Machine Translation between paired Single Cell Multi Omics Data

Xabier Martinez-de-Morentin[1], Sumeer A. Khan[2], Robert Lehmann[2],
Jesper Tegner[2], David Gomez-Cabrero[1,2]

[1]Navarrabiomed, Complejo Hospitalario de Navarra (CHN), Universidad Pública de Navarra (UPNA), IdiSNA, Pamplona, Spain.

[2]Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

Abstract

Single-cell multi-omics technologies enable profiling of several data-modalities from the same cell. We designed LIBRA, a Neural Network based framework, for learning translations between paired multi-omics profiles into a shared latent space. We demonstrate LIBRA to be state-of-the-art for multi-omics clustering. In addition, LIBRA is more robust with decreasing cell-numbers compared with existing tools. Training LIBRA on paired data-sets, LIBRA predicts multi-omic profiles using only a single data-modality from the same biological system.

Main Text

Single Cell Genomics technologies set the stage for unraveling the intrinsic complex organization at different levels within cells. Such investigations require simultaneous profiling of several layers of transcriptional regulation at the resolution of single cells[1]. Recent multi-omic single-cell technologies enable profiling of joint "chromatin accessibility & mRNA profiles" (e.g., SNARE-seq[2], sci-CAR[3]) and "mRNA profiles & protein antibody-derived tags" (CITE-seq[4]). New methodologies for data-analysis are emerging, such as Seurat3[5] and MOFA+[6]. However, those methodologies do not use the *paired* data (*profiles derived from the same cell*). Very recently, though, Seurat4[7] has been presented as the first tool exploiting such *paired* multi-omic profiles. Yet, since Seurat is not formulated as a machine learning tool, scalability and robustness are two potentially limiting factors. Consequently, Seurat4 does not *learn* paired-based derived predictive models for imputation across data-modalities. Inspired by the ideas from Neural Machine Translation[8], we introduce LIBRA, an encoder-decoder architecture using AutoEncoders (AE). LIBRA integrates single-cell multi-omic data by leveraging and balancing information of *paired* single-cell omics data. LIBRA improves accuracy for detecting cell subtypes and their associated markers and is robust when considering fewer cells. Unlike other methods, predictive LIBRA models can be used for imputation for those samples where only one omic profile is available. Furthermore, LIBRA is generalizable to any pair of omics.

LIBRA, inspired by the neural machine translation efforts[8], "*translates*" between omics. Implemented using Autoencoders, LIBRA encodes one omic and decodes the other omics to and from a reduced space. Here the decoder minimizes the distance to a second and paired data type (*joined translation and projection*). Briefly, LIBRA consists of two neural networks (NN) (Fig1(a)); the first NN is designed similarly to an Autoencoder, but the difference is that input (dt1) and output (dt2) data correspond to two different paired multi-modal datasets (Supp. Fig1(a)). The idea is to identify a shared latent space (SLS) for two data-types. The second NN is used to generate a mapping from the *dt2 to the shared projected space* (Supp. Fig1(b)).

To evaluate the performance of LIBRA, we designed several quality metrics. The first metric measures the quality of the first NN (referred to as Q1), while the second, Q2, measures the accuracy of the dt2 projection (Fig.1(a)) separately for the cells used in the training or validation in NN1 (Supp. Fig1(c)). To evaluate the additional value of the SLS, we designed the Preserved Pairwise Jacard Index (PPJI), a non-symmetric distance metric aimed to investigate the added value (finer granularity) of clustering B (multi-omic) in relation to cluster A (single-omic) as shown in Fig.1(b). Briefly, when comparing RNA derived clustering, and SLS derived clusterings, the PPJI computes the Jaccard Index between every pair of clusters from A and B and summarizes the value (e.g., sum) over the clusters of B (Fig.1(c)). Unless the two clusterings have the same number of clusters, the outcome will not be symmetric. Using these three quality measurements (Q1, Q2, PPJI) and the SNARE-seq[2] adult brain mouse dataset, we selected the hyperparameters for LIBRA: (i) Autoencoder-type configuration=AE-based framework, (ii) number of dimensions of the projected space=10, (iii) peak derived information for ATAC-seq, (iv) the ordering (dt1=ATAC-seq and dt2=RNA-seq), (v) to consider most variable features only and (vi) the number of hidden layers=2; see. Supp. Fig 1(d). In all cases, we favored the added value of a finer granularity (maximal PPJI) and secondly, the quality of the neural networks (minimal Q2, Q1).

Next, we compared LIBRA using a paired adult mouse brain single-cell RNAseq and ATACseq dataset[2] against Seurat3 and Seurat4 using PPJI measure. We observe that both Seurat4 and LIBRA outperform Seurat3, and both methodologies are of similar quality (Fig.1(d)). We consider that improvement is the outcome of using paired information. Additionally, as a proof-of-concept, we compared LIBRA with a clustering based on concatenating both data-type matrices and running a classic Autoencoder (Fig.1(d)). We also evaluate the methodologies' robustness by reducing the number of cells by randomly select a % of cells and compute PPJI for each case. As expected, Fig.1(e), a decrease in the number of cells diminished both Seurat4 and LIBRA's performance. Interestingly, when reducing the number of cells, LIBRA significantly outperforms Seurat4. This result suggests better robustness using a machine learning architecture as compared to an anchoring methodology.

To further assess performance and biological relevance, we compared the integrated clusters resulting from LIBRA, Seurat 3, and 4. Both Seurat4 and LIBRA were comparable (Fig1.f). Furthermore, as shown in Fig.1(g), both clusterings are comparable for the majority of the clusters identified, whereas such similarity is not maintained with Seurat3 (Supp. Fig.1(e)). To further interrogate biological relevance, we investigated the cluster-specific markers from Seurat4 and LIBRA. First, when taking Seurat4 as a reference, the top markers identified at Seurat4 are also recognized by LIBRA (Supp. Fig. 2(a)); additionally, LIBRA also identified additional markers (Fig.2(a)). We conclude that both methodologies can recover a similar level of resolution for cell subtypes and their associated markers.

While Seurat4 and LIBRA's outcomes on the clustering are comparable, the LIBRA framework's added value is its predictive power. Once a LIBRA-model is generated for a paired dataset, it can predict profiles from single-omic single-cell data of the same biological system. Considering the dt1=ATAC and dt2=RNA, we quantified the predictive power for RNA profiles, predRNA, as the correlation between known and predicted profiles; for the SNARE-seq[1] dataset is 0.72. We also observed that those values are consistent among the different integrated-based identified clusters, where predRNA ranges between 0.65 and 0.75 (Supp Fig.2(b)); we also did not observe a significant Spearman correlation between imputation quality and the number of cells per cluster (p-value=0.36). We also investigated using the adult brain mouse LIBRA model to predict scRNA profiles in embryonic mouse brain[2]. Here we obtained a predRNA value of 0.63, which is – as expected – lower than when applied to adult brain cells but relevant for applying LIBRA derived models to close-related systems.

To validate our results in additional datasets, we compared Seurat4 and LIBRA in a PBMC data[7] (DataSet2); the results showed, similarly to DataSet1, that both methodologies have similar power to identify clusters based on PPJI values (Fig.2(b,c)). In contrast, better PPJI values are associated with Seurat4. The comparison between integrated-based clusters is 0.71 between Seurat4 and LIBRA. To investigate if the results depended on the normalization protocol, we conducted the same analysis with the scTransform normalization protocol. We observed that the number of clusters obtained by Seurat4 was different depending on the protocol (Fig.2(b)). Hence, the granularity is also impacted by an integrative analysis. Most importantly, we observed that scTransform reduced the predictive power from predRNA=0.69 to predRNA=0.45; such a decrease in the quality is likely associated with the regress-out type of normalization procedure.

Finally, we designed LIBRA as a general tool for any type of paired single-cell integration. To validate the generalizability of LIBRA, we analyzed a CITE-seq data-set[4]. We observed that both LIBRA and Seurat4 improved the integrated clustering with slightly better results for Seurat4 (Fig.2(d,e)). However, both were returning the same number of clusters when comparing both clusterings, and both were highly similar (0.70 PPJI). Interestingly, the predRNA value (predicting RNA-seq from ADT profiles) returned a correlation of 0.79, which was of similar quality for all clusters identified and not significantly associated with the cluster size (Supp. Fig.2(c)).

In summary, LIBRA is a tool that leverages paired-single-cell information to generate an accurate cell-subtype identification to the same degree as the reference tool Seurat4. Besides, LIBRA allows for single-cell multi-omic imputation to a high degree of accuracy and robustness (Supp. Fig3(a,b,c)). LIBRA is a generalizable alternative to any pair of omics. The trade-off for its predictive power is the computational effort associated with computing the models. The LIBRA model's simplicity allows easy improvement in future versions that scale up current performance (e.g., reduce computational burden) and the extension to be applied in more than two omic integration analyses. Other tasks, such as identifying markers, have shown great power when identifying robust markers for biological interpretation.

Figure 1. LIBRA design and optimization. (a) Visual description of the LIBRA framework. (b) Example of clustering resolution in the integrated space. Two left upper panels to denote the UMAP projection and their clustering for RNA and ATAC, respectively. The right panel shows the UMAP projection and clustering of cells in the integrated space (in the LIBRA optimized model). Finally, the two left bottom panels project the clustering information derived from the integrated space in the UMAP projections for RNA and ATAC, respectively. (c) PPJI graphical description. Left (right) panels show the PPJI estimation for the RNA (ATAC) vs. integrated clustering. Jaccard Pairwase Index is computed between every pair of clusters. A sum is then calculated for each cluster identified in RNA (ATAC), and the average of the value is calculated for all clusters. The final value, PPJI, shows how well integrated-based clustering adds granularity to the single-data derived clustering. (d) PPJI compares single-cell omic projections versus four integrated approaches. Unpaired AE denotes the concatenation of RNA and ATAC profiles into a single-matrix and running a classical Autoencoder. (e) PPJI values between clusterings are derived from single-omic and integrative approaches when the number of cells is reduced by a predefined percentage. (f) PPJI is comparing the three integrated projections. (g) Jaccard Index between clustering derived from LIBRA and Seurat4, both using paired information.

Figure 2. LIBRA validation. (a) For each cluster identified in Seurat4 (rows), each line denotes a gene and, if it is identified in the Seurat4 cluster, by LIBRA or by both. (b,d) PPJI comparing the single-omic and integrated derived clusterings for both normalization protocols. scT added the scTransform analysis. #clu denotes the number of clusters for the references and the clustering that is compared against. (c) Example of clustering resolution in the integrated space for the PBMC data-set[7]. Two left upper panels to denote the UMAP projection and their clustering for RNA and ATAC, respectively. The right panel shows the UMAP projection and clustering of cells in the integrated space (in the LIBRA optimized model). Finally, the two left bottom panels project the clustering information derived from the integrated space in the UMAP projections for RNA and ATAC, respectively. (d) Same as (c) but with CITE-seq dataset[4].

REFERENCES

1. Schier, A. F. Single-cell biology: beyond the sum of its parts. *Nat. Methods* **17**, 17–20 (2020).

2. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).

3. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (80-.).* **361**, 1380–1385 (2018).

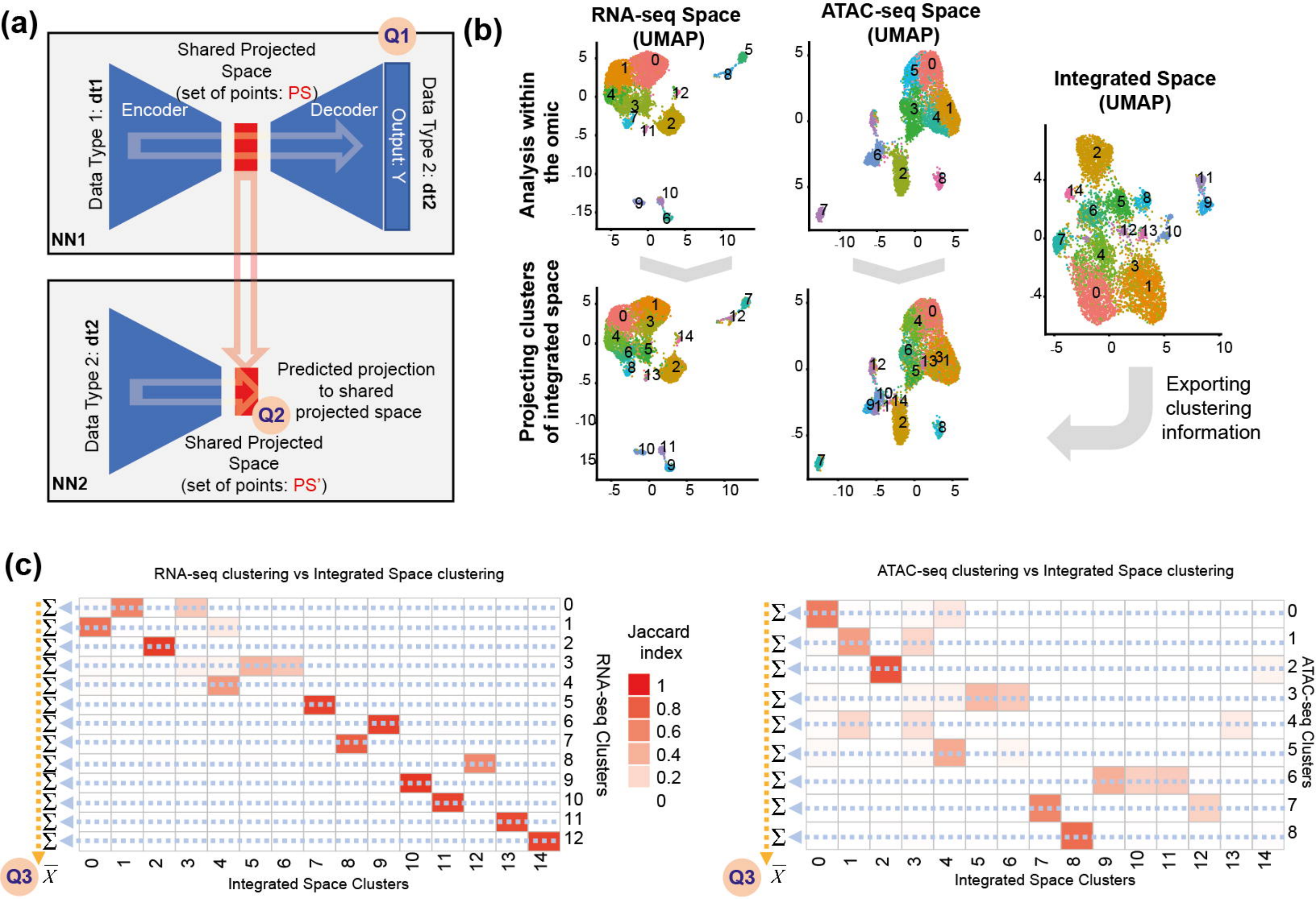
4. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

5. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

6. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).

7. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *bioRxiv* 2020.10.12.335331 (2020) doi:10.1101/2020.10.12.335331.

8. Kyunghyun, C. *et al.* Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* (2014).


(a)
Q1
Shared Projected Space (set of points: PS)
Data Type 1: dt1
Encoder
Decoder
Output: Y
Data Type 2: dt2
NN1
Data Type 2: dt2
Predicted projection to shared projected space
Q2
Shared Projected Space (set of points: PS')
NN2
(b)
RNA-seq Space (UMAP)
ATAC-seq Space (UMAP)
Integrated Space (UMAP)
Analysis within the omic
Projecting clusters of integrated space
Exporting clustering information
(c)
RNA-seq clustering vs Integrated Space clustering
ATAC-seq clustering vs Integrated Space clustering
Σ
Q3
$\overline{X}$
Integrated Space Clusters
RNA-seq Clusters
ATAC-seq Clusters
Jaccard index
1
0.8
0.6
0.4
0.2
0

(d)

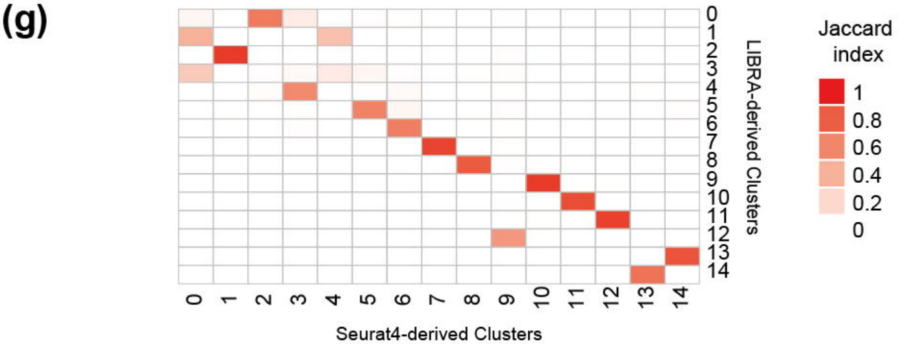
	PPJI	
	Ref: RNA	Ref: ATAC
Ref to Seurat3	0.67	0.74
Ref to Seurat4	0.87	0.80
Ref to LIBRA	0.85	0.78
Ref to unpaired AE	0.48	0.67

(e)

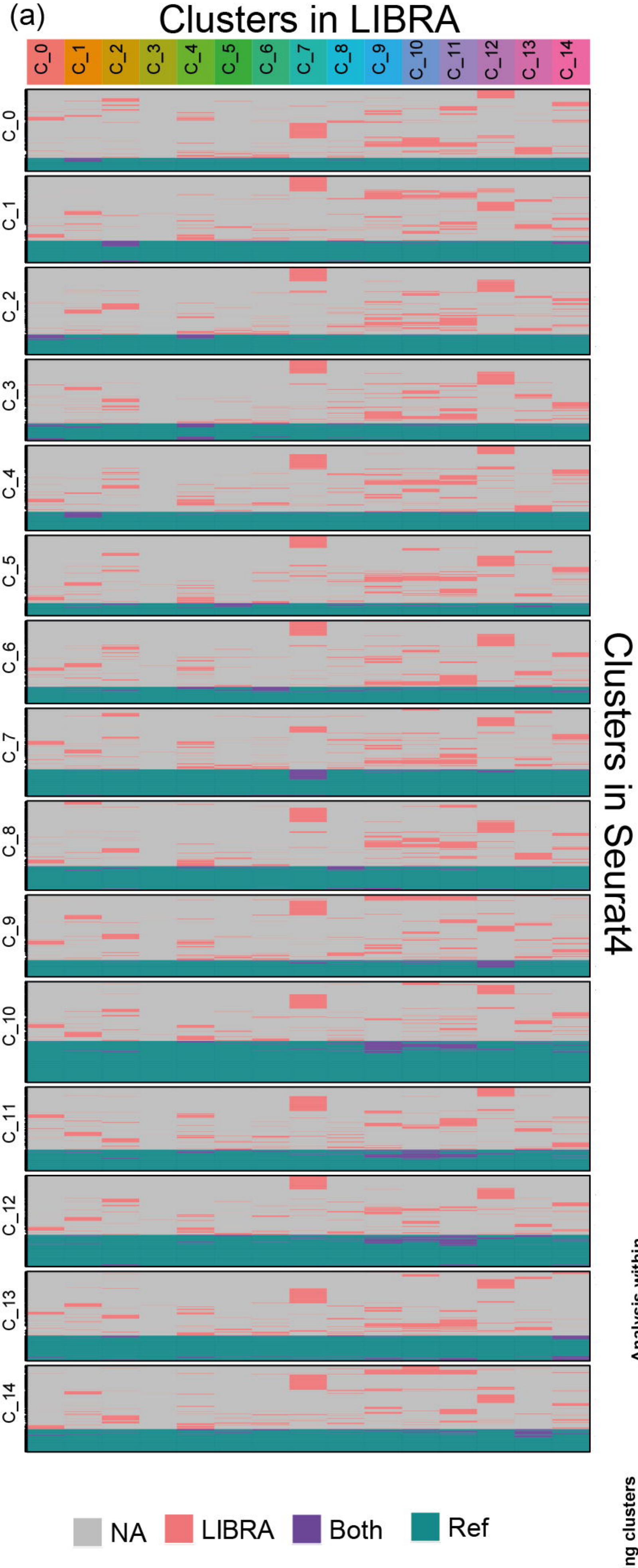
		100%	80%	60%	40%	20%
Ref RNA vs	Seurat4	**0.87**	0.70	0.54	0.37	0.19
	LIBRA	0.85	**0.76**	**0.64**	**0.58**	**0.53**
Ref ATAC vs	Seurat4	**0.80**	0.66	0.52	0.35	0.19
	LIBRA	0.78	**0.76**	**0.77**	**0.77**	**0.74**

(f)

	PPJI
Seurat3 vs Seurat4	0.60
Seurat3 vs LIBRA	0.55
Seurat4 vs LIBRA	0.69



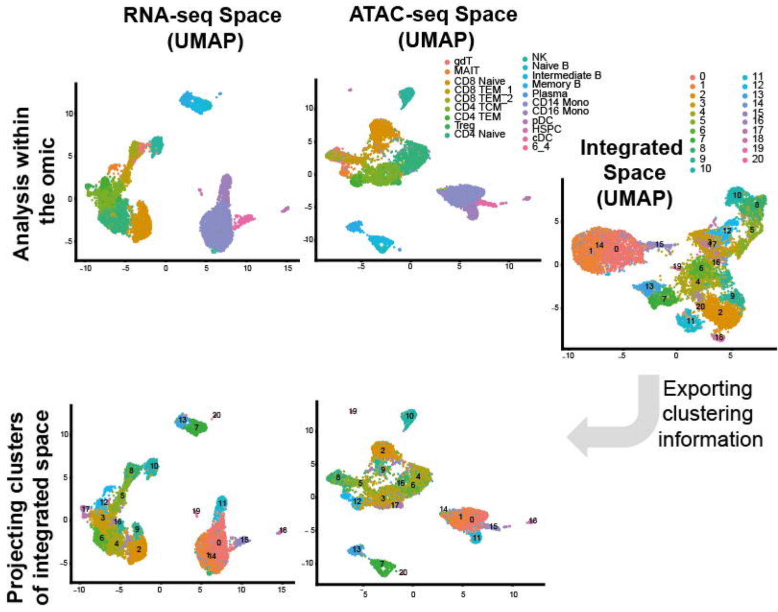
(a)



(b)

		Regular		scT	
Reference	vs	PPJI	#clu	PPJI	#clu
RNA	Seurat4	0.82	18/23	0.72	21/21
ATAC	Seurat4	0.75	18/23	0.77	18/21
RNA	LIBRA	0.75	18/18	0.76	21/21
ATAC	LIBRA	0.67	18/18	0.70	18/21
Seurat4	LIBRA	0.71	23/18	0.71	23/21

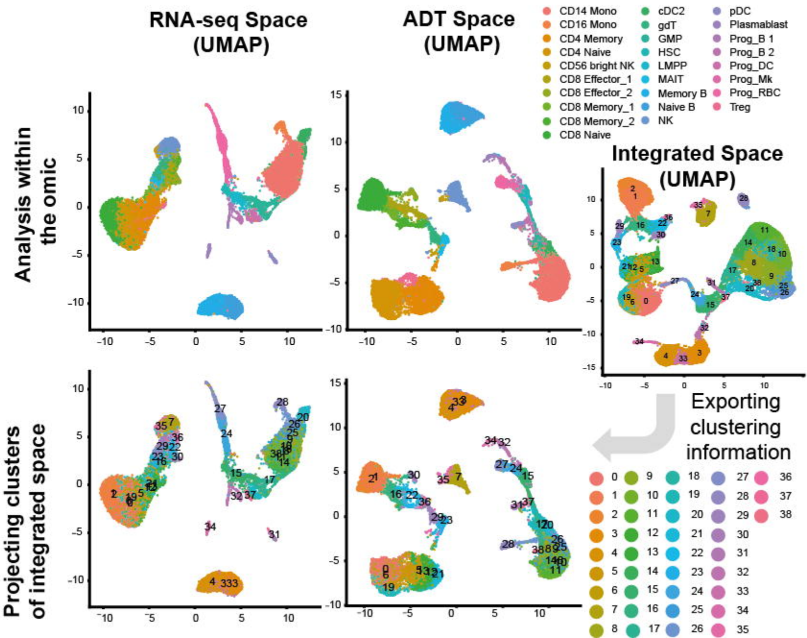
(c)



(d)

Reference	vs	PPJI
ADT	Seurat4	0.63
RNA	Seurat4	0.63
ADT	LIBRA	0.62
RNA	LIBRA	0.62

(e)


RNA-seq Space (UMAP)
ADT Space (UMAP)
Analysis within the omic
Projecting clusters of integrated space
Integrated Space (UMAP)
Exporting clustering information
CD14 Mono
CD16 Mono
CD4 Memory
CD4 Naive
CD56 bright NK
CD8 Effector_1
CD8 Effector_2
CD8 Memory_1
CD8 Memory_2
CD8 Naive
cDC2
gdT
GMP
HSC
LMPP
MAIT
Memory B
Naive B
NK
pDC
Plasmablast
Prog_B 1
Prog_B 2
Prog_DC
Prog_Mk
Prog_RBC
Treg