 doi:10.20944/preprints202109.0504.v1

Article

Sparsely Connected Autoencoders: a multi-purpose tool for single cell OMICs analysis

Luca Alessandri [1,2], Maria Luisa Ratto [1], Sandro Gepiro Contaldo [2], Marco Beccuti [2], Francesca Cordero [2], Maddalena Arigoni [1] and Raffaele A. Calogero [1,*]

[1] Department of Molecular Biotechnology and Health Sciences, University of Torino, Italy; alessandri.luca1991@gmail.com, maria.ratto@edu.unito.it, Maddalena.arigoni@unito.it, raffaele.calogero@unito.it

[2] Department of Computer Science, University of Torino, Italy; marco.beccuti@unito.it, francesca.cordero@unito.it, alessandri.luca1991@gmail.com, sandro.contaldo@edu.unito.it

* Correspondence: raffaele.calogero@unito.it; Tel.: +39 0116706454

Abstract: Background: Biological processes are based on complex networks of cells and molecules. Single cell multi-OMICs is a new tool aiming to provide new incites in the complex network of events controlling the functionality of the cell.; Methods: Since single cell technologies provide many sample measurements, they are the ideal environment for the application of deep learning and machine learning approaches. An autoencoder (AE) is composed of an encoder and a decoder sub-model. AE are very powerful in data compression and noise removal. However, the decoder model remains a black box from which is impossible to depict the contribution of the single input elements. We have recently developed a new class of autoencoders, called Sparsely Connected Autoencoders (SCA), which have the advantage of providing a controlled association among the input layer and the decoder module. This new architecture has the benefit that the decoder model is no anymore a black box and it can be used to depict new biologically interesting features from single cell data; Results: In this paper, we show that SCA hidden layer can grab new information usually hidden in single cell data, like as providing clustering on meta-features difficult, i.e. transcription factors expression, or impossible, miRNA expression, to depict in single cell RNAseq data. Furthermore, a SCA representation of cell clusters has the advantage of simulating a conventional bulk RNAseq, which is a data transformation allowing the identification of similarity among independent experiments; Conclusions: In our opinion, SCA represent the bioinformatics version of a "Swiss Army knife" for the extraction of hidden knowledgeable features from single cell OMICs data.

Keywords: single cell RNAseq; single cell ATACseq, sparsely connected autoencoders, gene regulatory network, transcription factor, miRNA, pseudo-bulk.

1. Introduction

Single cell RNAseq (scRNA) [1] together with CITEseq [2], scATACseq [3] and single cell spatial transcriptomics [4] represent the frontier of biological and medical research. Grasping biological knowledge from single cell OMICs [5] is today a mandatory issue. Since single cell technologies provide a large number of sample measurements, e.g. single cell gene expression level, single cell chromatin accessibility measurements, etc., these methods produce the ideal input data for deep learning based analyses [6-8].

In 2019 Gold and coworkers [9] and more recently us [10] presented sparsely connected autoencoder (SCA) as a tool for projecting gene-level data onto gene sets. SCA encoding/decoding functions consists of a single sparse layer [10], with connections based on known biological knowledge. Each node represents a known biological relationship,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0504.v1

e.g. transcription factor (TF) targets or miRNA targets. SCA receives inputs only from gene nodes associated with a biological relationship. In our recent paper [10], we described the use of SCA to uncover hidden features associated with scRNAseq data. In Alessandri et al. paper [10], we showed that SCA represents a new instrument to identify key elements for cells aggregation by single cell whole transcriptome analysis [10]. In this paper , we present extension of SCA to scRNAseq and scATAC seq. Specifically, we show that SCA is a bioinformatics "Swiss Army knife" for the extraction of hidden knowledgeable features from single cell OMICs data.

2. Results

To describe the peculiarities of SCA we used two scRNAseq datasets respectively made of a mixture of three (RNA-3c) and five (RNA-5c) human lung adenocarcinoma cell lines [11], and the scATACseq dataset encompassing the same five (ATAC-5c) human lung adenocarcinoma cell lines described in [11],. Both scRNAseq and scATACseq datasets were specifically developed as tools for benchmarking single cell data analysis methods [11].

2.1. Building clusters' specific pseudo-bulk using SCA.

Since single-cell expression data are zero-inflated [12], to depict differences among clusters of a scRNAseq experiment, a straightforward solution is to create "pseudo" bulk RNA-seq data, by adding up the fragment counts of a gene across cells for each cluster, and then applying methods designed to inspect differences among samples using bulk RNAseq data, e.g. hierarchical clustering. However, such pseudo-bulk solution reduces the distribution of gene expression across cells to a single vector. Thus, this pseudo-bulk representation loses a valuable amount of information. Due to the lack of replicates, it is impossible to perform conventional RNAseq differential expression analysis among clusters. Thus, we have investigated the possibility to generate multi-samples pseudo-bulk experiments using SCA. In our previous work [10], we have shown that SCA can be used to aggregate transcription factor (TF) target genes expression in a pseudo-value describing the importance of a TF in controlling its putative targets. Here, the same concept was applied to cells and their belonging cluster. The overall idea of this approach is that genes characterizing a specific cells' cluster will be the non-noise signal catch by the hidden layer of SCA (Figure 1A).

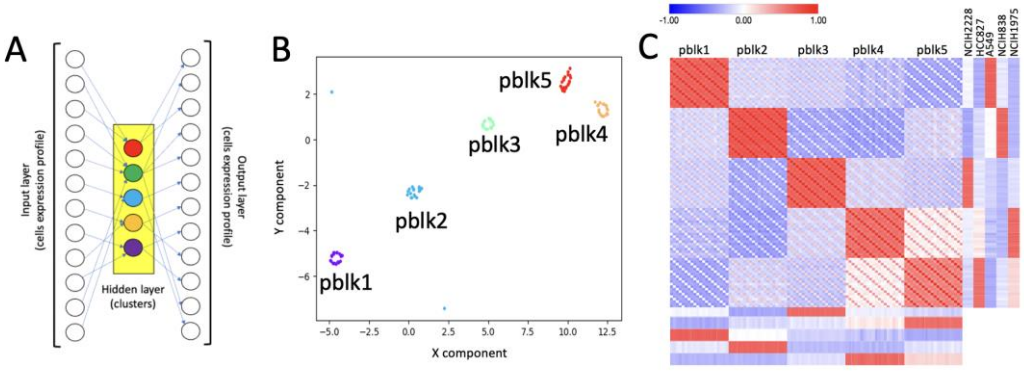


Figure 1. Pseudo-bulks generated by SCA. A) Structure of the SCA. The pseudo-bulks are generated using the hidden layer data, repeating multiple times the SCA runs. B) t-Sne output of 20 runs of SCA. C) Row-mean centered CPM expression for pseudo-bulks was combined with row-mean centered TPM expression of the RNAseq for H2228, H1975, A549, H838 and HCC827 cell lines and Pearson correlation matrix was built.

In this way, by running multiple times the SCA we can build pseudo-bulk experiments representing pseudo-replicates for each of the clusters' gene expression. We

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0504.v1

applied this SCA analysis to the RNA-5c dataset. RNA-5c was partitioned in clusters using Seurat clustering [13], implemented in rCASC [14], with the resolution parameter set to 0.1. This analysis, yielded 5 clusters, Figure 2A.

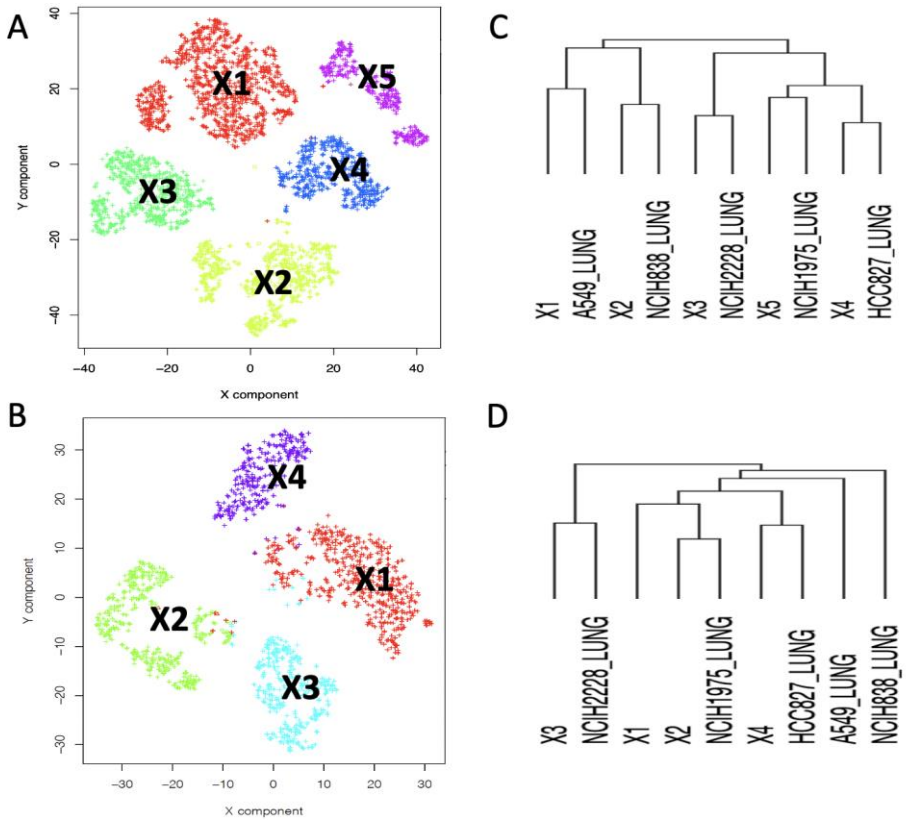


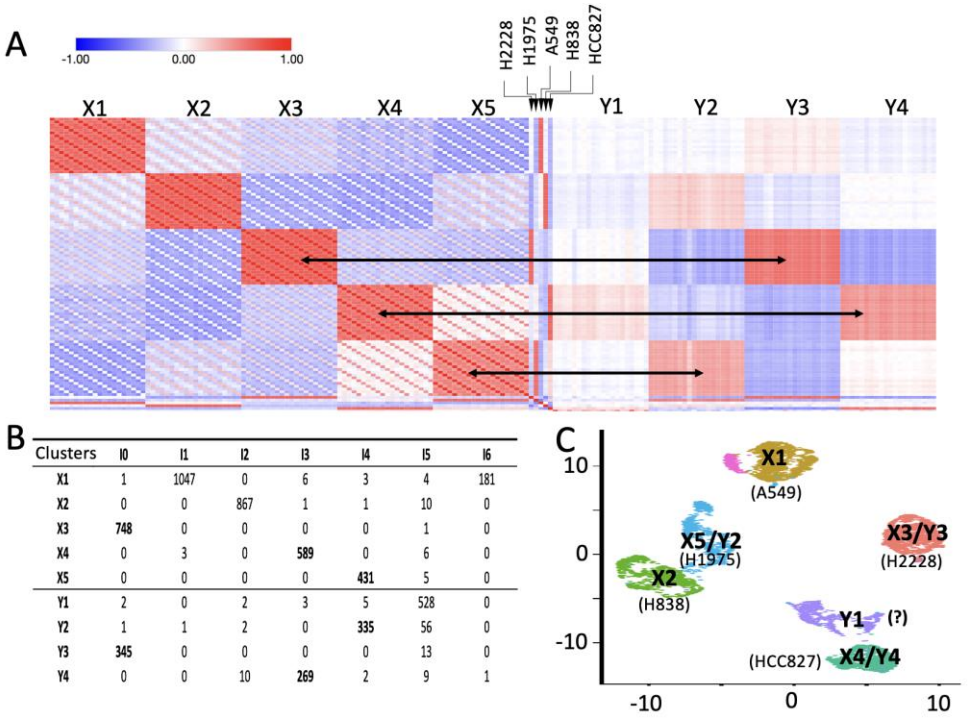
Figure 2: Assignment of cell line type to clusters generated with Seurat, implemented in rCASC. A) RNA-5c clustering, five clusters generated with Seurat (resolution=0.1), using 2500 genes selected as the most variant within the 5000 most expressed (rCASC topx function). B) RNA-3c clustering, four clusters generated with Seurat (resolution=0.1), using the 2500 genes selected for RNA-5c. C) RNA-5c hierarchical clustering (Euclidean distance, average linkage) of log_2CPM clusters' pseudo-bulk expression (rCASC bulkClusters function), row-mean centered, and CCLE lung cell lines A449, NCIH838, NCIH2228, NCIH1975 and HCC827 log_2TPM row-mean centered. D) RNA-3c hierarchical clustering (Euclidean distance, average linkage) of log_2CPM clusters' pseudo-bulk expression (rCASC bulkClusters function), row-mean centered, and CCLE lung cell lines A449, NCIH838, NCIH2228, NCIH1975 and HCC827 log_2TPM row-mean centered.

The clusters were assigned to the corresponding bulk cell line by mean of hierarchical clustering comparing bulk transcriptome of H2228, H1975, A549, H838 and HCC827 cell lines with the summary of the cell expression of each cluster, see Figure 2C. Subsequently, we apply the clustering results, Figure 2A, to the SCA shown in Figure 1A. The results of pseudo-bulk analysis are summarized in Figure 1B and C. In Figure 1B, it is shown the t-Sne representation of pseudo-clusters made of 20 runs each. Pseudo-bulk clusters result to be nicely well separated to each other. In Figure 1C, pseudo-bulks well correlate with the bulk RNAseq of their corresponding cell line. Thus, demonstrating the good correlation existing between SCA-pseudo-bulks and the corresponding cell line bulk RNAseq transcriptome, retrieved from CCLE database [15].

2.2. Depicting clusters correspondence among independent experiments using SCA pseudo-bulks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0504.v1

The single cell technology is becoming every day more used to investigate complex questions. Thus, it is becoming important to integrate the results of multiple experiments. One of the approaches used to integrate experiments is the one implemented in Seurat [13], where "*anchors*" are exploited to harmonize data from different experiments. Other options are based on experiment batch-effect correction [16] to aggregate independent single-cell experiments. Aggregation might distort single-cell data; thus, our idea is that it could be useful to simply identify clusters in common among experiments. Due to the results described in the above paragraph, we evaluated the possibility to use SCA pseudo-bulks to correlate the clustering results obtained in different experiments. Specifically, we compared RNA-5c and RNA-3c, which are single cell experiments performed independently and constituted respectively by the following cell lines: H2228, H1975, A549, H838 and HCC827 (RNA-5c) and H2228, H1975 and HCC827 (RNA-3c). We independently built the rCASC clustering for the two datasets using Seurat clustering [13], implemented in rCASC [14], using as resolution parameter 0.1. From this, we obtained five clusters for RNA-5c and four clusters for RNA-3c. RNA-3c cell line assignment, Figure 2B and D, was done and described for RNA-5c, in the previous paragraph. We independently generated SCA pseudo-bulks for the clusters of the two datasets. Pseudo-bulk counts were log_2CPM transformed and genes expression was centered on the gene's mean expression. The two datasets were also integrated using Seurat integration tool [13]. The results of the analysis are shown in Figure 3.



Clusters	I0	I1	I2	I3	I4	I5	I6
X1	1	1047	0	6	3	4	181
X2	0	0	867	1	1	10	0
X3	748	0	0	0	0	1	0
X4	0	3	0	589	0	6	0
X5	0	0	0	0	431	5	0
Y1	2	0	2	3	5	528	0
Y2	1	1	2	0	335	56	0
Y3	345	0	0	0	0	13	0
Y4	0	0	10	269	2	9	1

Figure 3. Comparing independent datasets using SCA pseudo-bulk. A) Pearson similarity matrix generated comparing RNA-5c (X1-5) and RNA-3c (Y1-4) clusters, together with the bulk cell lines transcriptome. Black arrows associate clusters on the higher similarity depicted among clusters. B) Seurat integration table. On the columns are shown the integration clusters and on the rows the number of cells from each RNA-5c and RNA-3c cluster present in the integration clusters. C) UMAP plot of the Seurat integrated clusters. Cell line association is given by the hierarchical clustering shown in Figure 2. Y1 cell line association is indicated with a question mark since by Figure 2D, Y1 seems to be associated with H1975, as instead, by Seurat integration and SCA, pseudo-bulk Y1 is more similar to HCC827 than to H1975.

Horizontal arrows in Figure 3A indicate the association between RNA-5c and RNA-3c. Pseudo-bulk similarity matrix also shows that RNA-3c Y1 seems to belong HCC827

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0504.v1

instead of H1975, as suggested by the hierarchical clustering in Figure 2D. Moreover, cluster Y2 shows a lighter similarity to cluster X2 and Y3 to X1. Notably, Seurat integration provides a superimposable picture to that of the SCA pseudo-bulk analysis, see Figure 3B and C. In Seurat integration, as in SCA pseudo-bulk analysis, Y1 cluster is more near to HCC827 cell cluster that to H1975 cluster, to which it was assigned based on the hierarchical cluster in Figure 2D.

2.3. SCA pseudo-clusters as tool in multi-modal analysis.

Since SCA pseudo-bulks seem to effectively depict the similarity among independent scRNAseq experiments, we tested their ability in identifying similarities in a multi-modal setting. In the GEO repository are present two scATACseq experiments (GSM4224432 and GSM4224433) including about 37000 cells each were H2228, H1975, A549, H838 and HCC827 cell lines are mixed in a one-to-one ratio as for the scRNAseq experiment GSM3022245.

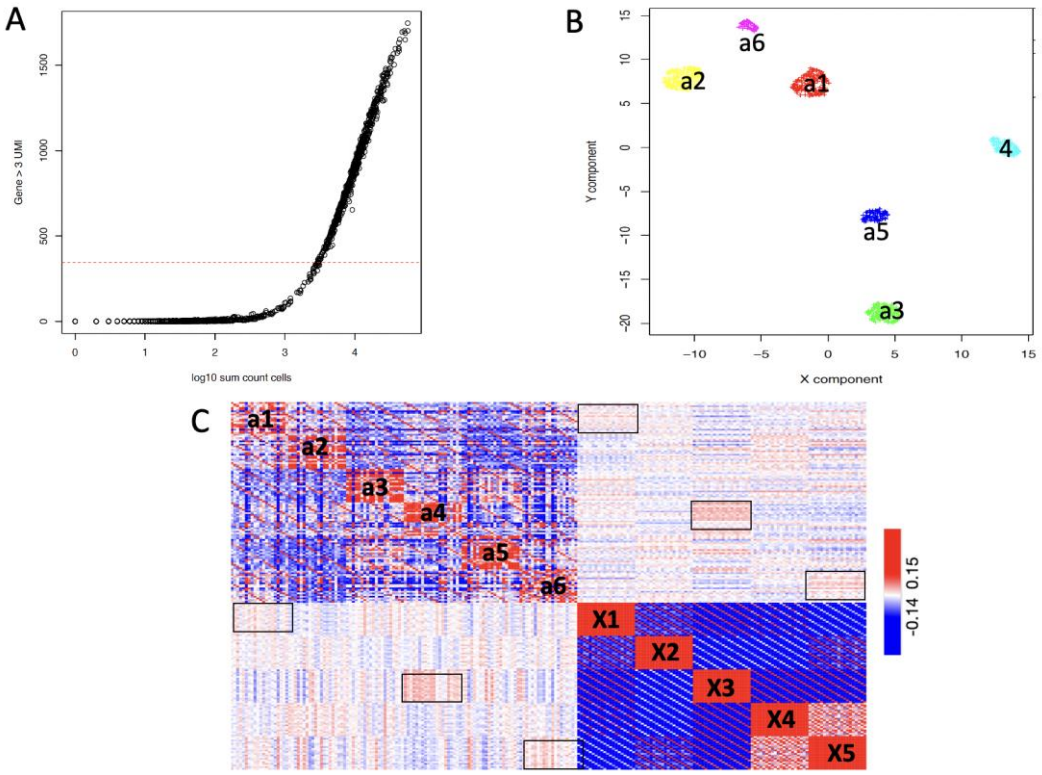


Figure 4. ATAC-5c versus RNA-5c. ATACseq regions are associated with genes; counts associated with each gene indicate the number of reads detected by ATACseq within the genomic coordinates of each gene. A) Cells were represented on the basis of the number of genes supported by at least 3 UMI and plotted with respect to the total number of UMI in each cell. B) Seurat clustering (resolution: 0.6) of the cells' ATACseq counts. C) SCA pseudo-bulk Pearson similarity among ATACseq clusters (a1-6) and scRNAseq (X1-5, Figure 2A). Square boxes indicate the highest similarity among clusters: a1/X1, a4/X3 and a6/X5.

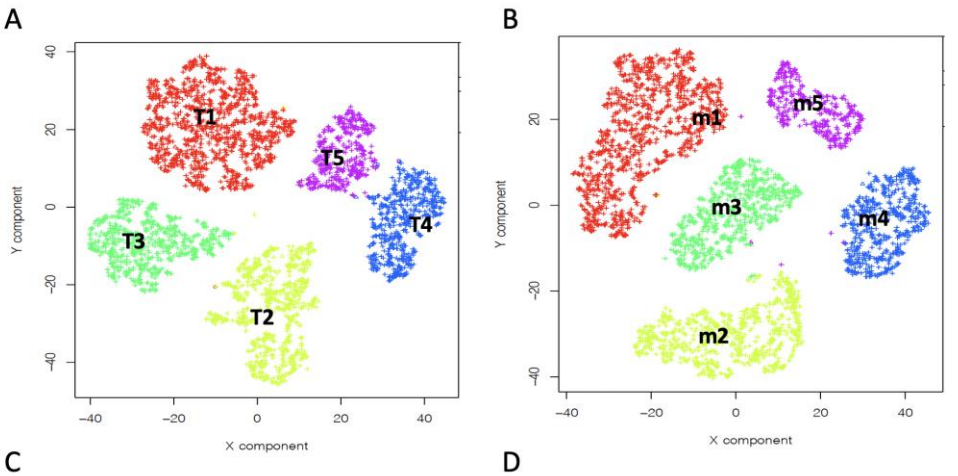
Unfortunately, when the scATACseq regions were annotated to gene loci, only a few hundred cells (GSM4224433, ATAC-5c) were characterized by the presence of at least 3 UMI/gene locus for at least 400 genes, see Figure 4A. Having such a limited number of cells, harboring gene associated chromatin conformation information, does not guarantee that all the five cell lines are equally presented. These cells were clustered with Seurat embedded in rCASC and we detected 6 clusters, see Figure 4B. SCA pseudo clusters were generated for the filtered cells from ATAC-5c, and similarity of such clusters with respect

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0504.v1

to RNA-5c was calculated, see Figure 4C. With respect to independent scRNAseq experiments as RNA-5c and RNA-3c the similarity between RNA-5c and ATAC-5c data is quite blur, Figure 4C; however, we can still depict a weak similarity among a1/X1, a4/X3 and a6/X5 clusters.

2.4. SCA as tool for the detection of regulatory gene hubs.

In our previous paper on SCA [10], we used the output of SCA based on TFs or miRNA targets to reconstruct at least some of the clusters generated using the full gene matrix. A SCA based on TFs or miRNA targets, generates a hidden layer made of metagenes pseudo-expression. Each meta-gene is a value derived by the level of expression of the genes, which expressions are controlled by a specific TF or miRNA. Thus, in case a SCA pseudo-expression matrix can reconstruct a cluster, previously generated using all genes' expression values, this is a strong indication of how the meta-features, i.e. TFs or miRNA, are important players in the formation of that specific cluster. Here, we show that we can grab more information using for clustering the output of a SCA based on TFs or miRNA targets. The peculiarity of the autoencoders is that the hidden layer encodes the most informative parts of a dataset. Thus, running multiple times a SCA and performing the cumulative sum of the output results will produce the progressive increase of important feature signals as instead the not informative portion will remain near to the background signal. Using such an approach we clustered the output of TFs and miRNA SCA, Figure 5A and B, for RNA-5c.



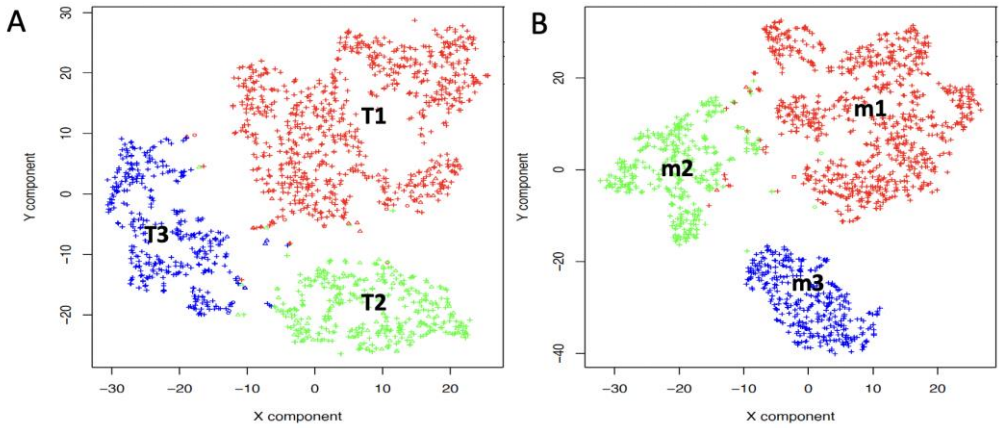
Cell lines	T1	T2	T3	T4	T5
A549	1228	0	2	7	5
H1975	1	3	1	0	431
H2228	0	2	745	0	2
H838	5	856	9	3	6
HCC827	3	0	0	589	6

Cell lines	m1	m2	m3	m4	m5
A549	1226	1	1	8	6
H1975	1	1	1	0	433
H2228	0	1	745	0	3
H838	5	863	4	7	0
HCC827	3	2	1	586	6

Figure 5. RNA-5c, clustering of cells using the SCA meta-genes. A) SCA using TFs targets, B) SCA using miRNA targets, C) clusters to cell association for A, D) clusters to cell association for B.

In Figure 2A, we can note that in clustering based on SCA meta-genes. cells are more homogeneously aggregated in clusters with respect to the analysis performed using the full gene set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0504.v1



C

Cell lines	T1	T2	T3
?	521	14	5
H1975	8	24	363
H2228	14	341	3
HCC827	288	1	2

D

Cell lines	m1	m2	m3
?	535	3	2
H1975	24	370	1
H2228	9	3	346
HCC827	286	5	0

Figure 6. RNA-3c, clustering of cells using the SCA meta-genes. A) SCA using TFs targets, B) SCA using miRNA targets, C) clusters to cell association for A, D) clusters to cell association for B.

We also ran the TF and miRNA SCA analysis on RNA-3c. In this case, we obtained three homogeneous clusters. The clusters Y1, Figure 2D and Figure 3D, which show a discrepancy in cell type assignment, was embedded in cluster T1 and m1, Figure 6A and B. This confirms that it belongs to HCC827 cell line.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0504.v1

3. Discussion

An autoencoder is composed of an encoder and a decoder sub-model. The encoder condenses the input, and the decoder tries to reconstruct the input from the condensed version provided by the encoder. Usually, after training, the encoder model is saved, and the decoder is discarded. Gold and coworkers [9] proposed a new type of autoencoders, Shallow Sparsely-Connected Autoencoder (SSCA) and Variational Autoencoder (SSCVA), in which the decoder is not fully connected, but it is made of meta-features, showing specific relation with respect to the input genes. Gold indicated SSCA/SSCVA as promising tools, which can be exploited in the identification of transcription factors with differential activity among conditions or cell types. More recently, we [10] showed that Sparsely Connected Autoencoder (SCA) can be used to reconstruct at least some of the clusters, previously depicted by clustering analysis of the full genes set of single cell RNAseq data (scRNAseq). The peculiarity of SCA [10] is that the decoder model is not discarded, but it is used to grab more functional knowledge from the input dataset. This is possible because, due to the specific relations existing among input genes and decoder meta-features, the decoder is not any more a black box. Thus, SCA [10] can provide new incites in the biological meaning of the reconstructed single cell RNAseq clusters. Notably, SCA has been recently applied to functional Magnetic Resonance Imaging (fMRI) for Detecting Autism Spectrum Disorder [12].

In this paper, we extend the mining strength of SCA. Specifically, we show in Figure 1 that SCA can be used to reconstruct multiple pseudo-replicates representing the average expression of scRNAseq clusters. These cluster pseudo replicates are a modelled average representation of the expression of the cell genes present in any cluster, previously defined by a scRNAseq analysis on the full gene set. SCA pseudo-bulks have the advantage of not being zero inflated [17], making them ideal to depict differentially expressed genes among different scRNAseq clusters using conventional RNAseq tools [18, 19]. Furthermore, we observed that SCA pseudo-bulks work effectively in depicting relationship among scRNAseq data coming from different experiments. Specifically, we showed in Figure 3 that SCA pseudo-bulks have the same integrative power of Seurat [20]. The use of SCA pseudo-bulks in multi-modal data analysis, Figure 4, requires further effort, due to the limited quality of the ATAC-5c dataset. We need to run it on a dataset in which scRNAseq and scATACseq data are collected from the same cell, to estimate how strong is the correlation between transcriptomics and open-chromatin. We have recently found on GEO the dataset GSE151302 [21], which is made by five single nucleus ATAC (snATAC-seq) and RNA (snRNA-seq) sequencing to generate paired, cell-type-specific chromatin accessibility and transcriptional profiles of the adult human kidney. We are in progress to analyze these data.

We have also extended the functional mining power of the SCA developed in [10]. In the present implementation, the decoder layer is based on transcription factors (TFs) or miRNAs, i.e. each node of the hidden layer represents a TF or a miRNA. Each node of the decoder layer is connected only to the corresponding input target genes. SCA is executed multiple times and the resulting decoder layers are sum Because the hidden layer of an autoencoder can provide a representation of the input data in which noise is at least partially discarded, summing multiple runs of the hidden layer results will highlight genes playing an important role in cell subpopulations with respect to genes that are non-specifically modulated. We observed that the clustering of TFs meta-features, Figure 4A, provides more homogeneous cells subpopulations than those depictable using the raw data, Figure 2A. Furthermore, the artifactual identification of four clusters for three cell lines, in Figure 2B, is completely solved in Figure 5, where both the number of clusters and the cell lines are perfectly matching. Furthermore, the results in Figure 5A, perfectly agree with the observation resulting from the integration with pseudo-clusters, Figure 3A, and with Seurat, Figure 3B and C. Both pseudo-bulk and Seurat integration highlight the association of cluster Y1, Figure 2B, with cell line HCC827 instead to H1975, observed from the clustering of raw data in Figure 2D. Thus, SCA meta-features provide a more

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2021 doi:10.20944/preprints202109.0504.v1

robust representation of the cell subpopulations depictable by clustering of single cell data.

Taken together our data highlight some interesting a useful feature of SCA as data mining tools.

4. Materials and Methods

scRNAseq benchmark preprocessing. Preprocessing was performed with the rCASC package [14]. Count matrix for scRNAseq, produced by a mixture of three (H2228, H1975 and HCC827) and five human lung adenocarcinoma cell lines [11] (H2228, H1975, A549, H838 and HCC827) were respectively retrieved from GSM3022245 and GSM3618014 datasets available at GEO database [22]. Counts were adjusted by SAVER [23], to provide accurate expression estimates for all genes. Subsequently, count tables were filtered to remove low quality cells, i.e. those cells with less than 250 genes called "*present*" (a gene is called present if supported by at least 3 UMIs, we use the 3 UMIs/gene as threshold because it allow to moderate the effect of sequencing errors in UMI counting, when up to two UMIs are used to call present a gene) for the five cell line dataset, and those cells with less than 100 genes called present for the three cell line dataset.

Then, for the five cell lines dataset, we retained only the 2500 most variant genes out of the 5000 most expressed. The resulting dataset made of 3904 cells and 2500 genes will be called from now RNA-5c. The three cell lines dataset were filtered to retain the same 2500 genes selected in RNA-5c, from now this dataset is called RNA-3c.

scRNAseq cell type association. TPM expression data for H2228, H1975, A549, H838 and HCC827 cell lines were retrieved from the CCLE repository (https://depmap.org/portal/download/). Bulk expression cell lines were row-mean centered.

RNA-5c and RNA-3c were independently clustered using Seurat, implemented in rCASC. The analysis was done using the Seurat resolution parameter set at 0.1 to generate a number of clusters as much like the expected cell line number. From Seurat clustering, we obtained five clusters for RNA-5c and four clusters for RNA-3c. Subsequently, each cluster was converted in pseudo-bulk CPM expression and the CPM data were row-mean centered.

Mean centered bulk cell lines and cluster pseudo-bulks were combined and clustered using Morpheus (https://software.broadinstitute.org/morpheus/), hierarchical clustering was performed using Euclidean distance and average linkage, see Figure 2.

Integration of RNA-5c and RNA-3c was performed using Seurat integration implemented in rCASC (Seurat resolution=0.1).

scATACseq benchmark preprocessing. The scATACseq processed data for the same mixture of five cell lines used in RNA-5c were retrieved from the GEO repository: GSE142285. GSE142285 includes two samples GSM4224432 and GSM4224433. Open chromatin regions for samples GSM4224432 and GSM4224433 were associated with human genes, defined as the gene locus in ENSEMBL hg38 version 100 GTF file. Subsequently, cells characterized with at least 400 genes, i.e. a gene is called present if supported by at least 3 UMIs. Respectively 98 and 597 cells out of 37000 passed the above filter in GSM4224432 and GSM4224433. The 597 cells from the GSM4224433, now named ATAC-5c, were subsequently analyzed by Seurat clustering implemented in rCASC.

Software implementation. All software described in the present paper was implemented in the rCASC package [14].

SCA-pseudo-clusters generation. SCA pseudo-clusters, Figure 1 and 3 were generated using the rCASC function *autoencoder4pseudoBulk* (parameters: permutation=20, nEpochs=1000).

Seurat clustering. Seurat clustering, Figure 2 A,B, were done using the Seurat clustering function embedded in rCASC: *seuratBootstrap* (parameters: resolution=0.1, nPerm=90, PCADIM=10, seed=111). The optimal number of PCA components to be used

in the clustering, i.e. PCADIM parameter, was depicted using the rCASC function *seuratPCAEval*.

Seurat clustering, Figure 4B, was done using the Seurat clustering function embedded in rCASC: *seuratBootstrap* (parameters: resolution=0.6, nPerm=90, PCADIM=10, seed=111).

Hierarchical clustering. Hierarchical clustering, Figure 2C and D was performed using online Morpheus tool from Broad Institute (https://software.broadinstitute.org/morpheus/). Figures 3A and 4C Pearson similarity was performed using Morpheus tool.

Seurat Integration. Seurat integration, Figure 3B and C, was performed using the Seurat embedded in rCASC, using the *seuratIntegration* function (parameters: seed=111, k=0.1).

SCA-metagenes. SCA TF and miRNA metagenes, Figure 4 and 5, were build using the function *autoencoder4clustering* (parameters: permutation=100, nEpochs=1000). Clustering on the matrix resulting from the sum of the dense spaces was done with *seuratBootstrap* (parameters: resolution=0.1, nPerm=40, PCADIM=5, seed=111).

5. Conclusions

Sparsely Connected Autoencoders (SCA) represent a new way of looking at deep learning tools. Specifically, SCAs offer the opportunity to transform data in a controlled way to grasp from single cell data hidden biological information like relations among cell sub-populations and transcription factors or miRNAs. Furthermore, the peculiar ability of autoencoder to retain only the important part of a signal can help in discriminating between true differences among cell sub-populations and clusterisation overfitting.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, L.A. and R.A.C.; methodology, L.A. and M.B.; software, S.G.C, L.A. and M.R.; validation, F.C., M.A. and M.R.; data curation, R.A.C.; writing—original draft preparation, R.A.C. and L.A.; writing—review and editing, , M.B. and F.C.; supervision, R.A.C..; project administration, R.A.C.. All authors have read and agreed to the published version of the manuscript."

Data Availability Statement: All data used for the generation of the figures shown in this paper are available at figshare.com: https://figshare.com/projects/Sparsely_Connected_Autoencoders_a_multi-purpose_tool_for_single_cell_OMICs_analysis/123226

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflict of interest.".

References

1. Gao, C., M. Zhang, and L. Chen, *The Comparison of Two Single-cell Sequencing Platforms: BD Rhapsody and 10x Genomics Chromium.* Curr Genomics, 2020. **21**(8): p. 602-609.
2. Stoeckius, M., et al., *Simultaneous epitope and transcriptome measurement in single cells.* Nat Methods, 2017. **14**(9): p. 865-868.
3. Sinha, S., et al., *Profiling Chromatin Accessibility at Single-cell Resolution.* Genomics Proteomics Bioinformatics, 2021.
4. Saviano, A., N.C. Henderson, and T.F. Baumert, *Single-cell genomics and spatial transcriptomics: Discovery of novel cell states and cellular interactions in liver physiology and disease biology.* J Hepatol, 2020. **73**(5): p. 1219-1230.
5. Chappell, L., A.J.C. Russell, and T. Voet, *Single-Cell (Multi)omics Technologies.* Annu Rev Genomics Hum Genet, 2018. **19**: p. 15-41.
6. Ji, Y., et al., *Machine learning for perturbational single-cell omics.* Cell Syst, 2021. **12**(6): p. 522-537.
7. Yan, R., et al., *Potential applications of deep learning in single-cell RNA sequencing analysis for cell therapy and regenerative medicine.* Stem Cells, 2021. **39**(5): p. 511-521.

8. Rai, V., et al., *Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures.* Mol Metab, 2020. **32**: p. 109-121.
9. Gold, M.P., A. LeNail, and E. Fraenkel, *Shallow Sparsely-Connected Autoencoders for Gene Set Projection.* Pac Symp Biocomput, 2019. **24**: p. 374-385.
10. Alessandri, L., et al., *Sparsely-connected autoencoder (SCA) for single cell RNAseq data mining.* NPJ Syst Biol Appl, 2021. **7**(1): p. 1.
11. Tian, L., et al., *Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments.* Nat Methods, 2019. **16**(6): p. 479-487.
12. Alessandri, L., M. Arigoni, and R. Calogero, *Differential Expression Analysis in Single-Cell Transcriptomics.* Methods Mol Biol, 2019. **1979**: p. 425-432.
13. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
14. Alessandri, L., et al., *rCASC: reproducible classification analysis of single-cell sequencing data.* Gigascience, 2019. **8**(9).
15. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
16. Tran, H.T.N., et al., *A benchmark of batch-effect correction methods for single-cell RNA sequencing data.* Genome Biol, 2020. **21**(1): p. 12.
17. Pierson, E. and C. Yau, *ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.* Genome Biol, 2015. **16**: p. 241.
18. Nikolayeva, O. and M.D. Robinson, *edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology.* Methods Mol Biol, 2014. **1150**: p. 45-79.
19. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
20. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13): p. 3573-3587 e29.
21. Muto, Y., et al., *Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney.* Nat Commun, 2021. **12**(1): p. 2190.
22. Clough, E. and T. Barrett, *The Gene Expression Omnibus Database.* Methods Mol Biol, 2016. **1418**: p. 93-110.
23. Huang, M., et al., *SAVER: gene expression recovery for single-cell RNA sequencing.* Nat Methods, 2018. **15**(7): p. 539-542.