datasets have perfect match if MR = 0 and are unrelated if MR = 1.

929 UMAP views of the co-embeddings generated by bindSC, CCA, Seurat, LIGER, and Harmony **(b)**. From

930 top to bottom are results with MR = 0.1, 0.5, and 0.9, respectively. Each point denotes one cell that is

931 colored based on its true cell type label (red, green, or cyan).

932

933   **Fig.3 Benchmarking bindSC performance on the DEX-treated A549 cell-line data.** UMAP of cells

934   from DEX-treated A549 cell-line data for bindSC, Seurat, LIGER and Harmony respectively, colored by

935   collection time (red: 0 hour, green: 1 hour and blue: 3 hour) on the top panel and by technologies (grey: sci-

936   RNA and brown: sci-ATAC) on the bottom panel **(a)**. Comparison of the 4 methods based on Silhouette

937   score (top), alignment mixing score (middle) and anchoring distance (bottom) **(b)**. Also included for

938   comparison are metrics from randomly selected cells. Accuracy of *in silico* multiomics profile in pseudo-

939   cells **(c)**. The TF-gene correlation is quantified in each pseudo-cell (top panel) by calculating for each TF-

940   gene pair (1.8 million pairs total) a Spearman's rank-order correlation coefficient (SRCC) between the TF

941   activity level, inferred based on motif enrichment in the ATAC data, and the gene expression level in the

942   RNA data. The peak-gene correlation is quantified by calculating for each peak-gene pair a Spearman's

943   rank-order correlation coefficients (SRCC) between a normalized ATAC peak level and a gene expression

944   levels for the cis-peaks (middle panel, 7,833 pairs) and the trans-peaks (bottom panel, 118.7 million pairs),

945   respectively. X-axes are the SRCCs estimated from the co-assayed cells, which serve as the gold standard,

946   while Y-axes are the SRCCs estimated from the pseudo-cells generated by each method. The overall

947   concordance between X and Y are further quantified using a single SRCC shown on the up-left corner of

948   each subfigure. Cis is defined as gene bodies plus 2,000 bps upstream. Reconstructing the gene expression

949   and the TF activity level (Y-axes) of *NR3C1* using bindSC pseudo-cells **(d)**. X-axis is the averaged

950   treatment time of the cells in each pseudo-cell. A genome browser view showing putative regulatory

951   relations between an accessible distal site chr2:201770437-201770992 and the gene *CFLAR* **(e)**. The 6

952   tracks at the top show ATAC peak levels and gene expression levels at six time points. The track in the

953   middle shows chromatin interactome from published Hi-C data. The bottom track shows the *NR3C1* binding

954   targets (ChIP-Seq) peaks published in an independent study [28].

955

956   **Fig. 4 Integrating single-cell RNA-seq and ATAC-seq on a mouse retinal cell atlas.** UMAP views of

957   9,383 mouse retina cells based on gene expression levels in the RNA-seq data **(a)**, chromatin accessibility

958   peak profiles in the ATAC-seq data **(b)**, gene-level collapsed chromatin accessibility profiles **(c)**. The cells

42

959   are colored by cell types annotated based on RNA expression levels (**Supplementary Fig. 8**). BindSC co-

960   embeddings for the cells in the RNA-seq data **(d)** and those in the ATAC-seq data **(e)**. Anchoring distances

961   resulting from bindSC, Seurat, LIGER and Harmony **(f)**. The median anchoring distance for each cell type

962   was highlighted as a bold horizonal bar in the middle of each box in each panel. The dotted line denotes the

963   anchoring distance based on random guesses. Zoomed out UMAP views for the BC cells in the co-

964   embeddings generated by bindSC **(g)**, Seurat **(h)**, LIGER **(i)**, and Harmony **(j)**. Integration results for all

965   the cell types can be seen in **Supplementary Fig. 9**. RGC: retinal ganglion cells; HC: horizontal cells; BC:

966   bipolar cells; AC: amacrine cells; MG: Müller glial cell.

967

968   **Fig. 5 Integrating spatially resolved transcriptomic (ST) data with scRNA-seq data from mouse**

969   **frontal cortex**. UMAPs of the gene expression levels for the 14,249 cells profiled by SMART-Seq2 and

970   for the 1,072 spots profiled by the 10x Visium technology **(a)**. Cell type labels are from the original

971   publication[57]. Predicted locations of each cell type in the histological images **(b)**. Color gradient

972   corresponds to the probability score of a cell being mapped to a particular spatial location. Comparison of

973   cell type frequencies estimated from the ST data (Y-axis) to those estimated from the scRNA-seq data (X-

974   axis) **(c)**. Correlation coefficients (R) and P values are calculated based on Pearson's correlation analysis.

975   Each dot corresponds to a cell-type (labeled in different colors). The blue line and the grey shade represent

976   regression lines and 95% confidence intervals from performing linear regressions.

977

978   **Fig. 6 Integrating single-cell RNA with protein data produced by a CITE-seq assay.** The UMAP of

979   30,672 human bone marrow cells based on 25 surface protein levels **(a)**, 3,000 highly variable gene

980   expression levels **(b)** and 25 protein-homologous gene expression levels **(c)**. The cell type labels are from

981   the original study [17]. UMAP of the protein **(d)** and the RNA **(e)** expression data in the co-embedding

982   generated by bindSC. Comparison of anchoring distances generated by bindSC, Seurat, LIGER and

983   Harmony **(f)**. The red dotted line denotes the anchoring distance from random guesses. Anchoring distances

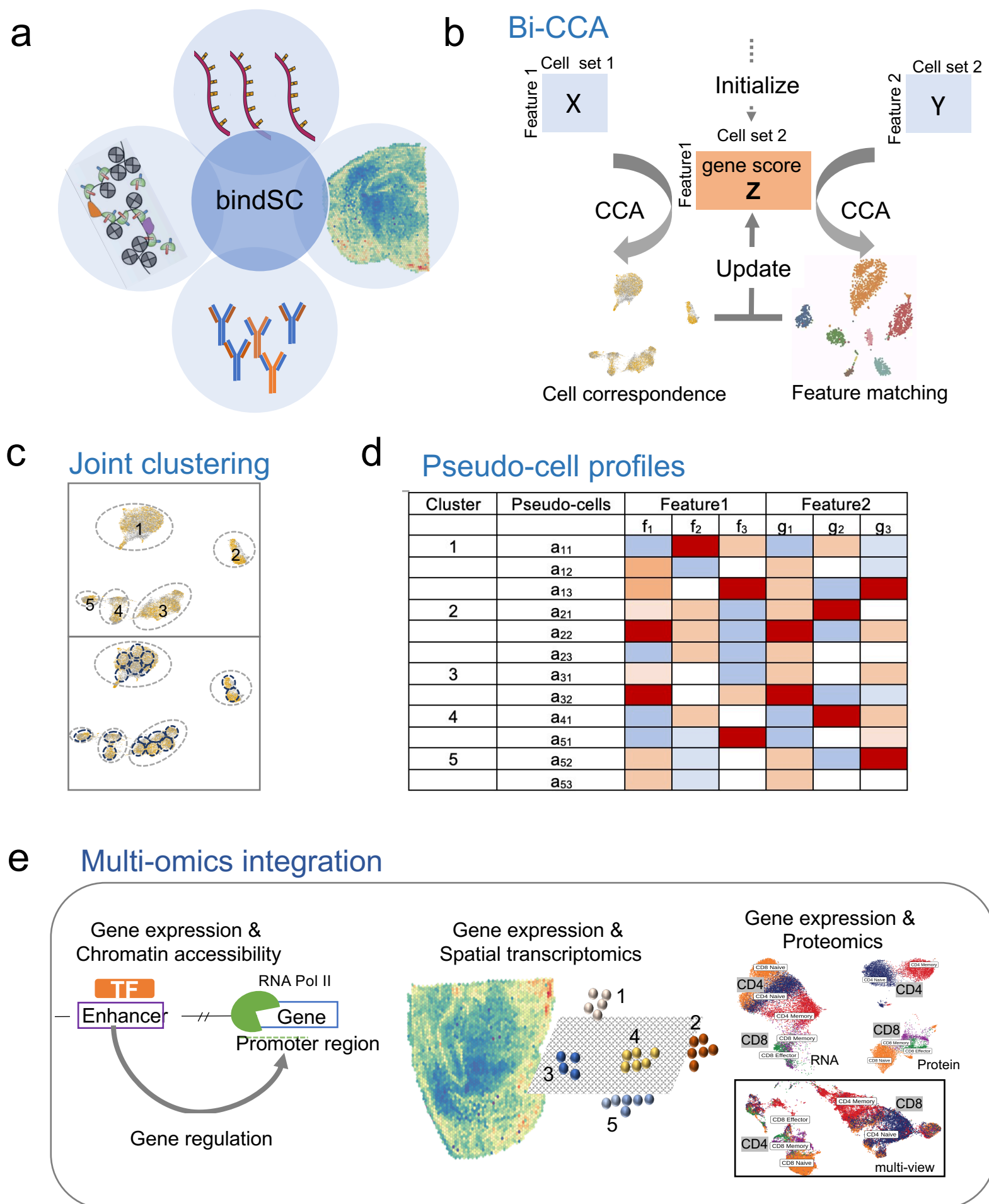984   for each cell type in the bindSC co-embedding **(g)**.

Fig. 1

a

b Bi-CCA

c Joint clustering

d Pseudo-cell profiles

Cluster	Pseudo-cells	Feature1			Feature2		
		f_1	f_2	f_3	g_1	g_2	g_3
1	a_{11}						
	a_{12}						
	a_{13}						
2	a_{21}						
	a_{22}						
	a_{23}						
3	a_{31}						
	a_{32}						
4	a_{41}						
	a_{51}						
5	a_{52}						
	a_{53}						

e Multi-omics integration

Fig. 1 Overview of bindSC. BindSC takes as input two data matrices produced by different modalities from the same cell population (**a**). The modalities may include transcriptomes, epigenomes, spatial transcriptomes and proteomes. Bi-order integration of two modalities (**X** and **Y**) with unpaired cells and unmatched features using Bi-CCA algorithm (**b**). In the data matrices, each row represents one gene/locus, and each column represents one cell. The gene score matrix **Z** that links the first modality with the second is initialized by prior gene activity modeling (see **Methods**). Bi-CCA algorithm aims to update gene score matrix **Z** iteratively by maximizing the correlation of between **X** and **Z** and between **Y** and **Z** simultaneously. Based on canonical correlation vectors (CCVs) in the derived latent space, K-nearest neighbor (KNNs) clustering is performed to define cell types in both modalities (**c**). Within each cell-type cluster, KNN clustering is further performed at a higher resolution to define pseudo-cells consisting of 10s cells from both modalities. *In silico* multimodal profiles are constructed from cells assigned to the same pseudo-cell (**d**). The color in each box indicates the relative level of each feature, with white corresponding to missing values. The multiomics feature profiles enable us to 1) link genes to regulatory elements, 2) map RNA expressions to spatial locations and 3) delineate cells by both RNA and protein signatures (**e**).
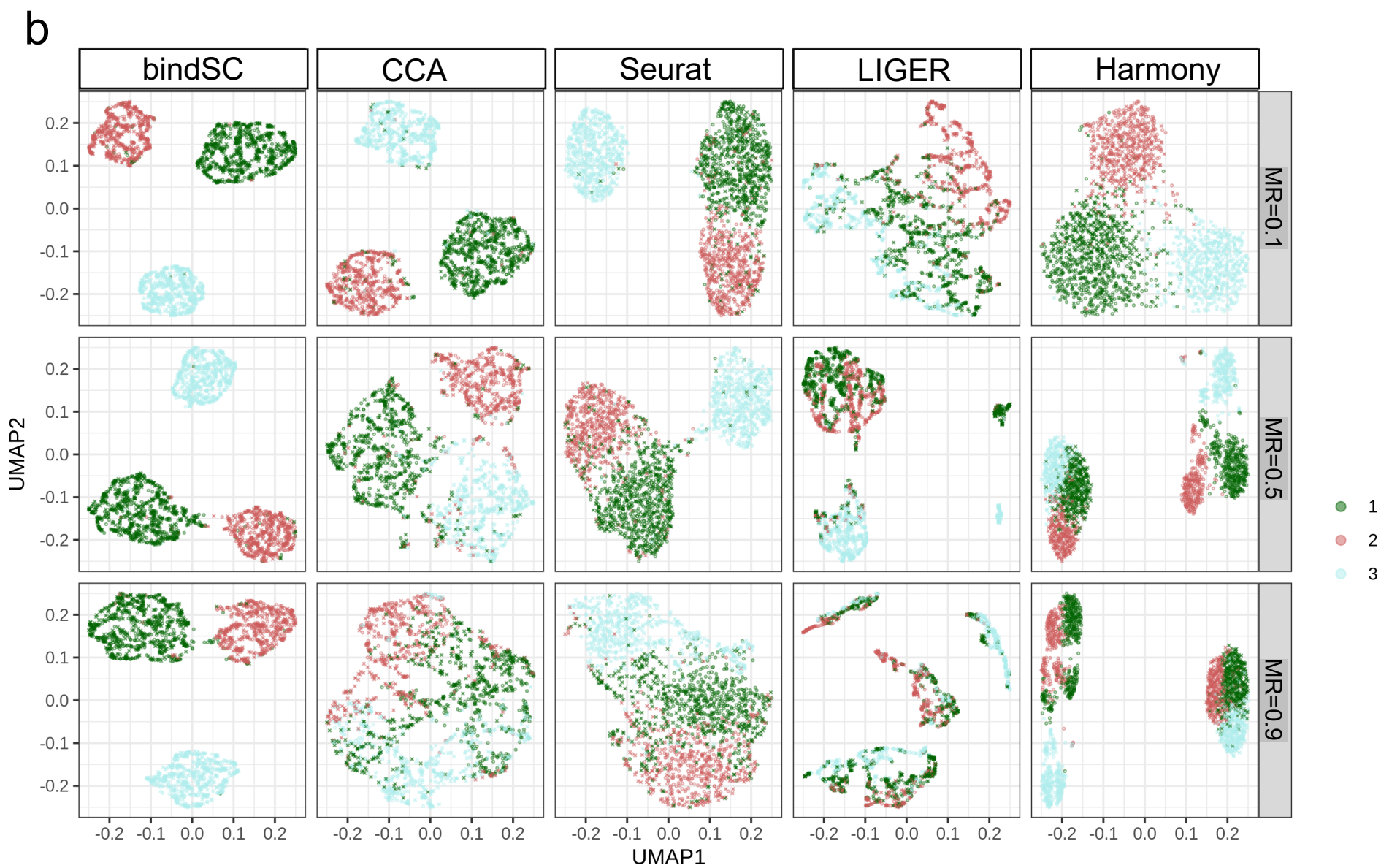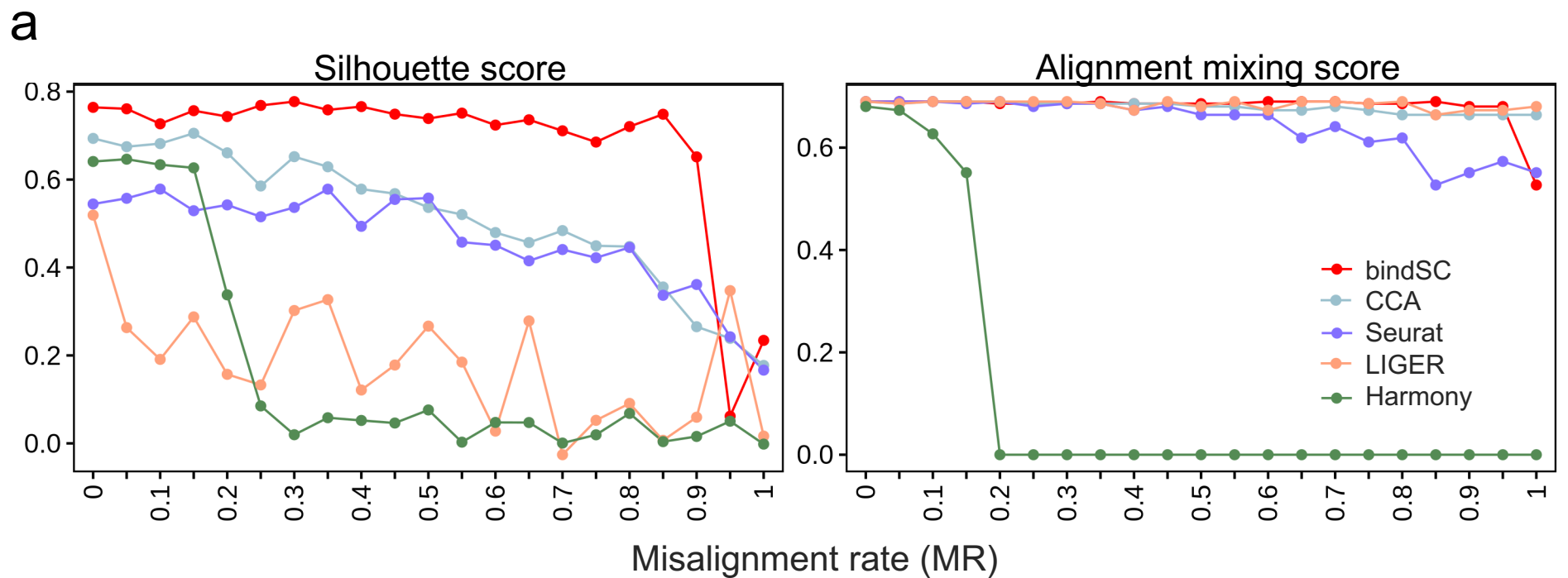
Fig. 2

a

b

Fig. 2 Benchmarking bindSC performance on simulation datasets. Comparison of bindSC to CCA, Seurat, LIGER, and Harmony based on Silhouette score and alignment mixing score **(a)**. The dataset contains 1,000 genes and 1,000 cells equally distributed in 3 cell types. Signal-to-noise ratio (SNR) was set at 0.25. X-axes denote the misalignment rates (MR) between features in the two datasets, which ranges from 0 to 1. The features between two datasets have perfect match if MR = 0 and are unrelated if MR = 1. UMAP views of the co-embeddings generated by bindSC, CCA, Seurat, LIGER, and Harmony **(b)**. From top to bottom are results with MR = 0.1, 0.5, and 0.9, respectively. Each point denotes one cell that is colored based on its true cell type label (red, green, or cyan).
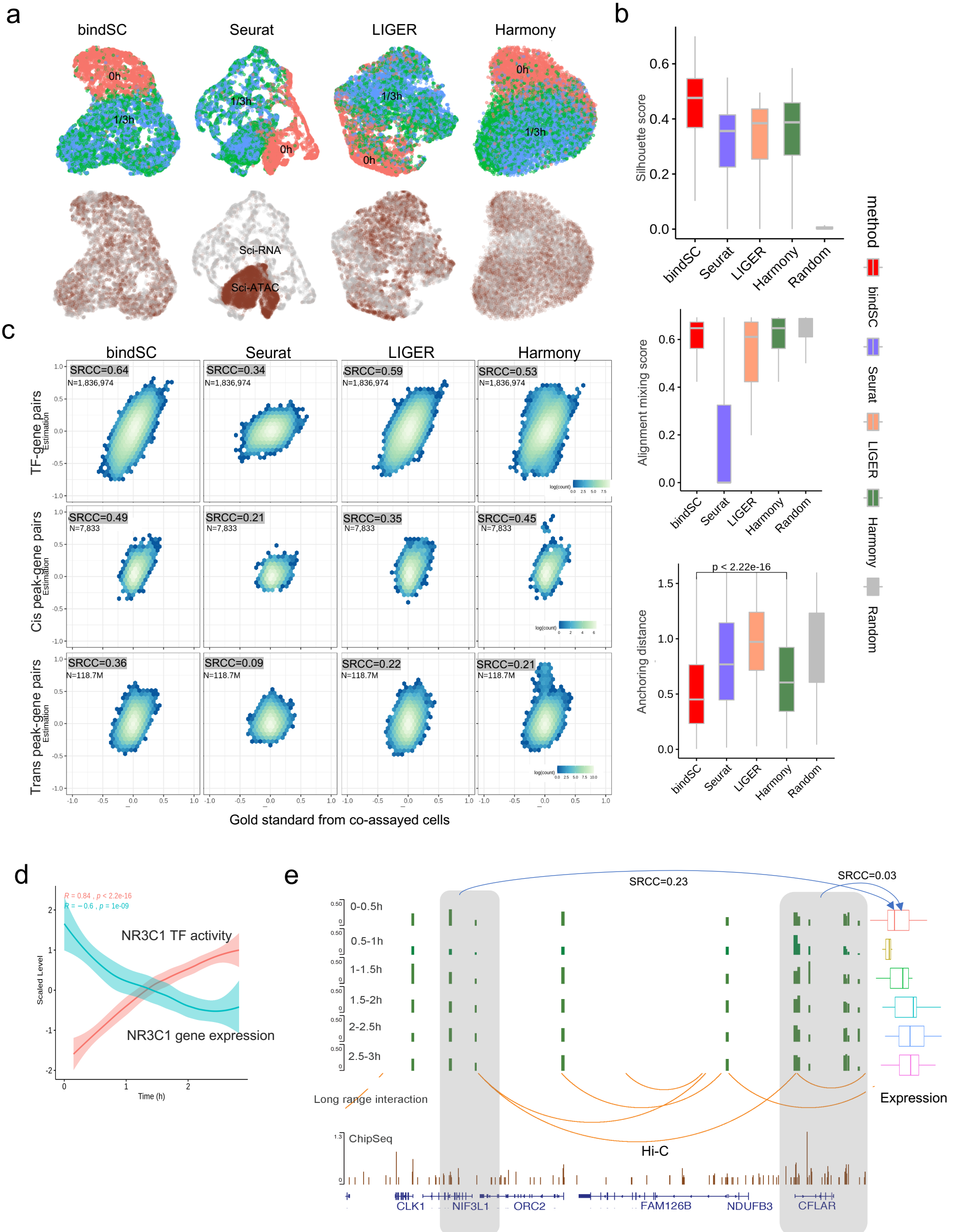
Fig. 3

Fig.3 Benchmarking bindSC performance on the DEX-treated A549 cell-line data. UMAP of cells from DEX-treated A549 cell-line data for bindSC, Seurat, LIGER and Harmony respectively, colored by collection time (red: 0 hour, green: 1 hour and blue: 3 hour) on the top panel and by technologies (grey: sci-RNA and brown: sci-ATAC) on the bottom panel **(a)**. Comparison of the 4 methods based on Silhouette score (top), alignment mixing score (middle) and anchoring distance (bottom) **(b)**. Also included for comparison are metrics from randomly selected cells. Accuracy of *in silico* multiomics profile in pseudo-cells **(c)**. The TF-gene correlation is quantified in each pseudo-cell (top panel) by calculating for each TF-gene pair (1.8 million pairs total) a Spearman's rank-order correlation coefficient (SRCC) between the TF activity level, inferred based on motif enrichment in the ATAC data, and the gene expression level in the RNA data. The peak-gene correlation is quantified by calculating for each peak-gene pair a Spearman's rank-order correlation coefficients (SRCC) between a normalized ATAC peak level and a gene expression levels for the cis-peaks (middle panel, 7,833 pairs) and the trans-peaks (bottom panel, 118.7 million pairs), respectively. X-axes are the SRCCs estimated from the co-assayed cells, which serve as the gold standard, while Y-axes are the SRCCs estimated from the pseudo-cells generated by each method. The overall concordance between X and Y are further quantified using a single SRCC shown on the up-left corner of each subfigure. Cis is defined as gene bodies plus 2,000 bps upstream. Reconstructing the gene expression and the TF activity level (Y-axes) of *NR3C1* using bindSC pseudo-cells **(d)**. X-axis is the averaged treatment time of the cells in each pseudo-cell. A genome browser view showing putative regulatory relations between an accessible distal site chr2:201770437-201770992 and the gene *CFLAR* **(e)**. The 6 tracks at the top show ATAC peak levels and gene expression levels at six time points. The track in the middle shows chromatin interactome from published Hi-C data. The bottom track shows the *NR3C1* binding targets (ChIP-Seq) peaks published in an independent study [28].
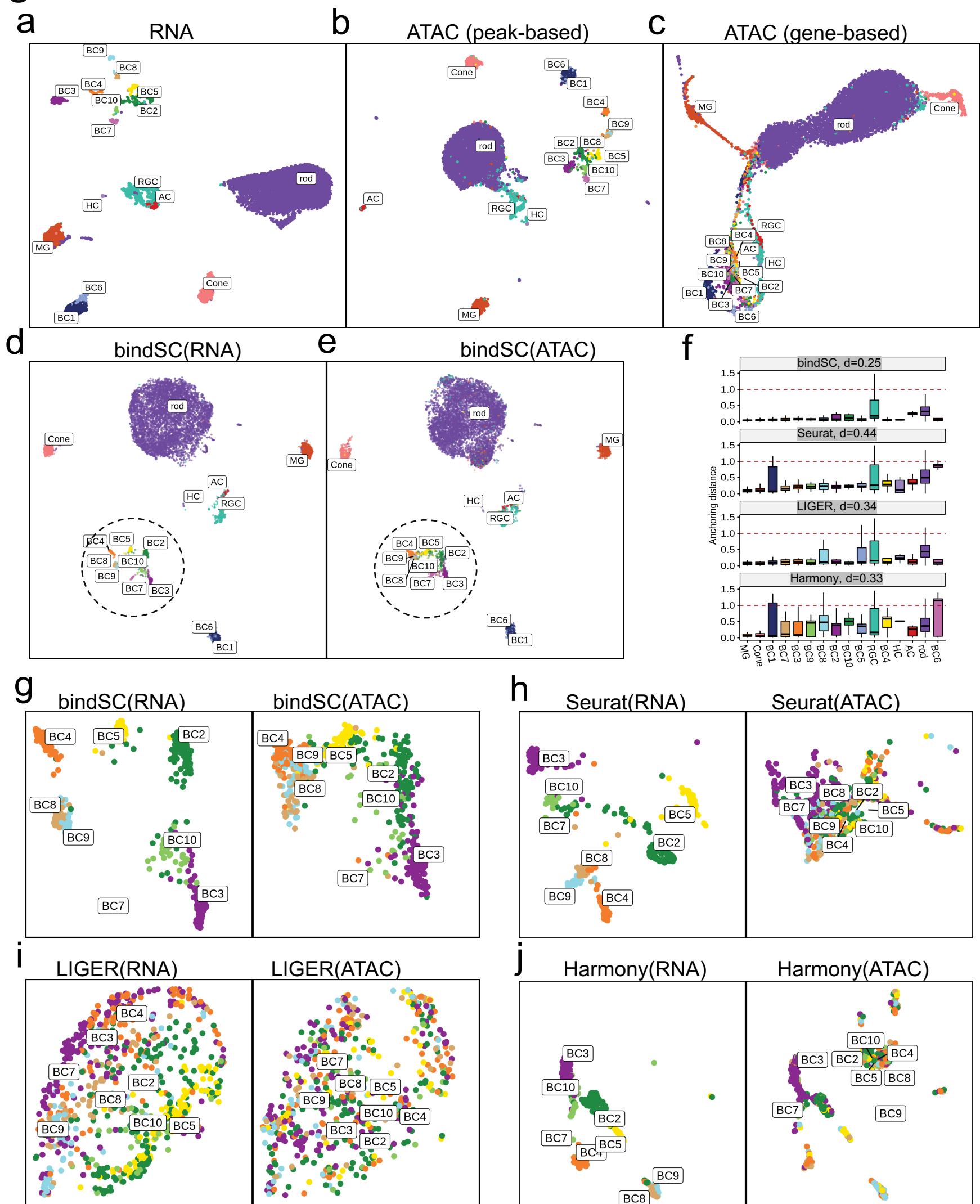
Fig. 4

Fig. 4 Integrating single-cell RNA-seq and ATAC-seq on a mouse retinal cell atlas. UMAP views of 9,383 mouse retina cells based on gene expression levels in the RNA-seq data **(a)**, chromatin accessibility peak profiles in the ATAC-seq data **(b)**, gene-level collapsed chromatin accessibility profiles **(c)**. The cells are colored by cell types annotated based on RNA expression levels (**Supplementary Fig. 8**). BindSC co-embeddings for the cells in the RNA-seq data **(d)** and those in the ATAC-seq data **(e)**. Anchoring distances resulting from bindSC, Seurat, LIGER and Harmony **(f)**. The median anchoring distance for each cell type was highlighted as a bold horizonal bar in the middle of each box in each panel. The dotted line denotes the anchoring distance based on random guesses. Zoomed out UMAP views for the BC cells in the co-embeddings generated by bindSC **(g)**, Seurat **(h)**, LIGER **(i)**, and Harmony **(j)**. Integration results for all the cell types can be seen in **Supplementary Fig. 9**. RGC: retinal ganglion cells; HC: horizontal cells; BC: bipolar cells; AC: amacrine cells; MG: Müller glial cell.
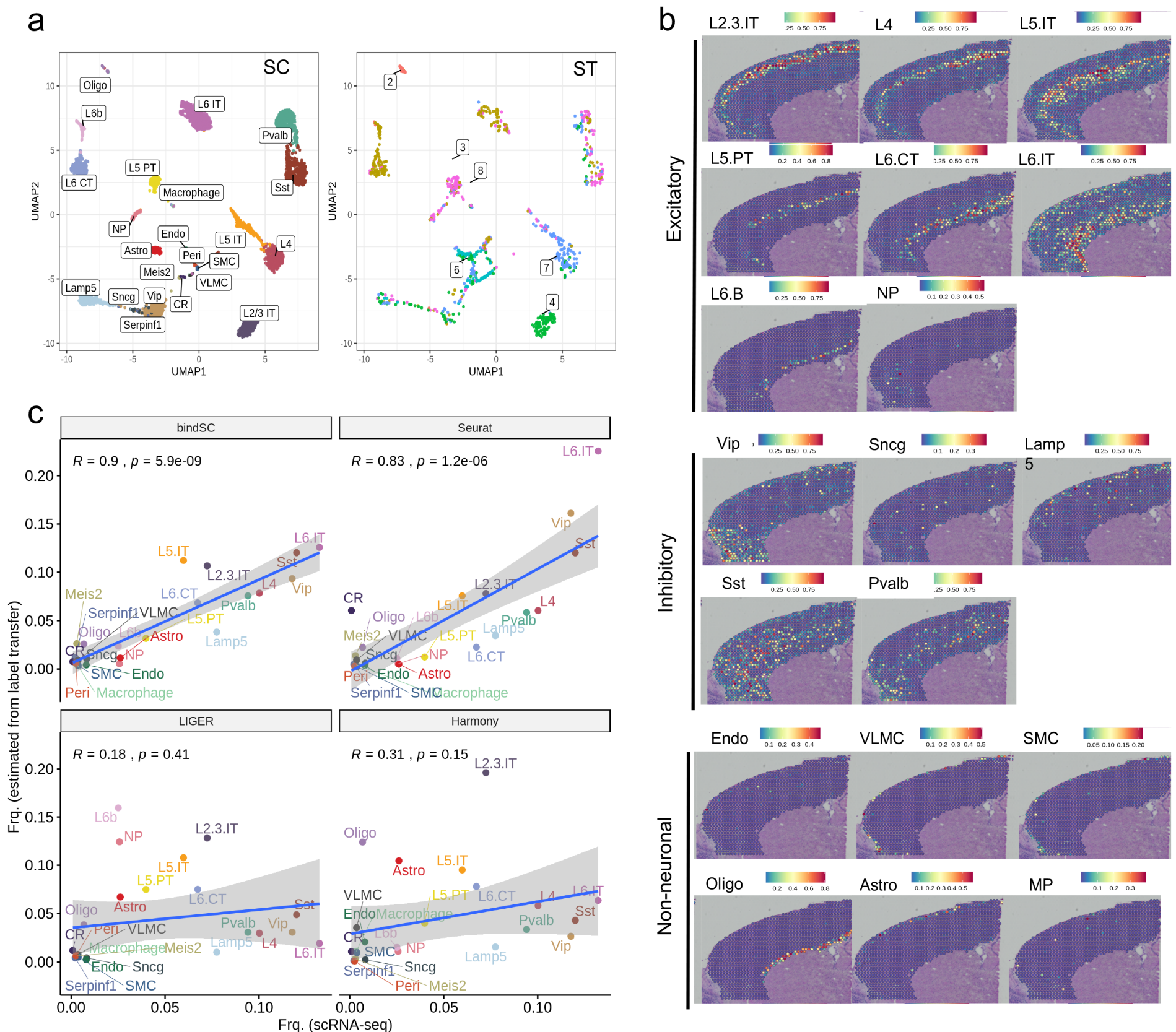
Fig. 5

Fig. 5 Integrating spatially resolved transcriptomic (ST) data with scRNA-seq data from mouse frontal cortex. UMAPs of the gene expression levels for the 14,249 cells profiled by SMART-Seq2 and for the 1,072 spots profiled by the 10x Visium technology **(a)**. Cell type labels are from the original publication[57]. Predicted locations of each cell type in the histological images **(b)**. Color gradient corresponds to the probability score of a cell being mapped to a particular spatial location. Comparison of cell type frequencies estimated from the ST data (Y-axis) to those estimated from the scRNA-seq data (X-axis) **(c)**. Correlation coefficients (R) and P values are calculated based on Pearson's correlation analysis. Each dot corresponds to a cell-type (labeled in different colors). The blue line and the grey shade represent regression lines and 95% confidence intervals from performing linear regressions.
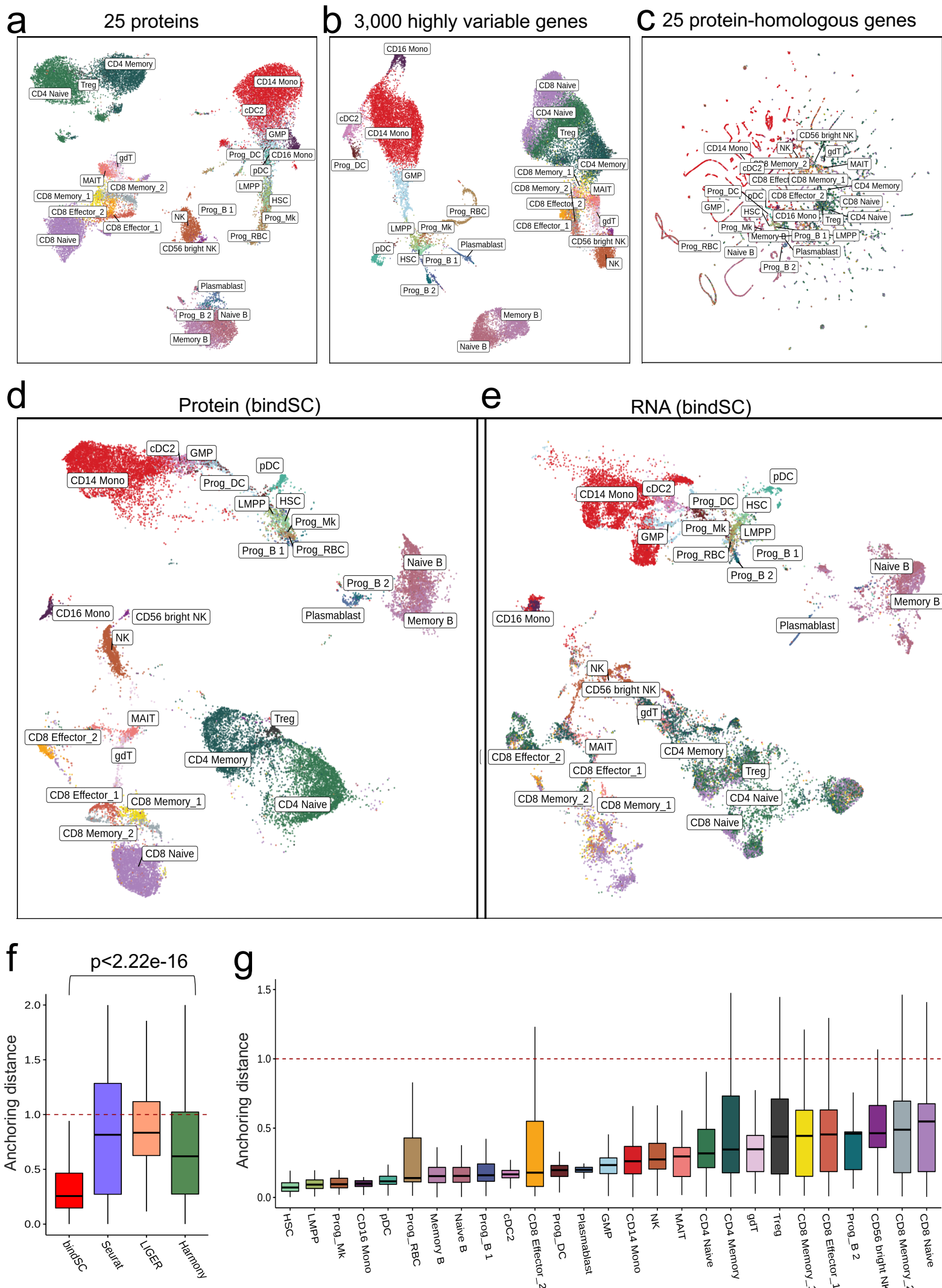
Fig. 6

Fig. 6 Integrating single-cell RNA with protein data produced by a CITE-seq assay. The UMAP of 30,672 human bone marrow cells based on 25 surface protein levels **(a)**, 3,000 highly variable gene expression levels **(b)** and 25 protein-homologous gene expression levels **(c)**. The cell type labels are from the original study [17]. UMAP of the protein **(d)** and the RNA **(e)** expression data in the co-embedding generated by bindSC. Comparison of anchoring distances generated by bindSC, Seurat, LIGER and Harmony **(f)**. The red dotted line denotes the anchoring distance from random guesses. Anchoring distances for each cell type in the bindSC co-embedding **(g)**.